Section 1 Disorders of the Alimentary Tract

314 Approach to the Patient with Gastrointestinal Disease

William L. Hasler, Chung Owyang

ANATOMIC CONSIDERATIONS

The gastrointestinal (GI) tract extends from the mouth to the anus and is composed of several organs with distinct functions. Specialized sphincters that assist in gut compartmentalization separate the organs. The gut wall is organized into distinct layers that contribute to regional activities. The mucosa is a barrier to luminal contents or a site for fluid and nutrient transfer. Gut smooth muscle in association with the enteric nervous system mediates propulsion from one region to the next. Many GI organs possess a serosal layer that provides a supportive foundation and permits external input.

Interactions with other systems serve the needs of the gut and the body. Pancreaticobiliary conduits deliver bile and enzymes into the duodenum. The vascular supply is modulated by GI activity. Lymphatic channels assist in gut immune activities. Intrinsic nerves provide the controls for propulsion and fluid regulation. Extrinsic neural input provides volitional or involuntary control that is specific for each gut region.

FUNCTIONS OF THE GI TRACT

The GI tract serves two main functions—assimilating nutrients and eliminating waste. In the mouth, food is processed, mixed with salivary amylase, and delivered to the gut lumen. The esophagus propels the bolus into the stomach; the lower esophageal sphincter prevents oral reflux of gastric contents. The squamous esophageal mucosa protects against significant diffusion or absorption. Aboral esophageal contractions coordinate with relaxation of the upper and lower esophageal sphincters on swallowing.

The stomach triturates and mixes the food bolus with pepsin and acid. Gastric acid also sterilizes the upper gut. The proximal stomach serves a storage function by relaxing to accommodate the meal. Phasic contractions in the distal stomach propel food residue against the pylorus, where it is ground and thrust proximally for further mixing before it is emptied into the duodenum. The stomach secretes intrinsic factor for vitamin B_{12} absorption.

Most nutrient absorption occurs in the small intestine. The intestinal mucosal villus architecture provides maximal surface area for absorption and is endowed with specialized enzymes and transporters. Triturated food from the stomach mixes with pancreatic juice and bile in the duodenum. Pancreatic juice contains enzymes for carbohydrate, protein, and fat digestion as well as bicarbonate to optimize the pH for enzyme activation. Bile secreted by the liver and stored in the gallbladder is essential for lipid digestion. The proximal intestine is optimized for rapid absorption of most nutrients and minerals, whereas the ileum is better suited for absorbing vitamin B_{12} and bile acids. Bile contains by-products of erythrocyte degradation, toxins, medications, and cholesterol for fecal evacuation. Small intestinal motor function delivers indigestible residue into the colon for processing. The ileocecal junction is a sphincteric structure that prevents coloileal reflux, maintaining small-intestinal sterility.

The colon prepares waste for evacuation. The colonic mucosa dehydrates the stool, decreasing daily volumes of 1000–1500 mL in the ileum to 100–200 mL expelled from the rectum. The colon possesses a dense bacterial colonization that ferments undigested carbohydrates and short-chain fatty acids. Additional roles for the gut microbiome include modulation of immune and physiologic activity. Transit in the esophagus takes seconds and times in the stomach and small intestine range from minutes to a few hours, but colon propagation requires more than 1 day in most individuals. Colon contractions exhibit a to-and-fro character that promotes fecal desiccation. The proximal colon mixes and absorbs fluid, while the distal colon exhibits peristaltic contractions and mass movements to expel the stool. The colon terminates in the anus, which possesses volitional and involuntary controls to permit fecal retention until it can be released in a convenient setting.

EXTRINSIC MODULATION OF GUT FUNCTION

GI function is modified by influences outside the gut. Unlike other organs, the gut is in continuity with the outside environment. Protective mechanisms are vigilant against damage from foods, medications, toxins, and infectious organisms. Mucosal immune mechanisms include epithelial and lamina propria lymphocyte and plasma cell populations supported by lymph node chains to prevent noxious agents from entering the circulation. Antimicrobial peptides secreted by intestinal Paneth cells also defend against luminal pathogens. All drugs and toxins absorbed into the bloodstream are filtered and detoxified in the liver via the portal venous circulation. Although intrinsic nerves control most basic gut activities, extrinsic neural input modulates many functions. Many GI reflexes involve extrinsic vagus or splanchnic nerve pathways. The brain-gut axis alters function in regions not under volitional regulation. As an example, stress has potent effects on gut motor, secretory, and sensory functions.

OVERVIEW OF GI DISEASES

GI diseases develop as a result of abnormalities within or outside of the gut and range in severity from those that produce mild symptoms and no long-term morbidity to those with intractable symptoms or adverse outcomes. Diseases may be localized to one organ or exhibit diffuse involvement at many sites.

CLASSIFICATION OF GI DISEASES

GI diseases are manifestations of alterations in nutrient assimilation or waste evacuation or in the activities supporting these main functions.

Impaired Digestion and Absorption Diseases of the stomach, intestine, biliary tree, and pancreas can disrupt digestion and absorption. The most common intestinal maldigestion syndrome, lactase deficiency, produces gas and diarrhea after ingestion of dairy products and has no adverse outcomes. Other intestinal enzyme deficiencies produce similar symptoms after ingestion of other simple sugars. Conversely, celiac disease, bacterial overgrowth, infectious enteritis, Crohn's ileitis, and radiation damage, which affect digestion and/or absorption more diffusely, produce anemia, dehydration, electrolyte disorders, or malnutrition. Gastric hypersecretory conditions such as Zollinger-Ellison syndrome damage the intestinal mucosa, impair pancreatic enzyme activation, and accelerate transit due to excess gastric acid. Biliary obstruction from stricture or neoplasm impairs fat digestion. Impaired pancreatic enzyme release in chronic pancreatitis or pancreatic cancer decreases intraluminal digestion and can lead to malnutrition.

Altered Secretion Selected GI diseases result from dysregulation of gut secretion. Gastric acid hypersecretion occurs in Zollinger-Ellison syndrome, G cell hyperplasia, retained antrum syndrome, and some individuals with duodenal ulcers. Conversely, patients with atrophic gastritis or pernicious anemia release little or no gastric acid. Inflammatory and infectious small-intestinal and colonic diseases produce fluid loss through impaired absorption or enhanced secretion. Common intestinal and colonic hypersecretory conditions cause diarrhea and include acute bacterial or viral infection, chronic *Giardia* or cryptosporidia infections, small-intestinal bacterial overgrowth, bile salt diarrhea,

microscopic colitis, diabetic diarrhea, and abuse of certain laxatives. Less common causes include large colonic villus adenomas and endocrine neoplasias with tumor overproduction of secretagogue transmitters like vasoactive intestinal polypeptide.

Altered Gut Transit Impaired gut transit may be secondary to mechanical obstruction. Esophageal occlusion most often results from stricture (due to acid exposure or eosinophilic esophagitis) or neoplasm. Gastric outlet obstruction develops from peptic ulcer disease or gastric cancer. Small-intestinal obstruction most commonly results from adhesions but may also occur with Crohn's disease, radiation- or drug-induced strictures, and less likely malignancy. The most common cause of colonic obstruction is colon cancer, although inflammatory strictures develop in patients with inflammatory bowel disease (IBD), after certain infections such as diverticulitis, or with some drugs.

Retardation of propulsion also develops from disordered motor function. Achalasia is characterized by impaired esophageal body peristalsis and incomplete lower esophageal sphincter relaxation. Gastroparesis is the symptomatic delay in gastric emptying of meals due to impaired gastric motility. Intestinal pseudoobstruction causes marked delays in small-bowel transit due to enteric nerve or intestinal smooth-muscle injury. Slow-transit constipation is produced by diffusely impaired colonic propulsion. Constipation also is produced by outlet abnormalities such as rectal prolapse, intussusception, or dyssynergia—a failure of anal or puborectalis relaxation upon attempted defecation.

Disorders of rapid propulsion are less common than those with delayed transit. Rapid gastric emptying occurs in postvagotomy dumping syndrome, with gastric hypersecretion, and in some cases of functional dyspepsia and cyclic vomiting syndrome. Exaggerated intestinal or colonic motor patterns may be responsible for diarrhea in irritable bowel syndrome (IBS). Accelerated transit with hyperdefecation is noted in hyperthyroidism.

Immune Dysregulation Many inflammatory GI conditions are consequences of altered gut immune function. The mucosal inflammation of celiac disease results from dietary ingestion of gluten-containing grains. Some patients with food allergy also exhibit altered immune populations. Eosinophilic esophagitis and eosinophilic gastroenteritis are inflammatory disorders with prominent mucosal eosinophils. Ulcerative colitis and Crohn's disease are disorders of uncertain etiology that produce mucosal injury primarily in the lower gut. The microscopic colitides, lymphocytic and collagenous colitis, exhibit colonic subepithelial infiltrates without visible mucosal damage. Bacterial, viral, and protozoal organisms may produce ileitis or colitis in selected patient populations. Furthermore, alterations in the gut microbiome (termed dysbiosis) are postulated to trigger flares of IBD, celiac disease, and IBS.

Impaired Gut Blood Flow Different GI regions are at variable risk for ischemic damage from impaired blood flow. Rare cases of gastroparesis result from blockage of the celiac and superior mesenteric arteries. More commonly encountered are intestinal and colonic ischemia that are consequences of arterial embolus, arterial thrombosis, venous thrombosis, or hypoperfusion from dehydration, sepsis, hemorrhage, or reduced cardiac output. These may produce mucosal injury, hemorrhage, or even perforation. Chronic ischemia may result in intestinal stricture. Some cases of radiation enterocolitis exhibit reduced mucosal blood flow.

Neoplastic Degeneration All GI regions are susceptible to malignant degeneration to varying degrees. In the United States, colorectal cancer is most common and usually presents after age 50 years. Worldwide, gastric cancer is prevalent especially in certain Asian regions. Esophageal cancer develops with chronic acid reflux or after an extensive alcohol or tobacco use history. Small-intestinal neoplasms are rare and occur with underlying inflammatory disease. Anal cancers arise after prior anal infection or inflammation. Pancreatic and biliary cancers elicit severe pain, weight loss, and jaundice and have poor prognoses. Hepatocellular carcinoma usually arises in the setting of chronic viral hepatitis or cirrhosis secondary to other causes. Most GI cancers exhibit carcinomatous histology; however, lymphomas and other cell types also are observed.

Disorders without Obvious Organic Abnormalities The most common GI disorders show no abnormalities on biochemical or structural testing and include IBS, functional dyspepsia, and functional heartburn. These disorders exhibit altered gut motor function; however, the pathogenic relevance of these abnormalities is uncertain. Exaggerated visceral sensory responses to noxious stimulation may cause discomfort in these disorders. Symptoms in other patients result from altered processing of visceral pain sensations in the central nervous system. Functional bowel patients with severe symptoms may exhibit significant emotional disturbances on psychometric testing. Subtle immunologic defects may contribute to functional symptoms as well.

Genetic Influences Although many GI diseases result from environmental factors, others exhibit hereditary components. Family members of IBD patients show a genetic predisposition to disease development themselves. Colonic, esophageal, and pancreatic malignancies arise in certain inherited disorders. Rare genetic dysmotility syndromes are described. Familial clustering is observed in the functional bowel disorders, although this may be secondary learned familial illness behavior rather than a true hereditary factor.

SYMPTOMS OF GI DISEASE

Symptoms of GI disease include abdominal pain, heartburn, nausea and vomiting, altered bowel habits, GI bleeding, jaundice, and other manifestations (Table 314-1).

Abdominal Pain Abdominal pain results from GI disease and extra-intestinal conditions involving the genitourinary tract, abdominal wall, thorax, or spine. Visceral pain generally is midline in location and vague in character, whereas parietal pain is localized and precisely described. Painful inflammatory diseases include peptic ulcer,

TABLE 314-1 Common Causes of Common Gastrointestinal (GI) Symptoms

ABDOMINAL PAIN	NAUSEA AND VOMITING	DIARRHEA	GI BLEEDING	OBSTRUCTIVE JAUNDICE
Appendicitis	Medications	Infection	Ulcer disease	Bile duct stones
Gallstone disease	GI obstruction	Poorly absorbed sugars	Esophagitis	Cholangiocarcinoma
Pancreatitis	Motor disorders	Inflammatory bowel disease	Varices	Cholangitis
Diverticulitis	Functional bowel disorder	Microscopic colitis	Vascular lesions	Sclerosing cholangitis
Ulcer disease	Enteric infection	Functional bowel disorder	Neoplasm	Ampullary stenosis
Esophagitis	Pregnancy	Celiac disease	Diverticula	Ampullary carcinoma
GI obstruction	Endocrine disease	Pancreatic insufficiency	Hemorrhoids	Pancreatitis
Inflammatory bowel disease	Motion sickness	Hyperthyroidism	Fissures	Pancreatic tumor
Functional bowel disorder	Central nervous system disease	Ischemia	Inflammatory bowel disease	
Vascular disease		Endocrine tumor	Infectious colitis	
Gynecologic causes				
Renal stone				

appendicitis, diverticulitis, IBD, pancreatitis, cholecystitis, and infectious enterocolitis. Noninflammatory visceral sources include biliary colic, mesenteric ischemia, and neoplasia. The most common causes of abdominal pain are IBS and functional dyspepsia.

Heartburn Heartburn, a burning substernal sensation, is reported intermittently by 40% of the population. Classically, heartburn results from excess gastroesophageal acid reflux, but some cases exhibit normal esophageal acid exposure and are caused by reflux of nonacidic material or heightened sensitivity of esophageal nerves.

Nausea and Vomiting Nausea and vomiting are caused by GI diseases, medications, toxins, infection, endocrine disorders, labyrinthine conditions, and central nervous system disease. Mechanical obstructions of the upper gut are commonly excluded as causes of chronic nausea and vomiting, but disorders of propulsion including gastroparesis and intestinal pseudoobstruction elicit similar symptoms. Nausea and vomiting also are commonly reported by patients with IBS and functional disorders of the upper gut (including chronic nausea vomiting syndrome and cyclic vomiting syndrome).

Altered Bowel Habits Altered bowel habits are common complaints of patients with GI disease. Constipation may be reported as infrequent defecation, straining with defecation, passage of hard stools, or a sense of incomplete fecal evacuation and is caused by obstruction, colonic motor disorders, medications, and endocrine diseases like hypothyroidism and hyperparathyroidism. Diarrhea may be reported as frequent defecation, passage of loose or watery stools, fecal urgency, or a similar sense of incomplete evacuation. The differential diagnosis of diarrhea includes infections, inflammatory causes, malabsorption, and medications. IBS produces constipation, diarrhea, or an alternating bowel pattern. Fecal mucus is common in IBS, whereas pus and blood characterize IBD. Steatorrhea develops with malabsorption.

GI Bleeding Hemorrhage may develop from any gut organ. Upper GI bleeding presents with melena or hematemesis, whereas lower GI bleeding produces passage of bright red or maroon stools. However, briskly bleeding upper sites can elicit voluminous red rectal bleeding, whereas slowly bleeding ascending colon sites may produce melena. Chronic occult GI bleeding may present with iron deficiency anemia. Causes of upper GI bleeding include ulcer disease, gastroduodenitis, esophagitis, portal hypertensive etiologies, malignancy, tears across the gastroesophageal junction, and vascular lesions. Lower GI sources of hemorrhage include hemorrhoids, anal fissures, diverticula, ischemic colitis, neoplasm, IBD, infectious colitis, drug-induced colitis, arteriovenous malformations, and other vascular lesions.

Jaundice Jaundice results from prehepatic, intrahepatic, or posthepatic disease. Posthepatic causes of jaundice include biliary diseases, like choledocholithiasis, acute cholangitis, primary sclerosing cholangitis, other strictures, and neoplasm, and pancreatic disorders, like acute and chronic pancreatitis, stricture, and malignancy.

Other Symptoms Other symptoms are manifestations of GI disease. Dysphagia, odynophagia, and unexplained chest pain suggest esophageal disease. A globus sensation is reported with esophagopharyngeal conditions, but also occurs with functional GI disorders. Weight loss, anorexia, and fatigue present with neoplastic, inflammatory, motility, pancreatic, and psychiatric conditions. IBD is associated with hepatobiliary dysfunction, skin and eye lesions, and arthritis. Celiac disease may present with dermatitis herpetiformis. Jaundice can produce pruritus. Conversely, systemic diseases have GI consequences. Systemic lupus may cause gut ischemia, presenting with pain or bleeding. Severe burns may lead to gastric ulcer formation.

EVALUATION OF THE PATIENT WITH GI DISEASE

Evaluation of the patient with suspected GI disease begins with a careful history and examination. Subsequent investigation with tools to test gut structure or function and luminal constituents are indicated in selected cases. In patients with normal findings on diagnostic testing, validated symptom profiles are used to confidently diagnose a functional bowel disorder.

HISTORY

The history in suspected GI disease has several components. Symptom timing, patterns, and duration suggest specific etiologies. Short duration symptoms commonly result from acute infection or inflammation, toxin exposure, or ischemia. Long-standing symptoms point to chronic inflammation, neoplasia, or functional bowel disorders. Symptoms from mechanical obstruction, ischemia, IBD, and functional bowel disorders are worsened by meals, while ulcer symptoms may be relieved by eating or antacids. Ulcer pain occurs intermittently over weeks to months, whereas biliary colic has a sudden onset and lasts up to several hours. Acute pancreatitis pain is severe and persists for days to weeks. Meals elicit diarrhea while defecation relieves discomfort in some cases of IBD and IBS. Functional bowel disorders are exacerbated by stress. Sudden awakening from sound sleep by pain suggests organic rather than functional disease. Diarrhea from malabsorption usually improves with fasting, whereas secretory diarrhea persists without oral intake.

Symptom relation to other factors narrows the list of diagnostic possibilities. Obstructive symptoms with prior abdominal surgery raise concern for adhesions. Loose stools after gastrectomy or cholecystectomy suggest dumping syndrome or postcholecystectomy diarrhea. Symptom onset after travel prompts consideration of infection. Medications produce pain, altered bowel habits, or GI bleeding. Celiac disease is prevalent in people of northern European descent, whereas IBD is more common in Jewish populations. A sexual history may raise concern for infection or immunodeficiency.

For nearly 40 years, working groups have devised symptom criteria to improve diagnosis of functional bowel disorders and to minimize the numbers of unnecessary diagnostic tests performed. The best accepted symptom-based criteria are the Rome criteria. However, when tested against findings of structural investigations in IBS and functional dyspepsia, the Rome criteria exhibit sensitivities and specificities of only 55–75% indicating the need for careful test selection in patients at high risk of organic disease.

PHYSICAL EXAMINATION

The physical examination complements information from the history. Abnormal vital signs provide diagnostic clues and determine the need for acute intervention. Fever suggests inflammation or neoplasm. Orthostasis is produced by significant blood loss, dehydration, sepsis, or autonomic neuropathy. Skin, eye, or joint findings may point to specific diagnoses. Neck examination with swallowing assessment evaluates dysphagia. Lung and cardiac examinations evaluate for cardiopulmonary disease as causes of abdominal pain or nausea. Pelvic examination tests for a gynecologic source of abdominal pain. Rectal examination may detect blood, indicating mucosal injury or neoplasm or a palpable inflammatory mass in appendicitis. Metabolic conditions and gut motor disorders have associated peripheral neuropathy.

Abdominal inspection may reveal distention from obstruction, tumor, or ascites or vascular abnormalities with liver disease. Ecchymoses develop with severe pancreatitis. Auscultation detects bruits or friction rubs from vascular disease or hepatic tumors. Loss of bowel sounds signifies ileus, whereas high-pitched, hyperactive sounds characterize intestinal obstruction. Percussion assesses liver size and detects shifting dullness from ascites. Palpation assesses for hepatosplenomegaly and neoplastic or inflammatory masses. Intestinal ischemia elicits severe pain but little tenderness. Patients with visceral pain may exhibit generalized discomfort, whereas those with parietal pain or peritonitis have localized pain with involuntary guarding, rigidity, or rebound. Patients with musculoskeletal abdominal wall pain may note tenderness exacerbated by Valsalva or leg lift maneuvers.

TOOLS FOR PATIENT EVALUATION

Laboratory, radiographic, and functional tests assist in diagnosis of suspected GI disease. The GI tract also is amenable to internal evaluation

using endoscopy and to examination of luminal contents. Histopathologic examinations of GI tissues complement these tests.

Laboratory Laboratory tests facilitate diagnosis of GI disease. Iron-deficiency anemia suggests mucosal blood loss, whereas vitamin B_{12} deficiency results from intestinal, gastric, or pancreatic disease. Either can result from inadequate oral intake. Leukocytosis and increased sedimentation rates and C-reactive proteins are found in inflammation, whereas leukopenia is seen in viremic illness. Severe vomiting or diarrhea elicits electrolyte disturbances, acid-base abnormalities, and elevated blood urea nitrogen. Pancreaticobiliary or liver disease produce elevated pancreatic or liver chemistries. Thyroid chemistries, cortisol, and calcium levels evaluate for endocrinologic causes of symptoms. Pregnancy testing is considered for women with unexplained nausea. Serologic tests screen for celiac disease, IBD, connective tissue diseases, and paraneoplastic dysmotility syndromes. Hormone levels are obtained for suspected endocrine neoplasia. Intraabdominal malignancies produce tumor markers including the carcinoembryonic antigen CA 19-9 and α-fetoprotein. Blood testing also monitors medication therapy, as with thiopurine metabolite levels in IBD. Pharmacogenetic methods are being adopted to determine optimal patient populations for GI medication use. In areas including IBD, research into novel biomarkers is being conducted to predict longitudinal course and treatment response. Other body fluids are sampled under certain circumstances. Ascitic fluid is analyzed for infection, malignancy, or findings of portal hypertension. Urine samples screen for carcinoid, porphyria, and heavy metal intoxication.

Luminal Contents Luminal contents can provide diagnostic clues. Stool samples are cultured for bacterial pathogens, examined for leukocytes and parasites, or tested for *Giardia* antigen. Duodenal aspirates can be examined for parasites or cultured for bacterial overgrowth. Fecal fat is quantified in possible malabsorption. Elevations in fecal calprotectin or lactoferrin are found in inflammatory conditions like IBD. Stool electrolytes can be measured in diarrheal conditions. Laxative screens are performed for suspected laxative abuse. Fecal immunochemical and DNA tests are assuming emerging roles in colon cancer screening in low risk populations. Gastric acid is quantified to exclude gastrinoma. Esophageal pH testing is done for refractory symptoms of acid reflux, whereas impedance techniques quantify nonacidic reflux. Pancreatic juice is analyzed for enzyme or bicarbonate content to exclude exocrine insufficiency.

Endoscopy The gut is accessible with endoscopy, which can diagnose causes of bleeding, pain, nausea and vomiting, weight loss, altered bowel function, and fever. Table 314-2 lists common indications for endoscopic procedures. Upper endoscopy evaluates the esophagus, stomach, and duodenum, whereas colonoscopy assesses the colon and distal ileum. Upper endoscopy is advocated as the initial test performed for suspected ulcer disease, esophagitis, neoplasm, malabsorption, and Barrett's metaplasia because of its abilities to visualize and biopsy any abnormality. Colonoscopy is the preferred procedure for colon cancer screening and surveillance and to biopsy colitis secondary to IBD, infection, ischemia, and radiation. Sigmoidoscopy examines the colon to the splenic flexure and excludes distal inflammation or obstruction in young patients not at significant risk for colon cancer. For elusive GI bleeding from arteriovenous malformations or superficial ulcers, small-intestinal examination is performed with push enteroscopy, capsule endoscopy, or double-balloon enteroscopy. Capsule endoscopy also visualizes small-intestinal Crohn's disease in individuals with negative radiography. Endoscopic retrograde cholangiopancreaticography (ERCP) provides diagnoses of pancreatic and biliary disease. Endoscopic ultrasound (EUS) diagnoses and stages GI malignancy, excludes choledocholithiasis, evaluates pancreatitis, and assesses anal continuity.

A newer area involves development of novel imaging protocols which permit optical biopsies to define mucosal histology and detect dysplasia in selected settings. Methods employed include narrow band imaging and chromoendoscopy in colitis and confocal laser endomicroscopy and optical coherence tomography in Barrett's esophagus and gastric cancer surveillance.

Radiography/Nuclear Medicine Radiographic tests evaluate gut diseases and extraluminal structures. Contrast radiography with barium provides mucosal definition and can assess gut transit and pelvic floor dysfunction. Barium swallow is the initial procedure to exclude subtle rings, strictures, or achalasia as causes of dysphagia, whereas small-bowel contrast radiology detects intestinal tumors and Crohn's ileitis. Contrast enemas are performed when colonoscopy is unsuccessful or contraindicated. Ultrasound and computed tomography (CT) evaluate regions not accessible by endoscopy or contrast studies, including the liver, pancreas, gallbladder, kidneys, and retroperitoneum and are useful for diagnosing mass lesions, fluid collections, organ enlargement, and, in the case of ultrasound, gallstones. CT and magnetic resonance (MR) colonography have been

TABLE 314-2 Common Indications for Endoscopy

UPPER ENDOSCOPY	COLONOSCOPY	ENDOSCOPIC RETROGRADE CHOLANGIOPANCREATOGRAPHY	ENDOSCOPIC ULTRASOUND	CAPSULE ENDOSCOPY	DOUBLE-BALLOON ENDOSCOPY
Dyspepsia despite treatment	Cancer screening	Jaundice	Staging of malignancy	Obscure gastrointestinal (GI) bleeding	Ablation of small-intestinal bleeding sources
Dyspepsia with signs of organic disease	Lower GI bleeding	Postbiliary surgery complaints	Characterize and biopsy submucosal mass	Suspected Crohn's disease of the small intestine	Biopsy of suspicious small-intestinal masses/ulcers
Refractory vomiting	Anemia	Cholangitis	Bile duct stones		
Dysphagia	Diarrhea	Gallstone pancreatitis	Chronic pancreatitis		
Upper GI bleeding	Polypectomy	Pancreatic/biliary/ampullary tumor	Drain pseudocyst		
Anemia	Obstruction	Unexplained pancreatitis	Anal continuity		
Weight loss	Biopsy radiologic abnormality	Pancreatitis with unrelenting pain			
Malabsorption	Cancer surveillance: family history prior polyp/cancer, colitis	Fistulas			
Biopsy radiologic abnormality		Biopsy radiologic abnormality			
Polypectomy	Palliate neoplasm	Pancreaticobiliary drainage			
Place gastrostomy	Remove foreign body	Sample bile			
Barrett's surveillance	Place stent across stenosis	Sphincter of Oddi manometry			
Palliate neoplasm					
Sample duodenal tissue/fluid					
Remove foreign body					
Endoscopic mucosal resection or ablation of dysplastic Barrett's mucosa					
Place stent across stenosis					

considered as alternatives to colonoscopy for colon cancer screening. MR methods image the pancreaticobiliary ducts to exclude neoplasm, stones, and sclerosing cholangitis, and the liver to characterize benign and malignant tumors. Specialized CT or MR enterography quantifies IBD intensity. Angiography excludes mesenteric ischemia and determines spread of malignancy. Angiographic techniques also access the biliary tree in obstructive jaundice. CT and MR techniques screen for mesenteric occlusion, thereby limiting exposure to angiographic dyes. Positron emission tomography can distinguish malignant from benign disease in several organ systems.

Scintigraphy evaluates structural abnormalities and quantifies luminal transit. Radionuclide scans localize bleeding sites in patients with brisk hemorrhage to direct therapy with endoscopy, angiography, or surgery. Radiolabeled leukocyte scans search for intraabdominal abscesses not visualized on CT. Biliary scintigraphy complements ultrasound in assessing for cholecystitis. Scintigraphy to quantify esophageal and gastric emptying is well established, whereas techniques to measure small-intestinal or colonic transit are less widely used.

Histopathology Endoscopic mucosal biopsies evaluate for inflammatory, infectious, and neoplastic disease. Deep rectal biopsies facilitate diagnosis of Hirschsprung's disease or amyloid. Liver biopsy is performed for abnormal liver chemistries, in unexplained jaundice, following liver transplant to exclude rejection, and to characterize inflammation in chronic viral hepatitis prior to initiating antiviral therapy. Biopsies obtained during CT or ultrasound evaluate for intraabdominal conditions not accessible by endoscopy.

Functional Testing Tests of gut function provide important data when structural testing is nondiagnostic. Functional testing of motor activity is provided by newer high resolution manometric techniques. Esophageal manometry is useful for suspected achalasia, whereas small-intestinal manometry tests for pseudoobstruction and colon manometry evaluates for colonic inertia. A wireless motility capsule measures transit and contractile activity in the stomach, small intestine, and colon in a single test. Anorectal manometry with balloon expulsion testing is used for unexplained incontinence or constipation from outlet dysfunction. Biliary manometry tests for sphincter of Oddi dysfunction with unexplained biliary pain. A novel endoluminal functional lumen imaging probe is available to measure heightened distensibility in the lower esophageal sphincter in achalasia and pylorus in gastroparesis. Measurement of breath hydrogen while fasting and after oral mono- or oligosaccharide challenge can screen for carbohydrate intolerance and small-intestinal bacterial overgrowth. Urea breath testing assesses for persistent *Helicobacter pylori* infection, while a recently approved gastric emptying breath test is an alternative to scintigraphy for gastroparesis diagnosis.

TREATMENT

Gastrointestinal Disease

Management options for GI diseases depend on the cause of symptoms. Available treatments include modifications in dietary intake, medications, treatment of gut dysbiosis, interventional endoscopy or radiology techniques, surgery, and therapies directed to external influences. Given the hereditary predisposition of many GI diseases, genetic testing may be indicated in some patients.

NUTRITIONAL MANIPULATION

Dietary modifications for GI disease include treatments that only reduce symptoms, therapies correct pathologic defects, or replace normal food intake with enteral or parenteral formulations. Changes that improve symptoms but do not reverse organic abnormalities include lactose restriction for lactase deficiency, liquid meals in gastroparesis, carbohydrate restrictions with dumping syndrome, and low-FODMAP (fermentable oligo-di-monosaccharides and polyols) diets in IBS. The gluten-free diet for celiac disease exemplifies a primary therapy to reduce mucosal inflammation. Likewise, elimination diets may improve histology in some cases of eosinophilic esophagitis. Medium-chain triglycerides replace normal fats in short-gut syndrome or severe ileal disease. Perfusing liquid meals through a gastrostomy is performed in those who cannot swallow safely. Enteral jejunostomy feedings are considered for gastric dysmotility syndromes that preclude feeding into the stomach. Intravenous hyperalimentation is used for generalized gut malfunction which does not permit enteral nutrition.

PHARMACOTHERAPY

Several medications can treat GI diseases. Considerable resources are expended on over-the-counter remedies. Many prescription drug classes are offered as short-term or continuous therapy of GI illness. Alternative treatments have gained popularity in conditions for which traditional therapies provide incomplete relief.

Over-the-Counter Agents Over-the-counter agents are reserved for mild GI symptoms. Antacids and histamine H_2 antagonists decrease symptoms in gastroesophageal reflux disease (GERD) and dyspepsia. Potent acid inhibitors are available over the counter for treatment of more persistent GERD. Fiber supplements, stool softeners, enemas, and laxatives are used for constipation. Laxatives are categorized as stimulants, osmotic agents (including isotonic preparations containing polyethylene glycol), and poorly absorbed sugars. Nonprescription antidiarrheal agents include bismuth subsalicylate, kaolin-pectin combinations, and loperamide. Supplemental enzymes include lactase pills for lactose intolerance and bacterial α-galactosidase to treat excess gas. Capsules containing peppermint oil are available over the counter for treating discomfort in IBS and dyspepsia, while antiflatulents and adsorbents reduce gaseous symptoms. In general, using a nonprescription preparation for more than a short time for chronic persistent symptoms should be supervised by a health care provider.

Prescription Drugs Prescription drugs are approved for a broad range of GI diseases. Higher dose prescription proton pump inhibitors are advocated for GERD when over-the-counter preparations are inadequate. Cytoprotective agents are available for upper gut ulcers but are less frequently prescribed. Prokinetic drugs stimulate GI propulsion in gastroparesis and pseudoobstruction. Prosecretory drugs are prescribed for constipation refractory to other agents, while peripheral opiate antagonists are offered for opiate-induced constipation. Prescription antidiarrheals include opiate drugs, anticholinergic antispasmodics, tricyclics, bile acid binders, and serotonin antagonists. Antispasmodics and antidepressants also are useful for functional GI disorders, whereas narcotics are used for pain control in organic conditions such as disseminated malignancy and chronic pancreatitis. Antiemetics reduce nausea and vomiting. Potent pancreatic enzymes decrease malabsorption and pain from pancreatic disease. Antisecretory drugs such as the somatostatin analogue octreotide treat hypersecretory states. Antibiotics treat *Helicobacter pylori*-induced ulcers, infectious diarrhea, diverticulitis, intestinal bacterial overgrowth, and Crohn's disease. Anti-inflammatory and immunomodulatory drugs are used in IBD, microscopic colitis, refractory celiac disease, and gut vasculitis. Over the past decade, several newer biologic agents including those with anti-tumor necrosis factor activity have had dramatic impact in Crohn's disease and ulcerative colitis. Chemotherapy with or without radiotherapy is offered for GI malignancies. Most GI carcinomas respond poorly to such therapy, whereas lymphomas may be cured with such intervention.

Complementary and Alternative Medicine Treatments Alternative treatments are marketed to treat selected GI symptoms. Ginger, acupressure, and acustimulation have been advocated for nausea, whereas pyridoxine has been investigated for nausea of first-trimester pregnancy. Herbal preparations like STW 5 (Iberogast, a mixture of nine herbs) are useful in cases of functional dyspepsia and IBS. Low-potency pancreatic enzyme preparations are sold as general digestive aids but have little evidence to support their efficacy.

THERAPIES TARGETING GUT DYSBIOSIS

Some cases of diarrhea predominant-IBS respond to nonabsorbable antibiotics. Oral antibiotics also are the mainstay of managing intestinal bacterial overgrowth. Probiotics containing active bacterial cultures and prebiotics that selectively nourish non-noxious commensal bacteria are used as adjuncts in some cases of infectious diarrhea and IBS. Transplantation of donor feces into the colon by colonoscopy or enema has become accepted and effective treatment for recurrent, refractory *Clostridium difficile* colitis.

INTERVENTIONAL ENDOSCOPY AND RADIOLOGY

Gut luminal intubations are performed in some situations. Nasogastric tube suction decompresses the upper gut in ileus or mechanical obstruction. Nasogastric lavage of saline or water in the patient with upper GI hemorrhage determines the rate of bleeding and helps evacuate blood before therapeutic endoscopy. Enteral feedings can be delivered through nasogastric or nasoenteric tubes. Enemas relieve fecal impaction or assist in gas evacuation in acute colonic pseudoobstruction. A rectal tube can be placed to vent the distal colon in colonic pseudoobstruction and other colonic distention disorders.

In addition to its diagnostic role, endoscopy has therapeutic capabilities in many settings. Cautery techniques and injection of vasoconstrictor substances can stop hemorrhage from ulcers and vascular malformations. Endoscopic encirclement of varices and hemorrhoids with constricting bands stops hemorrhage from these sites, whereas endoscopically placed clips can occlude arterial bleeding sites. Cyanoacrylate and hemostatic powder sprays have been evaluated for abilities to stop brisk GI bleeding. Endoscopy can remove polyps or debulk lumen-narrowing malignancies. Colonoscopy is used to withdraw luminal gas in some cases of acute colonic pseudoobstruction. Endoscopic mucosal resection, submucosal dissection, and radiofrequency techniques can ablate some cases of Barrett's esophagus with dysplasia or superficial cancer and early gastric malignancies. Obstructions of the gut lumen and pancreaticobiliary tree are relieved by endoscopic dilation or placing plastic or expandable metal stents. Endoscopic sphincterotomy of the ampulla of Vater relieves symptoms of choledocholithiasis. Cholangioscopy can help with stone lithotripsy in the common bile duct, ablation of small ductal tumors, and placement of gallbladder stents to facilitate drainage in non-operative candidates. Endoscopic methods have been developed for pancreatic cyst gastrostomy, pancreatic necrosectomy, and placement of fiducial markers to direct pancreatic and rectal radiotherapy. Endoscopy is commonly used to insert gastric feeding tubes. Peroral endoscopic myotomy is now being performed on the lower esophageal sphincter in achalasia and on the pylorus in gastroparesis by selected endoscopists. Endoscopic treatments for acid reflux including radiofrequency therapy, transoral fundoplication, endoscopic stapling, and antireflux mucosectomy have been devised. Similarly, endoscopic bariatric methodologies including intragastric balloons, aspiration therapy, gastroplasty, and duodenal bypass are in use or in development.

Radiologic measures also are useful in GI disease. Angiographic embolization or vasoconstriction decreases bleeding from gut sites not amenable to endoscopic intervention. Dilatation or stenting with fluoroscopic guidance relieves luminal strictures. Contrast enemas can reduce volvulus and evacuate air in acute colonic pseudoobstruction. CT and ultrasound help drain abdominal fluid collections, in many cases obviating the need for surgery. Percutaneous transhepatic cholangiography relieves biliary obstruction when ERCP is contraindicated. Transjugular intrahepatic portosystemic shunts are commonly performed by interventional radiologists for variceal hemorrhage not amenable to endoscopic therapy. Lithotripsy can fragment gallstones in patients who are not candidates for surgery. In some instances, radiologic approaches offer advantages over endoscopy for gastroenterostomy placement. Finally, central venous catheters for parenteral nutrition may be placed using radiographic techniques.

SURGERY

Surgery is performed to cure disease, control symptoms without cure, maintain nutrition, or palliate unresectable neoplasm. Medication-unresponsive ulcerative colitis, diverticulitis, cholecystitis, appendicitis, and intraabdominal abscess are curable with surgery, whereas symptom control without cure is only possible with Crohn's disease. Surgery is mandated for ulcer complications such as bleeding, obstruction, or perforation and intestinal obstructions that persist after conservative care. Fundoplication of the gastroesophageal junction is performed for severe ulcerative esophagitis and drug-refractory symptomatic acid reflux. Achalasia responds to operations to reduce lower esophageal sphincter tone. Operations for motor disorders have been introduced including implanted electrical stimulators for gastroparesis and electrical devices and artificial sphincters for fecal incontinence. Surgery may be needed to place a jejunostomy for long-term enteral feedings. The threshold for performing surgery depends on the clinical setting. In all cases, the benefits of operation must be weighed against the potential for postoperative complications.

THERAPY DIRECTED TO EXTERNAL INFLUENCES

In some conditions, GI symptoms respond to treatments directed outside the gut. Psychological therapies including psychotherapy, behavior modification, and hypnosis, have shown efficacy in functional bowel disorders. Patients with significant psychological dysfunction and those with little response to treatments targeting the gut are likely to benefit from this form of therapy. Biofeedback methods administered by physical therapies are accepted for treating refractory fecal incontinence or constipation secondary to dyssynergia.

FURTHER READING

Beg S et al: The use of optical imaging techniques in the gastrointestinal tract. Frontline Gastroenterol 7:207, 2016.

Goodman RP, Chung DC: Clinical genetic testing in gastroenterology. Clin Transl Gastroenterol 7:e167, 2016.

Panes J et al: Advances in use of endoscopy, radiology, and biomarkers to monitor inflammatory bowel diseases. Gastroenterology 152:362, 2017.

Peery AF et al: Burden of gastrointestinal, liver, and pancreatic diseases in the United States. Gastroenterology 149:1731, 2015.

Schreuders EH et al: Advances in fecal tests for colorectal cancer screening. Curr Treat Options Gastroenterol 14:152, 2016.

Weilert F, Binmoeller KF: New endoscopic technologies and procedural advances for endoscopic hemostasis. Clin Gastroenterol Hepatol 14:1234, 2016.

315 Gastrointestinal Endoscopy

Louis Michel Wong Kee Song, Mark Topazian

Gastrointestinal endoscopy has been attempted for over 200 years, but the introduction of semirigid gastroscopes in the middle of the twentieth century marked the dawn of the modern endoscopic era. Since then rapid advances in endoscopic technology have led to dramatic changes in the diagnosis and treatment of many digestive diseases. Innovative endoscopic devices and new endoscopic treatment modalities continue to expand the use of endoscopy in patient care.

Flexible endoscopes provide an electronic video image generated by a charge-coupled device in the tip of the endoscope. Operator controls permit deflection of the endoscope tip; fiberoptic bundles or light-emitting diodes provide light at the tip of the endoscope; and working channels allow washing, suctioning, and the passage of instruments **(Fig. 315-1)**. Progressive changes in the diameter and stiffness of endoscopes have improved the ease and patient tolerance of endoscopy.

FIGURE 315-1 Gastrointestinal endoscope. Shown here is a conventional colonoscope with control knobs for tip deflection, push buttons for suction and air insufflation (*single arrows*), and a working channel for passage of accessories *(double arrows)*.

ENDOSCOPIC PROCEDURES

UPPER ENDOSCOPY

Upper endoscopy, also referred to as esophagogastroduodenoscopy (EGD), is performed by passing a flexible endoscope through the mouth into the esophagus, stomach, and duodenum. The procedure is the best method for examining the upper gastrointestinal mucosa **(Fig. 315-2)**. While the upper gastrointestinal radiographic series has similar accuracy for diagnosis of duodenal ulcer **(Fig. 315-3)**, EGD is superior for detection of gastric ulcers **(Fig. 315-4)** and flat mucosal lesions, such as Barrett's esophagus **(Fig. 315-5)**, and it permits directed biopsy and endoscopic therapy. Intravenous conscious sedation is given to most patients in the United States to ease the anxiety and discomfort of the procedure, although in many countries EGD is routinely performed with topical pharyngeal anesthesia only. Patient tolerance of unsedated EGD is improved by the use of an ultrathin, 5-mm diameter endoscope that can be passed transorally or transnasally.

COLONOSCOPY

Colonoscopy is performed by passing a flexible colonoscope through the anal canal into the rectum and colon. The cecum is reached in >95% of cases and the terminal ileum **(Fig. 315-6)** can often be examined. Colonoscopy is the gold standard for imaging the colonic mucosa **(Fig. 315-7)**. Colonoscopy has greater sensitivity than barium enema for colitis **(Fig. 315-8)**, polyps **(Fig. 315-9)**, and cancer **(Fig. 315-10)**. CT colonography rivals the accuracy of colonoscopy for detection of some polyps and cancer, although it is not as sensitive for the detection of flat lesions, such as serrated polyps **(Fig. 315-11)**. Conscious sedation is usually given before colonoscopy in the United States, although a willing patient and a skilled examiner can complete the procedure without sedation in many cases.

FLEXIBLE SIGMOIDOSCOPY

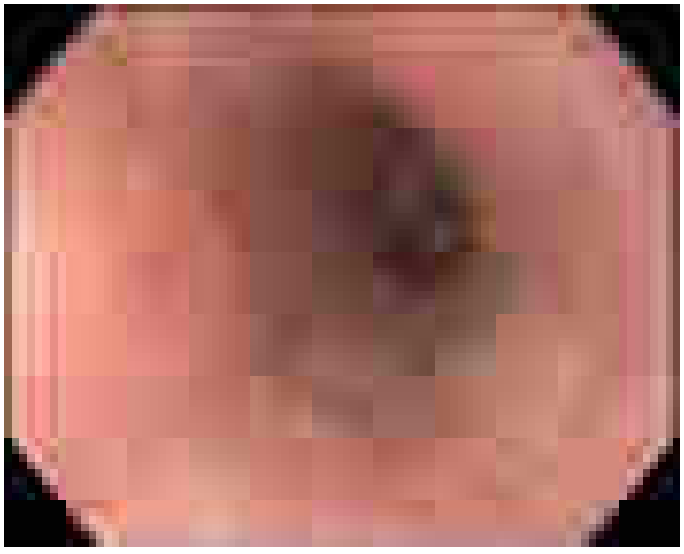
Flexible sigmoidoscopy is similar to colonoscopy, but it visualizes only the rectum and a variable portion of the left colon, typically to 60 cm from the anal verge. This procedure causes abdominal cramping, but it is brief and is usually performed without sedation. Flexible sigmoidoscopy is primarily used for evaluation of diarrhea and rectal outlet bleeding.

SMALL BOWEL ENDOSCOPY

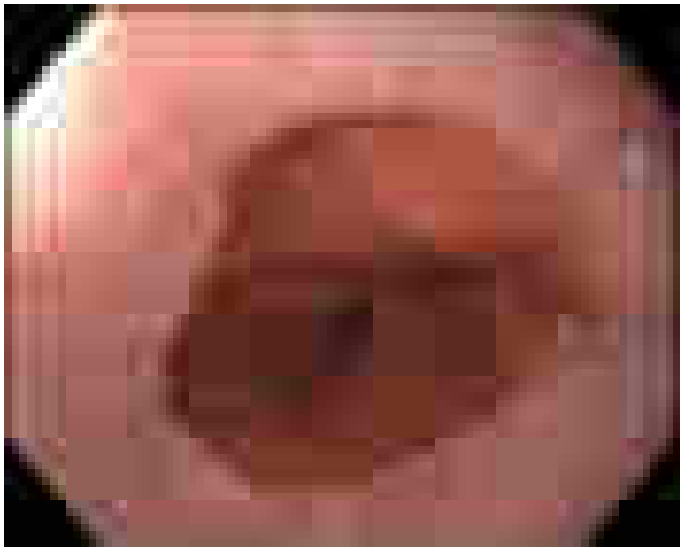
Three endoscopic techniques are currently used to evaluate the small intestine, most often in patients presenting with presumed small bowel bleeding. For *capsule endoscopy*, the patient swallows a disposable capsule that contains a complementary metal oxide silicon (CMOS) chip camera. Color still images **(Fig. 315-12)** are transmitted wirelessly to an external receiver at several frames per second until the capsule's battery is exhausted or it is passed into the toilet. Capsule endoscopy enables visualization of the small bowel mucosa beyond the reach of a conventional endoscope, and at present it is solely a diagnostic procedure. Patients with a history of prior intestinal surgery or Crohn's disease are at risk for capsule retention at the site of a clinically unsuspected small bowel stricture, and ingestion of a "patency capsule" composed of radiologically opaque biodegradable material may be indicated prior to capsule endoscopy in such patients.

Push enteroscopy is performed with a long endoscope similar in design to an upper endoscope. The enteroscope is pushed down the small bowel, sometimes with the help of a stiffening overtube that extends from the mouth to the small intestine. The proximal to mid-jejunum is usually reached, and the instrument channel of the endoscope allows for biopsy or endoscopic therapy.

Deeper insertion into the small bowel can be accomplished by *device-assisted enteroscopy,* which may utilize inflatable balloons at the tip of the enteroscope and/or an overtube (*single-* or *double-balloon enteroscopy*) or a rotating, screw-like overtube (*spiral enteroscopy*) to pleat the small intestine onto the endoscope **(Fig. 315-13, Video V5-1)**. Using device-assisted enteroscopy the entire small intestine can be visualized

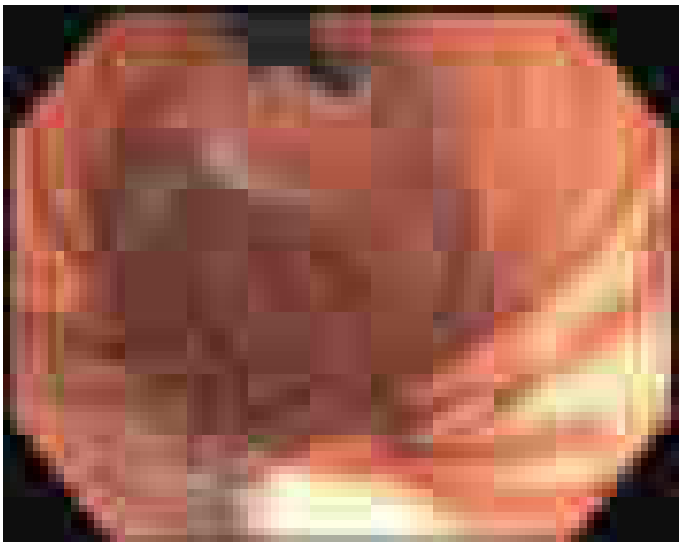


A

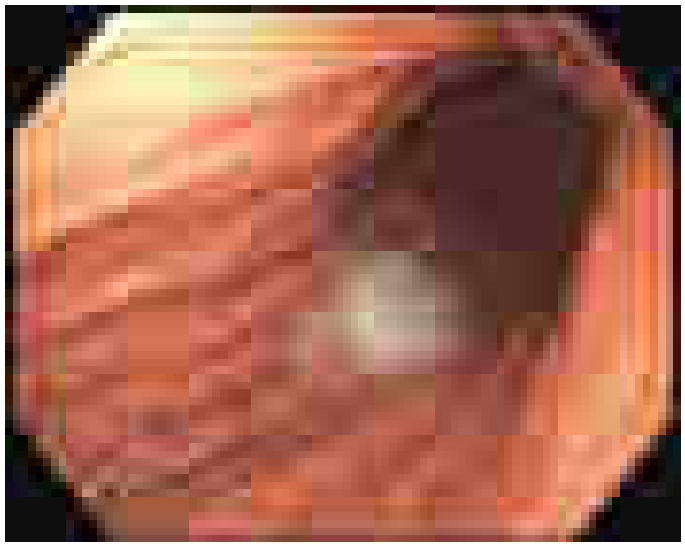


B

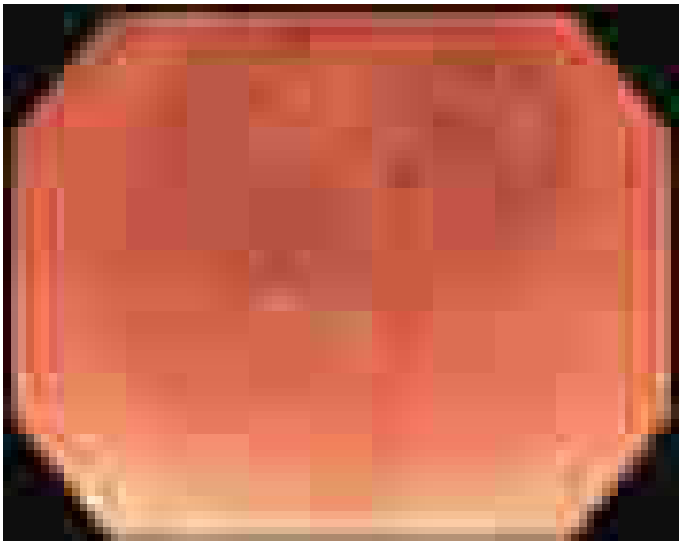
FIGURE 315-2 Normal upper endoscopic examination. *A.* Esophagus. ***B.*** Gastroesophageal junction. ***C.*** Gastric fundus. ***D.*** Gastric body. ***E.*** Gastric antrum. ***F.*** Pylorus. ***G.*** Duodenal bulb. ***H.*** Second portion of the duodenum.



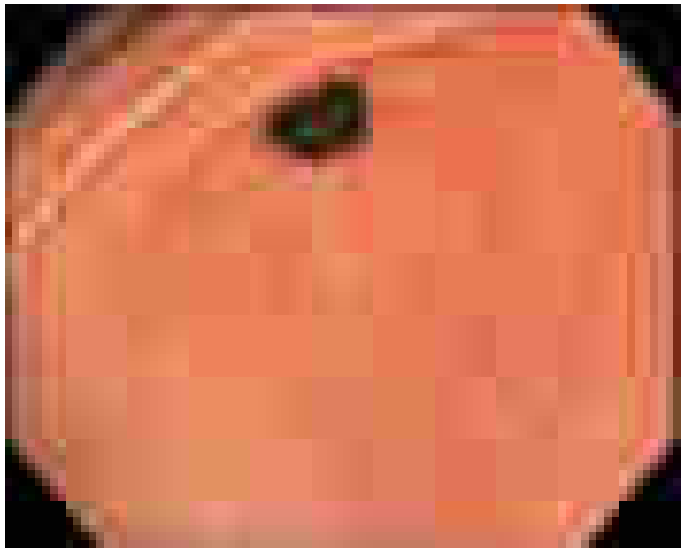
C



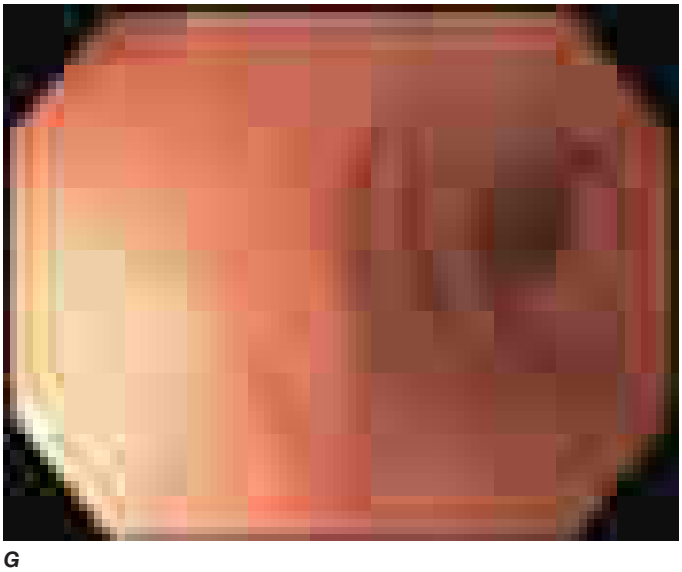
D



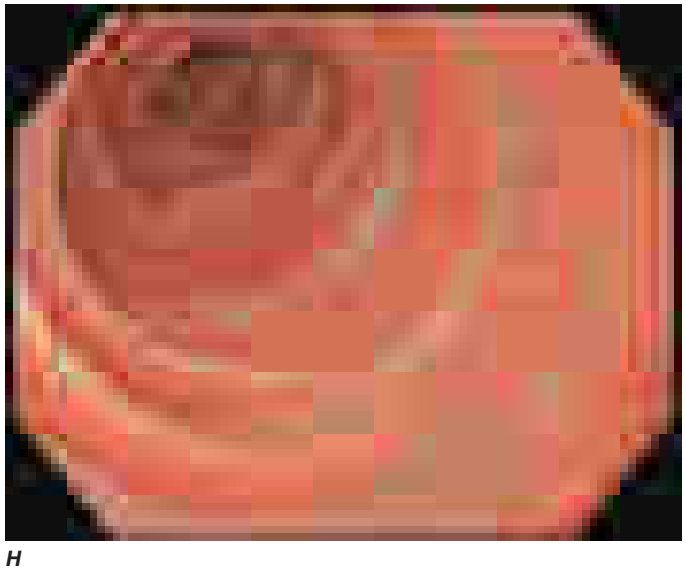
E



F



G



H

FIGURE 315-2 (*Continued*)

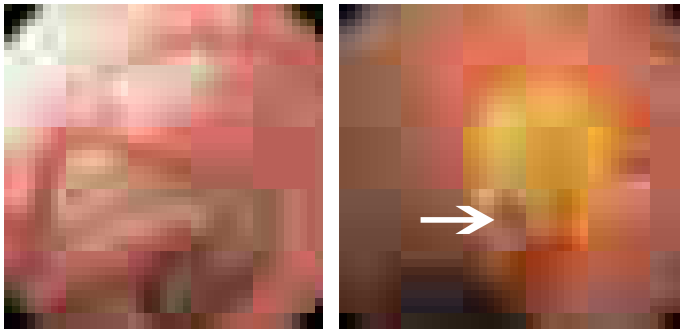


FIGURE 315-3 Duodenal ulcers. *A.* Ulcer with a clean base. ***B.*** Ulcer with a visible vessel *(arrow)* in a patient with recent hemorrhage.

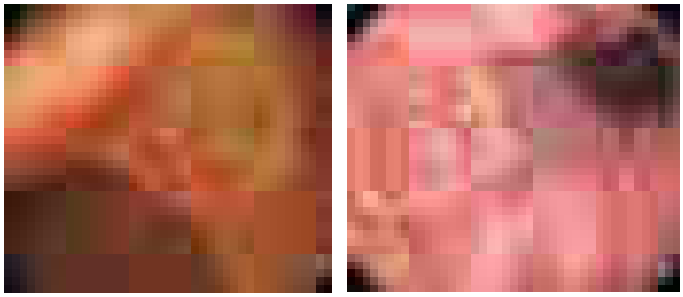


FIGURE 315-4 Gastric ulcers. *A.* Benign gastric ulcer. ***B.*** Malignant gastric ulcer involving greater curvature of stomach.

in some patients when both the oral and anal routes of insertion are used. Biopsies and endoscopic therapy can be performed throughout the visualized small bowel **(Fig. 315-14)**.

ENDOSCOPIC RETROGRADE CHOLANGIOPANCREATOGRAPHY (ERCP)

During ERCP a side-viewing endoscope is passed through the mouth to the duodenum, the ampulla of Vater is identified and cannulated with a thin plastic catheter, and radiographic contrast material is injected into the bile duct and pancreatic duct under fluoroscopic guidance **(Fig. 315-15)**. When indicated, the major papilla can be opened using the technique of endoscopic sphincterotomy **(Fig. 315-16)**. Stones can be retrieved from the ducts, biopsies can be performed, and strictures can be dilated and/or stented **(Fig. 315-17)**, and ductal leaks can be treated **(Fig. 315-18)**. ERCP is usually performed for therapy but is also important diagnostically, and facilitates tissue sampling of biliary or pancreatic ductal strictures.

ENDOSCOPIC ULTRASOUND (EUS)

EUS utilizes ultrasound transducers incorporated into the tip of a flexible endoscope. Ultrasound images are obtained of the gut wall and adjacent organs, vessels, lymph nodes, and other structures. High-resolution images are obtained by bringing a high-frequency ultrasound transducer close to the area of interest via endoscopy. EUS provides the most accurate preoperative local staging of esophageal, pancreatic, and rectal malignancies **(Fig. 315-19)**, but it does not detect most distant metastases. EUS is also useful for diagnosis of bile duct stones, gallbladder disease, subepithelial gastrointestinal lesions, and chronic pancreatitis. Fine-needle aspirates and core biopsies of organs, masses and lymph nodes in the posterior mediastinum, abdomen, pancreas, retroperitoneum, and pelvis can be obtained under EUS guidance **(Fig. 315-20)**. EUS-guided therapeutic procedures are increasingly performed, including drainage of abscesses, pseudocysts, and pancreatic necrosis into the gut lumen **(Video V5-2)**, celiac plexus neurolysis for treatment of pancreatic pain, ethanol ablation of pancreatic neuroendocrine tumors, treatment of gastrointestinal hemorrhage, and drainage of obstructed biliary and pancreatic ducts.

NATURAL ORIFICE TRANSLUMINAL ENDOSCOPIC SURGERY (NOTES)

NOTES is an evolving collection of endoscopic methods that entail passage of an endoscope or its accessories into or through the wall of the gastrointestinal tract to perform diagnostic or therapeutic interventions. Some NOTES procedures, such as percutaneous endoscopic gastrostomy (PEG) or endoscopic necrosectomy of pancreatic necrosis, are well-established clinical procedures **(Video V5-2)**; others such as peroral endoscopic myotomy (POEM) for achalasia **(Fig. 315-21)**, peroral endoscopic tumorectomy (POET) **(Fig. 315-22)**, and endoscopic full-thickness resection (EFTR) of gastrointestinal mural lesions **(Fig. 315-23, Video V5-3)**, are emerging as minimally invasive therapeutic options. NOTES is an area of continuing innovation and endoscopic research.

ENDOSCOPIC RESECTION AND CLOSURE TECHNIQUES

Endoscopic mucosal resection (EMR) **(Video V5-4)** and endoscopic submucosal dissection (ESD) **(Fig. 315-24, Video V5-5)** are the two commonly used techniques for the resection of benign and early-stage malignant gastrointestinal neoplasms. In addition to providing larger specimens for more accurate histopathological assessment and diagnosis, these techniques can be potentially curative for certain dysplastic lesions and focal intramucosal carcinomas involving the esophagus, stomach, and colon. Several devices are available for closure of EMR

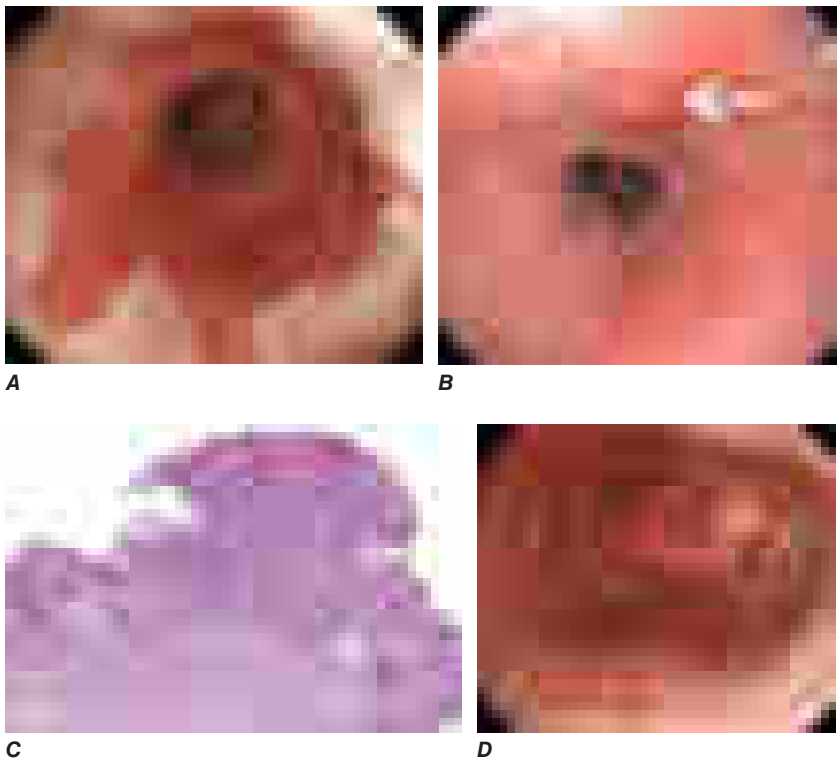
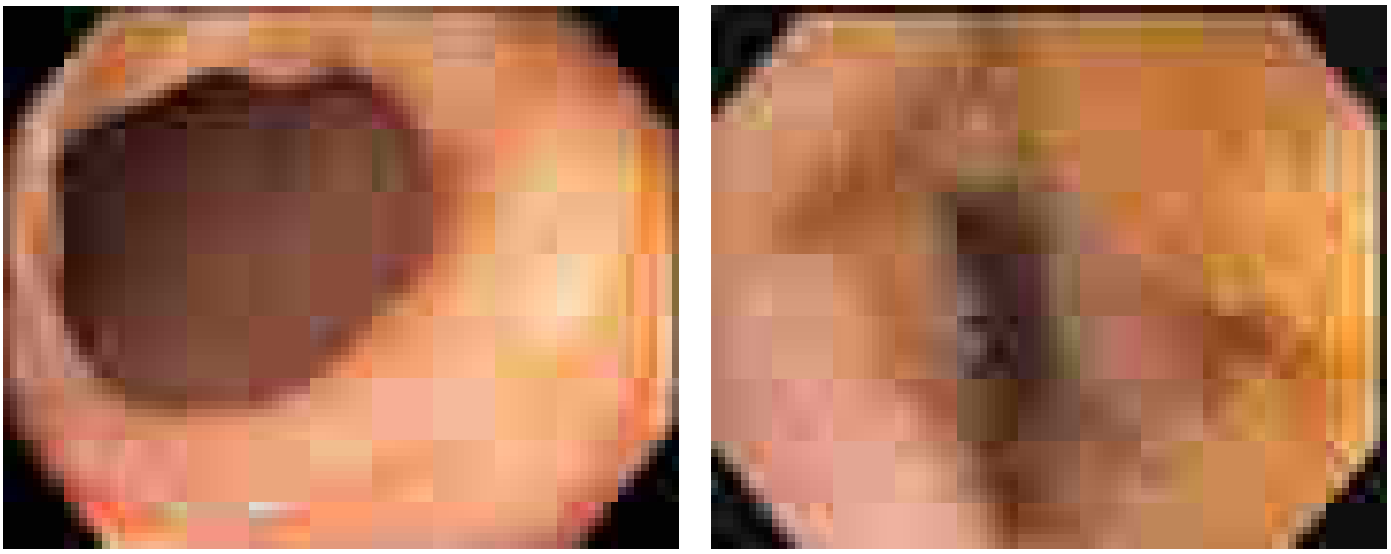
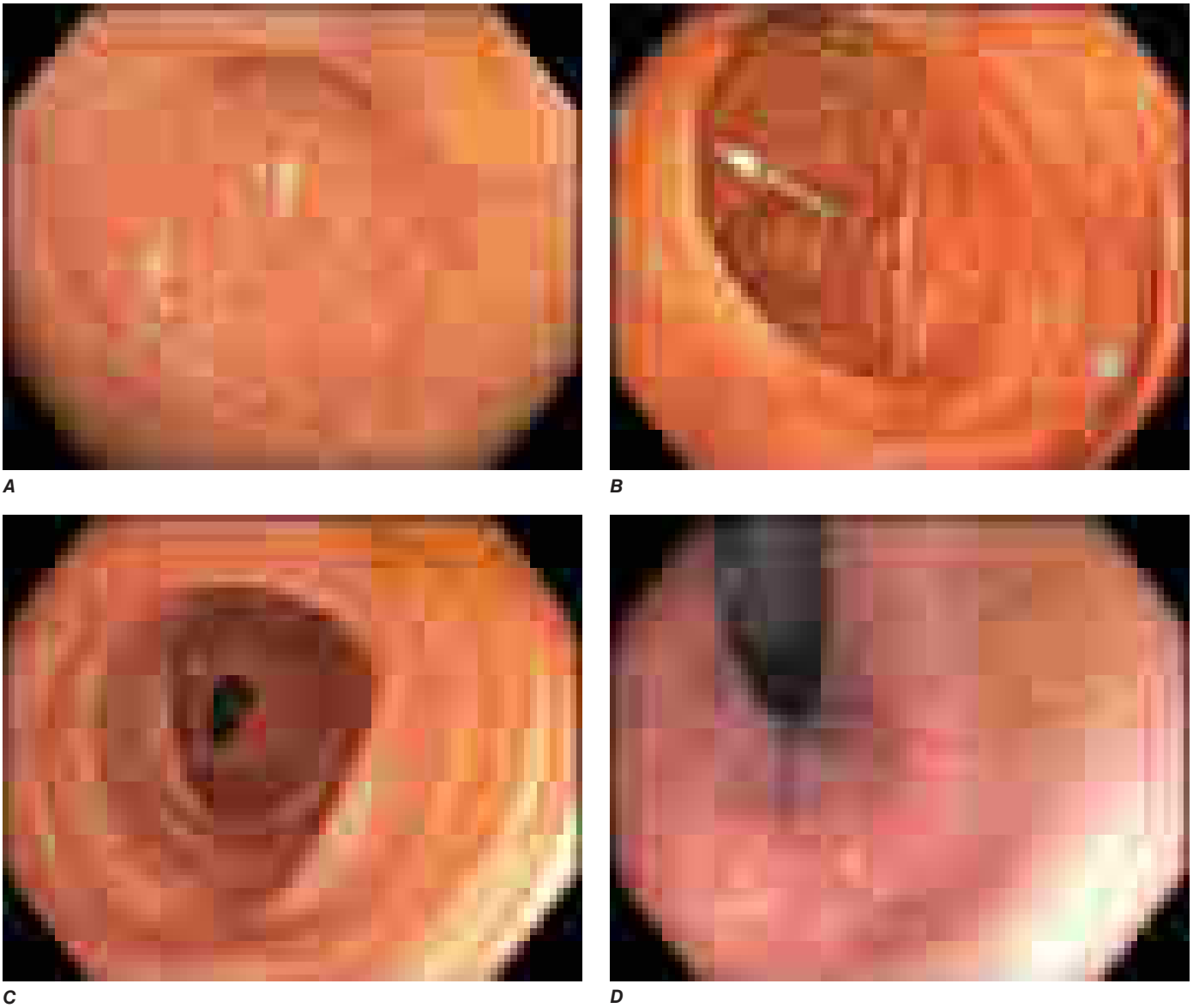


FIGURE 315-5 Barrett's esophagus. *A.* Pink tongues of Barrett's mucosa extending proximally from the gastroesophageal junction. ***B.*** Barrett's esophagus with a suspicious nodule *(arrow)* identified during endoscopic surveillance. ***C.*** Histologic finding of intramucosal adenocarcinoma in the endoscopically resected nodule. Tumor extends into the esophageal submucosa *(arrow)*. ***D.*** Barrett's esophagus with locally advanced adenocarcinoma.



A *B*

FIGURE 315-6 Colonoscopic view of terminal ileum. ***A.*** Normal-appearing terminal ileum (TI). ***B.*** View of normal villi of TI enhanced by examination under water immersion.



A *B*

C *D*

FIGURE 315-7 Normal colonoscopic examination. ***A.*** Cecum with view of appendiceal orifice. ***B.*** Ileocecal valve. ***C.*** Normal-appearing colon. ***D.*** Rectum (retroflexed view).

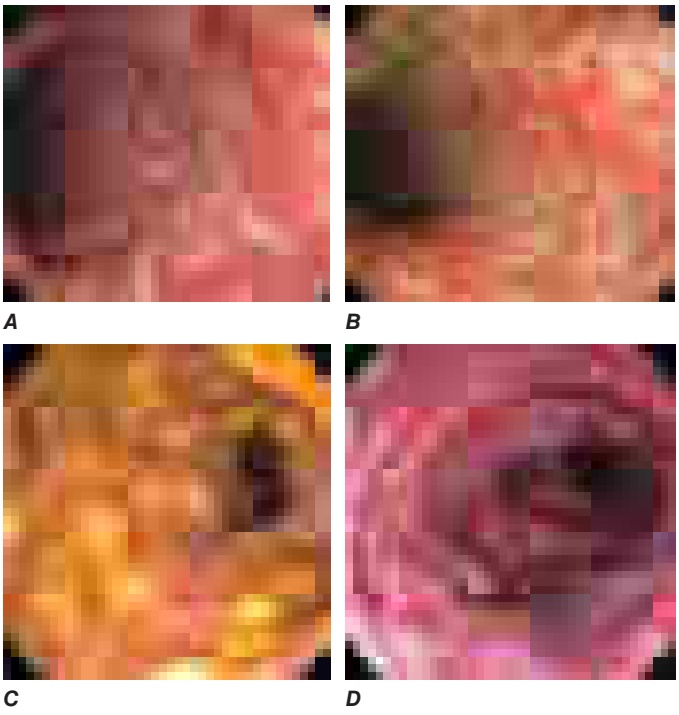
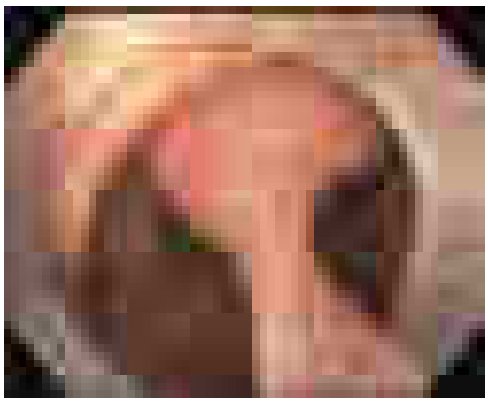


FIGURE 315-8 Causes of colitis. ***A.*** Chronic ulcerative colitis with diffuse ulcerations and exudates. ***B.*** Severe Crohn's colitis with deep ulcers. ***C.*** Pseudomembranous colitis with yellow, adherent pseudomembranes. ***D.*** Ischemic colitis with patchy mucosal edema, subepithelial hemorrhage, and cyanosis.



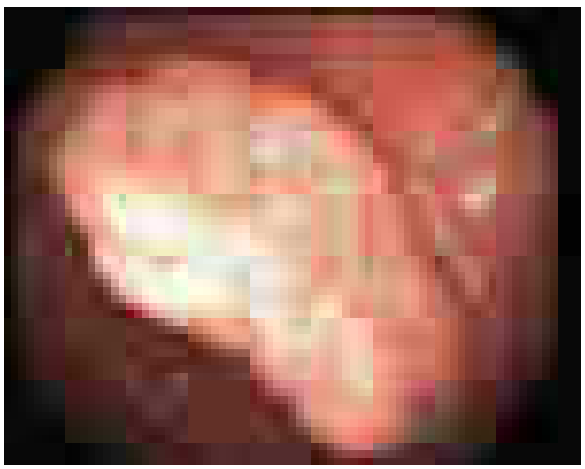


FIGURE 315-9 Colonic polyps. ***A.*** Pedunculated polyp on a stalk. ***B.*** Sessile polyp.

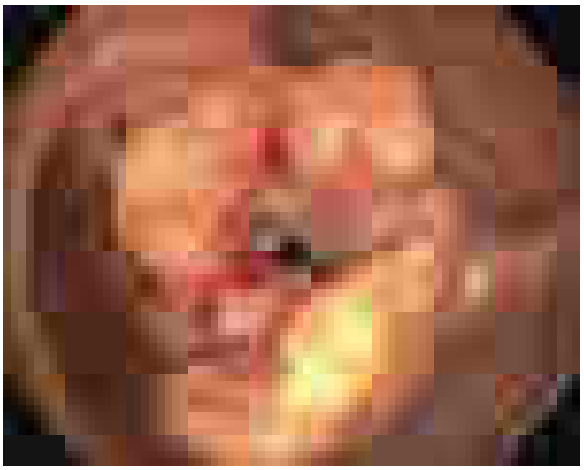
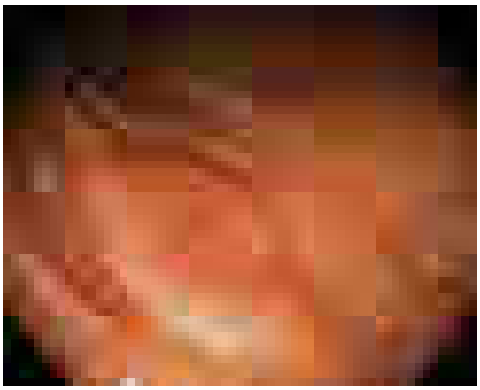
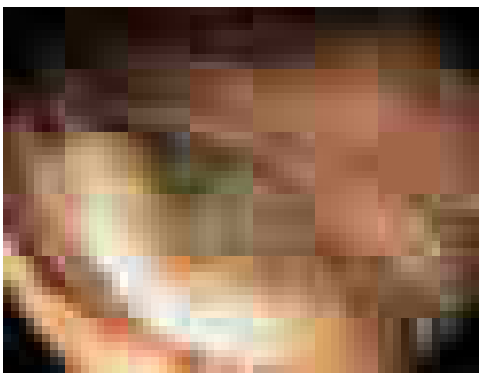


FIGURE 315-10 Ulcerated colon adenocarcinoma narrowing the colonic lumen.





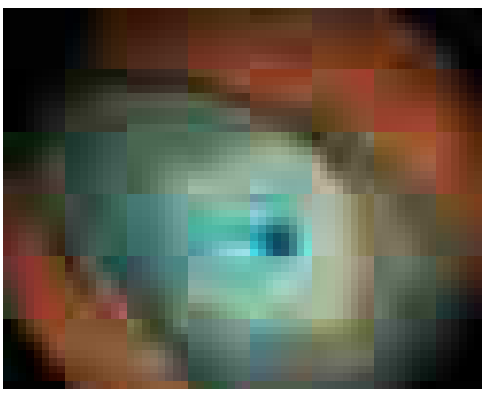


FIGURE 315-11 Flat serrated polyp in the cecum. ***A.*** Appearance of the lesion under conventional white-light imaging. ***B.*** Mucosal patterns and boundary of the lesion enhanced with narrow band imaging. ***C.*** Submucosal lifting of the lesion with dye (methylene blue) injection prior to resection.

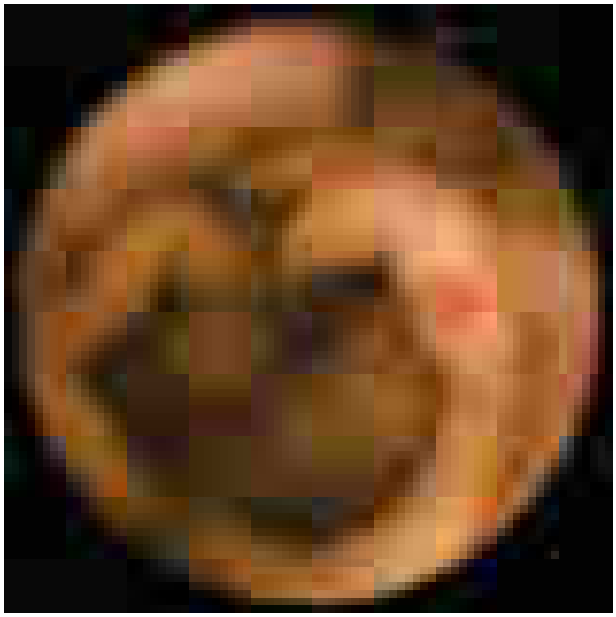


FIGURE 315-12 Capsule endoscopy image of jejunal vascular ectasia.

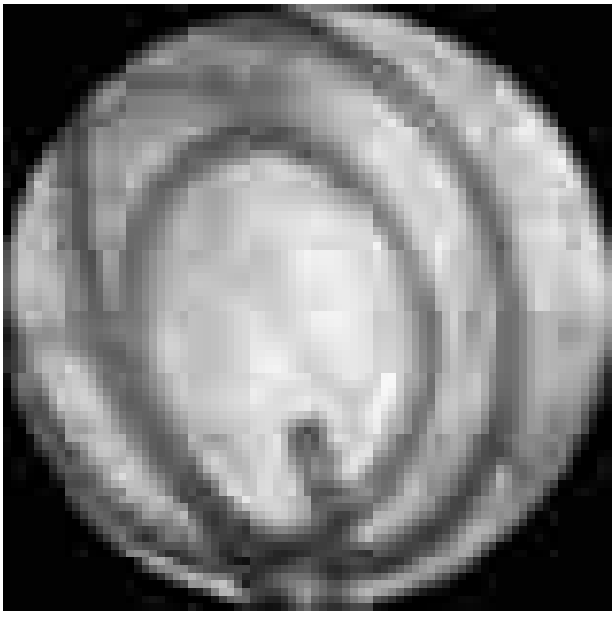
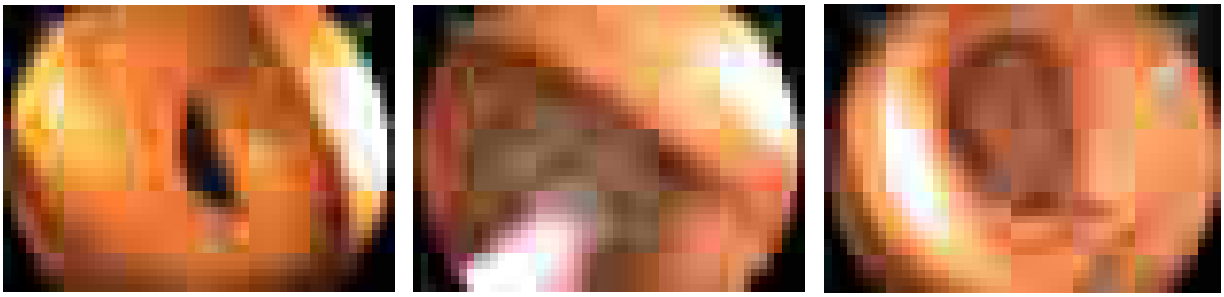
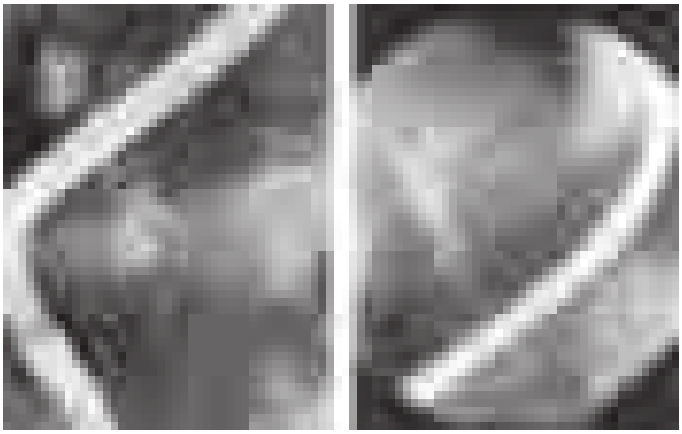


FIGURE 315-13 Radiograph of a double-balloon enteroscope in the small intestine.



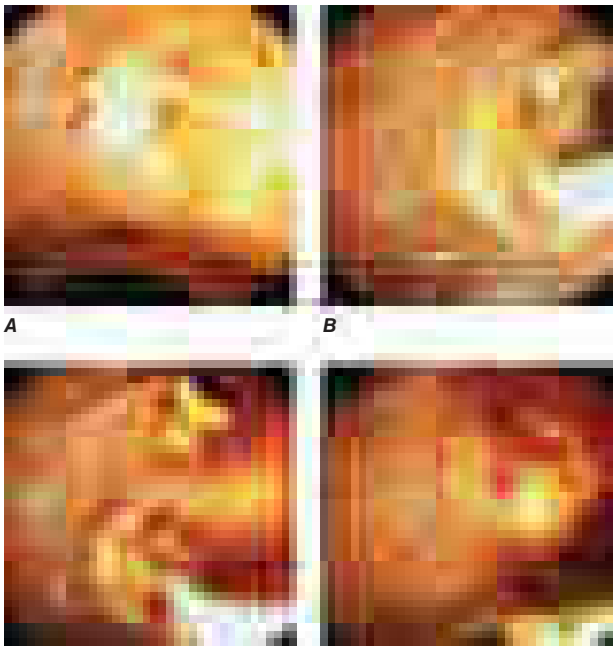
A ***B*** ***C***

FIGURE 315-14 Nonsteroidal anti-inflammatory drug (NSAID)-induced proximal ileal stricture diagnosed by double-balloon endoscopy. *A.* Ileal stricture causing obstructive symptoms. ***B.*** Balloon dilation of the ileal stricture. ***C.*** Appearance of stricture after dilation.



A ***B***

FIGURE 315-15 Endoscopic retrograde cholangiopancreatography (ERCP) for bile duct stones with cholangitis. *A.* Faceted bile duct stones are demonstrated in the common bile duct. ***B.*** After endoscopic sphincterotomy, the stones are extracted with a Dormia basket. A small abscess communicates with the left hepatic duct.



A ***B*** ***C*** ***D***

FIGURE 315-16 Endoscopic sphincterotomy. *A.* A normal-appearing ampulla of Vater. ***B.*** Sphincterotomy is performed with electrosurgery. ***C.*** Bile duct stones are extracted with a balloon catheter. ***D.*** Final appearance of the sphincterotomy.

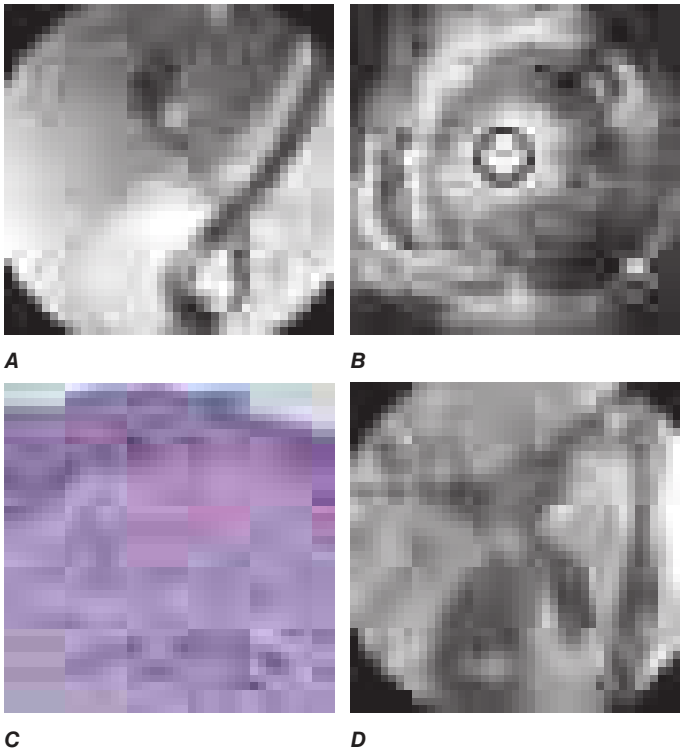


FIGURE 315-17 Endoscopic diagnosis, staging, and palliation of hilar cholangiocarcinoma. ***A.*** Endoscopic retrograde cholangiopancreatography (ERCP) in a patient with obstructive jaundice demonstrates a malignant-appearing stricture of the biliary confluence extending into the left and right intrahepatic ducts. ***B.*** Intraductal ultrasound of the biliary stricture demonstrates marked bile duct wall thickening due to tumor (T) with partial encasement of the hepatic artery *(arrow)*. ***C.*** Intraductal biopsy obtained during ERCP demonstrates malignant cells infiltrating the submucosa of the bile duct wall *(arrow)*. ***D.*** Endoscopic placement of bilateral self-expanding metal stents *(arrow)* relieves the biliary obstruction. GB, gallbladder. *(Image C courtesy of Dr. Thomas Smyrk; with permission.)*

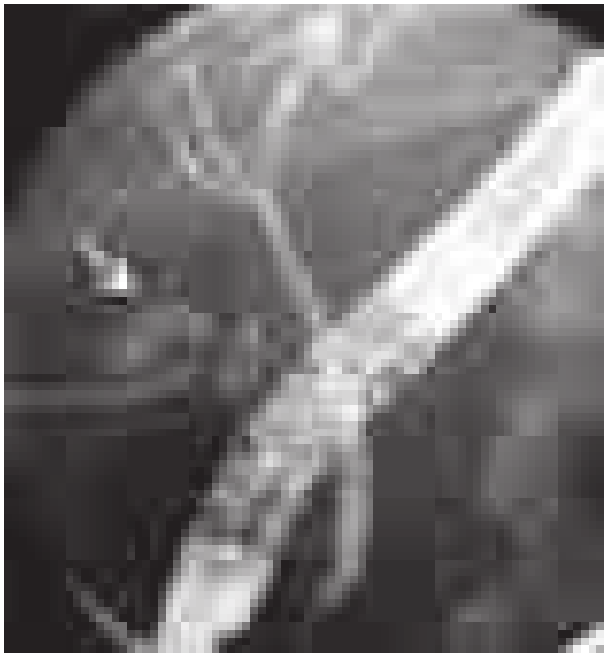


FIGURE 315-18 Bile leak *(arrow)* from a duct of Luschka after laparoscopic cholecystectomy. Contrast leaks from a small right intrahepatic duct into the gallbladder fossa, then flows into the pigtail of a percutaneous drainage catheter.

and ESD defects as well as gastrointestinal fistulas and perforations. Endoscopic clips deployed through the working channel of an endoscope have been used for many years to treat bleeding lesions, and the development of larger over-the-scope clips has facilitated endoscopic closure of gastrointestinal fistulas and perforations not previously amenable to endoscopic therapy **(Video V5-6)**. Endoscopic suturing can be used to close perforations and large defects **(Fig. 315-25)**, anastomotic leaks, and fistulas. Other potential indications for endoscopic suturing include stent fixation to prevent migration **(Fig. 315-26, Video V5-7)** and endoscopic bariatric procedures. These technologies are likely to have an expanding role in patient care.

RISKS OF ENDOSCOPY

Medications used during conscious sedation may cause respiratory depression or allergic reactions. All endoscopic procedures carry some risk of bleeding and gastrointestinal perforation. The risk is small with diagnostic upper endoscopy, flexible sigmoidoscopy, and colonoscopy (<1:1000 procedures), but it ranges from 0.5 to 5% when therapeutic procedures such as polypectomy, EMR, ESD, control of hemorrhage, or stricture dilation are performed. The risk of adverse events for diagnostic EUS (without needle aspiration) is similar to that for diagnostic upper endoscopy.

Infectious complications are uncommon with most endoscopic procedures. Some procedures carry a higher incidence of postprocedure bacteremia, and prophylactic antibiotics may be indicated **(Table 315-1)**. Management of antithrombotic agents prior to endoscopic procedures should take into account the procedural risk of hemorrhage, the agent, and the patient condition, as summarized in **Table 315-2**.

ERCP carries additional risks. Pancreatitis occurs in ~5% of patients undergoing the procedure and in up to 30% of patients with sphincter of Oddi dysfunction. Young anicteric patients with normal ducts are at increased risk. Post-ERCP pancreatitis is usually mild and self-limited, but it may result in prolonged hospitalization, surgery, diabetes, or

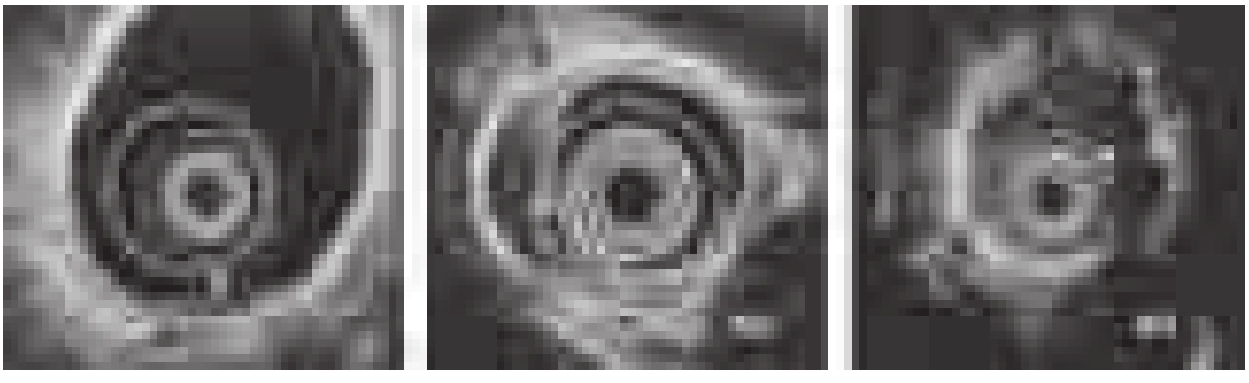


FIGURE 315-19 Local staging of gastrointestinal cancers with endoscopic ultrasound. In each example the white arrowhead marks the primary tumor and the black arrow indicates the muscularis propria (mp) of the intestinal wall. ***A.*** T1 gastric cancer. The tumor does not invade the mp. ***B.*** T2 esophageal cancer. The tumor invades the mp. ***C.*** T3 esophageal cancer. The tumor extends through the mp into the surrounding tissue, and focally abuts the aorta. AO, aorta.

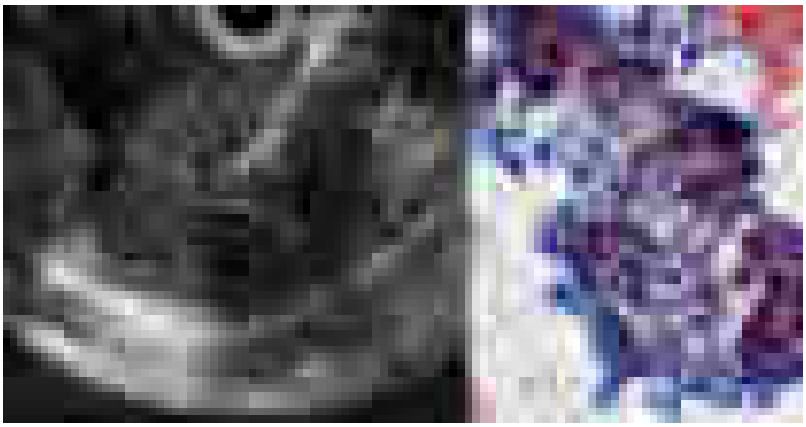


FIGURE 315-20 Endoscopic ultrasound (EUS)-guided fine-needle aspiration (FNA). ***A.*** Ultrasound image of a 22-gauge needle passed through the duodenal wall and positioned in a hypoechoic pancreatic head mass. ***B.*** Micrograph of aspirated malignant cells. *(Image B courtesy of Dr. Michael R. Henry; with permission.)*

death when severe. Significant bleeding occurs after endoscopic sphincterotomy in ~1% of cases. Ascending cholangitis, pseudocyst infection, duodenal perforation, and abscess formation may occur as a result of ERCP.

Percutaneous gastrostomy tube placement during EGD is associated with a 10–15% incidence of adverse events, most often wound infections. Fasciitis, pneumonia, bleeding (Fig. 315-27), buried bumper syndrome, and colonic injury may result from gastrostomy tube placement.

URGENT ENDOSCOPY

ACUTE GASTROINTESTINAL HEMORRHAGE

Endoscopy is the primary diagnostic and therapeutic technique for patients with acute gastrointestinal hemorrhage. Although gastrointestinal bleeding stops spontaneously in most cases, some patients will have persistent or recurrent hemorrhage that may be life-threatening.

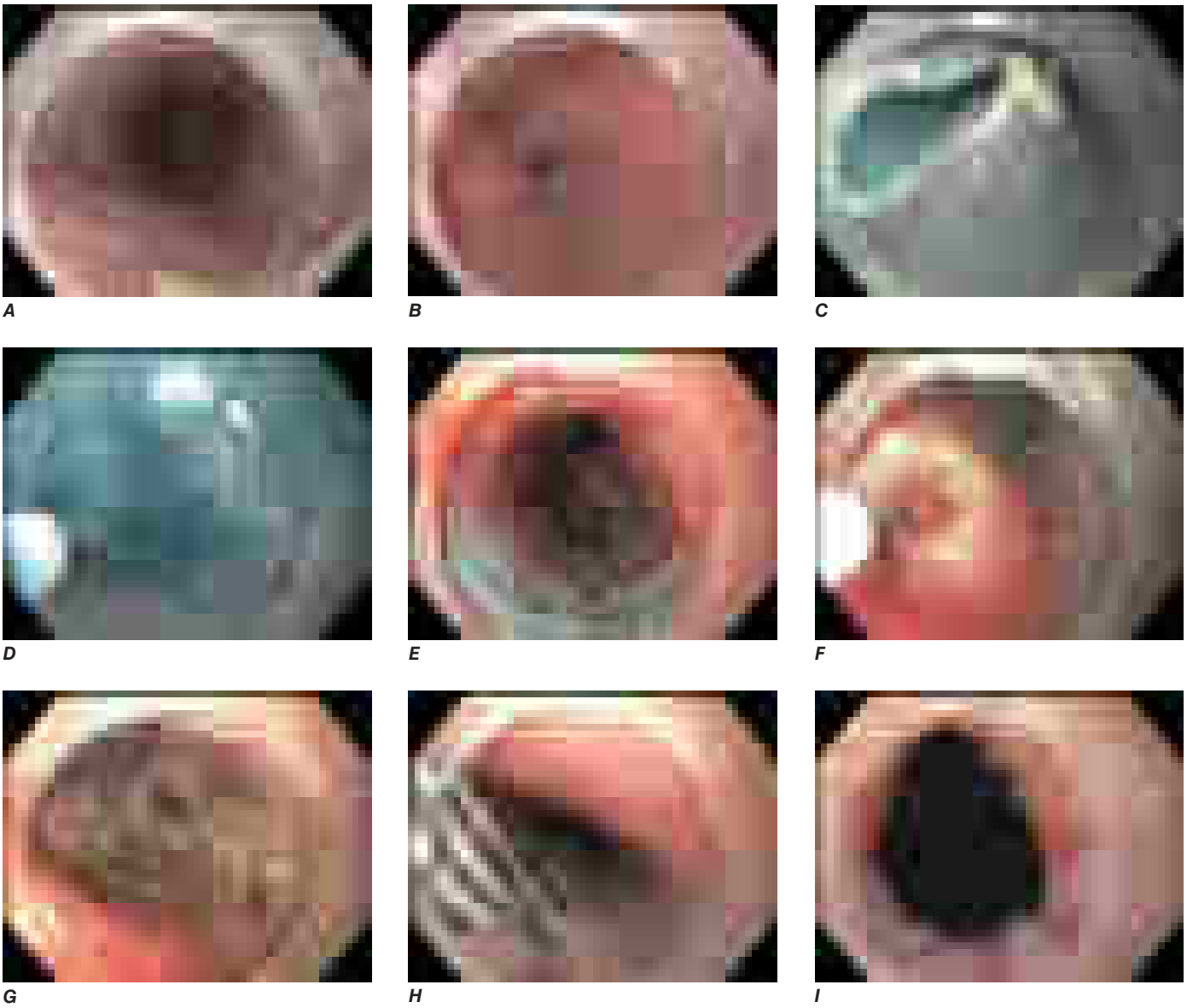


FIGURE 315-21 Peroral endoscopic myotomy (POEM) for achalasia. ***A.*** Dilated aperistaltic esophagus with retained secretions. ***B.*** Hypertonic lower esophageal sphincter (LES) region. ***C.*** Mucosal incision (mucosotomy) 10 cm proximal to the LES. ***D.*** Submucosal dissection using an electrosurgical knife following endoscope entry through the mucosotomy site into the submucosal space. ***E.*** Completion of submucosal tunnel to the cardia. ***F.*** Initiation of myotomy of the muscularis propria distal to the mucosotomy site. ***G.*** Completion of myotomy to the cardia. ***H.*** Closure of mucosotomy site with clips. ***I.*** Patulous gastroesophageal junction following myotomy.

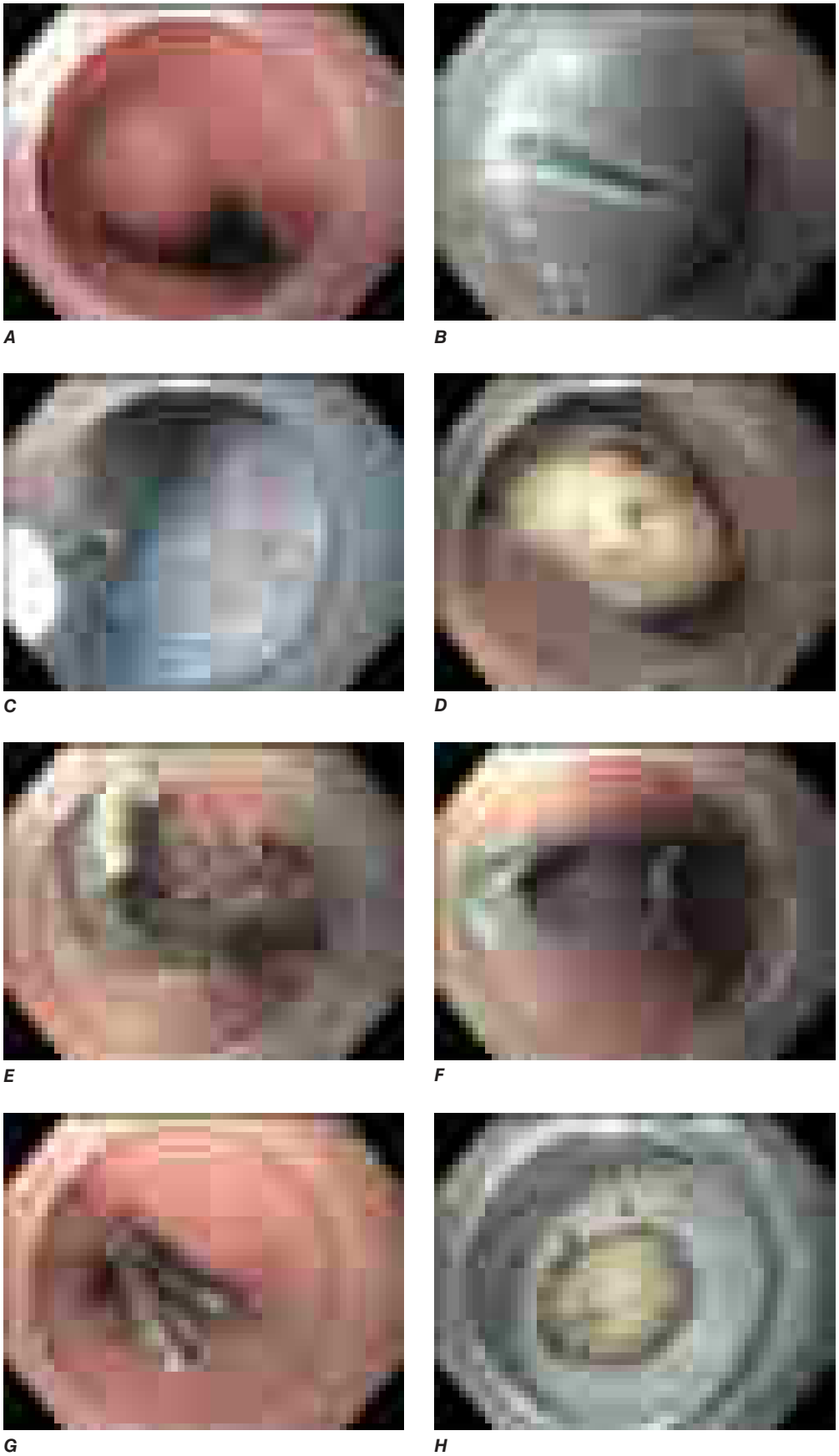


FIGURE 315-22 Peroral endoscopic tumorectomy (POET). ***A.*** Midesophageal subepithelial lesion. ***B.*** Mucosal incision (mucosotomy) 5 cm proximal to the lesion. ***C.*** Submucosal dissection and tunneling to the site of the lesion. ***D.*** Dissection of the lesion from its attachment to the muscularis propria. ***E.*** Postresection defect through the muscularis propria. ***F.*** Mucosotomy site. ***G.*** Closure of mucosotomy site with clips. ***H.*** Resected specimen (leiomyoma).

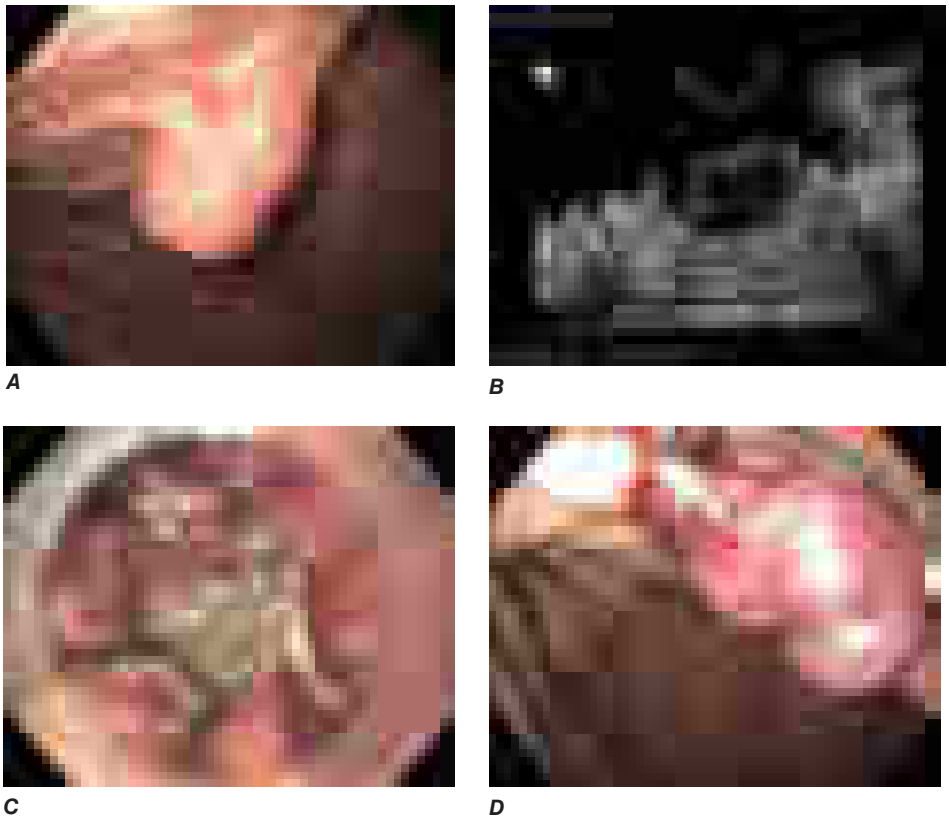


FIGURE 315-23 Endoscopic full-thickness resection (EFTR) of a gastrointestinal stromal tumor. ***A.*** Subepithelial lesion in the proximal stomach. ***B.*** Hypoechoic lesion arising from the fourth layer (muscularis propria) at endoscopic ultrasound. ***C.*** Full-thickness resection defect. ***D.*** Closure of defect using an over-the-scope clip.

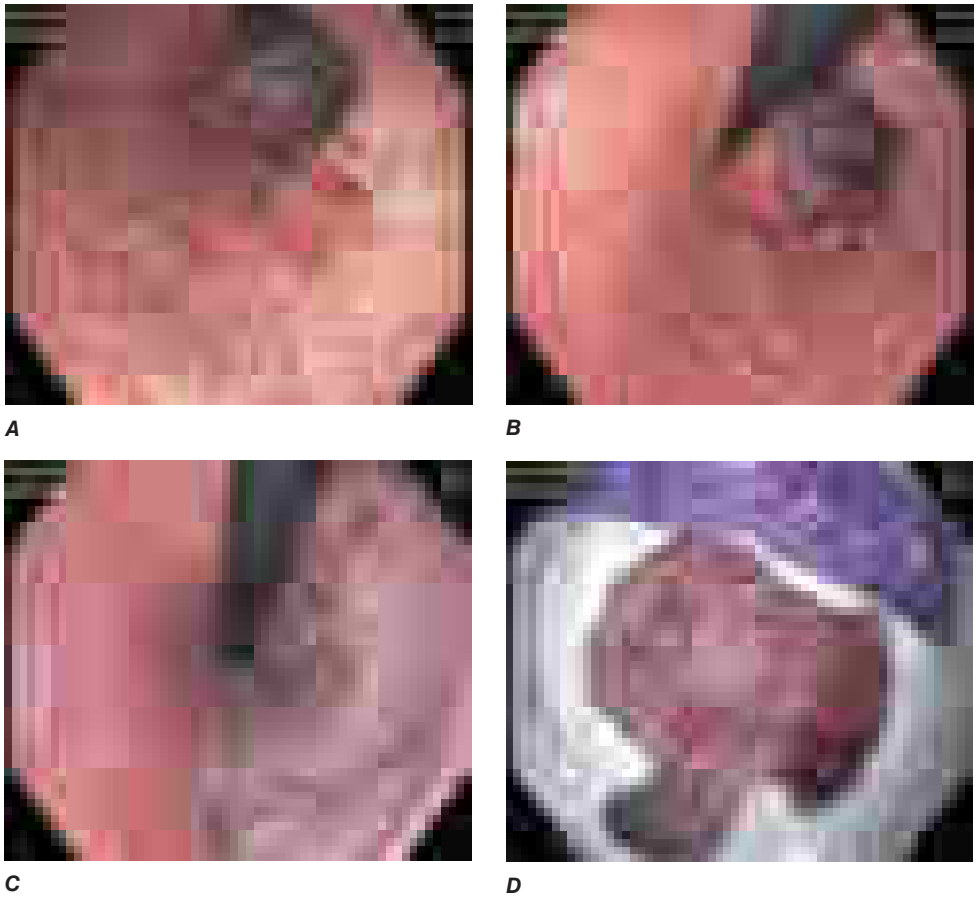


FIGURE 315-24 Endoscopic submucosal dissection (ESD). ***A.*** Large flat distal rectal adenoma with central lobulation. ***B.*** Marking the periphery of the lesion with coagulation dots. ***C.*** Rectal defect following ESD. ***D.*** Specimen resected en bloc.

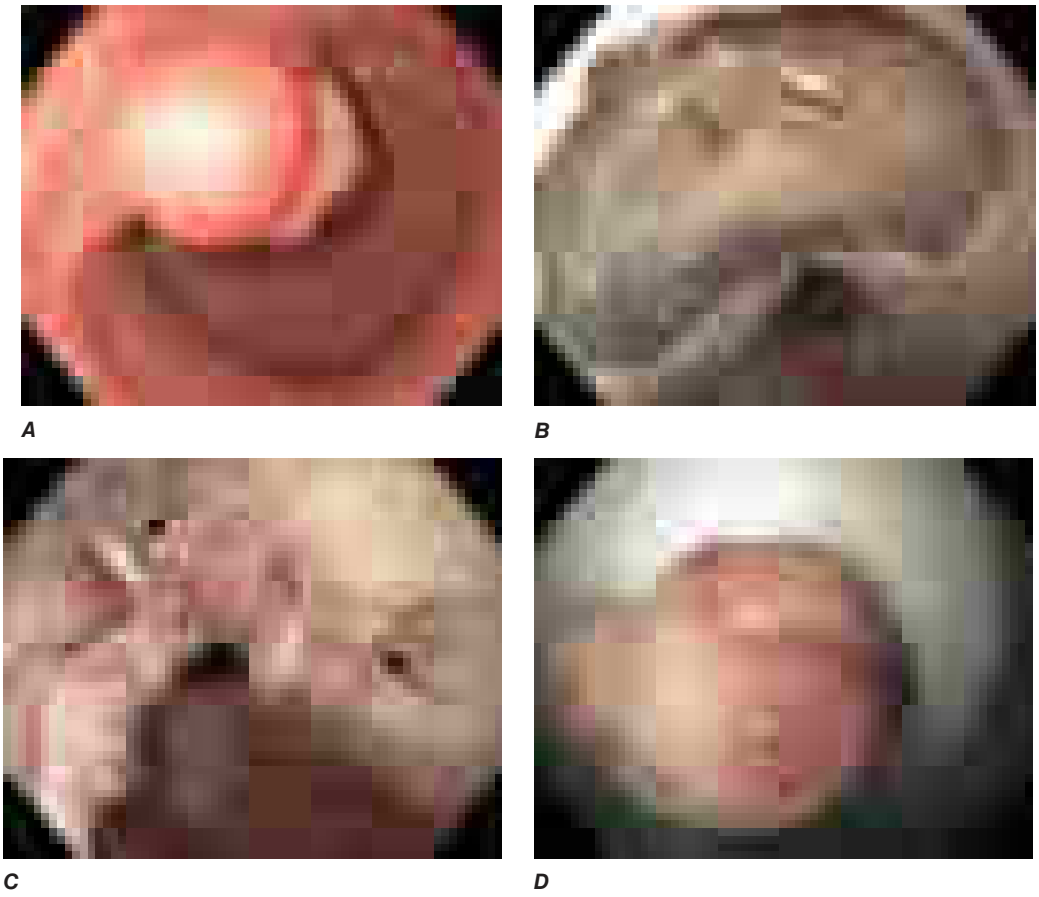


FIGURE 315-25 Closure of large defect using an endoscopic suturing device. ***A.*** Ulcerated inflammatory fibroid polyp in the antrum. ***B.*** Large defect following endoscopic submucosal dissection of the lesion. ***C.*** Closure of the defect using endoscopic sutures *(arrows)*. ***D.*** Resected specimen.

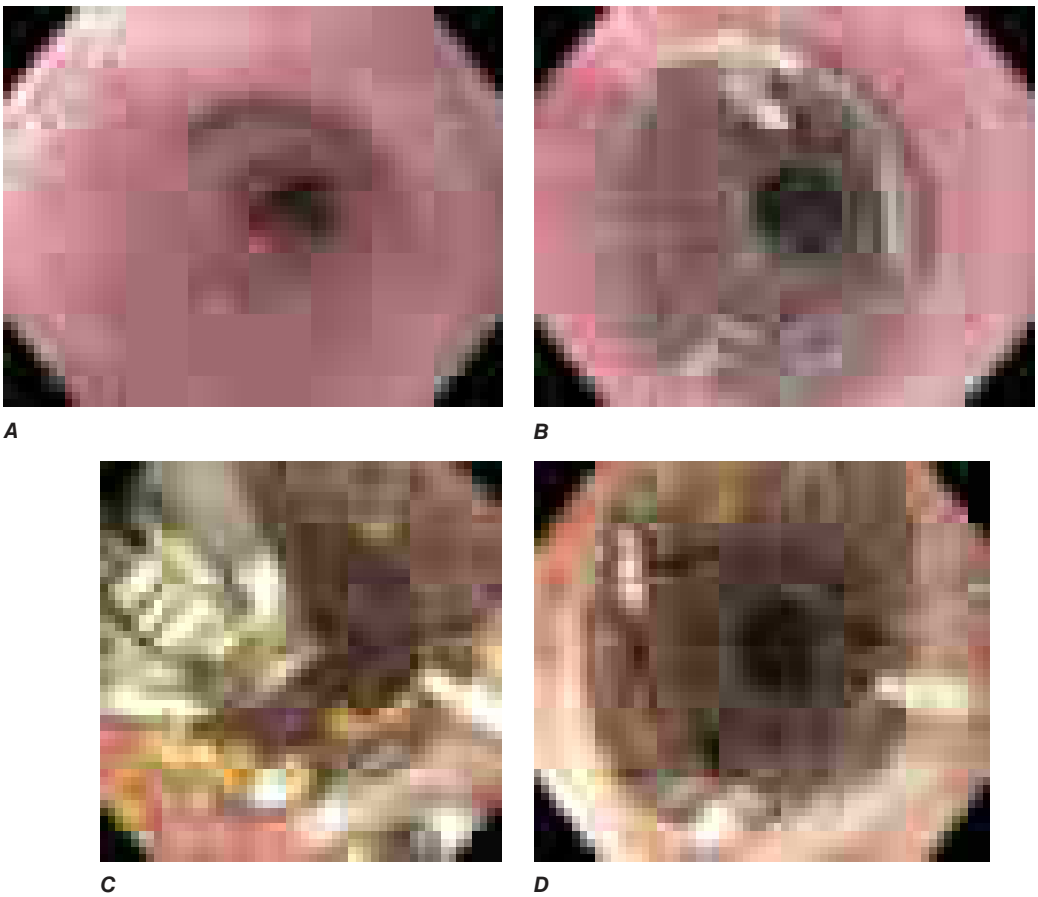


FIGURE 315-26 Prevention of stent migration using endoscopic sutures. ***A.*** Esophagogastric anastomotic stricture refractory to balloon dilation. ***B.*** Temporary placement of a covered esophageal stent. ***C.*** Endoscopic suturing device to anchor the stent to the esophageal wall. ***D.*** Stent fixation with endoscopic sutures *(arrows)*.

TABLE 315-1 Antibiotic Prophylaxis for Endoscopic Procedures

PATIENT CONDITION	PROCEDURE CONTEMPLATED	GOAL OF PROPHYLAXIS	PERIPROCEDURAL ANTIBIOTIC PROPHYLAXIS
All cardiac conditions	Any endoscopic procedure	Prevention of infective endocarditis	Not indicated
Bile duct obstruction in the absence of cholangitis	ERCP with complete drainage	Prevention of cholangitis	Not recommended
Bile duct obstruction in absence of cholangitis	ERCP with anticipated incomplete drainage (e.g., sclerosing cholangitis, hilar strictures)	Prevention of cholangitis	Recommended; continue antibiotics after the procedure
Sterile pancreatic fluid collection (e.g., pseudocyst, necrosis), which communicates with pancreatic duct	ERCP	Prevention of cyst infection	Recommended; continue antibiotics after the procedure
Sterile pancreatic fluid collection	Transmural drainage	Prevention of cyst infection	Recommended
Solid lesion along upper GI tract	EUS-FNA	Prevention of local infection	Not recommended[a]
Solid lesion along lower GI tract	EUS-FNA	Prevention of local infection	Not recommended[a]
Cystic lesions along GI tract (including mediastinum and pancreas)	EUS-FNA	Prevention of cyst infection	Recommended
All patients	Percutaneous endoscopic feeding tube placement	Prevention of peristomal infection	Recommended[b]
Cirrhosis with acute GI bleeding	Required for all such patients, regardless of endoscopic procedures	Prevention of infectious complications and reduction of mortality	Recommended, upon admission[c]
Continuous peritoneal dialysis	Lower GI tract endoscopy	Prevention of bacterial peritonitis	Recommended
Synthetic vascular graft and other nonvalvular cardiovascular devices	Any endoscopic procedure	Prevention of graft and device infection	Not recommended[d]
Prosthetic joints	Any endoscopic procedure	Prevention of septic arthritis	Not recommended[e]

[a]Low rates of bacteremia and local infection. [b]Cefazolin or an antibiotic with equivalent coverage of oral and skin flora. [c]Risk for bacterial infection associated with cirrhosis and GI bleeding is well established; ceftriaxone or a quinolone antibiotic recommended. [d]No reported cases of infection associated with endoscopy. [e]Very low risk of infection.

Abbreviations: ERCP, endoscopic retrograde cholangiopancreatography; EUS-FNA, endoscopic ultrasound–fine-needle aspiration; GI, gastrointestinal.

Source: Adapted from MA Kashab et al: Gastrointest Endosc 81:81, 2015; with permission from Elsevier.

Clinical predictors of rebleeding help identify patients most likely to benefit from urgent endoscopy and endoscopic, angiographic, or surgical hemostasis.

Initial Evaluation The initial evaluation of the bleeding patient focuses on the severity of hemorrhage as reflected by the presence of supine hypotension or tachycardia, postural vital sign changes, and the frequency of hematemesis or melena. Decreases in hematocrit and hemoglobin lag behind the clinical course and are not reliable gauges of the magnitude of acute bleeding. Nasogastric tube aspiration and lavage can also be used to judge the severity of bleeding, but these are no longer routinely performed for this purpose. The bedside initial evaluation, completed well before the bleeding source is confidently identified, guides immediate supportive care of the patient, triage to the ward or intensive care unit, and timing of endoscopy. The severity of the initial hemorrhage is the most important indication for urgent endoscopy, since a large initial bleed increases the likelihood of ongoing or recurrent bleeding. Patients with resting hypotension or orthostatic change in vital signs, repeated hematemesis, bloody nasogastric aspirate that does not clear with large volume lavage, or those requiring blood transfusions should be considered for urgent endoscopy. In addition, patients with cirrhosis, coagulopathy, respiratory or renal failure, and those over 70 years of age are more likely to have significant rebleeding and to benefit from prompt evaluation and treatment.

Bedside evaluation also suggests an upper or lower gastrointestinal source of bleeding in most patients. Over 90% of patients with melena are bleeding proximal to the ligament of Treitz, and ~85% of patients with hematochezia are bleeding from the colon. Melena can result from bleeding in the small bowel or right colon, especially in older patients with slow colonic transit. Conversely, some patients with massive hematochezia may be bleeding from an upper gastrointestinal source, such as a gastric Dieulafoy lesion or duodenal ulcer, with rapid intestinal transit. Early upper endoscopy should be considered in such patients.

Endoscopy should be performed after the patient has been resuscitated with intravenous fluids and transfusions, as necessary. Marked coagulopathy or thrombocytopenia is usually treated before endoscopy, since correction of these abnormalities may lead to resolution of bleeding, and techniques for endoscopic hemostasis are limited in such patients. Metabolic derangements should also be addressed. Tracheal intubation for airway protection should be considered before upper endoscopy in patients with repeated recent hematemesis, encephalopathy and suspected variceal hemorrhage. A single dose of erythromycin (3–4 mg/kg or 250 mg) administered intravenously 30–90 min prior to upper endoscopy increases gastric emptying and may clear blood and clots from the stomach to improve endoscopic visualization.

Most patients with hematochezia who are otherwise stable can undergo semielective colonoscopy. Controlled trials have not shown a benefit to urgent colonoscopy in patients hospitalized with hematochezia, although patients with massive or recurrent large-volume episodes of hematochezia should undergo urgent colonoscopy after a rapid colonic purge with a polyethylene glycol solution. Colonoscopy has a higher diagnostic yield than radionuclide bleeding scans or angiography in lower gastrointestinal bleeding, and endoscopic therapy can be applied in some cases. Urgent colonoscopy can be hindered by poor visualization due to persistent vigorous bleeding with recurrent hemodynamic instability, and other techniques (such as angiography or even emergent subtotal colectomy) must be employed. In such patients, massive bleeding originating from an upper gastrointestinal source should also be considered and excluded promptly by upper endoscopy. The anal and rectal mucosa should also be visualized endoscopically early in the course of massive rectal bleeding, as bleeding lesions in or close to the anal canal may be identified that are amenable to endoscopic or surgical transanal hemostatic techniques.

Peptic Ulcer The endoscopic appearance of peptic ulcers provides useful prognostic information and guides the need for endoscopic therapy in patients with acute hemorrhage **(Fig. 315-28)**. A clean-based ulcer is associated with a low risk (3–5%) of rebleeding; patients with melena and a clean-based ulcer are often discharged home from the emergency room or endoscopy suite if they are young, reliable, and otherwise healthy. Flat pigmented spots and adherent clots covering

TABLE 315-2 Management of Antithrombotic Drugs Prior to Endoscopic Procedures

DRUG	BLEEDING RISK OF PROCEDURE	MANAGEMENT	INTERVAL BETWEEN LAST DOSE AND PROCEDURE	COMMENTS
Warfarin	Low[a]	Continue	N/A	Ensure that INR is not supratherapeutic
	High[b]	Discontinue	3–7 days (usually 5), INR should be ≤1.5 for procedure	Consider bridging therapy with heparin;[c] usually safe to resume warfarin on the same or next day
Dabigatran, rivaroxaban, apixaban, edoxaban	Low[a]	Continue or hold morning dose on day of procedure	N/A	
Dabigatran	High[b]	Discontinue	2–3 days if GFR is ≥50 mL/min, 3–4 days if GFR is 30–49 mL/min	Bridging therapy not recommended; resume drug when bleeding risk is low
Rivaroxaban, apixaban, edoxaban	High[a]	Discontinue	2 days if GFR is ≥60 mL/min, 3 days if GFR is 30–59 mL/min, 4 days if GFR is <30 mL/min	Bridging therapy not recommended; resume drug when bleeding risk is low
Heparin	Low[a]	Continue	N/A	
	High[b]	Discontinue	4–6 h for unfractionated heparin	Skip one dose if using low-molecular-weight heparin
Aspirin	Any	Continue	N/A	Low-dose aspirin does not substantially increase the risk of endoscopic procedures
Aspirin with dipyridamole	Low[a]	Continue	N/A	
	High[b]	Discontinue	2–7 days	Consider continuing aspirin monotherapy
P2Y12 receptor antagonists (clopidogrel, prasugrel, ticlopidine, ticagrelor, cangrelor	Low[a]	Continue	N/A	
	High[b]	Coronary stent in place: discuss with cardiologist	5 days (clopidogrel or ticagrelor), 7 days (prasugrel), 10–4 days (ticlopidine)	Risk of stent thrombosis for 12 months after insertion of drug-eluting coronary stent or 1 month after insertion of bare metal coronary stent
		No coronary stent: Discontinue, consider substituting to aspirin		

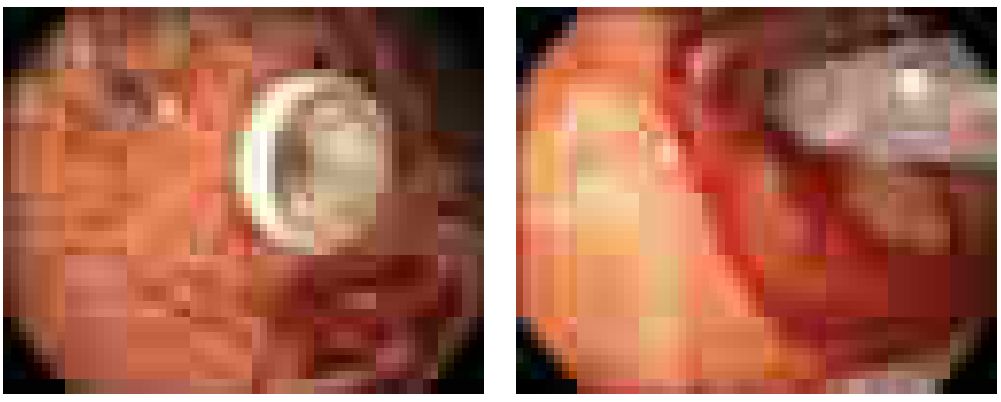
[a]Low-risk endoscopic procedures include EGD or colonoscopy with or without biopsy, EUS without FNA, ERCP with stent exchange. [b]High-risk endoscopic procedures include EGD or colonoscopy with dilation, polypectomy, or thermal ablation; PEG; EUS with FNA; ERCP with sphincterotomy or pseudocyst drainage. [c]Bridging therapy with low-molecular-weight heparin should be considered for patients discontinuing warfarin who are at high risk for thromboembolism, including those with (1) atrial fibrillation with a CHA2DS2-VASc score ≥2, mechanical valves, or history of stroke; (2) mechanical mitral valve; (3) mechanical aortic valve with other thromboembolic risk factors or older-generation mechanical aortic valve; (4) venous thromboembolism within the past 3 months.

Abbreviations: GFR, glomerular filtration rate; INR, international normalized ratio; N/A, not applicable.

Source: Data from RD Acosta et al: Gastrointest Endosc 83:3, 2016; and AM Veitch et al: Gut 65:374, 2016.

the ulcer base have a 10% and 20% risk of rebleeding, respectively. Endoscopic therapy may be considered for an ulcer with an adherent clot. When a fibrin plug is seen protruding from a vessel wall in the base of an ulcer (so-called sentinel clot or visible vessel), the risk of rebleeding from the ulcer is 40%. This finding generally leads to endoscopic therapy to decrease the rebleeding rate. When active spurting from an ulcer is seen, there is a 90% risk of ongoing bleeding without therapy.

Endoscopic therapy of ulcers with high-risk stigmata typically lowers the rebleeding rate to 5–10%. Several hemostatic techniques are available, including injection of epinephrine or a sclerosant into and around the vessel **(Fig. 315-29)**, "coaptive coagulation" of the vessel in the base of the ulcer using a thermal probe that is pressed against the site of bleeding **(Fig. 315-30)**, placement of through-the-scope hemoclips **(Fig. 315-31)** or an over-the-scope clip **(Fig. 315-32)**, or a combination of these modalities **(Video V5-8)**. In conjunction with endoscopic



A ***B***

FIGURE 315-27 Bleeding from percutaneous endoscopic gastrostomy (PEG) tube placement. *A.* Patient with melena from a recently placed PEG tube. ***B.*** Loosening of the internal disk bumper of the PEG tube revealed active bleeding from within the PEG tract.

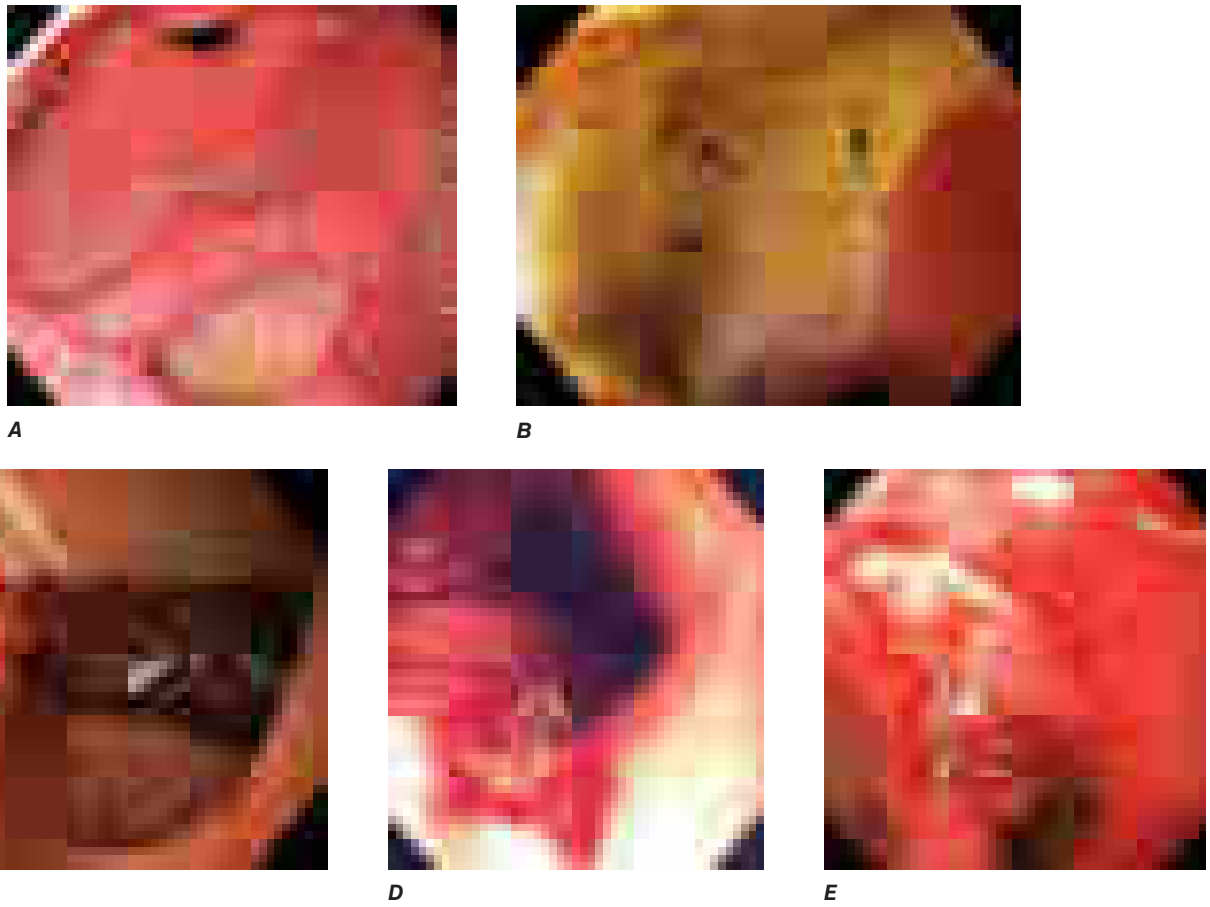


FIGURE 315-28 Stigmata of hemorrhage in peptic ulcers. ***A.*** Gastric antral ulcer with a clean base. ***B.*** Duodenal ulcer with flat pigmented spots *(arrows)*. ***C.*** Duodenal ulcer with a dense adherent clot. ***D.*** Gastric ulcer with a pigmented protuberance/visible vessel. ***E.*** Duodenal ulcer with active spurting *(arrow)*.

therapy, the administration of a proton pump inhibitor decreases the risk of rebleeding and improves patient outcome.

Varices Two complementary strategies guide therapy of bleeding varices: local treatment of the bleeding varices and treatment of the underlying portal hypertension. Local therapies, including endoscopic variceal band ligation, endoscopic variceal sclerotherapy (EVS), stent placement and balloon tamponade with a Sengstaken-Blakemore tube, effectively control acute hemorrhage in most patients, although therapies that decrease portal pressure (pharmacologic treatment, surgical shunts, or radiologically placed intrahepatic portosystemic shunts) also play an important role.

Endoscopic variceal ligation (EVL) is indicated for the prevention of a first bleed (primary prophylaxis) from large esophageal varices **(Fig. 315-33)**, particularly in patients in whom nonselective beta blockers are contraindicated or not tolerated. EVL is also the preferred endoscopic therapy for control of active esophageal variceal bleeding and for subsequent eradication of esophageal varices (secondary prophylaxis). During EVL a varix is suctioned into a cap fitted on the end of the endoscope, and a rubber band is released from the cap, ligating the varix **(Fig. 315-34, Video V5-9)**. EVL controls acute hemorrhage in up to 90% of patients. Complications of EVL, such as postligation ulcer bleeding and esophageal stenosis, are uncommon. EVS involves the injection of a sclerosing, thrombogenic solution into or next to esophageal varices. EVS also controls acute hemorrhage in most patients, but it is generally used as salvage therapy when band ligation fails because of its higher complication rate. Bleeding from large gastric fundic varices **(Fig. 315-35)** is best treated with endoscopic cyanoacrylate ("glue") injection **(Video V5-10)**, since EVL or EVS of these varices is associated with a high rebleeding rate. Complications of cyanoacrylate injection include infection and glue embolization to other organs, such as the lungs, brain, and spleen.

After treatment of the acute hemorrhage, an elective course of endoscopic therapy can be undertaken with the goal of eradicating esophageal varices and preventing rebleeding months to years later. However, this chronic therapy is less successful, preventing long-term rebleeding in ~50% of patients. Pharmacologic therapies that decrease

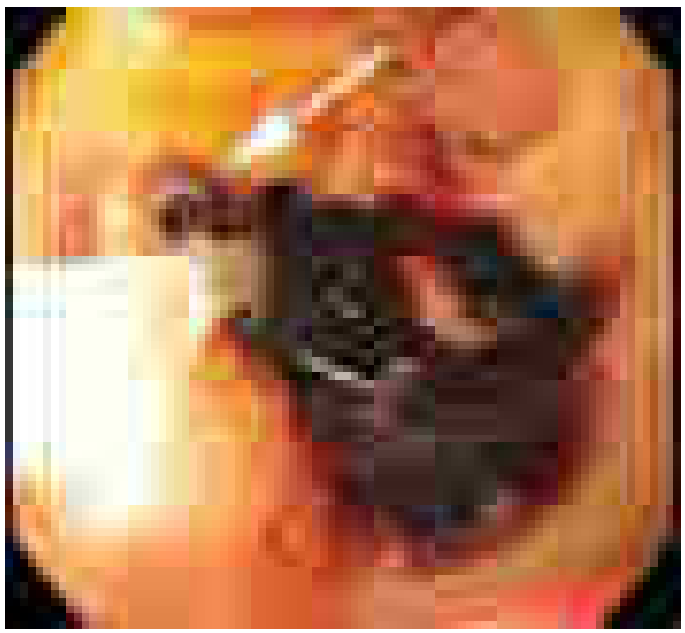
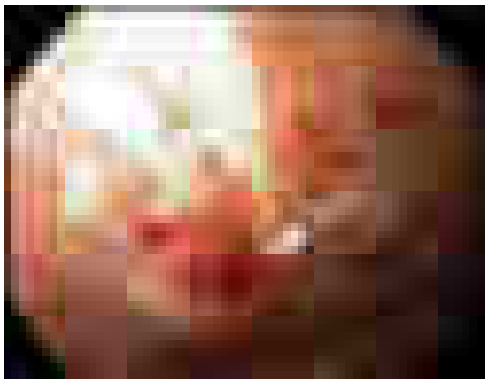


FIGURE 315-29 Epinephrine injection into a duodenal ulcer with visible vessel *(arrow)* and adherent clot.

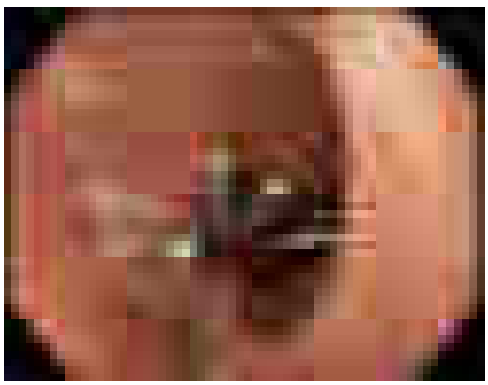
FIGURE 315-30 Treated duodenal ulcer by contact coagulation with a thermal probe.

portal pressure have similar efficacy, and the two modalities are usually combined.

Dieulafoy's Lesion This lesion, also called persistent caliber artery, is a large-caliber arteriole that runs immediately beneath the gastrointestinal mucosa and bleeds through a focal mucosal erosion (Fig. 315-36). Dieulafoy's lesion is seen most commonly on the lesser curvature of the proximal stomach, causes impressive arterial hemorrhage, and may be difficult to diagnose when not actively bleeding; it is often recognized only after repeated endoscopy for recurrent bleeding.

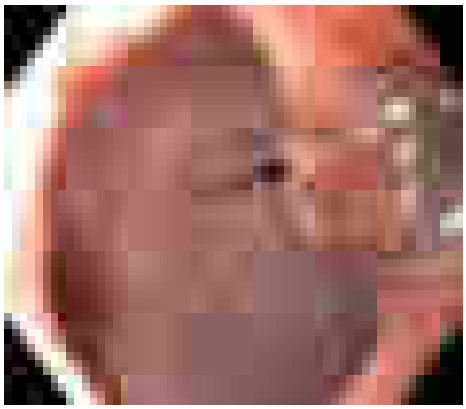


A

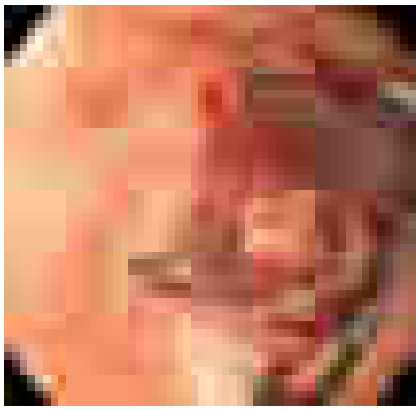


B

FIGURE 315-31 Ulcer hemostasis using through-the-scope clips. *A.* Superficial duodenal ulcer with visible vessel *(arrow)*. ***B.*** Hemostasis secured following placement of multiple clips.



A



B

FIGURE 315-32 Endoscopic hemostasis of ulcer bleeding. *A.* Pyloric channel ulcer with visible vessel *(arrow)*. ***B.*** Ulcer hemostasis with placement of an over-the-scope clip.

Endoscopic therapy, such as thermal coagulation, band ligation, or endoscopic suturing, is typically effective for control of bleeding and sealing of the underlying vessel once the lesion has been identified (Video V5-11). Rescue therapies, such as angiographic embolization or surgical oversewing, are considered in situations where endoscopic therapy has failed.

Mallory-Weiss Tear A Mallory-Weiss tear is a linear mucosal rent near or across the gastroesophageal junction that is often associated with retching or vomiting (Fig. 315-37). When the tear disrupts a submucosal arteriole, brisk hemorrhage may result. Endoscopy is the best method for diagnosis, and an actively bleeding tear can be treated endoscopically with epinephrine injection, coaptive coagulation, band ligation, or hemoclips (Video V5-12). Unlike peptic ulcer, a Mallory-Weiss tear with a nonbleeding sentinel clot in its base rarely rebleeds and thus does not necessitate endoscopic therapy.

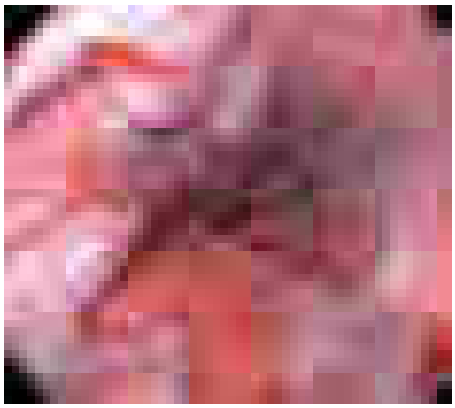
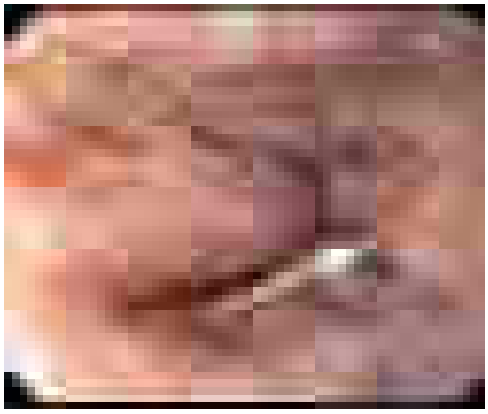
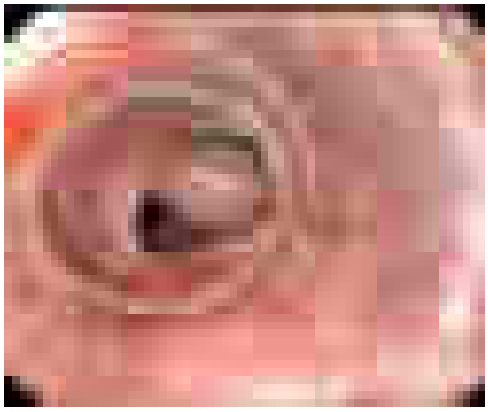


FIGURE 315-33 Esophageal varices.

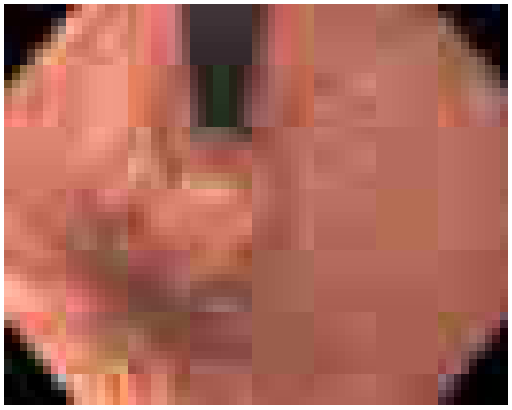


A

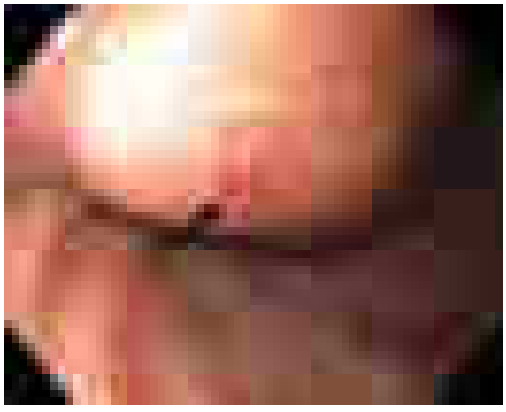


B

FIGURE 315-34 Endoscopic band ligation of esophageal varices. ***A.*** Large esophageal varices with stigmata of recent bleeding characterized by a fibrin plug *(arrow)*. ***B.*** Band ligation of varices.



A



B

FIGURE 315-35 Gastric varices. ***A.*** Large gastric fundal varices. ***B.*** Stigmata of recent bleeding from the same gastric varices *(arrow)*.

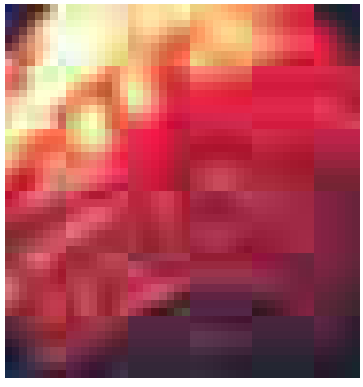
Vascular Ectasias Vascular ectasias are flat mucosal vascular anomalies that are best diagnosed by endoscopy. They usually cause slow intestinal blood loss and occur either in a sporadic fashion or in a well-defined pattern of distribution (e.g., gastric antral vascular ectasia [GAVE] or "watermelon stomach") **(Fig. 315-38)**. Cecal vascular ectasias, GAVE, and radiation-induced rectal ectasias are often responsive to local endoscopic ablative therapy, such as argon plasma coagulation **(Video V5-13)**. Patients with diffuse small bowel vascular ectasias (associated with chronic renal failure and with hereditary hemorrhagic telangiectasia) may continue to bleed despite endoscopic treatment of easily accessible lesions by conventional endoscopy. These patients may benefit from device-assisted enteroscopy with endoscopic therapy or pharmacologic treatment with octreotide or estrogen/progesterone.

Colonic Diverticula Diverticula form where nutrient arteries penetrate the muscular wall of the colon en route to the colonic mucosa **(Fig. 315-39)**. The artery found in the base of a diverticulum may bleed, causing painless and impressive hematochezia. Colonoscopy is indicated in patients with hematochezia and suspected diverticular hemorrhage, since other causes of bleeding (such as vascular ectasias, colitis, and colon cancer) must be excluded. In addition an actively bleeding diverticulum may be seen and treated during colonoscopy **(Fig. 315-40, Video V5-14)**.

GASTROINTESTINAL OBSTRUCTION AND PSEUDOOBSTRUCTION

Endoscopy is useful for evaluation and treatment of some forms of gastrointestinal obstruction. An important exception is small-bowel obstruction due to surgical adhesions, which is generally not diagnosed or treated endoscopically. Esophageal, gastroduodenal, and colonic obstruction or pseudoobstruction can all be diagnosed and often managed endoscopically.

Acute Esophageal Obstruction Esophageal obstruction by impacted food **(Fig. 315-41)** or an ingested foreign body **(Fig. 315-42)** is a potentially life-threatening event and represents an endoscopic emergency. Left untreated, the patient may develop esophageal ulceration, ischemia, and perforation. Patients with persistent esophageal obstruction often have hypersalivation and are usually unable to swallow water. Sips of a carbonated beverage, sublingual nifedipine or nitrates, or intravenous glucagon may resolve an esophageal food impaction, but in most patients an underlying web, ring, or stricture is present and endoscopic removal of the obstructing food bolus is necessary.



A

B

FIGURE 315-36 Dieulafoy's lesion. ***A.*** Actively spurting jejunal Dieulafoy's lesion. There is no underlying mucosal lesion. ***B.*** Histology of a gastric Dieulafoy's lesion. A persistent caliber artery *(arrows)* is present in the gastric submucosa, immediately beneath the mucosa.

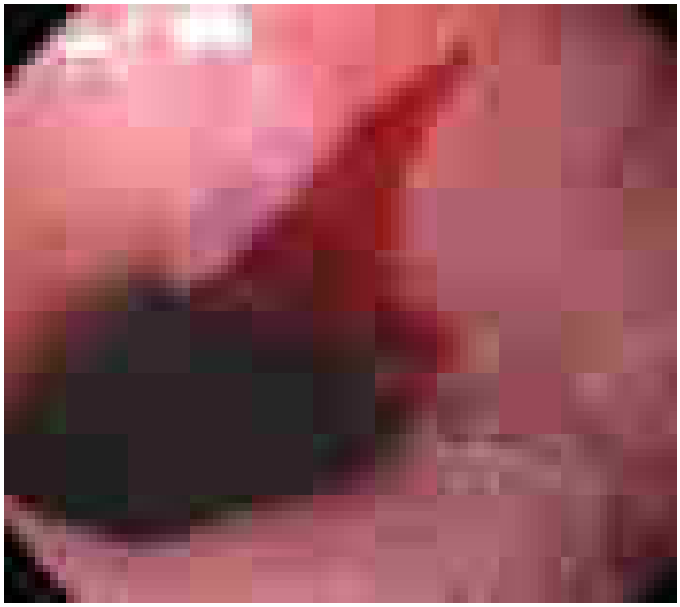


FIGURE 315-37 Mallory-Weiss tear at the gastroesophageal junction.

Endoscopy is generally the best initial test in such patients since endoscopic removal of the obstructing material is usually possible, and the presence of an underlying esophageal pathology can often be determined. Radiographs of the chest and neck should be considered before endoscopy in patients with fever, obstruction for ≥24 h, or ingestion of a sharp object, such as a fishbone. Radiographic contrast studies interfere with subsequent endoscopy and are not advisable in most patients with a clinical picture of esophageal obstruction.

Gastric Outlet Obstruction Obstruction of the gastric outlet is commonly caused by gastric, duodenal, or pancreatic malignancy, or chronic peptic ulceration with stenosis of the pylorus **(Fig. 315-43)**. Patients vomit partially digested food many hours after eating. Gastric decompression with a nasogastric tube and subsequent lavage for removal of retained material is the first step in treatment. The diagnosis can then be confirmed with a saline load test, if desired. Endoscopy is useful for diagnosis and treatment. Patients with benign pyloric stenosis may be treated with endoscopic balloon dilation of the pylorus, and a course of endoscopic dilation results in long-term relief of symptoms in ~50% of patients. Removable, fully covered lumen-apposing metal stents (LAMS) may also be used to treat benign pyloric stenosis **(Video V5-15)**. Malignant gastric outlet obstruction can be relieved with endoscopically placed expandable stents in patients with inoperable malignancy **(Video V5-16)**.

Colonic Obstruction and Pseudoobstruction These conditions both present with abdominal distention and discomfort, tympany, and a dilated colon on plain abdominal radiography. The radiographic appearance may be characteristic of a particular condition, such as sigmoid volvulus **(Fig. 315-44)**. Both obstruction and pseudoobstruction may lead to colonic perforation if left untreated. Acute colonic pseudoobstruction is a form of colonic ileus that is usually attributable to electrolyte disorders, narcotic and anticholinergic medications, immobility (as after surgery), or retroperitoneal hemorrhage or mass. Multiple causative factors are often present. Colonoscopy, water-soluble contrast enema, or CT may be used to assess for an obstructing lesion and differentiate obstruction from pseudoobstruction. One of these diagnostic studies should be strongly considered if the patient does not have clear risk factors for pseudoobstruction, if radiographs do not show air in the rectum, or if the patient fails to improve when underlying causes of pseudoobstruction have been addressed. The risk of cecal perforation in pseudoobstruction rises when the cecal diameter exceeds 12 cm, and decompression of the colon may be achieved using intravenous neostigmine or via colonoscopic decompression **(Fig. 315-45)**. Most patients should receive a trial of conservative therapy (with correction of electrolyte disorders, removal of offending medications, and increased mobilization) before undergoing an invasive decompressive procedure for colonic pseudoobstruction.

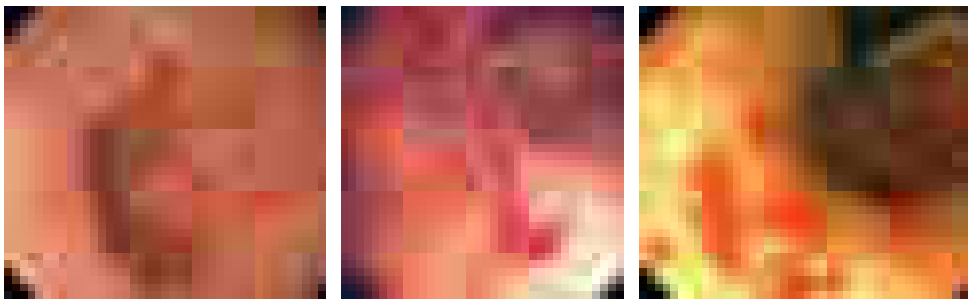
Colonic obstruction is an indication for urgent intervention. In the past, emergent diverting colostomy was usually performed with a subsequent second operation after bowel preparation to treat the underlying cause of obstruction. Colonoscopic placement of an expandable stent is an alternative treatment option that can relieve malignant colonic obstruction without emergency surgery and permit bowel preparation for an elective one-stage operation **(Fig. 315-46, Video V5-17)**.

ACUTE BILIARY OBSTRUCTION

The steady, severe pain that occurs when a gallstone acutely obstructs the common bile duct often brings patients to a hospital. The diagnosis of a ductal stone is suspected when the patient is jaundiced or when serum liver tests or pancreatic enzyme levels are elevated; it is confirmed by EUS, magnetic resonance cholangiography (MRCP), or direct cholangiography (performed endoscopically, percutaneously, or during surgery). ERCP is the primary means of treating common bile duct stones (Figs. 315-15 and 315-16), although they can also be removed by laparoscopic bile duct exploration at the time of cholecystectomy. Radiologic percutaneous biliary drainage may be required in some cases.

Bile Duct Imaging While transabdominal ultrasound diagnoses only a minority of bile duct stones, MRCP and EUS are >90% accurate and have an important role in diagnosis. Examples of these modalities are shown in **Fig. 315-47**.

If the suspicion for a bile duct stone is high and urgent treatment is required (as in a patient with obstructive jaundice and biliary sepsis), ERCP is the procedure of choice since it remains the gold standard for diagnosis and allows for immediate treatment **(Video V5-18)**. If a persistent bile duct stone is relatively unlikely (as in a patient with gallstone pancreatitis), ERCP may be supplanted by less invasive imaging techniques, such as EUS, MRCP, or intraoperative cholangiography



A ***B*** ***C***

FIGURE 315-38 Gastrointestinal vascular ectasias. ***A.*** Gastric antral vascular ectasia ("watermelon stomach") characterized by stripes of prominent flat or raised vascular ectasias. ***B.*** Cecal vascular ectasias. ***C.*** Radiation-induced vascular ectasias of the rectum in a patient previously treated for prostate cancer.

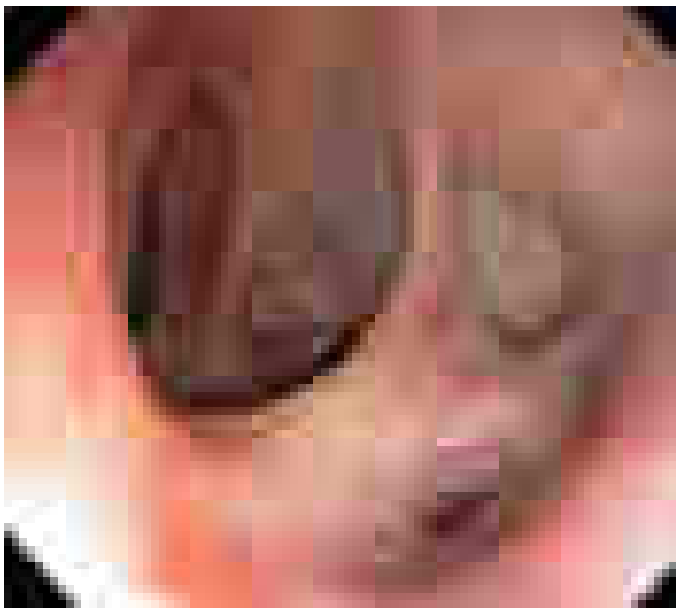


FIGURE 315-39 Colonic diverticula.

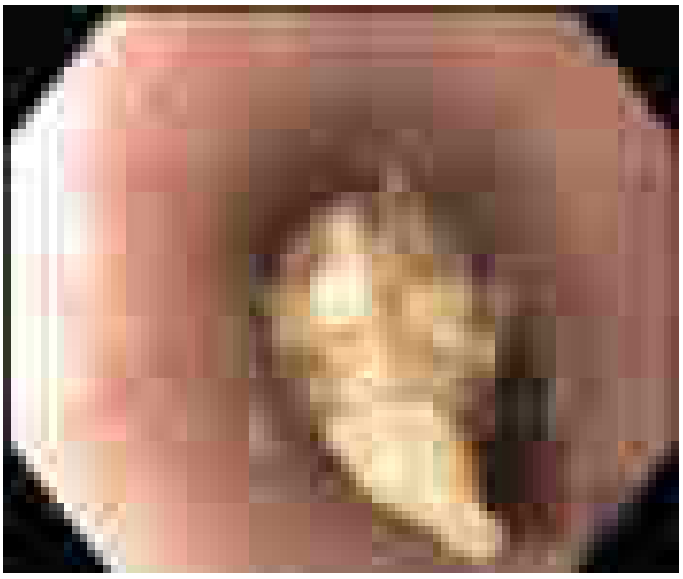


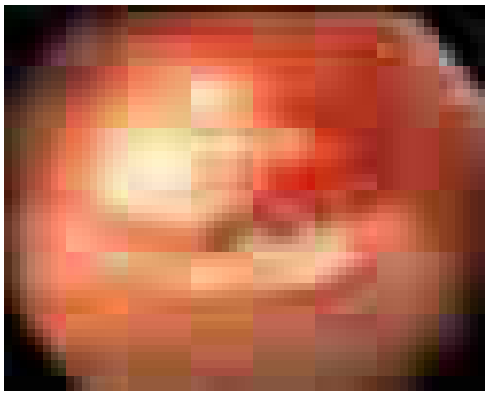
FIGURE 315-41 Esophageal food (meat) impaction.

performed during cholecystectomy, sparing some patients the risk and discomfort of ERCP.

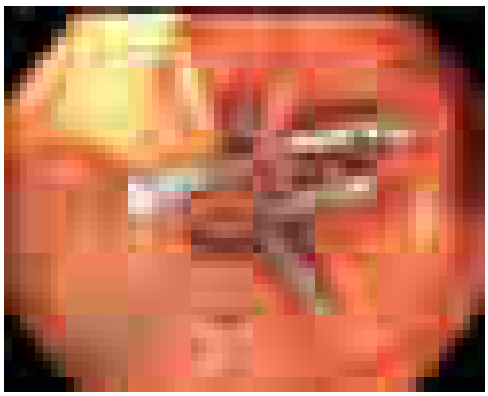
Ascending Cholangitis Charcot's triad of jaundice, abdominal pain, and fever is present in ~70% of patients with ascending cholangitis and biliary sepsis. These patients are managed initially with fluid resuscitation and intravenous antibiotics. Abdominal ultrasound is often performed to assess for gallbladder stones and bile duct dilation. However, the bile duct may not be dilated early in the course of acute biliary obstruction. Medical management usually improves the patient's clinical status, providing a window of ~24 h during which biliary drainage should be established, typically by ERCP. Undue delay can result in recrudescence of overt sepsis and increased morbidity and mortality rates. In addition to Charcot's triad, the additional presence of shock and confusion (Reynolds's pentad) is associated with a high mortality rate and should prompt urgent intervention to restore biliary drainage.

Gallstone Pancreatitis Gallstones may cause acute pancreatitis as they pass through the ampulla of Vater. The occurrence of gallstone pancreatitis usually implies passage of a stone into the duodenum, and only ~20% of patients harbor a persistent stone in the ampulla or the common bile duct. Retained stones are more common in patients with jaundice, rising serum liver tests following hospitalization, severe pancreatitis, or superimposed ascending cholangitis.

Urgent ERCP decreases the morbidity rate of gallstone pancreatitis in a subset of patients with retained bile duct stones. It is unclear whether the benefit of ERCP is mainly attributable to treatment and prevention of ascending cholangitis or to relief of pancreatic ductal obstruction. ERCP is warranted early in the course of gallstone pancreatitis if ascending cholangitis is suspected, especially in a jaundiced patient.



A



B

FIGURE 315-40 Diverticular hemorrhage. ***A.*** Actively bleeding sigmoid diverticulum. ***B.*** Hemostasis achieved using endoscopic clips.

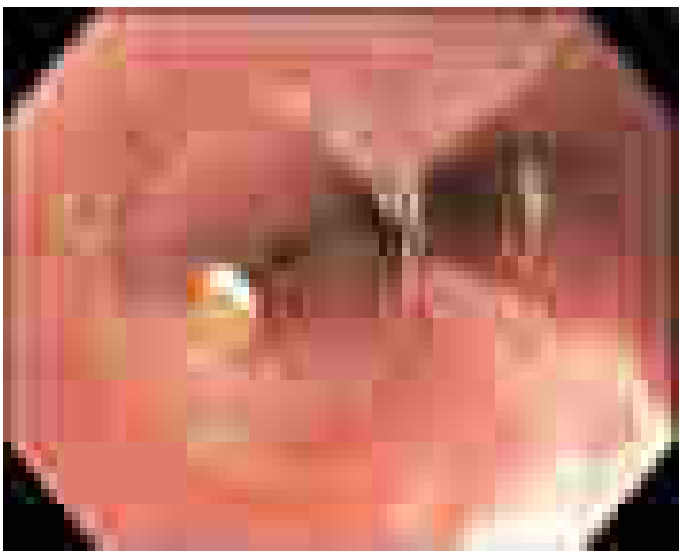
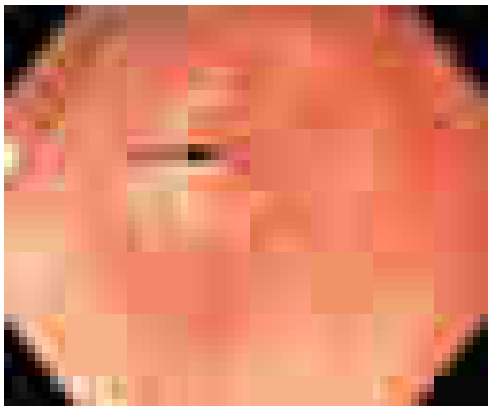
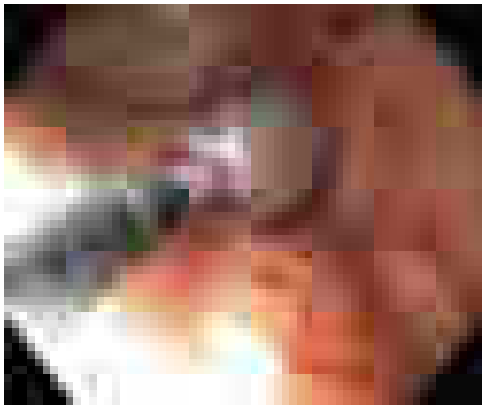


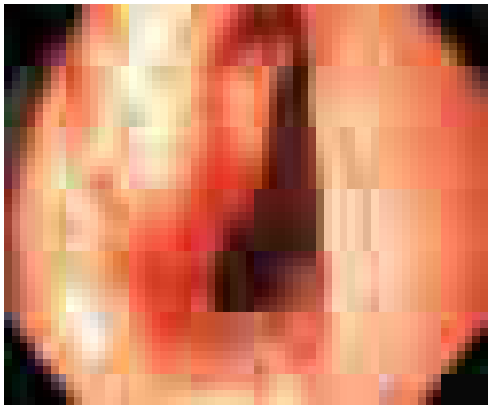
FIGURE 315-42 Impacted nail in the esophagus.



A



B



C

FIGURE 315-43 Gastric outlet obstruction due to pyloric stenosis. ***A.*** Nonsteroidal anti-inflammatory agent-induced ulcer disease with severe stenosis of the pylorus *(arrow)*. ***B.*** Balloon dilation of the stenosis. ***C.*** Appearance of pyloric ring post dilation.

Urgent ERCP may also benefit patients predicted to have severe pancreatitis using a clinical index of severity, such as the Glasgow or Ranson score. Since the benefit of ERCP is limited to patients with a retained bile duct stone, a strategy of initial MRCP or EUS for diagnosis decreases the utilization of ERCP in gallstone pancreatitis and improves clinical outcomes by limiting the occurrence of ERCP-related adverse events.

ELECTIVE ENDOSCOPY

DYSPEPSIA

Dyspepsia is a chronic or recurrent burning discomfort or pain in the upper abdomen that may be caused by diverse processes, such as gastroesophageal reflux, peptic ulcer disease, and "nonulcer dyspepsia,"

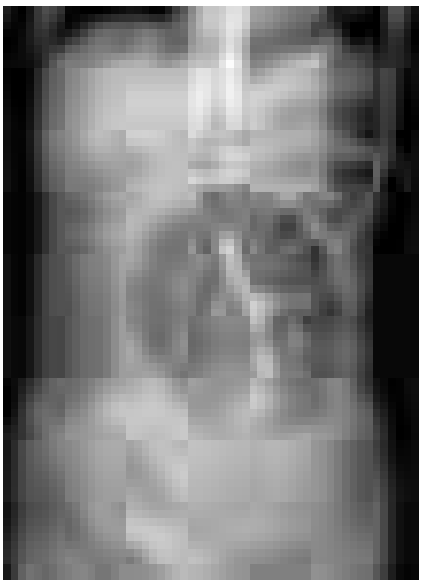


FIGURE 315-44 Sigmoid volvulus with the characteristic radiologic appearance of a "bent inner tube."

a heterogeneous category that includes disorders of motility, sensation and somatization. Gastric and esophageal malignancies are less common causes of dyspepsia. Careful history-taking allows accurate differential diagnosis of dyspepsia in only about half of patients. In the remainder, endoscopy can be a useful diagnostic tool, especially in patients whose symptoms are not resolved by *Helicobacter pylori* treatment or an empirical trial of acid-reducing therapy. Endoscopy should be performed at the outset in patients with dyspepsia and alarm features, such as weight loss or iron-deficiency anemia.

GASTROESOPHAGEAL REFLUX DISEASE (GERD)

When classic symptoms of gastroesophageal reflux are present, such as water brash and substernal heartburn, presumptive diagnosis and

A

B

FIGURE 315-45 Acute colonic pseudoobstruction. ***A.*** Acute colonic dilation occurring in a patient soon after knee surgery. ***B.*** Colonoscopic placement of decompression tube with marked improvement in colonic dilation.

FIGURE 315-46 Obstructing colonic carcinoma. ***A.*** Colonic adenocarcinoma causing marked luminal narrowing of the distal transverse colon. ***B.*** Endoscopic placement of a self-expandable metal stent. ***C.*** Radiograph of expanded stent across the obstructing tumor with a residual waist *(arrow)*.

empirical treatment are often sufficient. Endoscopy is a sensitive test for diagnosis of esophagitis (Fig. 315-48), but it will miss nonerosive reflux disease (NERD) since some patients have symptomatic reflux without esophagitis. The most sensitive test for diagnosis of GERD is 24-h ambulatory pH monitoring. Endoscopy is indicated in patients with reflux symptoms refractory to antisecretory therapy; in those with alarm symptoms, such as dysphagia, weight loss, or gastrointestinal bleeding; and in those with recurrent dyspepsia after treatment that is not clearly due to reflux on clinical grounds alone. Endoscopy should be considered in patients with long-standing (≥10 years) GERD, as they have a sixfold increased risk of harboring Barrett's esophagus compared to patients with <1 year of reflux symptoms.

Barrett's Esophagus Barrett's esophagus is specialized columnar metaplasia that replaces the normal squamous mucosa of the distal esophagus in some persons with GERD. Barrett's epithelium is a major risk factor for adenocarcinoma of the esophagus and is readily detected endoscopically, due to proximal displacement of the squamocolumnar junction (Fig. 315-5). A screening EGD for Barrett's esophagus should be considered in patients with a chronic (≥10 year) history of GERD symptoms. Endoscopic biopsy is the gold standard for confirmation of Barrett's esophagus and for dysplasia or cancer arising in Barrett's mucosa.

Periodic EGD with biopsies is recommended for surveillance of patients with Barrett's esophagus. Endoscopic resection (EMR or ESD) and/or ablation are performed when high-grade dysplasia or intramucosal cancer are found in the Barrett's mucosa. Although guidelines recommend observation and surveillance of low-grade dysplasia in Barrett's mucosa, recent evidence suggests that endoscopic treatment may be appropriate in select patients. Radiofrequency ablation (RFA) is the commonest ablative modality used for endoscopic treatment of Barrett's esophagus, and other modalities, such as cryotherapy, are also available.

PEPTIC ULCER

Peptic ulcer classically causes epigastric gnawing or burning, often occurring nocturnally and promptly relieved by food or antacids.

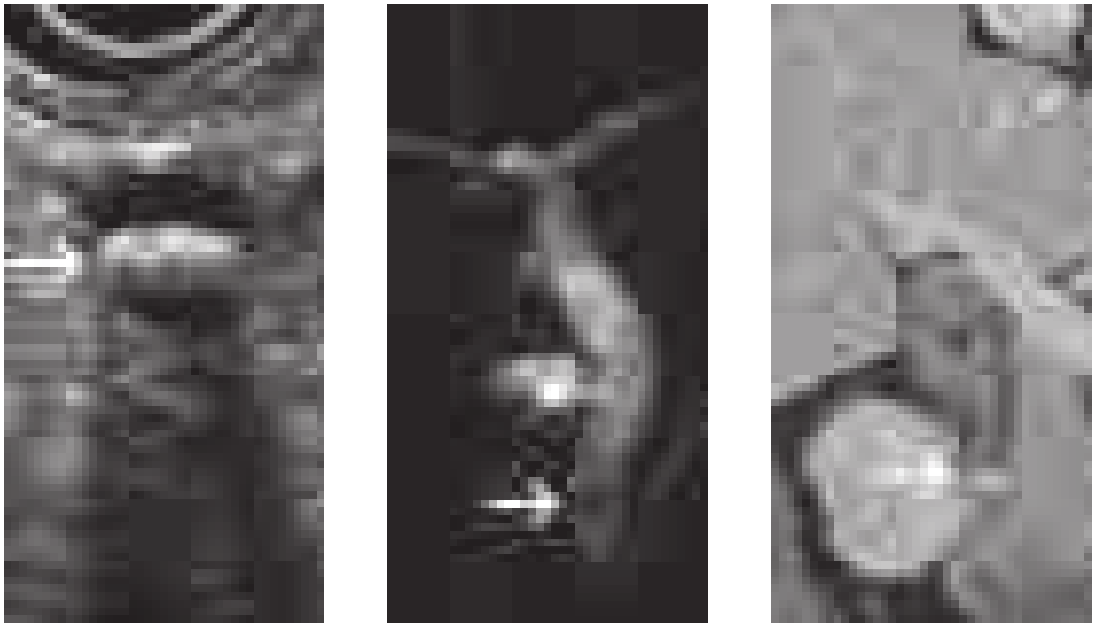


FIGURE 315-47 Methods of bile duct imaging. Arrows mark bile duct stones. ***A.*** Endoscopic ultrasound (EUS). ***B.*** Magnetic resonance cholangiopancreatography (MRCP). ***C.*** Helical computed tomography (CT).

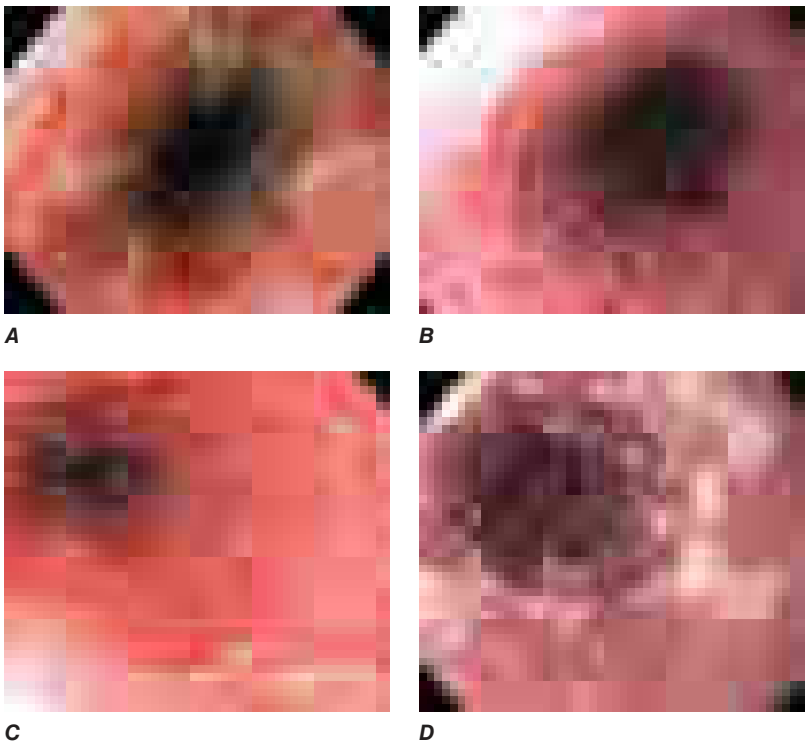


FIGURE 315-48 Causes of esophagitis. ***A.*** Severe reflux esophagitis with mucosal ulceration and friability. ***B.*** Cytomegalovirus esophagitis. ***C.*** Herpes simplex virus esophagitis with target-type shallow ulcerations. ***D.*** Candida esophagitis with white plaques adherent to the esophageal mucosa.

(*feline esophagus*) should raise suspicion for eosinophilic esophagitis, an increasingly recognized cause for recurrent dysphagia and food impaction **(Fig. 315-51)**. Blind or forceful passage of an endoscope may lead to perforation in a patient with stenosis of the cervical esophagus or a Zenker's diverticulum, but gentle passage of an endoscope under direct visual guidance is reasonably safe. Endoscopy can miss a subtle stricture or ring in some patients.

When transfer dysphagia is evident or an esophageal motility disorder is suspected, esophageal radiography and/or a video-swallow study are the best initial diagnostic tests. The oropharyngeal swallowing mechanism, esophageal peristalsis, and the lower esophageal sphincter can all be assessed. In some disorders, subsequent esophageal manometry may also be important for diagnosis.

Various causes of dysphagia are amenable to endoscopic therapy. Benign strictures, rings, and webs can be dilated using a through-the-scope balloon **(Fig. 315-52)** or a polyvinyl dilator passed over a guide wire. In some instances, thin fibrotic strictures may respond to needle-knife electroincision **(Fig. 315-53)** when they prove refractory to dilation. Esophageal covered stents can be used to palliate dysphagia from malignant obstruction **(Fig. 315-54)**, and flexible endoscopic myotomy is an option for Zenker's diverticulum **(Video V5-19)**. Recent advances in submucosal endoscopy have enabled the development of procedures, such as POEM **(Video V5-20)** and POET **(Video V5-21)** for the management of achalasia and select subepithelial esophageal tumors, respectively.

Although endoscopy is the most sensitive diagnostic test for peptic ulcer, it is not a cost-effective strategy in young patients with ulcer-like dyspeptic symptoms unless endoscopy is available at low cost. Patients with suspected peptic ulcer should be evaluated for *H. pylori* infection. Serology (past or present infection), urea breath testing (current infection), and stool tests are noninvasive and less costly than endoscopy with biopsy. Patients aged >50 and those with alarm symptoms or persistent symptoms despite treatment should undergo endoscopy to exclude malignancy.

NONULCER DYSPEPSIA

Nonulcer dyspepsia may be associated with bloating and, unlike peptic ulcer, tends not to remit and recur. Most patients describe persistent symptoms despite acid-reducing, prokinetic, or anti-*Helicobacter* therapy, and are referred for endoscopy to exclude a refractory ulcer and assess for other causes. Although endoscopy is useful for excluding other diagnoses, its impact on the treatment of patients with nonulcer dyspepsia is limited.

DYSPHAGIA

About 50% of patients presenting with difficulty swallowing have a mechanical obstruction; the remainder has a motility disorder, such as achalasia or diffuse esophageal spasm. Careful history-taking often points to a presumptive diagnosis and leads to the appropriate use of diagnostic tests. Esophageal strictures **(Fig. 315-49)** typically cause progressive dysphagia, first for solids, then for liquids; motility disorders often cause intermittent dysphagia for both solids and liquids. Some underlying disorders have characteristic historic features: Schatzki's ring **(Fig. 315-50)** causes episodic dysphagia for solids, typically at the beginning of a meal; oropharyngeal motor disorders typically present with difficulty initiating deglutition (*transfer dysphagia*) and nasal reflux or coughing with swallowing; and achalasia may cause nocturnal regurgitation of undigested food.

When mechanical obstruction is suspected, endoscopy is a useful initial diagnostic test, since it permits immediate biopsy and/or dilation of strictures, masses, or rings. The presence of linear furrows and multiple corrugated rings throughout a narrowed esophagus

ENDOSCOPIC TREATMENT OF OBESITY

The majority of Americans are overweight or obese, and obesity-associated diabetes has become a major public health problem. Bariatric surgery is the most effective weight-loss intervention, and it has been shown to decrease long-term mortality in obese persons, but many patients do not undergo surgery. Endoscopic treatments for obesity have been developed and include insertion of an intragastric balloon or duodenojejunal bypass liner, placement of a percutaneous gastric tube for aspiration of gastric contents after meals, or endoscopic sleeve gastroplasty, which utilizes endoscopic suturing to narrow the lumen of the gastric body **(Video V5-22)**. Prospective trials show that these treatments induce total body weight loss of 7–20% and varying degrees of glycemic control. Additional endoscopic modalities are undergoing

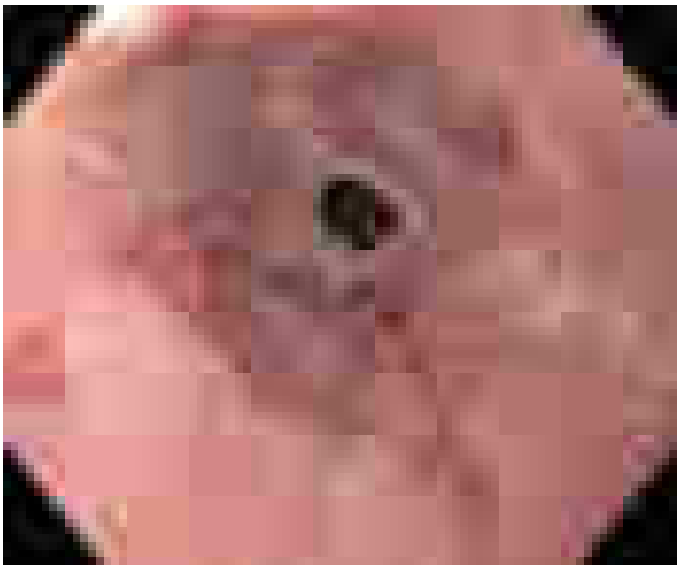


FIGURE 315-49 Peptic esophageal stricture associated with esophagitis.

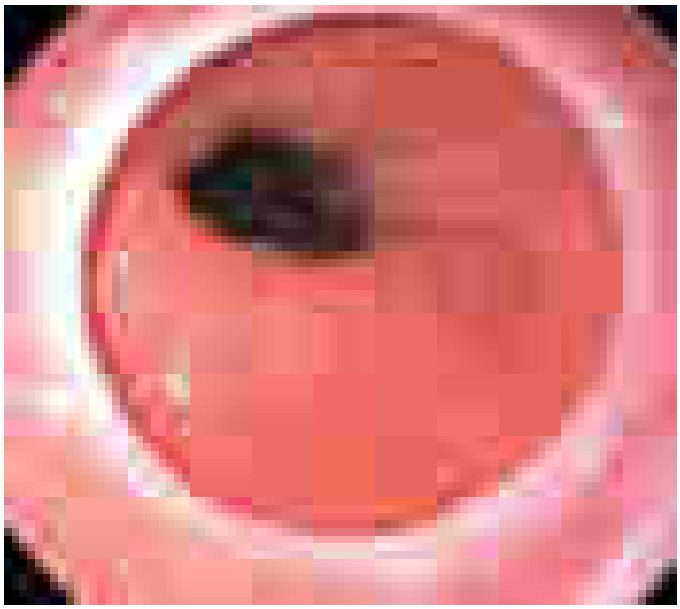


FIGURE 315-50 Schatzki's ring at the gastroesophageal junction.

initial clinical trials. The long-term efficacy of endoscopic bariatric treatment is currently unknown.

TREATMENT OF MALIGNANCIES

Endoscopy plays an important role in the treatment of gastrointestinal malignancies. Early-stage malignancies limited to the superficial layers of the gastrointestinal mucosa may be resected using the techniques of EMR (Video V5-4) or ESD (Video V5-5). Photodynamic therapy (PDT) and RFA are effective modalities for ablative treatment of high-grade dysplasia and intramucosal cancer in Barrett's esophagus **(Video V5-23)**. Gastrointestinal stromal tumors can be removed en bloc by EFTR (Video V5-3). In general, endoscopic techniques offer the advantage of a minimally invasive approach to treatment but rely on other imaging techniques (such as CT, MRI, positron emission tomography [PET], and EUS) to exclude distant metastases or locally advanced disease better treated by surgery or other modalities. The decision to treat an early-stage gastrointestinal malignancy endoscopically is often made in collaboration with a surgeon and/or oncologist.

Endoscopic palliation of gastrointestinal malignancies relieves symptoms and in many cases prolongs survival. Malignant obstruction can be relieved by endoscopic stent placement (**Figs.** 315-17, 315-54, and **315-55**; Videos V5-16, V5-17), and malignant gastrointestinal bleeding can often be palliated endoscopically as well. EUS-guided celiac plexus neurolysis may relieve pancreatic cancer pain.

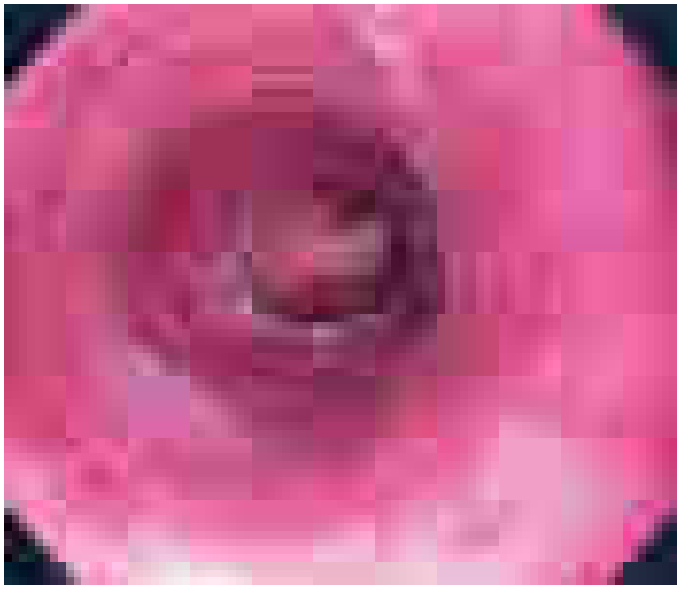


FIGURE 315-51 Eosinophilic esophagitis with multiple circular rings of the esophagus creating a corrugated appearance, and an impacted grape at the narrowed esophagogastric junction. The diagnosis requires biopsy with histologic finding of >15–20 eosinophils/high-power field.

ANEMIA AND OCCULT BLOOD IN THE STOOL

Iron-deficiency anemia may be attributed to poor iron absorption (as in celiac sprue) or, more commonly, chronic blood loss. Intestinal bleeding should be strongly suspected in men and postmenopausal women with iron-deficiency anemia, and colonoscopy is indicated in such patients, even in the absence of detectable occult blood in the stool. Approximately 30% will have large colonic polyps, 10% will have colorectal cancer, and a few patients will have colonic vascular lesions. When a convincing source of blood loss is not found in the colon, upper gastrointestinal endoscopy should be considered; if no lesion is found, duodenal biopsies should be obtained to exclude sprue **(Fig. 315-56)**. Small bowel evaluation with capsule endoscopy **(Fig. 315-57)**, CT or MR enterography, or device-assisted enteroscopy may be appropriate if both EGD and colonoscopy are unrevealing.

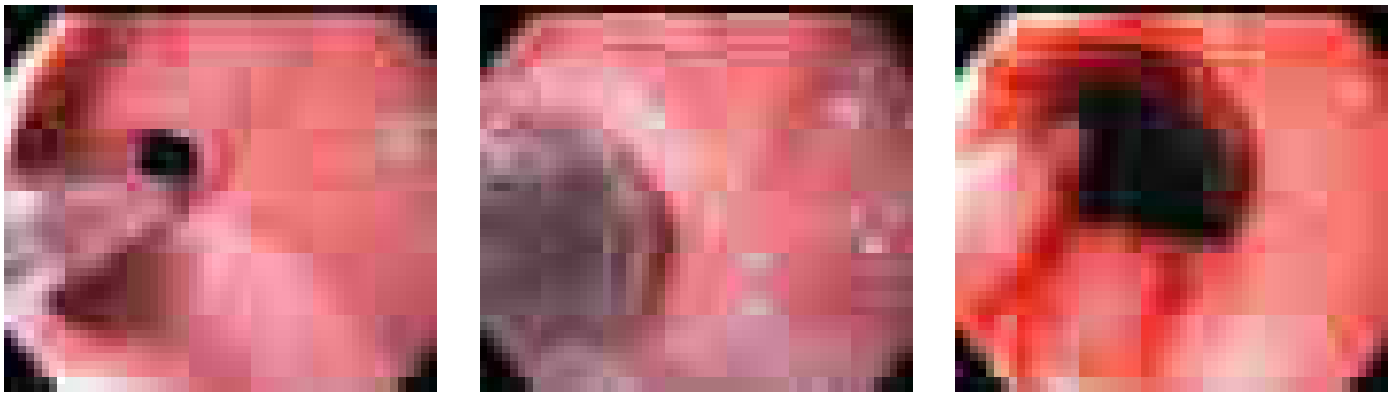
Tests for occult blood in the stool detect hemoglobin or the heme moiety and are most sensitive for colonic blood loss, although they will also detect larger amounts of upper gastrointestinal bleeding. Patients with occult blood in normal-appearing stool should undergo colonoscopy to diagnose or exclude colorectal neoplasia, especially if they are over 50 years of age or have a family history of colonic neoplasia. Whether upper endoscopy is also indicated depends on the patient's symptoms.

The small intestine may be the source of chronic intestinal bleeding, especially if colonoscopy and upper endoscopy are not diagnostic. The utility of small bowel evaluation varies with the clinical setting and is most important in patients in whom bleeding causes chronic or recurrent anemia. In contrast to the low diagnostic yield of small bowel radiography, positive findings on capsule endoscopy are seen in 50–70% of patients with suspected small intestinal bleeding. The most common finding is mucosal vascular ectasia. CT or MR enterography accurately detects small bowel masses and inflammation, and are also useful for initial small bowel evaluation. Deep enteroscopy may follow capsule endoscopy for biopsy of lesions or to provide specific therapy, such as argon plasma coagulation of vascular ectasias **(Fig. 315-58)**.

COLORECTAL CANCER SCREENING

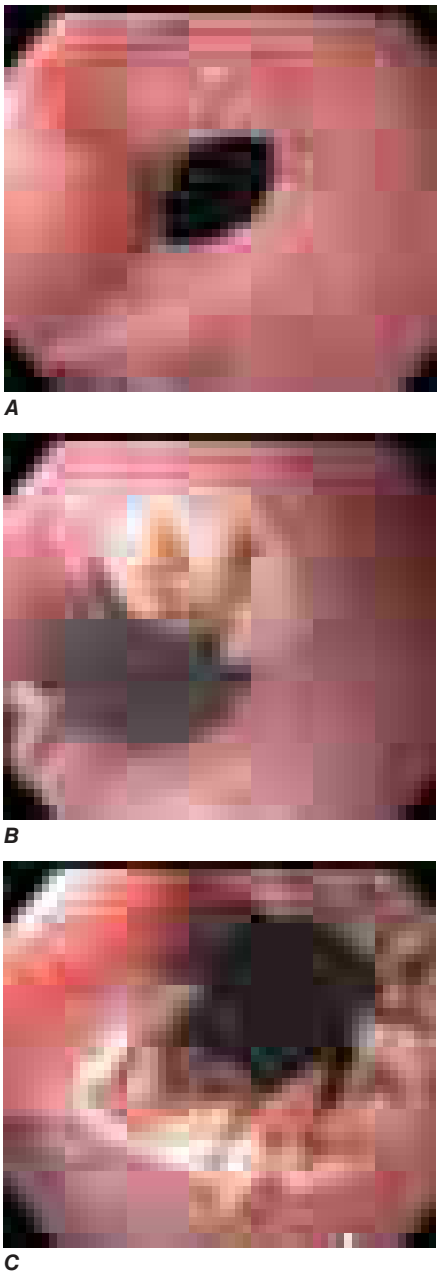
The majority of colon cancers develop from preexisting colonic adenomas, and colorectal cancer can be largely prevented by the detection and removal of adenomatous polyps **(Video V5-24)**. The choice of screening strategy for an asymptomatic person depends on personal and family history. Individuals with inflammatory bowel disease, a history of colorectal polyps or cancer, family members with adenomatous polyps or cancer, or certain familial cancer syndromes **(Fig. 315-59)** are at increased risk for colorectal cancer. An individual without these factors is generally considered at average risk.

Screening strategies are summarized in **Table 315-3**. While stool tests for occult blood have been shown to decrease mortality rate from colorectal cancer, they do not detect some cancers and many polyps, and direct visualization of the colon is a more effective screening strategy. Either sigmoidoscopy or colonoscopy may be used for cancer screening in asymptomatic average-risk individuals. The use of sigmoidoscopy was based on the historical finding that the majority of colorectal cancers occurred in the rectum and left colon, and that patients with right-sided colon cancers had left-sided polyps. Over the past several decades, however, the distribution of colon cancers has changed in the United States, with proportionally fewer rectal and left-sided cancers than in the past. Large American studies of colonoscopy for screening of average-risk individuals show that cancers are roughly



A B C

FIGURE 315-52 Endoscopic management of peptic stricture. ***A.*** Peptic stricture. ***B.*** Through-the-scope balloon dilation of stricture. ***C.*** Improvement in luminal diameter postdilation.



A B C

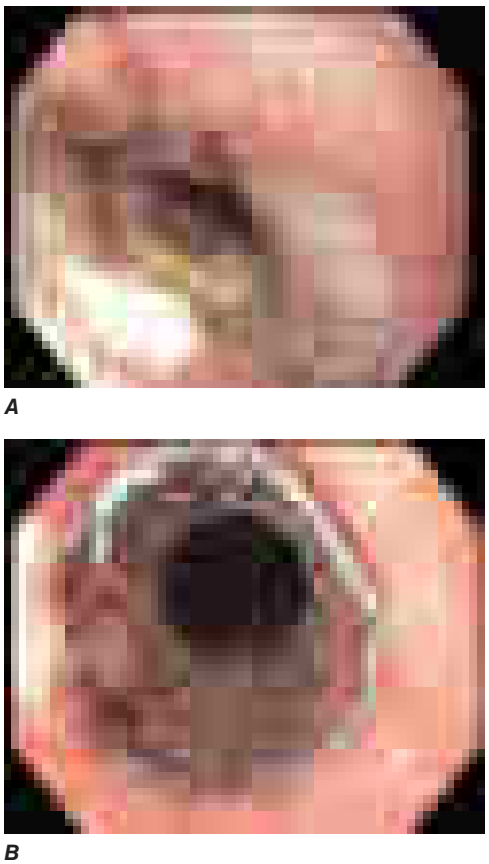
FIGURE 315-53 Endoscopic management of an esophagogastric anastomotic stricture. ***A.*** Recurrent anastomotic stricture despite periodic balloon dilation. ***B.*** Needle-knife electroincision of stricture. ***C.*** Improvement in luminal opening post therapy.

equally distributed between left and right colon and half of patients with right-sided lesions have no polyps in the left colon. Visualization of the entire colon thus appears to be the optimal strategy for colorectal cancer screening and prevention.

Virtual colonoscopy (VC) is a radiologic technique that images the colon with CT following rectal insufflation of the colonic lumen. Computer rendering of CT images generates an electronic display of a virtual "flight" along the colonic lumen, simulating colonoscopy (Fig. 315-60). Findings detected during VC often require subsequent conventional colonoscopy for confirmation and treatment.

DIARRHEA

Most cases of diarrhea are acute, self-limited, and due to infections or medication. Chronic diarrhea (lasting >6 weeks) is more often due to a



A B

FIGURE 315-54 Palliation of malignant dysphagia. ***A.*** Obstructing distal esophageal cancer. ***B.*** Palliative stent placement.

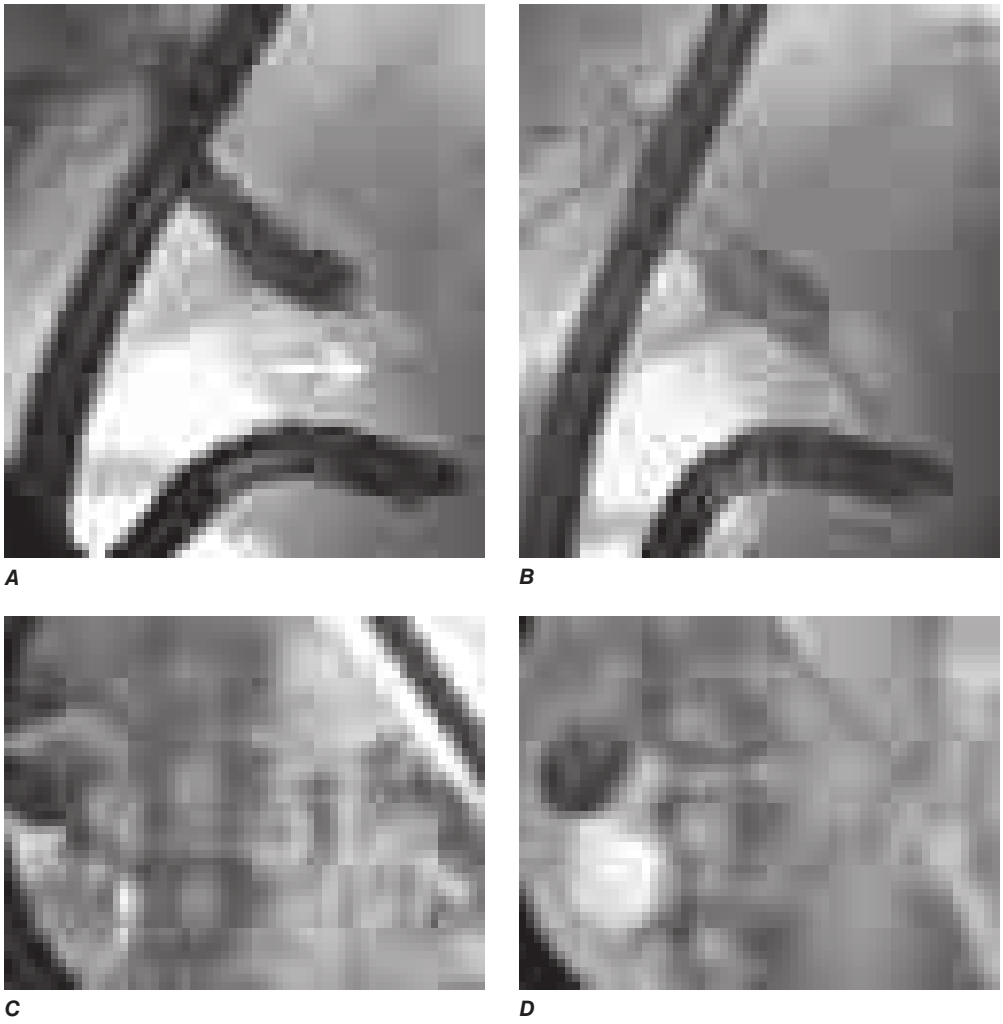


FIGURE 315-55 Biliary and duodenal self-expanding metal stents (SEMS) for obstruction caused by pancreatic cancer. ***A.*** Endoscopic retrograde cholangiopancreatography (ERCP) demonstrates a distal bile duct stricture *(arrow)*. ***B.*** A biliary SEMS is placed. ***C.*** Contrast injection demonstrates a duodenal stricture *(arrow)*. ***D.*** Biliary and duodenal SEMS in place.

primary inflammatory, malabsorptive, or motility disorder, is less likely to resolve spontaneously, and generally requires diagnostic evaluation. Patients with chronic diarrhea or severe, unexplained acute diarrhea often undergo endoscopy if stool tests for pathogens are unrevealing. The choice of endoscopic testing depends on the clinical setting.

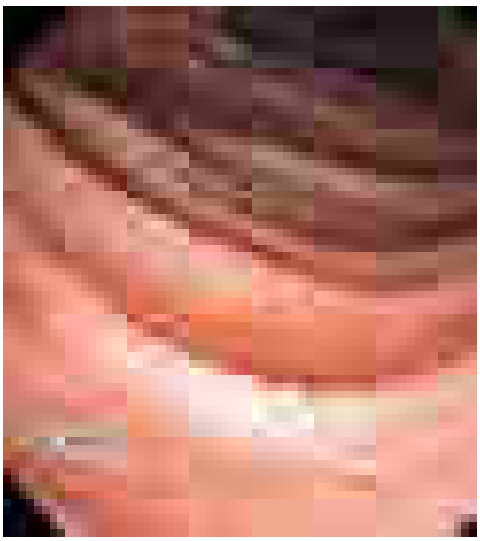


FIGURE 315-56 Scalloped duodenal folds in a patient with celiac sprue.

Patients with colonic symptoms and findings such as bloody diarrhea, tenesmus, fever, or leukocytes in stool generally undergo sigmoidoscopy or colonoscopy to assess for colitis (Fig. 315-8). Sigmoidoscopy is an appropriate initial test in most patients. Conversely, patients with symptoms and findings suggesting small bowel disease, such as large-volume watery stools, substantial weight loss, and malabsorption of iron, calcium, or fat, may undergo upper endoscopy with duodenal aspirates for assessment of bacterial overgrowth and biopsies for assessment of mucosal diseases, such as celiac sprue.

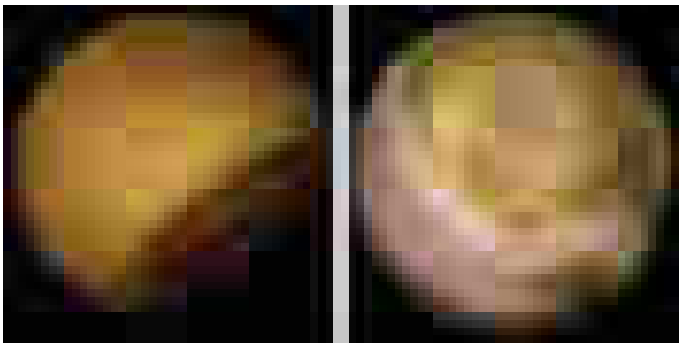
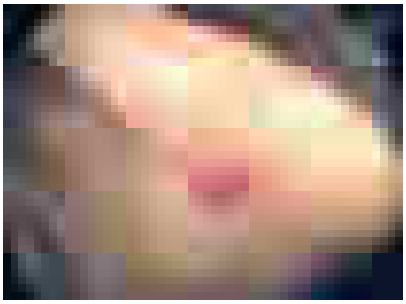
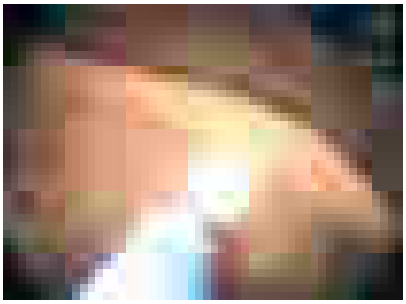


FIGURE 315-57 Capsule endoscopy images of a mildly scalloped jejunal fold (*left*) and an ileal tumor (*right*) in a patient with celiac sprue. *(Images courtesy of Dr. Elizabeth Rajan; with permission.)*



A



B

FIGURE 315-58 ***A.*** Mid-jejunal vascular ectasia identified by double-balloon endoscopy. ***B.*** Ablation of vascular ectasia with argon plasma coagulation.

Many patients with chronic diarrhea do not fit either of these patterns. In the setting of a long-standing history of alternating constipation and diarrhea dating to early adulthood, without findings such as blood in the stool or anemia, a diagnosis of irritable bowel syndrome may be made without direct visualization of the bowel. Steatorrhea and upper abdominal pain may prompt evaluation of the pancreas rather than the gut. Patients whose chronic diarrhea is not easily categorized often undergo initial colonoscopy to examine the entire colon and terminal ileum for inflammatory or neoplastic disease **(Fig. 315-61)**.

MINOR HEMATOCHEZIA

Bright red blood passed with or on formed brown stool usually has a rectal, anal, or distal sigmoid source **(Fig. 315-62)**. Patients with even trivial amounts of hematochezia should be investigated with flexible sigmoidoscopy and anoscopy to exclude polyps or cancers in the distal colon. Patients reporting red blood on the toilet tissue only, without blood in the toilet or on the stool, are generally bleeding from a lesion in the anal canal; careful external inspection, digital examination, and proctoscopy with anoscopy may be sufficient for diagnosis in such cases.

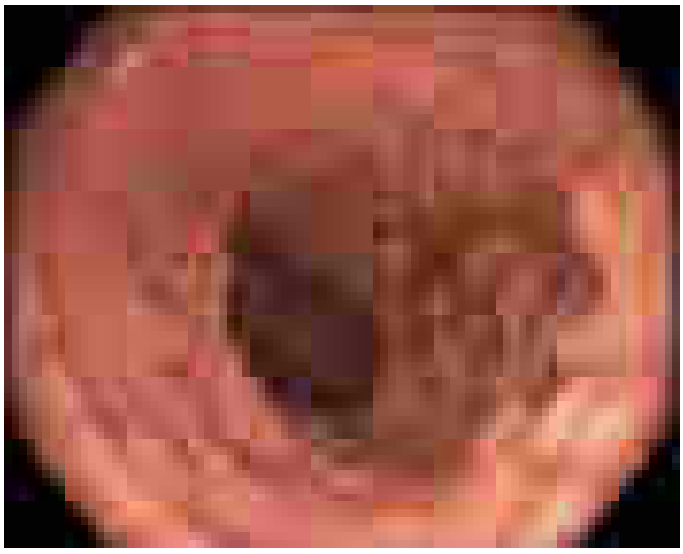


FIGURE 315-59 **Numerous colon polyps** in a patient with familial adenomatous polyposis syndrome.

PANCREATITIS

About 20% of patients with pancreatitis have no identified cause after routine clinical investigation (including a review of medication and alcohol use, measurement of serum triglyceride and calcium levels, abdominal ultrasonography, and CT or MR). Endoscopic assessment leads to a specific diagnosis in the majority of such patients, often altering clinical management. Endoscopic investigation is particularly appropriate if the patient has had more than one episode of pancreatitis.

Microlithiasis, or the presence of microscopic crystals in bile, is a leading cause of previously unexplained acute pancreatitis and is sometimes seen during abdominal ultrasonography as layering sludge or flecks of floating, echogenic material in the gallbladder. EUS may identify previously undetected microlithiasis.

Previously undetected chronic pancreatitis, pancreatic malignancy, or pancreas divisum may be diagnosed by either ERCP or EUS. Autoimmune pancreatitis is often suspected on the basis of CT, MR, or serologic findings, but it may first become apparent during EUS and may require EUS-guided pancreatic biopsy for histologic diagnosis.

Severe pancreatitis often results in pancreatic fluid collections. Symptomatic pseudocysts and areas of walled-off pancreatic necrosis can be drained into the stomach or duodenum endoscopically, using transpapillary and transmural endoscopic techniques. Pancreatic necrosis can be treated by direct endoscopic necrosectomy (Video V5-2) via an endoscopically created transmural drainage site.

CANCER STAGING

Local staging of esophageal, gastric, pancreatic, bile duct, and rectal cancers can be obtained with EUS (Fig. 315-19). EUS with fine-needle aspiration (Fig. 315-20) currently provides the most accurate preoperative assessment of local tumor and nodal staging, but it does not detect many distant metastases. Details of the local tumor stage can guide treatment decisions including resectability and need for neoadjuvant therapy. EUS with transesophageal needle biopsy may also be used to assess the presence of non-small-cell lung cancer in mediastinal nodes.

OPEN-ACCESS ENDOSCOPY

Direct scheduling of endoscopic procedures by primary care physicians without preceding gastroenterology consultation, or *open-access endoscopy*, is common. When the indications for endoscopy are clear-cut and appropriate, the procedural risks are low, and the patient understands what to expect, open-access endoscopy streamlines patient care and decreases costs.

Patients referred for open-access endoscopy should have a recent history, physical examination, and medication review. A copy of such an evaluation should be available when the patient comes to the endoscopy suite. Patients with unstable cardiovascular or respiratory conditions should not be referred directly for open-access endoscopy. Patients with particular conditions and undergoing certain procedures should be prescribed prophylactic antibiotics prior to endoscopy (Table 315-1). In addition, patients taking anticoagulants and/or antiplatelet drugs may require adjustment of these agents before endoscopy based on the procedural risk for bleeding and their underlying risk for a thromboembolic event (Table 315-2).

Common indications for open-access EGD include dyspepsia resistant to a trial of appropriate therapy, dysphagia, gastrointestinal bleeding, and persistent anorexia or early satiety. Open-access colonoscopy is often requested in men or postmenopausal women with iron-deficiency anemia, in patients age >50 with occult blood in the stool, in patients with a previous history of colorectal adenomatous

TABLE 315-3 Colorectal Cancer Screening Strategies

	CHOICES/RECOMMENDATIONS	COMMENTS
Average-Risk Patients		
Asymptomatic individuals ≥50 years of age (≥45 years of age for African Americans)	Colonoscopy every 10 years[a]	Preferred cancer prevention strategy
	Annual FIT or FOBT, multiple take-home specimen cards	Cancer detection strategy, does not detect most polyps; colonoscopy if results are positive
	CT colonography every 5 years	Colonoscopy if results are positive
	Flexible sigmoidoscopy every 5 years	Does not detect proximal colon polyps and cancers; colonoscopy if results are positive
	Double-contrast barium enema every 5 years	Less sensitive than colonoscopy or CT colonography, misses some cancers and polyps; colonoscopy if results are positive
	Stool DNA test every 3 years	Does not detect many polyps; colonoscopy if results are positive
Personal History of Polyps or CRC		
1 or 2 small (<1 cm) adenomas with low-grade dysplasia	Repeat colonoscopy in 5–10 years	Assuming complete polyp resection. Interval may vary based on prior history, family history
3–10 adenomas, or any high-risk adenoma[b]	Repeat colonoscopy in 3 years; subsequent colonoscopy based on findings	Assuming complete polyp resection
>10 adenomas	Repeat colonoscopy in <3 years based on clinical judgment	Consider evaluation for FAP or HNPCC; see recommendations below
Piecemeal removal of a sessile polyp	Exam in 2–6 months to verify complete removal	
Small (<1 cm) hyperplastic polyps of sigmoid and rectum	Repeat colonoscopy in 10 years	Those with hyperplastic polyposis syndrome merit more frequent follow-up
Sessile serrated adenoma/polyp <10 mm, without dysplasia	Repeat colonoscopy in 5 years	
Sessile serrated adenoma/polyp ≥10 mm or with dysplasia, or ≥2 serrated polyps	Repeat colonoscopy in 3 years	Serrated polyposis syndrome merits more frequent follow-up
Incompletely removed serrated polyp ≥1 cm	Exam in 2–6 months to verify complete removal	
Colon cancer	Evaluate entire colon around the time of resection, then repeat colonoscopy in 1 year	Subsequent colonoscopy in 3 years if the 1-year examination is normal
Inflammatory Bowel Disease		
Long-standing (>8 years) ulcerative pancolitis or Crohn's colitis, or left-sided ulcerative colitis of >15 years' duration	Colonoscopy with biopsies every 1–2 years	Consider chromoendoscopy or other advanced imaging techniques for detection of flat dysplasia during colonoscopy
Family History of Polyps or CRC		
First-degree relatives with only small tubular adenomas	Same as average risk	
Single first-degree relative with CRC or advanced adenoma at age ≥60 years	Colonoscopy every 10 years starting at age 40	
Single first-degree relative with CRC or advanced adenoma at age <60 years, OR two first-degree relatives with CRC or advanced adenomas at any age	Colonoscopy every 5 years beginning at age 40 years or 10 years younger than age at diagnosis of the youngest affected relative, whichever is earlier	
FAP	Sigmoidoscopy or colonoscopy annually, beginning at age 10–12 years	Consider genetic counseling and testing
HNPCC	Colonoscopy every 2 years beginning at age 20–25 years (or 10 years younger than the youngest affected first-degree relative) until age 40, then annually thereafter	Consider histologic evaluation for microsatellite instability in tumor specimens of patients who meet modified Bethesda criteria; consider genetic counseling and testing

[a]Assumes good colonic preparation and complete examination to cecum. [b]High-risk adenoma: any adenoma ≥ 1 cm in size, or containing high-grade dysplasia or villous features

Abbreviations: CRC, colorectal cancer; FAP, familial adenomatous polyposis; FIT, fecal immunochemical test; FOBT, fecal occult blood test; HNPCC, hereditary nonpolyposis colorectal cancer.

Source: Adapted from United States Preventative Services Task Force Guidelines released in 2016 *(https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/colorectal-cancer-screening2?ds=1&s=colorectal)* accessed on Jan 2, 2017, and American Cancer Society Guidelines updated in 2016 (*http://www.cancer.org/cancer/colonandrectumcancer/moreinformation/colonandrectumcancerearlydetection/colorectal-cancer-early-detection-acs-recommendations*) accessed Jan 2, 2017.

polyps or cancer, and for colorectal cancer screening. Flexible sigmoidoscopy is commonly performed as an open-access procedure.

When patients are referred for open-access colonoscopy, the primary care provider may need to choose a colonic preparation. Commonly used oral preparations include polyethylene glycol lavage solution, with or without citric acid. A "split-dose" regimen improves the quality of colonic preparation. Sodium phosphate purgatives may cause fluid and electrolyte abnormalities and renal toxicity, especially in patients with renal failure or congestive heart failure and those >70 years of age.

FIGURE 315-60 Virtual colonoscopy image of a colon polyp (*arrow*). *(Image courtesy of Dr. Jeff Fidler; with permission.)*

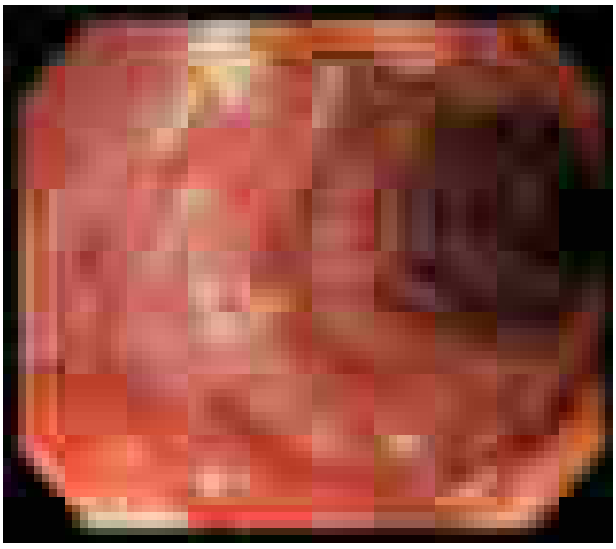


FIGURE 315-61 Crohn's ileitis.

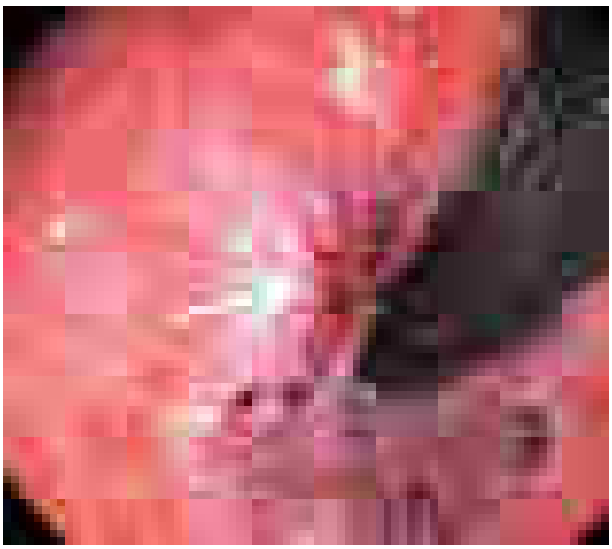


FIGURE 315-62 Internal hemorrhoids with bleeding *(arrow)* as seen on a retroflexed view of the rectum.

FURTHER READING

ASGE Standards of Practice Committee et al: Appropriate use of GI endoscopy. Gastrointest Endosc 75:1127, 2012.

ASGE Standards of Practice Committee et al: Antibiotic prophylaxis for GI endoscopy. Gastrointest Endosc 81:81, 2015.

ASGE Standards of Practice Committee et al: Open-access endoscopy. Gastrointest Endosc 81:1326, 2015.

Garcia-Tsao G et al: Portal hypertensive bleeding in cirrhosis: Risk stratification, diagnosis, and management: 2016 practice guidance by the American Association for the study of liver diseases. Hepatology 65:310, 2017.

Gralnek IM et al: Diagnosis and management of nonvariceal upper gastrointestinal hemorrhage: European Society of Gastrointestinal Endoscopy (ESGE) Guideline. Endoscopy 47:a1, 2015.

Lin JS et al: Screening for colorectal cancer: Updated evidence report and systematic review. JAMA 315:2576, 2016.

Shaheen NJ et al: ACG Clinical Guideline: Diagnosis and management of Barrett's esophagus. Am J Gastroenterol 111:30, 2016.

Strate LL et al: ACG Clinical Guideline: Management of patients with acute lower gastrointestinal bleeding. Am J Gastroenterol 111:459, 2016.

316 Diseases of the Esophagus

Peter J. Kahrilas, Ikuo Hirano

ESOPHAGEAL STRUCTURE AND FUNCTION

The esophagus is a hollow, muscular tube coursing through the posterior mediastinum joining the hypopharynx to the stomach with a sphincter at each end. It functions to transport food and fluid between these ends, otherwise remaining empty. The physiology of swallowing, esophageal motility, and oral and pharyngeal dysphagia are described in Chap. 40. Esophageal diseases can be manifested by impaired function or pain. Key functional impairments are swallowing disorders and excessive gastroesophageal reflux. Pain, sometimes indistinguishable from cardiac chest pain, can result from inflammation, infection, dysmotility, or neoplasm.

SYMPTOMS OF ESOPHAGEAL DISEASE

The clinical history remains central to the evaluation of esophageal symptoms. A thoughtfully obtained history will often expedite management. Important details include weight gain or loss, gastrointestinal bleeding, dietary habits including the timing of meals, smoking, and alcohol consumption. The major esophageal symptoms are heartburn, regurgitation, chest pain, dysphagia, odynophagia, and globus sensation.

Heartburn (pyrosis), the most common esophageal symptom, is characterized by a discomfort or burning sensation behind the sternum that arises from the epigastrium and may radiate toward the neck. Heartburn is an intermittent symptom, most commonly experienced after eating, during exercise, and while lying recumbent. The discomfort is relieved with drinking water or antacid but can occur frequently interfering with normal activities including sleep. The association between heartburn and gastroesophageal reflux disease (GERD) is so strong that empirical therapy for GERD has become accepted management. However, the term "heartburn" is often misused and/or referred to with other terms such as "indigestion" or "repeating," making it important to clarify the intended meaning.

Regurgitation is the effortless return of food or fluid into the pharynx without nausea or retching. Patients report a sour or burning fluid in the throat or mouth that may also contain undigested food particles. Bending, belching, or maneuvers that increase intraabdominal pressure can provoke regurgitation. A clinician needs to discriminate

among regurgitation, vomiting, and rumination. *Vomiting* is preceded by nausea and accompanied by retching. *Rumination* is a behavior in which recently swallowed food is regurgitated and then reswallowed repetitively for up to an hour. Although there is some linkage between rumination and cognitive deficiency, the behavior is also exhibited by unimpaired individuals.

Chest pain is a common esophageal symptom with characteristics similar to cardiac pain, sometimes making this distinction difficult. Esophageal pain is usually experienced as a pressure type sensation in the mid chest, radiating to the mid back, arms, or jaws. The similarity to cardiac pain is likely because the two organs share a nerve plexus and the nerve endings in the esophageal wall have poor discriminative ability among stimuli. Esophageal distention or even chemostimulation (e.g., with acid) will often be perceived as chest pain. Gastroesophageal reflux is the most common cause of esophageal chest pain.

Esophageal *dysphagia* **(Chap. 40)** is often described as a feeling of food "sticking" or even lodging in the chest. Important distinctions are between uniquely solid food dysphagia as opposed to liquid and solid, episodic versus constant dysphagia, and progressive versus static dysphagia. If the dysphagia is for liquids as well as solid food, it suggests a motility disorder such as achalasia. Conversely, uniquely solid food dysphagia is suggestive of a stricture, ring, or tumor. Of note, a patient's localization of food hang-up in the esophagus is notoriously imprecise. Approximately 30% of distal esophageal obstructions are perceived as cervical dysphagia. In such instances, the absence of concomitant symptoms generally associated with oropharyngeal dysphagia such as aspiration, nasopharyngeal regurgitation, cough, drooling, or obvious neuromuscular compromise should suggest an esophageal etiology.

Odynophagia is pain either caused by or exacerbated by swallowing. Although typically considered distinct from dysphagia, odynophagia may manifest concurrently with dysphagia. Odynophagia is more common with pill or infectious esophagitis than with reflux esophagitis and should prompt a search for these entities. When odynophagia does occur in GERD, it is likely related to an esophageal ulcer or extensive erosions.

Globus sensation, also known as globus pharyngeus, is the perception of a lump or fullness in the throat that is felt irrespective of swallowing. Although such patients are frequently referred for an evaluation of dysphagia, globus sensation is often relieved by the act of swallowing. As implied by its alternative name, "globus hystericus," globus sensation often occurs in the setting of anxiety or obsessive-compulsive disorders. Clinical experience teaches that it is often attributable to GERD.

Water brash is excessive salivation resulting from a vagal reflex triggered by acidification of the esophageal mucosa. This is not a common symptom. Afflicted individuals will describe the unpleasant sensation of the mouth rapidly filling with salty thin fluid, often in the setting of concomitant heartburn.

DIAGNOSTIC STUDIES

ENDOSCOPY

Endoscopy, also known as esophagogastroduodenoscopy (EGD), is the most useful test for the evaluation of the proximal gastrointestinal tract. Modern instruments produce high-quality, color images of the esophageal, gastric, and duodenal lumen. Endoscopes also have an instrumentation channel through which biopsy forceps, injection catheters for local delivery of therapeutic agents, balloon dilators, or hemostatic devices can be used. The key advantages of endoscopy over barium radiography are: (1) increased sensitivity for the detection of mucosal lesions, (2) vastly increased sensitivity for the detection of abnormalities mainly identifiable by color such as Barrett's metaplasia or vascular lesions, (3) the ability to obtain biopsy specimens for histologic examination of suspected abnormalities, and (4) the ability to dilate strictures during the examination. The main disadvantages of endoscopy are low sensitivity for detection of diffuse, non-focal esophageal strictures, cost, and the utilization of sedatives or anesthetics.

RADIOGRAPHY

Contrast radiography of the esophagus, stomach, and duodenum can demonstrate reflux of the contrast media, hiatal hernia, mucosal granularity, erosions, ulcerations, and strictures. The sensitivity of radiography compared with endoscopy for detecting reflux esophagitis reportedly ranges from 22 to 95%, with higher grades of esophagitis (i.e., ulceration or stricture) exhibiting greater detection rates. Conversely, the sensitivity of barium radiography for detecting esophageal strictures is greater than that of endoscopy, especially when the study is done in conjunction with a 13-mm barium tablet. Barium studies also provide an assessment of esophageal function and morphology that may be undetected on endoscopy. Tracheoesophageal fistula, altered postsurgical anatomy, and extrinsic esophageal compression are conditions where radiographic imaging complements endoscopic assessment. Hypopharyngeal pathology and disorders of the cricopharyngeus muscle are better appreciated on radiographic examination than with endoscopy, particularly with rapid sequence or video fluoroscopic recording. The major shortcoming of barium radiography is that it rarely obviates the need for endoscopy. Either a positive or a negative study is usually followed by an endoscopic evaluation either to obtain biopsies, provide therapy, or clarify findings in the case of a positive examination or to add a level of certainty in the case of a negative one.

ENDOSCOPIC ULTRASOUND

Endoscopic ultrasound (EUS) instruments combine an endoscope with an ultrasound transducer to create a transmural image of the tissue surrounding the endoscope tip. The key advantage of EUS over alternative radiologic imaging techniques is much greater resolution attributable to the proximity of the ultrasound transducer to the area being examined. Available devices can provide either radial imaging (360-degree, cross-sectional) or a curved linear image that can guide fine-needle aspiration of imaged structures such as lymph nodes or tumors. Major esophageal applications of EUS are to stage esophageal cancer, to evaluate dysplasia in Barrett's esophagus, and to assess submucosal lesions.

ESOPHAGEAL MANOMETRY

Esophageal manometry, or motility testing, entails positioning a pressure-sensing catheter within the esophagus and then observing the contractility following test swallows. The upper and lower esophageal sphincters (LESs) appear as zones of high pressure that relax on swallowing, while the intersphincteric esophagus exhibits peristaltic contractions. Manometry is used to diagnose motility disorders (achalasia, diffuse esophageal spasm [DES]) and to assess peristaltic integrity prior to the surgery for reflux disease. Technologic advances have enhanced esophageal manometry as high-resolution esophageal pressure topography **(Fig. 316-1)**. Manometry can also be combined with intraluminal impedance monitoring. Impedance recordings use a series of paired electrodes added to the manometry catheter. Esophageal luminal contents in contact with the electrodes decrease (liquid) or increase (air) the impedance signal, allowing detection of anterograde or retrograde esophageal bolus transit.

REFLUX TESTING

GERD is often diagnosed in the absence of endoscopic esophagitis, which would otherwise define the disease. This occurs in the settings of partially treated disease, an abnormally sensitive esophageal mucosa, or without obvious explanation. In such instances, reflux testing can demonstrate excessive esophageal exposure to refluxed gastric juice, the physiologic abnormality of GERD. This can be done by ambulatory 24- to 96-h esophageal pH recording using either a wireless pH-sensitive transmitter that is affixed to the esophageal mucosa or a transnasally positioned wire electrode with the tip stationed in the distal esophagus. Either way, the outcome is expressed as the percentage of the day that the pH was <4 (indicative of recent acid reflux), with values exceeding 5% indicative of GERD. Reflux testing is useful in the evaluation of patients presenting with atypical symptoms or an inexplicably poor response to therapy. Intraluminal impedance monitoring can be added to pH monitoring to detect reflux events irrespective of whether or not they are acidic, potentially increasing the sensitivity of the study.

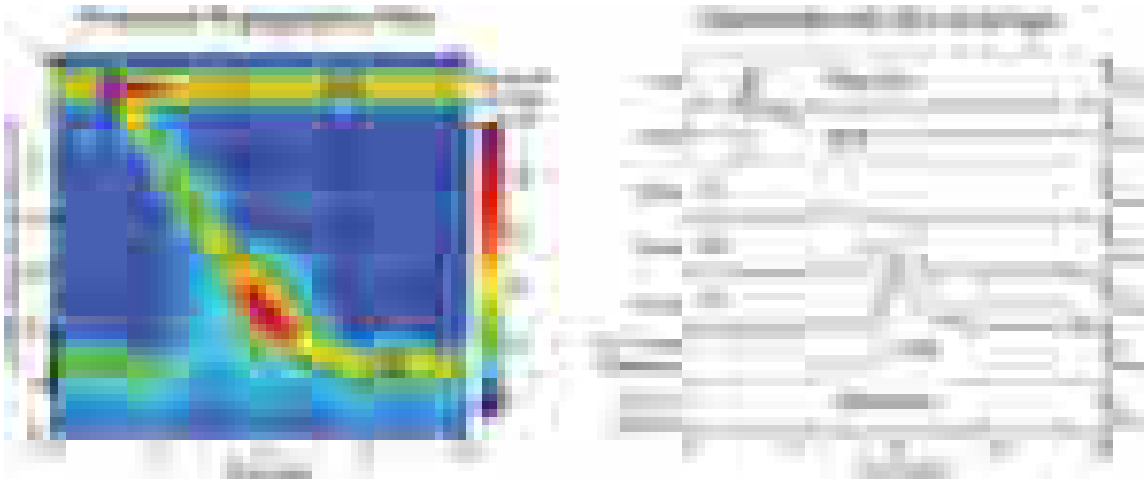


FIGURE 316-1 High-resolution esophageal pressure topography (*right*) and conventional manometry (*left*) of a normal swallow. E, esophageal body; LES, lower esophageal sphincter; UES, upper esophageal sphincter.

STRUCTURAL DISORDERS

HIATAL HERNIA

Hiatus hernia is a herniation of viscera, most commonly the stomach, into the mediastinum through the esophageal hiatus of the diaphragm. Four types of hiatus hernia are distinguished with type I, or sliding hiatal hernia, comprising at least 95% of the overall total. A sliding hiatal hernia is one in which the gastroesophageal junction and gastric cardia translocate cephalad as a result of weakening of the phrenoesophageal ligament attaching the gastroesophageal junction to the diaphragm at the hiatus and dilatation of the diaphragmatic hiatus. The incidence of sliding hernia increases with age. True to its name, sliding hernias enlarge with increased intraabdominal pressure, swallowing, and respiration. Conceptually, sliding hernias are the result of wear and tear: increased intraabdominal pressure from abdominal obesity, pregnancy, etc., along with hereditary factors predisposing to the condition. The main significance of sliding hernias is the propensity of affected individuals to have GERD.

Type II, III, and IV hiatal hernias are all subtypes of paraesophageal hernia in which the herniation into the mediastinum includes a visceral structure other than the gastric cardia. With type II and III paraesophageal hernias, the gastric fundus also herniates with the distinction being that in type II, the gastroesophageal junction remains fixed at the hiatus, whereas type III is a combined sliding and paraesophageal hernia. With type IV hiatal hernias, viscera other than the stomach herniate into the mediastinum, most commonly the colon. With type II and III paraesophageal hernias, the stomach inverts as it herniates and large paraesophageal hernias can lead to an "upside down stomach," gastric volvulus, and even strangulation of the stomach. Because of this risk, surgical repair is often advocated for large paraesophageal hernias particularly when they are symptomatic.

RINGS AND WEBS

A lower esophageal mucosal ring, also called a *B ring*, is a thin membranous narrowing at the squamocolumnar mucosal junction **(Fig. 316-2)**. Its origin is unknown, but B rings are demonstrable in about 10–15% of the general population and are usually asymptomatic. When the lumen diameter is <13 mm, distal rings are usually associated with episodic solid food dysphagia and are called *Schatzki rings*. Patients typically present older than 40 years, consistent with an acquired rather than congenital origin. Schatzki ring is one of the most common causes of intermittent food impaction, also known as "steakhouse syndrome" because meat is a typical instigator. Symptomatic rings are readily treated by dilation.

Web-like constrictions higher in the esophagus can be of congenital or inflammatory origin. Asymptomatic cervical esophageal webs are demonstrated in about 10% of people and typically originate along the anterior aspect of the esophagus. When circumferential, they can cause intermittent dysphagia to solids similar to Schatzki rings and are similarly treated with dilation. The combination of symptomatic proximal esophageal webs and iron-deficiency anemia in middle-aged women constitutes Plummer-Vinson syndrome.

DIVERTICULA

Esophageal diverticula are categorized by location with the most common being epiphrenic, hypopharyngeal (Zenker's), and midesophageal. Epiphrenic and Zenker's diverticula are false diverticula involving herniation of the mucosa and submucosa through the muscular layer of the esophagus. These lesions result from increased intraluminal pressure associated with distal obstruction. In the case of Zenker's, the obstruction is a stenotic cricopharyngeus muscle (upper esophageal sphincter), and the hypopharyngeal herniation most commonly occurs in an area of natural weakness proximal to the cricopharyngeus known as *Killian's triangle* **(Fig. 316-3)**. Small Zenker's diverticula are usually asymptomatic, but when they enlarge sufficiently to retain food and saliva they can be associated with dysphagia, halitosis, and aspiration. Treatment is by surgical diverticulectomy and cricopharyngeal

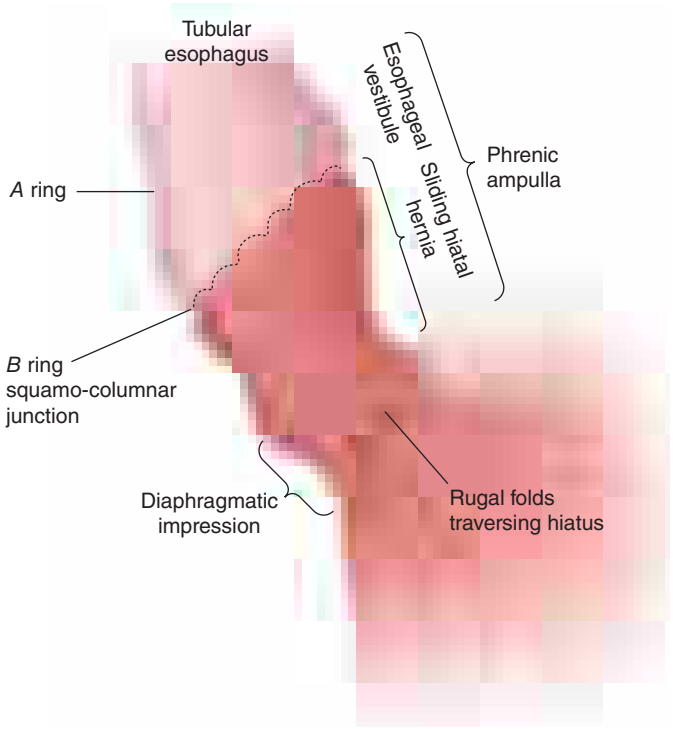


FIGURE 316-2 Radiographic anatomy of the gastroesophageal junction.

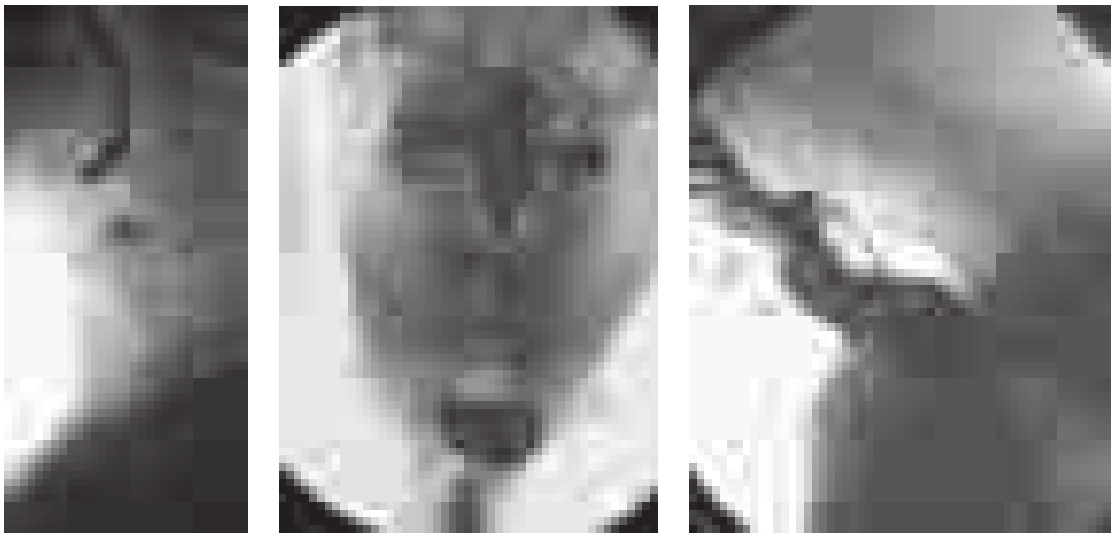


FIGURE 316-3 Examples of small (*A*) and large (*B*, *C*) Zenker's diverticula arising from Killian's triangle in the distal hypopharynx. Smaller diverticula are evident only during the swallow, whereas larger ones retain food and fluid.

myotomy or a marsupialization procedure in which an endoscopic stapling device is used to divide the cricopharyngeus.

Epiphrenic diverticula are often associated with achalasia, esophageal hypercontractile disorders, or a distal esophageal stricture. Midesophageal diverticula may be caused by traction from adjacent inflammation (classically tuberculosis) in which case they are true diverticula involving all layers of the esophageal wall, or by pulsion associated with esophageal motor disorders. Midesophageal and epiphrenic diverticula are usually asymptomatic until they enlarge sufficiently to retain food and cause dysphagia and regurgitation. Symptoms attributable to the diverticula tend to correlate more with the underlying esophageal disorder than the size of the diverticula. Large diverticula can be removed surgically, usually in conjunction with a myotomy if the underlying motility disorder is identified. Diffuse intramural esophageal pseudodiverticulosis is a rare entity that results from dilatation of the excretory ducts of submucosal esophageal glands **(Fig. 316-4)**. Esophageal candidiasis and proximal esophageal strictures are commonly found in association with this disorder.

TUMORS

Esophageal cancer occurs in about 4.5:100,000 people in the United States with the associated mortality being only slightly less at 4.4:100,000. It is about 10 times less common than colorectal cancer but kills about one-quarter as many patients. These statistics emphasize both the rarity and lethality of esophageal cancer. One notable trend is the shift of dominant esophageal cancer type from squamous cell to adenocarcinoma, strongly linked to reflux disease and Barrett's metaplasia. Other distinctions between cell types are the predilection for adenocarcinoma to affect the distal esophagus in white males and squamous cell to affect the more proximal esophagus in black males with the added risk factors of smoking, alcohol consumption, caustic injury, and human papilloma virus infection **(Chap. 76)**.

The typical presentation of esophageal cancer is of progressive solid food dysphagia and weight loss. Associated symptoms may include odynophagia, iron deficiency, and, with midesophageal tumors, hoarseness from left recurrent laryngeal nerve injury. Generally, these are indications of locally invasive or even metastatic disease manifest by tracheoesophageal fistulas and vocal cord paralysis. Even when detected as a small lesion, esophageal cancer has poor survival because of the abundant esophageal lymphatics leading to regional lymph node metastases.

Benign esophageal tumors are uncommon and usually discovered incidentally. In decreasing frequency of occurrence, cell types include leiomyoma, fibrovascular polyps, squamous papilloma, granular cell tumors, lipomas, neurofibromas, and inflammatory fibroid polyps. These generally become symptomatic only when they are associated with dysphagia and merit removal only under the same circumstances.

CONGENITAL ANOMALIES

The most common congenital esophageal anomaly is esophageal atresia, occurring in about 1 in 5000 live births. Atresia can occur in several permutations, the common denominator being developmental

FIGURE 316-4 Intramural esophageal pseudodiverticulosis associated with chronic obstruction. Invaginations of contrast into the esophageal wall outline deep esophageal glands.

failure of fusion between the proximal and distal esophagus associated with a tracheoesophageal fistula, most commonly with the distal segment excluded. Alternatively, there can be an H-type configuration in which esophageal fusion has occurred, but with a tracheoesophageal fistula. Esophageal atresia is usually recognized and corrected surgically within the first few days of life. Later life complications include dysphagia from anastomotic strictures or absent peristalsis and reflux, which can be severe. Less common developmental anomalies include congenital esophageal stenosis, webs, and duplications.

Dysphagia can also result from congenital abnormalities that cause extrinsic compression of the esophagus. In dysphagia lusoria, the esophagus is compressed by an aberrant right subclavian artery arising from the descending aorta and passing behind the esophagus. Alternatively, vascular rings may surround and constrict the esophagus.

Heterotopic gastric mucosa, also known as an esophageal inlet patch, is a focus of gastric type epithelium in the proximal cervical esophagus; the estimated prevalence is 4–5%. The inlet patch is thought to result from incomplete replacement of embryonic columnar epithelium with squamous epithelium. The majority of inlet patches are asymptomatic, but acid production can occur as most contain fundic type gastric epithelium with parietal cells.

ESOPHAGEAL MOTILITY DISORDERS

Esophageal motility disorders are diseases attributable to esophageal neuromuscular dysfunction commonly associated with dysphagia, chest pain, or heartburn. The major entities are achalasia, DES, and GERD. Motility disorders can also be secondary to broader disease processes as is the case with pseudoachalasia, Chagas' disease, and scleroderma. Not included in this discussion are diseases affecting the pharynx and proximal esophagus, the impairment of which is almost always part of a more global neuromuscular disease process.

ACHALASIA

Achalasia is a rare disease caused by loss of ganglion cells within the esophageal myenteric plexus with a population incidence estimated to be 1–3 per 100,000 and usually presenting between age 25 and 60. With long-standing disease, aganglionosis is noted. The disease involves both excitatory (cholinergic) and inhibitory (nitric oxide) ganglionic neurons. Functionally, inhibitory neurons mediate deglutitive LES relaxation and the sequential propagation of peristalsis. Their absence leads to impaired deglutitive LES relaxation and absent peristalsis. Increasing evidence suggests that the ultimate cause of ganglion cell degeneration in achalasia is an autoimmune process attributable to a latent infection with human herpes simplex virus 1 combined with genetic susceptibility.

Long-standing achalasia is characterized by progressive dilatation and sigmoid deformity of the esophagus with hypertrophy of the LES. Clinical manifestations may include dysphagia, regurgitation, chest pain, and weight loss. Most patients report solid and liquid food dysphagia. Regurgitation occurs when food, fluid, and secretions are retained in the dilated esophagus. Patients with advanced achalasia are at risk for bronchitis, pneumonia, or lung abscess from chronic regurgitation and aspiration. Chest pain is frequent early in the course of achalasia, thought to result from esophageal spasm. Patients describe a squeezing, pressure-like retrosternal pain, sometimes radiating to the neck, arms, jaw, and back. Paradoxically, some patients complain of heartburn that may be a chest pain equivalent. Treatment of achalasia is less effective in relieving chest pain than it is in relieving dysphagia or regurgitation.

The differential diagnosis of achalasia includes DES, Chagas' disease, and pseudoachalasia. Chagas' disease is endemic in areas of central Brazil, Venezuela, and northern Argentina and spread by the bite of the reduviid (kissing) bug that transmits the protozoan, *Trypanosoma cruzi.* The chronic phase of the disease develops years after infection and results from destruction of autonomic ganglion cells throughout the body, including the heart, gut, urinary tract, and respiratory tract. Tumor infiltration, most commonly seen with carcinoma in the gastric fundus or distal esophagus, can mimic idiopathic achalasia. The resultant "pseudoachalasia" accounts for up to 5% of suspected cases and is more likely with advanced age, abrupt onset of symptoms (<1 year), and weight loss. Hence, endoscopy is a necessary part of the evaluation of achalasia. When the clinical suspicion for pseudoachalasia is high and endoscopy nondiagnostic, computed tomography (CT) scanning or EUS may be of value. Rarely, pseudoachalasia can result from a paraneoplastic syndrome with circulating antineuronal antibodies.

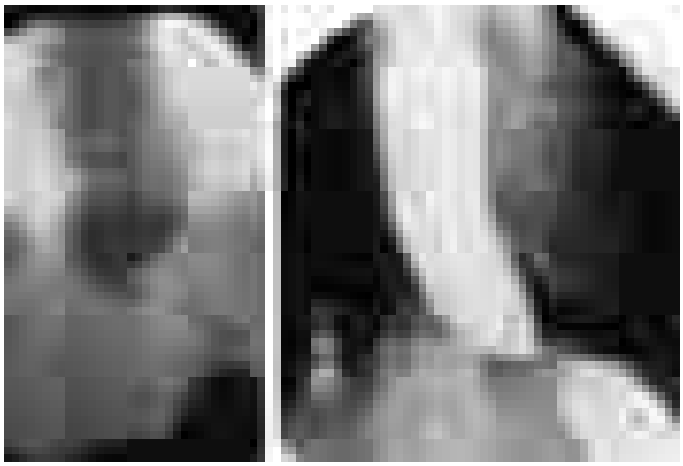


FIGURE 316-5 Achalasia with esophageal dilatation, tapering at the gastroesophageal junction, and an air-fluid level within the esophagus. The example on the *left* shows sigmoid deformity with very advanced disease.

Achalasia is diagnosed by barium swallow x-ray and/or esophageal manometry; endoscopy has a relatively minor role other than to exclude pseudoachalasia. The barium swallow x-ray appearance is of a dilated esophagus with poor emptying, an air-fluid level, and tapering at the LES giving it a beak-like appearance **(Fig. 316-5)**. Occasionally, an epiphrenic diverticulum is observed. In long-standing achalasia, the esophagus may assume a sigmoid configuration. The diagnostic criteria for achalasia with esophageal manometry are impaired LES relaxation and absent peristalsis. High-resolution manometry has somewhat advanced this diagnosis; three subtypes of achalasia are differentiated based on the pattern of pressurization in the nonperistaltic esophagus **(Fig. 316-6)**. Because manometry identifies early disease before esophageal dilatation and food retention, it is the most sensitive diagnostic test.

There is no known way of preventing or reversing achalasia. Therapy is directed at reducing LES pressure so that gravity and esophageal pressurization promote esophageal emptying. Peristalsis does not recover. However, in many instances, remnants of peristalsis masked by esophageal pressurization and dilatation prior to therapy are demonstrable following effective treatment. LES pressure can be reduced by pharmacologic therapy, pneumatic balloon dilation, or surgical myotomy. No large, controlled trials of the therapeutic alternatives exist, and the optimal approach is debated. Pharmacologic therapies are relatively ineffective but are often used as temporizing therapies. Nitrates or calcium channel blockers are administered before eating, advising caution because of their effects on blood pressure. Botulinum toxin, injected into the LES under endoscopic guidance, inhibits acetylcholine release from nerve endings and improves dysphagia in about 66% of cases for at least 6 months. Sildenafil and alternative phosphodiesterase inhibitors effectively decrease LES pressure, but practicalities limit their clinical use in achalasia.

The only durable therapies for achalasia are pneumatic dilation and Heller myotomy. Pneumatic dilation, with a reported efficacy ranging from 32 to 98%, is an endoscopic technique using a noncompliant, cylindrical balloon dilator positioned across the LES and inflated to a diameter of 3–4 cm. The major complication is perforation with a reported incidence of 0.5–5%. The most common surgical procedure for achalasia is laparoscopic Heller myotomy, usually performed in conjunction with an antireflux procedure (partial fundoplication); good to excellent results are reported in 62–100% of cases. A European-randomized controlled trial demonstrated an equivalent response rate of ~90% for both pneumatic dilation and laparoscopic Heller myotomy

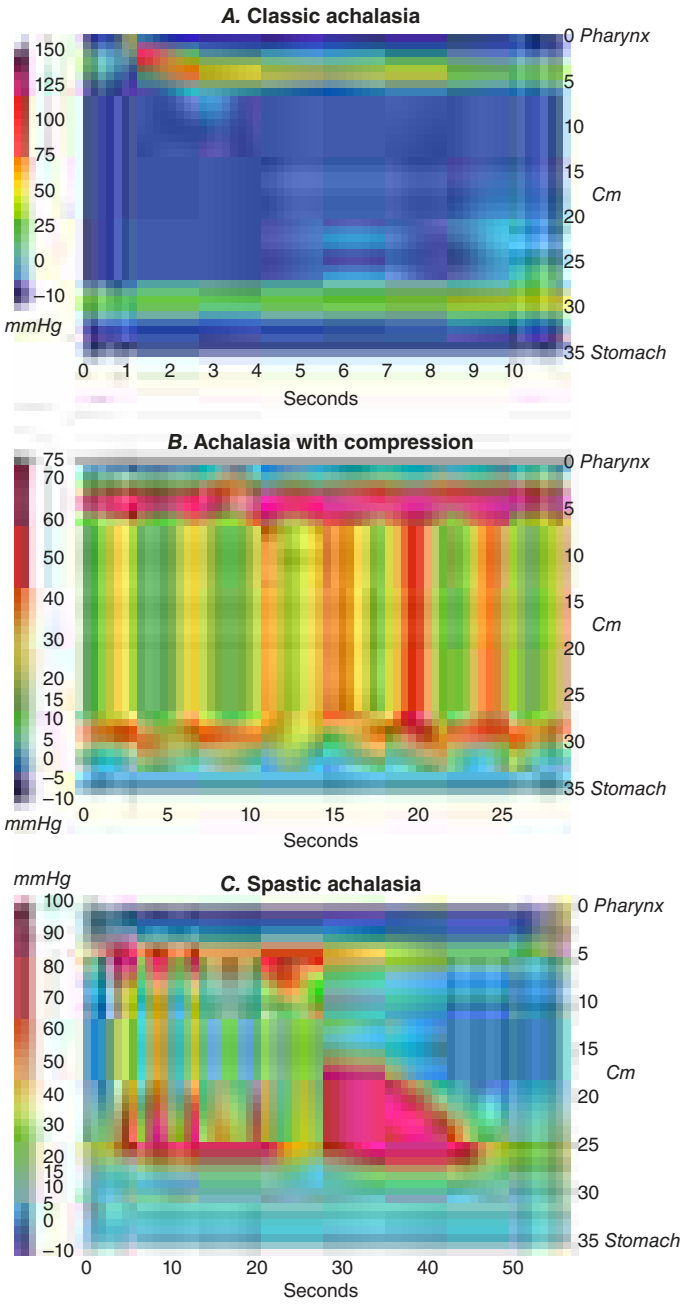


FIGURE 316-6 Three subtypes of achalasia: classic (*A*), with esophageal compression (*B*), and spastic achalasia (*C*) imaged with pressure topography. All are characterized by impaired lower esophageal sphincter (LES) relaxation and absent peristalsis. However, classic achalasia has minimal pressurization of the esophageal body, whereas substantial fluid pressurization is observed in achalasia with esophageal compression, and spastic esophageal contractions are observed with spastic achalasia.

at 5-year follow-up. Occasionally, patients with advanced disease fail to respond to pneumatic dilation or Heller myotomy. In such refractory cases, esophageal resection with gastric pull-up or interposition of a segment of transverse colon may be the only option other than gastrostomy feeding.

An endoscopic approach to LES myotomy has been introduced, referred to as per oral esophageal myotomy (POEM). This technique involves the creation of a submucosal tunnel within the esophageal wall through which the circular muscle of the LES and distal esophagus are transected with electrocautery. Short-term studies of efficacy have been favorable. Potential advantages over the conventional laparoscopic approach include avoidance of surgical disruption of the diaphragmatic hiatus and more rapid recovery.

In untreated or inadequately treated achalasia, esophageal dilatation predisposes to stasis esophagitis. Prolonged stasis esophagitis is the likely explanation for the association between achalasia and esophageal squamous cell cancer. Tumors develop after years of achalasia, usually in the setting of extreme esophageal dilatation with the overall squamous cell cancer risk increased seventeenfold compared to controls.

DIFFUSE ESOPHAGEAL SPASM

DES is manifested by episodes of dysphagia and chest pain attributable to abnormal esophageal contractions with normal deglutitive LES relaxation. Beyond that, there is little consensus. The pathophysiology and natural history of DES are ill defined. Radiographically, DES has been characterized by tertiary contractions or a "corkscrew esophagus" **(Fig. 316-7)**, but in many instances, these abnormalities are actually indicative of achalasia. Manometrically, a variety of defining features have been proposed including uncoordinated ("spastic") activity in the distal esophagus, spontaneous and repetitive contractions, or high-amplitude and prolonged contractions. The current consensus, derived from high-resolution manometry studies, is to define spasm by the occurrence of contractions in the distal esophagus with short latency relative to the time of the pharyngeal contraction, a dysfunction indicative of impairment of inhibitory myenteric plexus neurons. When defined in this restrictive fashion **(Fig. 316-8)**, DES is actually much less common than achalasia.

Esophageal chest pain closely mimics angina pectoris. Features suggesting esophageal pain include pain that is nonexertional, prolonged, interrupts sleep, meal-related, relieved with antacids, and accompanied by heartburn, dysphagia, or regurgitation. However, all of these features exhibit overlap with cardiac pain, which still must be the primary consideration. Furthermore, even within the spectrum of esophageal diseases, both chest pain and dysphagia are also characteristic of peptic or infectious esophagitis. Only after these more common entities have been excluded by evaluation and/or treatment should a diagnosis of DES be pursued.

Although DES is diagnosed by manometry, endoscopy is useful to identify alternative structural and inflammatory lesions that may cause chest pain. Radiographically, a "corkscrew esophagus," "rosary bead esophagus," pseudodiverticula, or curling can be indicative of DES, but these are also found with spastic achalasia. Given these vagaries of defining DES, and the resultant heterogeneity of patients identified for inclusion in therapeutic trials, it is not surprising that trial results have been disappointing. Only small, uncontrolled trials exist, reporting response to nitrates, calcium channel blockers, hydralazine, botulinum toxin, and anxiolytics. The only controlled trial showing efficacy was

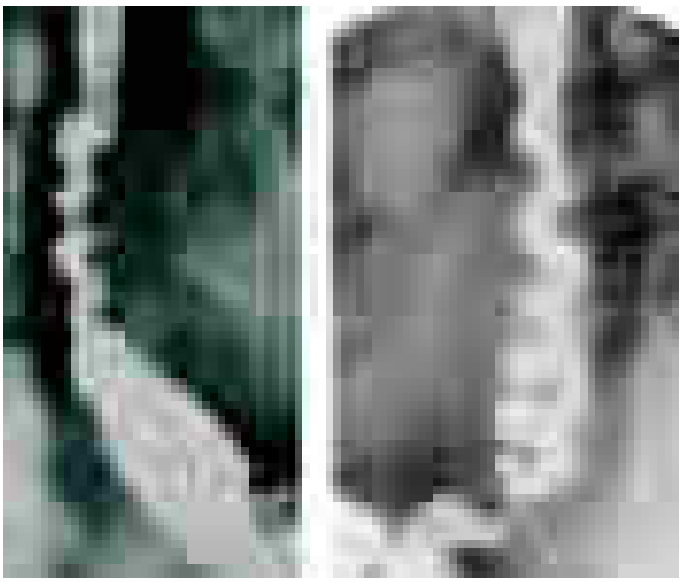


FIGURE 316-7 Diffuse esophageal spasm. The characteristic "corkscrew" esophagus results from spastic contraction of the circular muscle in the esophageal wall; more precisely, this is actually a helical array of muscle. These findings are also seen with spastic achalasia.

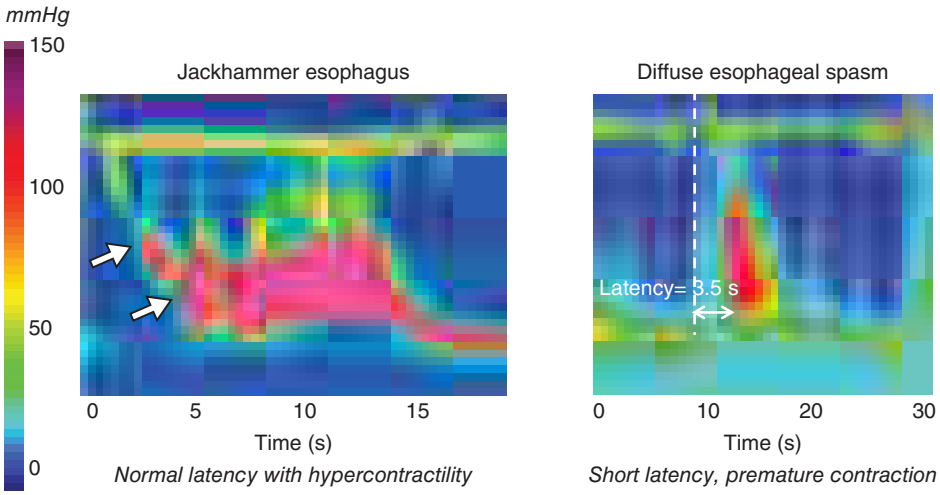


FIGURE 316-8 Esophageal pressure topography of the two major variants of esophageal spasm: jackhammer esophagus (*left*) and diffuse esophageal spasm (*right*). Jackhammer esophagus is defined by the extraordinarily vigorous and repetitive contractions with normal peristaltic onset and normal latency of the contraction. Diffuse esophageal spasm is similar but primarily defined by a short latency (premature) contraction.

with an anxiolytic. Surgical therapy (long myotomy or even esophagectomy) should be considered only with severe weight loss or unbearable pain. These indications are extremely rare.

NONSPECIFIC MANOMETRIC FINDINGS

Manometric studies done to evaluate chest pain and/or dysphagia often report minor abnormalities (e.g., hypertensive or hypotensive peristalsis, hypertensive LES) that are insufficient to diagnose either achalasia or DES. These findings are of unclear significance. Reflux and psychiatric diagnoses, particularly anxiety and depression, are common among such individuals. A lower visceral pain threshold and symptoms of irritable bowel syndrome are noted in more than half of such patients. Consequently, therapy for these individuals should either target the most common esophageal disorder, GERD, or more global conditions such as depression or somatization neurosis that are found to be coexistent.

GASTROESOPHAGEAL REFLUX DISEASE

The current conception of GERD is to encompass a family of conditions with the commonality that they are caused by gastroesophageal reflux resulting in either troublesome symptoms or an array of potential esophageal and extraesophageal manifestations. It is estimated that 10–15% of adults in the United States are affected by GERD, although such estimates are based only on population studies of self-reported chronic heartburn. With respect to the esophagus, the spectrum of injury includes esophagitis, stricture, Barrett's esophagus, and adenocarcinoma (Fig. 316-9). Of particular concern is the rising incidence of esophageal adenocarcinoma, an epidemiologic trend that parallels the increasing incidence of GERD. There were about 8000 incident cases of esophageal adenocarcinoma in the United States in 2013 (half of all esophageal cancers); it is estimated that this disease burden has increased two- to sixfold in the last 20 years.

PATHOPHYSIOLOGY

The best-defined subset of GERD patients, albeit a minority overall, have esophagitis. Esophagitis occurs when refluxed gastric acid and pepsin cause necrosis of the esophageal mucosa causing erosions and ulcers. Note that some degree of gastroesophageal reflux is normal, physiologically intertwined with the mechanism of belching (transient LES relaxation), but esophagitis results from excessive reflux, often accompanied by impaired clearance of the refluxed gastric juice. Restricting reflux to that which is physiologically intended depends on the anatomic and physiologic integrity of the esophagogastric junction, a complex sphincter comprised of both the LES and the surrounding crural diaphragm. Three dominant mechanisms of esophagogastric junction incompetence are recognized: (1) transient LES relaxations (a vagovagal reflex in which LES relaxation is elicited by gastric distention), (2) LES hypotension, or (3) anatomic distortion of the esophagogastric junction inclusive of hiatus hernia. Of note, the third factor, esophagogastric junction anatomic disruption, is both significant unto itself and also because it interacts with the first two mechanisms. Transient LES relaxations account for about 90% of reflux in normal subjects or GERD patients without hiatus hernia, but patients with hiatus hernia have a more heterogeneous mechanistic profile. Factors tending to exacerbate reflux regardless of mechanism are abdominal obesity, pregnancy, gastric hypersecretory states, delayed gastric emptying, disruption of esophageal peristalsis, and gluttony.

After acid reflux, peristalsis returns the refluxed fluid to the stomach and acid clearance is completed by titration of the residual acid by bicarbonate contained in swallowed saliva. Consequently, two causes of prolonged acid clearance are impaired peristalsis and reduced salivation. Impaired peristaltic emptying can be attributable to disrupted peristalsis or superimposed reflux associated with a hiatal hernia.

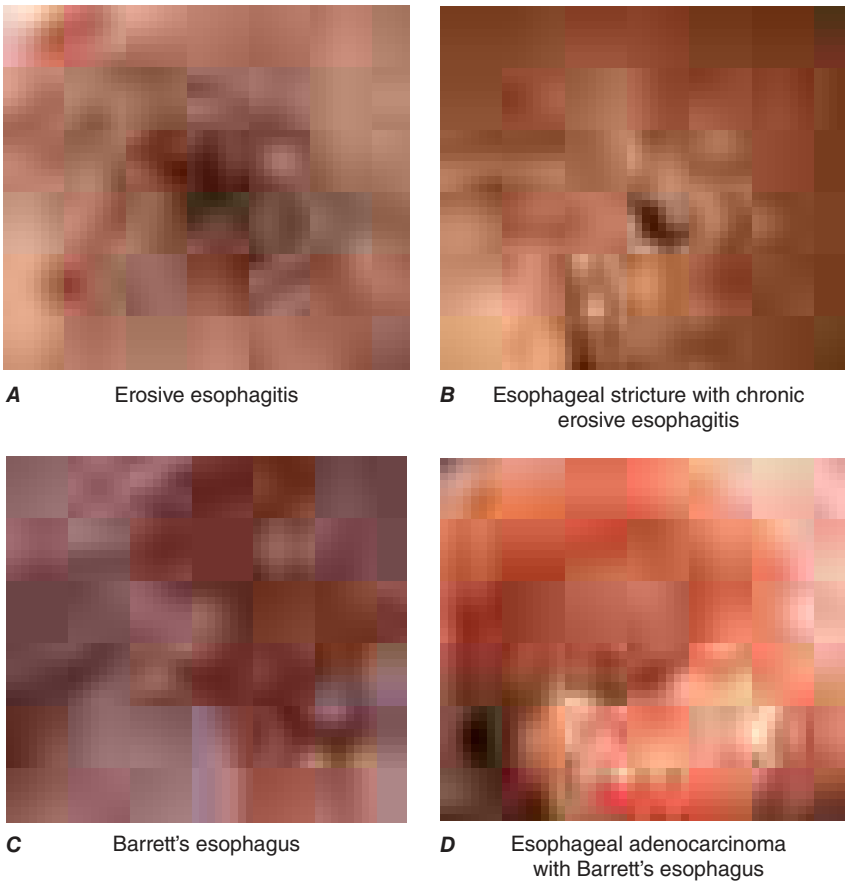


FIGURE 316-9 Endoscopic appearance of (*A*) peptic esophagitis, (***B***) a peptic stricture, (***C***) Barrett's metaplasia, and (***D***) adenocarcinoma developing within an area of Barrett's esophagus.

With superimposed reflux, fluid retained within a sliding hiatal hernia refluxes back into the esophagus during swallow-related LES relaxation, a phenomenon that does not normally occur.

Inherent in the pathophysiologic model of GERD is that gastric juice is harmful to the esophageal epithelium. However, gastric acid hypersecretion is usually not a dominant factor in the development of esophagitis. An obvious exception is with Zollinger-Ellison syndrome, which is associated with severe esophagitis in about 50% of patients. Another caveat is with chronic *Helicobacter pylori* gastritis, which may have a protective effect by inducing atrophic gastritis with concomitant hypoacidity. Pepsin, bile, and pancreatic enzymes within gastric secretions can also injure the esophageal epithelium, but their noxious properties are either lessened without an acidic environment or dependent on acidity for activation. Bile warrants attention because it persists in refluxate despite acid-suppressing medications. Bile can transverse the cell membrane, imparting severe cellular injury in a weakly acidic environment, and has also been invoked as a cofactor in the pathogenesis of Barrett's metaplasia and adenocarcinoma. Hence, the causticity of gastric refluxate extends beyond hydrochloric acid.

SYMPTOMS

Heartburn and regurgitation are the typical symptoms of GERD. Somewhat less common are dysphagia and chest pain. In each case, multiple potential mechanisms for symptom genesis operate that extend beyond the basic concepts of mucosal erosion and activation of afferent sensory nerves. Specifically, hypersensitivity and functional pain are increasingly recognized as cofactors. Nonetheless, the dominant clinical strategy is empirical treatment with acid inhibitors, reserving further evaluation for those who fail to respond. Important exceptions to this are patients with chest pain or persistent dysphagia, each of which may be indicative of more morbid conditions. With chest pain, cardiac disease must be carefully considered. In the case of persistent dysphagia, chronic reflux can lead to the development of a peptic stricture or adenocarcinoma, each of which benefits from early detection and/or specific therapy.

Extraesophageal syndromes with an established association to GERD include chronic cough, laryngitis, asthma, and dental erosions. A multitude of other conditions including pharyngitis, chronic bronchitis, pulmonary fibrosis, chronic sinusitis, cardiac arrhythmias, sleep apnea, and recurrent aspiration pneumonia have proposed associations with GERD. However, in both cases, it is important to emphasize the word *association* as opposed to *causation*. In many instances, the disorders likely coexist because of shared pathogenetic mechanisms rather than strict causality. Potential mechanisms for extraesophageal GERD manifestations are either regurgitation with direct contact between the refluxate and supraesophageal structures or via a vagovagal reflex wherein reflux activation of esophageal afferent nerves triggers efferent vagal reflexes such as bronchospasm, cough, or arrhythmias.

DIFFERENTIAL DIAGNOSIS

Although generally quite characteristic, symptoms from GERD need to be distinguished from symptoms related to infectious, pill, or eosinophilic esophagitis (EoE), peptic ulcer disease, dyspepsia, biliary colic, coronary artery disease, and esophageal motility disorders. It is especially important that coronary artery disease be given early consideration because of its potentially lethal implications. The remaining elements of the differential diagnosis can be addressed by endoscopy, upper gastrointestinal series, or esophageal manometry as appropriate. Erosive esophagitis at the esophagogastric junction is the endoscopic hallmark of GERD, but identified in only about one third of patients with GERD. The distinction among etiologies of esophagitis is readily made by endoscopic appearance but mucosal biopsies may be helpful to evaluate for infectious or eosinophilic inflammation. In terms of endoscopic appearance, the ulcerations seen in peptic esophagitis are usually solitary and distal, whereas infectious ulcerations are punctate and diffuse. EoE characteristically exhibits multiple esophageal rings, linear furrows, white punctate exudate, and strictures. Esophageal ulcerations from pill esophagitis are usually singular and deep at points of luminal narrowing, especially near the carina, with sparing of the distal esophagus.

COMPLICATIONS

The complications of GERD are related to chronic esophagitis (bleeding and stricture) and the relationship between GERD and esophageal adenocarcinoma. However, both erosive esophagitis and peptic strictures have become increasingly rare in the era of potent antisecretory medications. Conversely, the most severe histologic consequence of GERD is Barrett's metaplasia with the associated risk of esophageal adenocarcinoma, and the incidence of these lesions has increased, not decreased, in the era of potent acid suppression. Barrett's metaplasia, endoscopically recognized by tongues of salmon-colored mucosa extending proximally from the gastroesophageal junction (Fig. 316-9) or histopathologically by the finding of specialized columnar metaplasia, is associated with a significantly increased risk for development of esophageal adenocarcinoma.

Barrett's metaplasia can progress to adenocarcinoma through the intermediate stages of low- and high-grade dysplasia **(Fig. 316-10)**. Owing to this risk, areas of Barrett's and especially any included areas of mucosal irregularity should be carefully inspected and extensively biopsied. The rate of cancer development is estimated at 0.1–0.3% per year, but vagaries in definitional criteria and of the extent of Barrett's metaplasia requisite to establish the diagnosis have contributed to variability and inconsistency in this risk assessment. The group at greatest risk is obese white males in their sixth decade of life. However, despite common practice, the utility of endoscopic screening and surveillance programs intended to control the adenocarcinoma risk has not been established. Also of note, no high-level evidence confirms that aggressive antisecretory therapy or antireflux surgery causes regression of Barrett's esophagus or prevents adenocarcinoma.

Although the management of Barrett's esophagus remains controversial, the finding of dysplasia in Barrett's, particularly high-grade dysplasia, mandates further intervention. In addition to the high rate of progression to adenocarcinoma, there is also a high prevalence of unrecognized coexisting cancer with high-grade dysplasia. Treatment recommendations for Barrett's esophagus with high-grade dysplasia

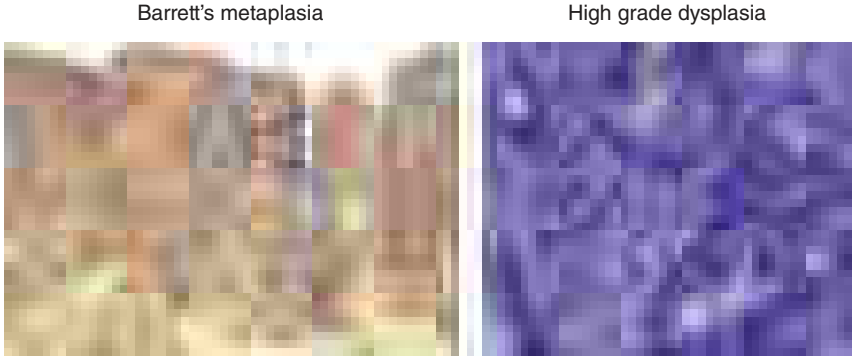


FIGURE 316-10 Histopathology of Barrett's metaplasia and Barrett's with high-grade dysplasia. H&E, hematoxylin and eosin.

have evolved over the past several years. Historically, esophagectomy was the gold standard treatment for high-grade dysplasia. However, esophagectomy has a mortality ranging from 3 to 10%, along with substantial morbidity and recent prospective studies have demonstrated the efficacy of mucosal ablation therapy with substantially less morbidity and essentially no mortality. Consequently, current societal guidelines endorse endoscopic mucosal ablation therapies for the management of high-grade dysplasia.

TREATMENT

Gastroesophageal Reflux Disease

Lifestyle modifications are routinely advocated as GERD therapy. Broadly speaking, these fall into three categories: (1) avoidance of foods that reduce LES pressure, making them "refluxogenic" (these commonly include fatty foods, alcohol, spearmint, peppermint, and possibly coffee and tea); (2) avoidance of acidic foods that are inherently irritating (citrus fruits, tomato-based foods); and (3) adoption of behaviors to minimize reflux and/or heartburn. In general, minimal evidence supports the efficacy of these measures. However, clinical experience dictates that subsets of patients are benefitted by specific recommendations, based on their individual history and symptom profile. A patient with sleep disturbance from nighttime heartburn is more likely to benefit from elevation of the head of the bed and avoidance of eating before retiring. The most broadly applicable recommendation is for weight reduction. Even though the benefit with respect to reflux cannot be assured, the strong epidemiologic relationship between body mass index and GERD and the secondary health gains of weight reduction is beyond dispute.

The dominant pharmacologic approach to GERD management is with inhibitors of gastric acid secretion, and abundant data support the effectiveness of this approach. Pharmacologically reducing the acidity of gastric juice does not prevent reflux, but it ameliorates reflux symptoms and allows esophagitis to heal. The hierarchy of effectiveness among pharmaceuticals parallels their antisecretory potency. Proton pump inhibitors (PPIs) are more efficacious than histamine$_2$ receptor antagonists (H_2RAs), and both are superior to placebo. No major differences exist among PPIs, and only modest gain is achieved by increased dosage.

Paradoxically, the perceived frequency and severity of heartburn correlate poorly with the presence or severity of esophagitis. When GERD treatments are assessed in terms of resolving heartburn, both efficacy and differences among pharmaceuticals are less clear-cut than with the objective of healing esophagitis. Although the same overall hierarchy of effectiveness exists, observed efficacy rates are lower and vary widely, likely reflecting patient heterogeneity.

Reflux symptoms tend to be chronic, irrespective of esophagitis. Thus, a common management strategy is indefinite treatment with PPIs or H_2RAs as necessary for symptom control. The side effects of PPI therapy are generally minimal. Rare cases of interstitial nephritis and severe, reversible hypomagnesemia have been reported. Vitamin B_{12} and iron absorption may be compromised and susceptibility to enteric infections, particularly *Clostridium difficile* colitis, increased with treatment. Observational data have also noted an association between PPI exposure and renal disease, dementia, and cardiovascular disease, but the hazard ratios reported in these studies were small and potential for unrecognized residual confounding bias was substantial. Population studies have also suggested a slight increased risk of bone fracture with chronic PPI use suggesting an impairment of calcium absorption, but prospective studies have failed to corroborate this. Nonetheless, as with any medication, PPI dosage should be minimized to that necessary for the clinical indication.

Laparoscopic Nissen fundoplication, wherein the proximal stomach is wrapped around the distal esophagus to create an antireflux barrier, is a surgical alternative to the management of chronic GERD. Just as with PPI therapy, evidence on the utility of fundoplication is strongest for treating esophagitis, and controlled trials suggest similar efficacy to PPI therapy. However, the benefits of fundoplication must be weighed against potential deleterious effects, including surgical morbidity and mortality, postoperative dysphagia, failure or breakdown requiring reoperation, an inability to belch, and increased bloating, flatulence, and bowel symptoms after surgery.

EOSINOPHILIC ESOPHAGITIS

EoE is increasingly recognized in adults and children around the world. Current prevalence estimates in the United States identified 4–6 cases per 10,000 with a predilection for white males between 30 and 40 years of age. The increasing prevalence of EoE is attributable to a combination of an increasing incidence and a growing recognition of the condition. There is also an incompletely understood, but important, overlap between EoE and GERD that may confound the diagnosis of the disease.

EoE is diagnosed based on the combination of typical esophageal symptoms and esophageal mucosal biopsies demonstrating squamous epithelial eosinophil-predominant inflammation. Alternative etiologies of esophageal eosinophilia include GERD, drug hypersensitivity, connective tissue disorders, hypereosinophilic syndrome, Crohn's disease, and infection. Current evidence indicates that EoE is an immunologic disorder induced by antigen sensitization in susceptible individuals. Dietary factors play an important role in both the pathogenesis and treatment of EoE. Aeroallergens may also contribute, but the evidence is weaker. The natural history of EoE is unclear, but an increased risk of esophageal stricture development paralleling the duration of untreated disease has been noted.

EoE should be strongly considered in children and adults with dysphagia and esophageal food impactions. In preadolescent children, symptom presentations of EoE include chest or abdominal pain, nausea, vomiting, and food aversion. Other symptoms in adults may include atypical chest pain and heartburn, particularly heartburn that is refractory to PPI therapy. An atopic history of food allergy, asthma, eczema, or allergic rhinitis is present in the majority of patients. Peripheral blood eosinophilia is demonstrable in up to 50% of patients, but the specificity of this finding is problematic in the setting of concomitant atopy. The characteristic endoscopically identified esophageal findings include loss of vascular markings (edema), multiple esophageal rings, longitudinally oriented furrows, and punctate exudate **(Fig. 316-11)**. Histologic confirmation is made with the demonstration of esophageal

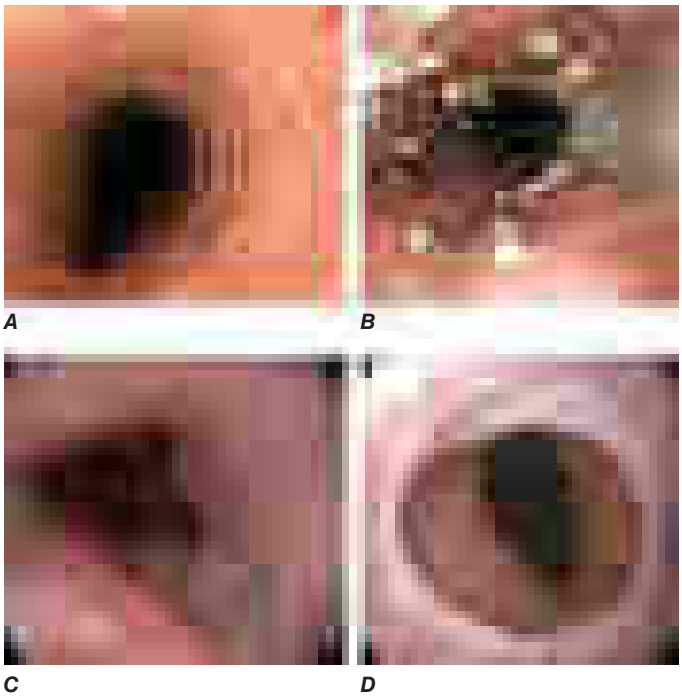


FIGURE 316-11 Endoscopic features of (***A***) eosinophilic esophagitis (EoE), (***B***) *Candida* esophagitis, (***C***) giant ulcer associated with HIV, and (***D***) a Schatzki ring.

FIGURE 316-12 Histopathology of eosinophilic esophagitis (EoE) showing infiltration of the esophageal squamous epithelium with eosinophils. Additional features of basal cell hyperplasia and lamina propria fibrosis are present. Eosinophilic inflammation can also be seen with gastroesophageal reflux disease.

mucosal eosinophilia (peak density ≥15 eosinophils per high-power field) **(Fig. 316-12)**. Complications of EoE include esophageal stricture, narrow-caliber esophagus, food impaction, and esophageal perforation.

The goals of EoE management are symptom control and the prevention of complications. Once esophageal eosinophilia is demonstrated, patients typically undergo a trial of PPI therapy as a means of excluding a contribution of GERD to the symptoms and esophageal mucosal inflammation. PPI-responsive esophageal eosinophilia, characterized by elimination of mucosal eosinophilia, occurs in 30–50% of cases of suspected EoE. Patients with persistent symptoms and eosinophilic inflammation following PPI therapy are subsequently considered for treatments such as elimination diets or swallowed topical glucocorticoids. Elemental formula diets are a highly effective therapy but are limited by palatability. Notably, allergy testing by means of either serum IgE or skin prick testing has demonstrated poor sensitivity and specificity in the identification of foods that incite the esophageal inflammatory response. Allergy testing combining skin prick and atopy patch testing has been effective in children with EoE, but additional validation is needed. Empiric elimination of common food allergies (milk, wheat, egg, soy, nuts, and seafood) followed by systematic reintroduction has been an effective diet therapy in both children and adults with EoE. The intent of the elimination diet approach is the identification of a single or a small number of food triggers. Swallowed, topical glucocorticoids (e.g., fluticasone propionate or budesonide) are highly effective, but recurrence of disease is common following the cessation of short-term therapy. Systemic glucocorticoids are reserved for severely afflicted patients refractory to less morbid treatments. Esophageal dilation is very effective at relieving dysphagia in patients with fibrostenosis. Dilation should be approached conservatively because of the risk of deep, esophageal mural laceration or perforation in the stiff-walled esophagus that is characteristic of the disease.

INFECTIOUS ESOPHAGITIS

With the increased use of immunosuppression for organ transplantation as well as chronic inflammatory diseases and chemotherapy along with the AIDS epidemic, infections with *Candida* species, herpesvirus, and cytomegalovirus (CMV) have become relatively common. Although rare, infectious esophagitis also occurs among the non-immunocompromised, with herpes simplex and *Candida albicans* being the most common pathogens. Among AIDS patients, infectious esophagitis becomes more common as the CD4 count declines; cases are rare with a CD4 count >200 and common when <100. HIV itself may also be associated with a self-limited syndrome of acute esophageal ulceration with oral ulcers and a maculopapular skin rash at the time of seroconversion. Additionally, some patients with advanced disease have deep, persistent esophageal ulcers treated with oral glucocorticoids or thalidomide. However, with the widespread use of highly effective anti-viral therapies, a reduction in these HIV complications has been noted.

Regardless of the infectious agent, odynophagia is a characteristic symptom of infectious esophagitis; dysphagia, chest pain, and hemorrhage are also common. Odynophagia is uncommon with reflux esophagitis, so its presence should always raise suspicion of an alternative etiology.

CANDIDA ESOPHAGITIS

Candida is normally found in the throat, but can become pathogenic and produce esophagitis in a compromised host; *C. albicans* is most common. *Candida* esophagitis also occurs with esophageal stasis secondary to esophageal motor disorders and diverticula. Patients complain of odynophagia and dysphagia. If oral thrush is present, empirical therapy is appropriate, but co-infection is common, and persistent symptoms should lead to prompt endoscopy with biopsy, which is the most useful diagnostic evaluation. *Candida* esophagitis has a characteristic appearance of white plaques with friability. Rarely, *Candida* esophagitis is complicated by bleeding, perforation, stricture, or systemic invasion. Oral fluconazole (200–400 mg on the first day, followed by 100–200 mg daily) for 14–21 days is the preferred treatment. Patients refractory to fluconazole may respond to voriconazole, or posaconazole. Alternatively, poorly responsive patients or those who cannot swallow medications can be treated with an intravenous echinocandin.

HERPETIC ESOPHAGITIS

Herpes simplex virus type 1 or 2 may cause esophagitis. Vesicles on the nose and lips may coexist and are suggestive of a herpetic etiology. Varicella-zoster virus can also cause esophagitis in children with chickenpox or adults with zoster. The characteristic endoscopic findings are vesicles and small, punched-out ulcerations. Because herpes simplex infections are limited to squamous epithelium, biopsies from the ulcer margins are most likely to reveal the characteristic ground-glass nuclei, eosinophilic Cowdry's type A inclusion bodies, and giant cells. Culture or polymerase chain reaction (PCR) assays are helpful to identify acyclovir-resistant strains. Acyclovir (200 mg orally five times a day for 7–10 days) can be used for immunocompetent hosts, although the disease is typically self-limited after a 1- to 2-week period in such patients. Immunocompromised patients are treated with acyclovir (400 mg orally five times a day for 14–21 days), famciclovir (500 mg orally three times a day), or valacyclovir (1 g orally three times a day). In patients with severe odynophagia, intravenous acyclovir, 5 mg/kg every 8 h for 7–14 days, reduces this morbidity.

CYTOMEGALOVIRUS

CMV esophagitis occurs primarily in immunocompromised patients, particularly organ transplant recipients. CMV is usually activated from a latent stage. Endoscopically, CMV lesions appear as serpiginous ulcers in an otherwise normal mucosa, particularly in the distal esophagus. Biopsies from the ulcer bases have the greatest diagnostic yield for finding the pathognomonic large nuclear or cytoplasmic inclusion bodies. Immunohistology with monoclonal antibodies to CMV and in situ hybridization tests are useful for early diagnosis. Data on therapy for CMV esophagitis are limited. Treatment studies of

CMV gastrointestinal disease have demonstrated effectiveness of both ganciclovir (5 mg/kg every 12 h intravenously) and valganciclovir (900 mg orally every 12 h). Therapy is continued until healing, which may take 3–6 weeks. Maintenance therapy may be needed for patients with relapsing disease.

MECHANICAL TRAUMA AND IATROGENIC INJURY

ESOPHAGEAL PERFORATION

Most cases of esophageal perforation are from instrumentation of the esophagus or trauma. Alternatively, forceful vomiting or retching can lead to spontaneous rupture at the gastroesophageal junction (Boerhaave's syndrome). More rarely, corrosive esophagitis or neoplasms lead to perforation. Instrument perforation from endoscopy or nasogastric tube placement typically occurs in the hypopharynx or at the gastroesophageal junction. Perforation may also occur at the site of a stricture in the setting of endoscopic food disimpaction or esophageal dilation. Esophageal perforation causes pleuritic retrosternal pain that can be associated with pneumomediastinum and subcutaneous emphysema. Mediastinitis is a major complication of esophageal perforation, and prompt recognition is key to optimizing outcome. CT of the chest is most sensitive in detecting mediastinal air. Esophageal perforation is confirmed by a contrast swallow, usually Gastrografin followed by thin barium. Treatment includes nasogastric suction and parenteral broad-spectrum antibiotics with prompt surgical drainage and repair in noncontained leaks. Conservative therapy with NPO status and antibiotics without surgery may be appropriate in cases of contained perforation that are detected early. Endoscopic clipping or stent placement may be indicated in nonoperated iatrogenic perforations or nonoperable cases such as perforated tumors.

MALLORY-WEISS TEAR

Vomiting, retching, or vigorous coughing can cause a nontransmural tear at the gastroesophageal junction that is a common cause of upper gastrointestinal bleeding. Most patients present with hematemesis. Antecedent vomiting is the norm, but not always evident. Bleeding usually abates spontaneously, but protracted bleeding may respond to local epinephrine or cauterization therapy, endoscopic clipping, or angiographic embolization. Surgery is rarely needed.

RADIATION ESOPHAGITIS

Radiation esophagitis can complicate treatment for thoracic cancers, especially breast and lung, with the risk proportional to radiation dosage. Radiosensitizing drugs such as doxorubicin, bleomycin, cyclophosphamide, and cisplatin also increase the risk. Dysphagia and odynophagia may last weeks to months after therapy. The esophageal mucosa becomes erythematous, edematous, and friable. Submucosal fibrosis and degenerative tissue changes and stricturing may occur years after the radiation exposure. Radiation exposure in excess of 5000 cGy has been associated with increased risk of esophageal stricture. Treatment for acute radiation esophagitis is supportive. Chronic strictures are managed with esophageal dilation.

CORROSIVE ESOPHAGITIS

Caustic esophageal injury from ingestion of alkali or, less commonly, acid can be accidental or from attempted suicide. Absence of oral injury does not exclude possible esophageal involvement. Thus, early endoscopic evaluation is recommended to assess and grade the injury to the esophageal mucosa. Severe corrosive injury may lead to esophageal perforation, bleeding, stricture, and death. Glucocorticoids have not been shown to improve the clinical outcome of acute corrosive esophagitis and are not recommended. Healing of more severe grades of caustic injury is commonly associated with severe stricture formation and often requires repeated dilation.

PILL ESOPHAGITIS

Pill-induced esophagitis occurs when a swallowed pill fails to traverse the entire esophagus and lodges within the lumen. Generally, this is attributed to poor "pill taking habits": inadequate liquid with the pill or lying down immediately after taking a pill. The most common location for the pill to lodge is in the mid-esophagus near the crossing of the aorta or carina. Extrinsic compression from these structures halts the movement of the pill or capsule. Since initially reported in 1970, more than 1000 cases of pill esophagitis have been reported, suggesting that this is not an unusual occurrence. A wide variety of medications are implicated with the most common being doxycycline, tetracycline, quinidine, phenytoin, potassium chloride, ferrous sulfate, nonsteroidal anti-inflammatory drugs (NSAIDs), and bisphosphonates.

Typical symptoms of pill esophagitis are the sudden onset of chest pain and odynophagia. Characteristically, the pain will develop over a period of hours or will awaken the individual from sleep. A classic history in the setting of ingestion of recognized pill offenders obviates the need for diagnostic testing in most patients. When endoscopy is performed, localized ulceration or inflammation is evident. Histologically, acute inflammation is typical. Chest CT imaging will sometimes reveal esophageal thickening consistent with transmural inflammation. Although the condition usually resolves within days to weeks, symptoms may persist for months and stricture can develop in severe cases. No specific therapy is known to hasten the healing process, but antisecretory medications are frequently prescribed to remove concomitant reflux as an aggravating factor. When healing results in stricture formation, dilation is indicated.

FOREIGN BODIES AND FOOD IMPACTION

Food or foreign bodies may lodge in the esophagus causing complete obstruction, which in turn can cause an inability to handle secretions (foaming at the mouth) and severe chest pain. Food impaction may occur due to peptic stricture, carcinoma, Schatzki ring, EoE, or simply inattentive eating. If it does not spontaneously resolve, impacted food can be removed endoscopically. Use of meat tenderizer enzymes to facilitate passage of a meat bolus is discouraged because of potential esophageal injury. Glucagon (1 mg IV) is sometimes tried before endoscopic dislodgement. After emergent treatment, patients should be evaluated for potential causes of the impaction with treatment rendered as indicated.

ESOPHAGEAL MANIFESTATIONS OF SYSTEMIC DISEASE

SCLERODERMA AND COLLAGEN VASCULAR DISEASES

Scleroderma esophagus (hypotensive LES and absent esophageal peristalsis) was initially described as a manifestation of scleroderma or other collagen vascular diseases and thought to be specific for these disorders. However, this nomenclature subsequently proved unfortunate and has been discarded because an estimated half of qualifying patients do not have an identifiable systemic disease, and reflux disease is often the only identifiable association. When scleroderma esophagus occurs as a manifestation of a collagen vascular disease, the histopathologic findings are of infiltration and destruction of the esophageal muscularis propria with collagen deposition and fibrosis. The pathogenesis of absent peristalsis and LES hypotension in the absence of a collagen vascular disease is unknown. Regardless of the underlying cause, the manometric abnormalities predispose patients to severe GERD due to inadequate LES barrier function combined with poor esophageal clearance of refluxed acid. Dysphagia may also be manifest but is generally mild and alleviated by eating in an upright position and using liquids to facilitate solid emptying.

DERMATOLOGIC DISEASES

A host of dermatologic disorders (pemphigus vulgaris, bullous pemphigoid, cicatricial pemphigoid, Behçet's syndrome, and epidermolysis bullosa) can affect the oropharynx and esophagus, particularly the proximal esophagus with blisters, bullae, webs, and strictures. Glucocorticoid treatment is usually effective. Erosive lichen planus, Stevens-Johnson syndrome, and graft-versus-host disease can also involve the esophagus. Esophageal dilation may be necessary to treat strictures.

FURTHER READING

Dellon ES et al: ACG Clinical Guideline: Evidence based approach to the diagnosis and management of esophageal eosinophilia and eosinophilic esophagitis. Am J Gastroenterol 108:679, 2013.

Furuta GT, Katzka DA: Eosinophilic esophagitis. N Engl J Med 373:1640, 2015.

Kahrilas PJ: Clinical practice. Gastroesophageal reflux disease. N Engl J Med 359:1700, 2008.

Kahrilas PJ, Boeckxstaens G: The spectrum of achalasia: Lessons from studies of pathophysiology and high-resolution manometry. Gastroenterology 145:954, 2013.

Kahrilas PJ et al: American Gastroenterological Association Institute technical review on the management of gastroesophageal reflux disease. Gastroenterology 135:1392, 2008.

Pandolfino JE, Gawron AJ: Achalasia: A systematic review. JAMA 313:1841, 2015.

Shaheen NJ et al: Diagnosis and Management of Barrett's Esophagus. Am J Gastroenterol 111:30, 2016.

Spechler SJ, Souza RF: Barrett's esophagus. N Engl J Med 371:836, 2014.

317 Peptic Ulcer Disease and Related Disorders

John Del Valle

PEPTIC ULCER DISEASE

A *peptic ulcer* is defined as disruption of the mucosal integrity of the stomach and/or duodenum leading to a local defect or excavation due to active inflammation. Although burning epigastric pain exacerbated by fasting and improved with meals is a symptom complex associated with peptic ulcer disease (PUD), it is now clear that >90% patients with this symptom complex (dyspepsia) do not have ulcers and that the majority of patients with peptic ulcers may be asymptomatic. Ulcers occur within the stomach and/or duodenum and are often chronic in nature. Acid peptic disorders are very common in the United States, with 4 million individuals (new cases and recurrences) affected per year. Lifetime prevalence of PUD in the United States is ~12% in men and 10% in women. PUD significantly affects quality of life by impairing overall patient well-being and contributing substantially to work absenteeism. Moreover, an estimated 15,000 deaths per year occur as a consequence of complicated PUD. The financial impact of these common disorders has been substantial, with an estimated burden on direct and indirect health care costs of ~$6 billion per year in the United States, with $3 billion spent on hospitalizations, $2 billion on physician office visits, and $1 billion in decreased productivity and days lost from work.

GASTRIC PHYSIOLOGY

Gastric Anatomy The gastric epithelial lining consists of rugae that contain microscopic gastric pits, each branching into four or five gastric glands made up of highly specialized epithelial cells. The makeup of gastric glands varies with their anatomic location. Glands within the gastric cardia comprise <5% of the gastric gland area and contain mucous and endocrine cells. The 75% of gastric glands are found within the oxyntic mucosa and contain mucous neck, parietal, chief, endocrine, enterochromaffin, and enterochromaffin-like (ECL) cells (Fig. 317-1). Pyloric glands contain mucous and endocrine cells (including gastrin cells) and are found in the antrum.

The parietal cell, also known as the oxyntic cell, is usually found in the neck, or isthmus, or in the oxyntic gland. The resting, or unstimulated, parietal cell has prominent cytoplasmic tubulovesicles and intracellular canaliculi containing short microvilli along its apical surface (Fig. 317-2). H^+,K^+-adenosine triphosphatase (ATPase) is expressed in the tubulovesicle membrane; upon cell stimulation, this membrane, along with apical membranes, transforms into a dense network of apical intracellular canaliculi containing long microvilli. Acid secretion, a process requiring high energy, occurs at the apical canalicular surface. Numerous mitochondria (30–40% of total cell volume) generate the energy required for secretion.

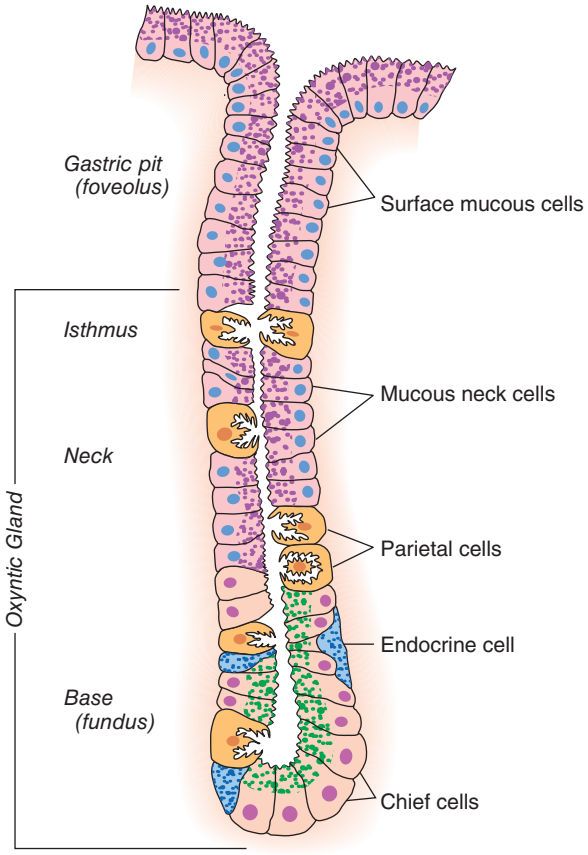


FIGURE 317-1 Diagrammatic representation of the oxyntic gastric gland. *(Adapted from S Ito, RJ Winchester: J Cell Biol 16:541, 1963. © 1963 Ito and Winchester.)*

Gastroduodenal Mucosal Defense The gastric epithelium is under constant assault by a series of endogenous noxious factors, including hydrochloric acid (HCl), pepsinogen/pepsin, and bile salts. In addition, a steady flow of exogenous substances such as medications, alcohol, and bacteria encounter the gastric mucosa. A highly

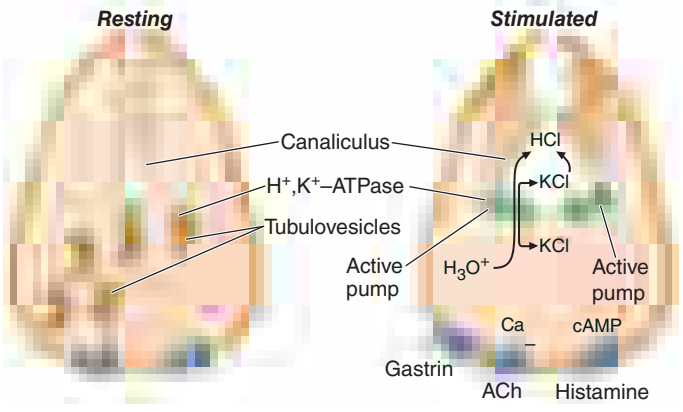


FIGURE 317-2 **Gastric parietal cell** undergoing transformation after secretagogue-mediated stimulation. cAMP, cyclic adenosine monophosphate. *(Adapted from SJ Hersey, G Sachs: Physiol Rev 75:155, 1995.)*

intricate biologic system is in place to provide defense from mucosal injury and to repair any injury that may occur.

The mucosal defense system can be envisioned as a three-level barrier, composed of preepithelial, epithelial, and subepithelial elements **(Fig. 317-3)**. The first line of defense is a mucus-bicarbonate-phospholipid layer, which serves as a physicochemical barrier to multiple molecules, including hydrogen ions. Mucus is secreted in a regulated fashion by gastroduodenal surface epithelial cells. It consists primarily of water (95%) and a mixture of phospholipids and glycoproteins (mucin). The mucous gel functions as a nonstirred water layer impeding diffusion of ions and molecules such as pepsin. Bicarbonate, secreted in a regulated manner by surface epithelial cells of the gastroduodenal mucosa into the mucous gel, forms a pH gradient ranging from 1 to 2 at the gastric luminal surface and reaching 6–7 along the epithelial cell surface.

Surface epithelial cells provide the next line of defense through several factors, including mucus production, epithelial cell ionic transporters that maintain intracellular pH and bicarbonate production, and intracellular tight junctions. Surface epithelial cells generate heat shock proteins that prevent protein denaturation and protect cells from certain factors such as increased temperature, cytotoxic agents, or oxidative stress. Epithelial cells also generate trefoil factor family peptides and cathelicidins, which also play a role in surface cell protection and regeneration. If the preepithelial barrier were breached, gastric epithelial cells bordering a site of injury can migrate to restore a damaged region (*restitution*). This process occurs independent of cell division and

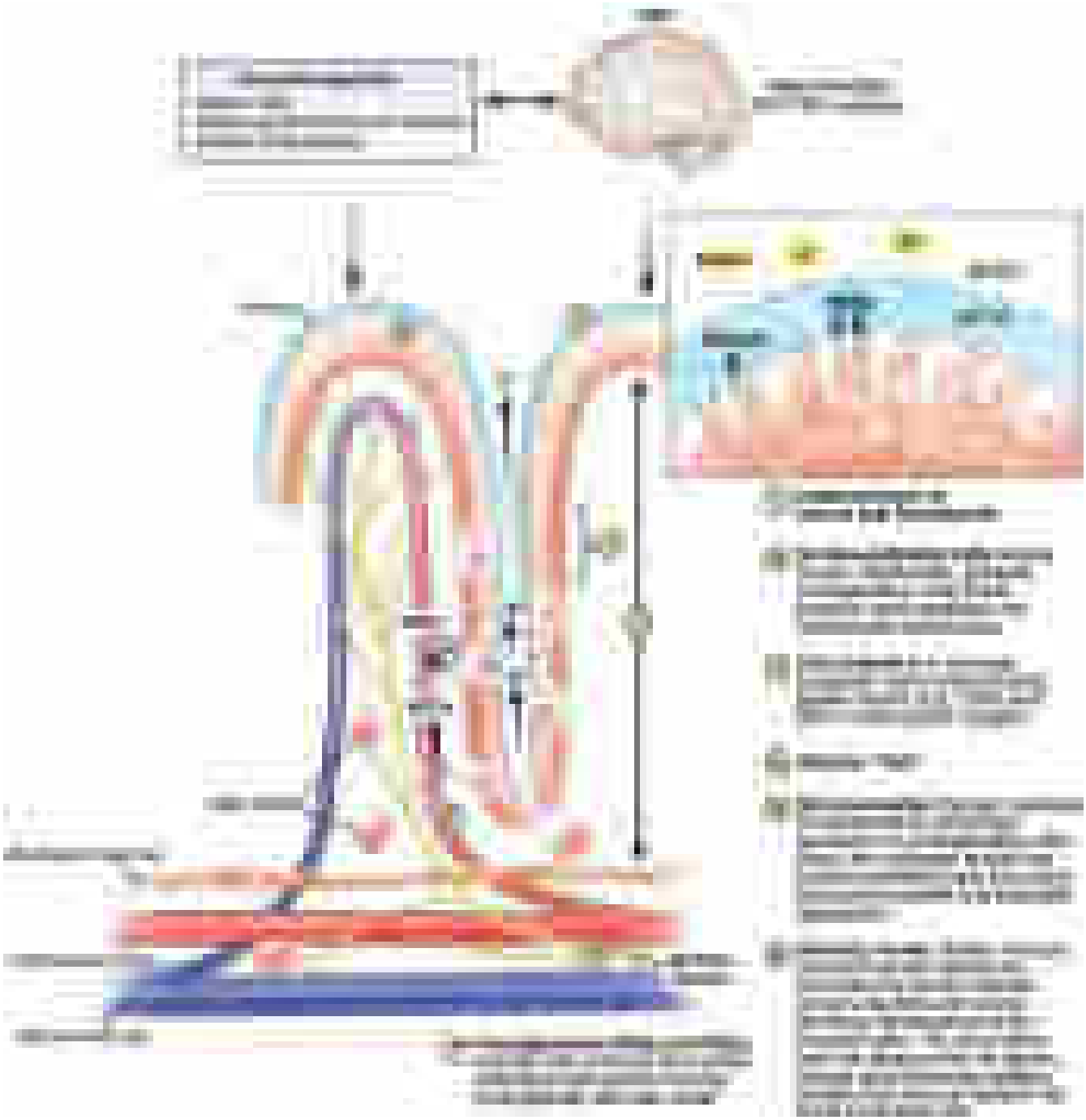


FIGURE 317-3 Components involved in providing gastroduodenal mucosal defense and repair. CCK, cholecystokinin; CRF, corticotropin-releasing factor; EGF, epidermal growth factor; HCl, hydrochloride; IGF, insulin-like growth factor; TGFα, transforming growth factor α; TRF, thyrotropin releasing factor. *(Modified and updated from A Tarnawski: Cellular and molecular mechanisms of mucosal defense and repair, in T Yoshikawa, T Arakawa [eds]: Bioregulation and Its Disorders in the Gastrointestinal Tract. Tokyo, Japan: Blackwell Science, 1998:3–17.)*

requires uninterrupted blood flow and an alkaline pH in the surrounding environment. Several growth factors, including epidermal growth factor (EGF), transforming growth factor (TGF) α, and basic fibroblast growth factor (FGF), modulate the process of restitution. Larger defects that are not effectively repaired by restitution require cell proliferation. Epithelial cell regeneration is regulated by prostaglandins and growth factors such as EGF and TGF-α. In tandem with epithelial cell renewal, formation of new vessels (*angiogenesis*) within the injured microvascular bed occurs. Both FGF and vascular endothelial growth factor (VEGF) are important in regulating angiogenesis in the gastric mucosa. In addition, the gastric peptide gastrin (see below) has been recently found to stimulate cell proliferation, migration, invasion and angiogenesis. It may also have the ability to regulate gastric stem cells.

An elaborate microvascular system within the gastric submucosal layer is the key component of the subepithelial defense/repair system, providing HCO_3^-, which neutralizes the acid generated by the parietal cell. Moreover, this microcirculatory bed provides an adequate supply of micronutrients and oxygen while removing toxic metabolic by-products. Several locally produced factors including nitric oxide (NO) (see below), hydrogen sulfide and prostacyclin contribute to the vascular protective pathway through vasodilation of the microcirculation.

Prostaglandins play a central role in gastric epithelial defense/repair **(Fig. 317-4)**. The gastric mucosa contains abundant levels of prostaglandins that regulate the release of mucosal bicarbonate and mucus, inhibit parietal cell secretion, and are important in maintaining mucosal blood flow and epithelial cell restitution. Prostaglandins are derived from esterified arachidonic acid, which is formed from phospholipids (cell membrane) by the action of phospholipase A_2. A key enzyme that controls the rate-limiting step in prostaglandin synthesis is cyclooxygenase (COX), which is present in two isoforms (COX-1, COX-2), each having distinct characteristics regarding structure, tissue distribution, and expression. COX-1 is expressed in a host of tissues, including the stomach, platelets, kidneys, and endothelial cells. This isoform is expressed in a constitutive manner and plays an important role in maintaining the integrity of renal function, platelet aggregation, and gastrointestinal (GI) mucosal integrity. In contrast, the expression of COX-2 is inducible by inflammatory stimuli, and it is expressed in macrophages, leukocytes, fibroblasts, and synovial cells. The beneficial effects of nonsteroidal anti-inflammatory drugs (NSAIDs) on tissue inflammation are due to inhibition of COX-2; the toxicity of these drugs (e.g., GI mucosal ulceration and renal dysfunction) is related to inhibition of the COX-1 isoform. The highly COX-2–selective NSAIDs have the potential to provide the beneficial effect of decreasing tissue inflammation while minimizing toxicity in the GI tract. Selective COX-2 inhibitors have had adverse effects on the cardiovascular (CV) system, leading to increased risk of myocardial infarction. Therefore, the U.S. Food and Drug Administration (FDA) has removed two of these agents (valdecoxib and rofecoxib) from the market (see below).

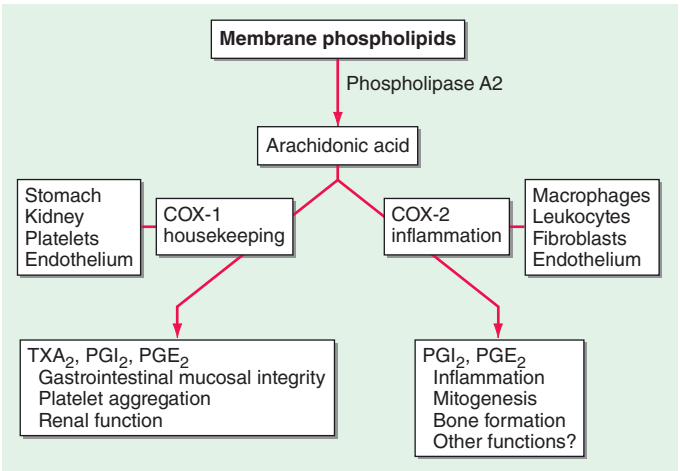


FIGURE 317-4 Schematic representation of the steps involved in synthesis of prostaglandin E_2 (PGE_2) and prostacyclin (PGI_2). Characteristics and distribution of the cyclooxygenase (COX) enzymes 1 and 2 are also shown. TXA_2, thromboxane A_2.

NO is important in the maintenance of gastric mucosal integrity. The key enzyme NO synthase is constitutively expressed in the mucosa and contributes to cytoprotection by stimulating gastric mucus, increasing mucosal blood flow, and maintaining epithelial cell barrier function. The central nervous system (CNS) and hormonal factors also play a role in regulating mucosal defense through multiple pathways (Fig. 317-3).

Since the discovery of *Helicobacter pylori* and its impact on gastric pathology, it has become clear that the stomach has an elaborate and complex inherent immunological system in place. Although a detailed description of the gastric immune system is beyond the scope of this chapter several features are worth highlighting. The gastric immune response to certain pathogens such as *H. pylori* (see below) involves extensive interplay between innate (dendritic cell, epithelial cells, neutrophils and macrophages) and adaptive (B and T cells) components. Helper T cells (Th and Th Reg cells) have been extensively studied and appear to play an important role in a broad array of gastric physiology extending from gastric secretion to epithelial cell turnover via production of a number of cytokines.

The discovery of *H. pylori* has also led to the understanding that the stomach, once thought to be devoid of microorganisms due to its highly adverse environment (acid and pepsin) can serve as host for bacterial communities consisting of hundreds of phylotypes, otherwise known as its microbiota. The conceptual framework of the microbiome has been receiving extensive attention in light of its importance in human health and disease. The overall relevance of the gastric microbiome and its impact on gastric pathology remains to be established but it is likely that alteration of microorganism homeostasis will play a role in aspects of certain disorders like PUD, gastritis and gastric cancer.

Physiology of Gastric Secretion HCl and pepsinogen are the two principal gastric secretory products capable of inducing mucosal injury. Gastric acid and pepsinogen play a physiologic role in protein digestion; absorption of iron, calcium, magnesium, and vitamin B_{12}; and killing ingested bacteria. Acid secretion should be viewed as occurring under basal and stimulated conditions. Basal acid production occurs in a circadian pattern, with highest levels occurring during the night and lowest levels during the morning hours. Cholinergic input via the vagus nerve and histaminergic input from local gastric sources are the principal contributors to basal acid secretion. Stimulated gastric acid secretion occurs primarily in three phases based on the site where the signal originates (cephalic, gastric, and intestinal). Sight, smell, and taste of food are the components of the cephalic phase, which stimulates gastric secretion via the vagus nerve. The gastric phase is activated once food enters the stomach. This component of secretion is driven by nutrients (amino acids and amines) that directly (via peptone and amino acid receptors) and indirectly (via stimulation of intramural gastrin releasing peptide neurons) stimulate the G cell to release gastrin, which in turn activates the parietal cell via direct and indirect mechanisms. Distention of the stomach wall also leads to gastrin release and acid production. The last phase of gastric acid secretion is initiated as food enters the intestine and is mediated by luminal distention and nutrient assimilation. A series of pathways that inhibit gastric acid production are also set into motion during these phases. The GI hormone somatostatin is released from endocrine cells found in the gastric mucosa (D cells) in response to HCl. Somatostatin can inhibit acid production by both direct (parietal cell) and indirect mechanisms (decreased histamine release from ECL cells, ghrelin release from Gr cells and gastrin release from G cells). Additional neural (central and peripheral) and humoral (amylin, atrial natriuretic peptide [ANP], cholecystokinin, ghrelin, interleukin 11 [IL-11], obestatin, secretin, and serotonin) factors play a role in counterbalancing acid secretion. Under physiologic circumstances, these phases occur simultaneously. Ghrelin, the appetite-regulating hormone expressed in Gr cells in the stomach, and its related peptide motilin (released from the duodenum) may increase gastric acid secretion through stimulation of histamine release from ECL cells, but this remains to be confirmed.

The acid-secreting parietal cell is located in the oxyntic gland, adjacent to other cellular elements (ECL cell, D cell) important in the gastric secretory process **(Fig. 317-5)**. This unique cell also secretes intrinsic factor (IF) and IL-11. The parietal cell expresses receptors for

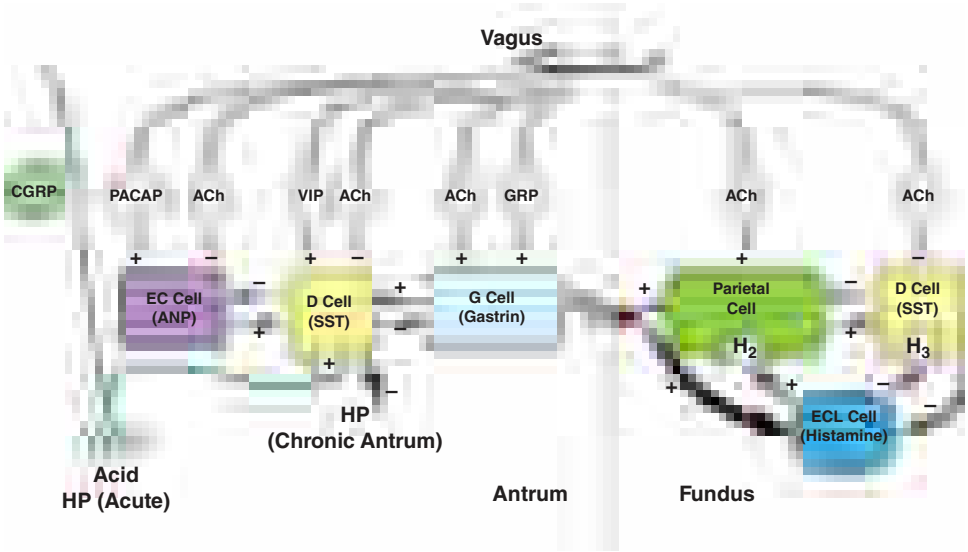


FIGURE 317-5 Regulation of gastric acid secretion at the cellular level. ACh, acetylcholine; ANP, atrial natriuretic peptide; CGRP, calcitonin gene-related peptide; EC, enterochromaffin; ECL, enterochromaffin-like; GRP, gastrin-releasing peptide; PACAP, pituitary adenylate-cyclase activating peptide; SST, somatostatin; VIP, vasoactive intestinal peptide.

several stimulants of acid secretion, including histamine (H_2), gastrin (cholecystokinin 2/gastrin receptor), and acetylcholine (muscarinic, M_3). Binding of histamine to the H_2 receptor leads to activation of adenylate cyclase and the phosphoinositol pathways in turn resulting in an increase in cyclic adenosine monophosphate (AMP) and intracellular calcium respectively. Activation of the gastrin and muscarinic receptors results in activation of the protein kinase C/phosphoinositide signaling pathway. Each of these signaling pathways in turn regulates a series of downstream kinase cascades that control the acid-secreting pump, H^+,K^+-ATPase. The discovery that different ligands and their corresponding receptors lead to activation of different signaling pathways explains the potentiation of acid secretion that occurs when histamine and gastrin or acetylcholine are combined. More importantly, this observation explains why blocking one receptor type (H_2) decreases acid secretion stimulated by agents that activate a different pathway (gastrin, acetylcholine). Parietal cells also express receptors for ligands that inhibit acid production (glucagon-like peptide-1, prostaglandins, somatostatin, and EGF). Histamine also stimulates gastric acid secretion indirectly by activating the histamine H_3 receptor on D-cells, which inhibits somatostatin release.

The enzyme H^+,K^+-ATPase is responsible for generating the large concentration of H^+. It is a membrane-bound protein that consists of two subunits, α and β. The active catalytic site is found within the α subunit; the function of the β subunit is unclear. This enzyme uses the chemical energy of adenosine triphosphate (ATP) to transfer H^+ ions from parietal cell cytoplasm to the secretory canaliculi in exchange for K^+. The H^+,K^+-ATPase is located within the secretory canaliculus and in nonsecretory cytoplasmic tubulovesicles. The tubulovesicles are impermeable to K^+, which leads to an inactive pump in this location. The distribution of pumps between the nonsecretory vesicles and the secretory canaliculus varies according to parietal cell activity (Fig. 317-2). Proton pumps are recycled back to the inactive state in cytoplasmic vesicles once parietal cell activation ceases. Ezrin (an actin binding protein), actin, myosin, soluble N-ethylmaleimide-sensitive factor attachment protein receptors (SNAREs), small G proteins of the Rab family and secretory carrier membrane proteins (SCAMPS) are postulated to participate in parietal cell membrane translocation. In addition, acid secretion requires a number of apical and basolateral parietal cell membrane chloride and potassium channels.

The chief cell, found primarily in the gastric fundus, synthesizes and secretes pepsinogen, the inactive precursor of the proteolytic enzyme pepsin. The acid environment within the stomach leads to cleavage of the inactive precursor to pepsin and provides the low pH (<2) required for pepsin activity. Pepsin activity is significantly diminished at a pH of 4 and irreversibly inactivated and denatured at a pH of ≥7. Many of the secretagogues that stimulate acid secretion also stimulate pepsinogen release. The precise role of pepsin in the pathogenesis of PUD remains to be established.

PATHOPHYSIOLOGIC BASIS OF PUD

PUD encompasses both gastric and duodenal ulcers (DUs). *Ulcers* are defined as breaks in the mucosal surface >5 mm in size, with depth to the submucosa. DUs and gastric ulcers (GUs) share many common features in terms of pathogenesis, diagnosis, and treatment, but several factors distinguish them from one another. *H. pylori* and NSAIDs are the most common risk factors for PUD, with estimated odds ratios in the United States of 3.7 and 3.3, respectively. Additional risk factors (odds ratio) include chronic obstructive lung disease (2.34), chronic renal insufficiency (2.29), current tobacco use (1.99), former tobacco use (1.55), older age (1.67), three or more doctor visits in a year (1.49), coronary heart disease (1.46), former alcohol use (1.29), African-American race (1.20), obesity (1.18), and diabetes (1.13). The mechanisms by which some of these risk factors lead to ulcer disease are highlighted below.

Epidemiology • DUODENAL ULCERS DUs are estimated to occur in 6–15% of the Western population. The incidence of DUs declined steadily from 1960 to 1980 and has remained stable since then. The death rates, need for surgery, and physician visits have decreased by >50% over the past 30 years. The reason for the reduction in the frequency of DUs is likely related to the decreasing frequency of *H. pylori.* Before the discovery of *H. pylori*, the natural history of DUs was typified by frequent recurrences after initial therapy. Eradication of *H. pylori* has reduced these recurrence rates by >80 %.

GASTRIC ULCERS GUs tend to occur later in life than duodenal lesions, with a peak incidence reported in the sixth decade. More than one-half of GUs occur in males and are less common than DUs, perhaps due to the higher likelihood of GUs being silent and presenting only after a complication develops. Autopsy studies suggest a similar incidence of DUs and GUs.

Pathology • DUODENAL ULCERS DUs occur most often in the first portion of the duodenum (>95%), with ~90% located within 3 cm of the pylorus. They are usually ≤1 cm in diameter but can occasionally reach 3–6 cm (giant ulcer). Ulcers are sharply demarcated, with depth at times reaching the muscularis propria. The base of the ulcer often consists of a zone of eosinophilic necrosis with surrounding fibrosis. Malignant DUs are extremely rare.

GASTRIC ULCERS In contrast to DUs, GUs can represent a malignancy and should be biopsied upon discovery. Benign GUs are most often found distal to the junction between the antrum and the acid secretory mucosa. Benign GUs are quite rare in the gastric fundus and are histologically similar to DUs. Benign GUs associated with *H. pylori* are also associated with antral gastritis. In contrast, NSAID-related GUs are not accompanied by chronic active gastritis but may instead have evidence of a chemical gastropathy, typified by foveolar hyperplasia, edema of the lamina propria, and epithelial regeneration in the absence of *H. pylori.* Extension of smooth-muscle fibers into the upper portions of the mucosa, where they are not typically found, may also occur.

Pathophysiology • DUODENAL ULCERS *H. pylori* and NSAID-induced injuries account for the majority of DUs. Many acid secretory abnormalities have been described in DU patients. Of these, average

basal and nocturnal gastric acid secretion appears to be increased in DU patients as compared to controls; however, the level of overlap between DU patients and control subjects is substantial. The reason for this altered secretory process is unclear, but *H. pylori* infection may contribute. Bicarbonate secretion is significantly decreased in the duodenal bulb of patients with an active DU as compared to control subjects. *H. pylori* infection may also play a role in this process (see below).

GASTRIC ULCERS As in DUs, the majority of GUs can be attributed to either *H. pylori* or NSAID-induced mucosal damage. Prepyloric GUs or those in the body associated with a DU or a duodenal scar are similar in pathogenesis to DUs. Gastric acid output (basal and stimulated) tends to be normal or decreased in GU patients. When GUs develop in the presence of minimal acid levels, impairment of mucosal defense factors may be present. GUs have been classified based on their location: Type I occur in the gastric body and tend to be associated with low gastric acid production; type II occur in the antrum and gastric acid can vary from low to normal; type III occur within 3 cm of the pylorus and are commonly accompanied by DUs and normal or high gastric acid production; and type IV are found in the cardia and are associated with low gastric acid production.

***H. PYLORI* AND ACID PEPTIC DISORDERS** Gastric infection with the bacterium *H. pylori* accounts for the majority of PUD **(Chap. 158)**. This organism also plays a role in the development of gastric mucosa-associated lymphoid tissue (MALT) lymphoma and gastric adenocarcinoma. Although the entire genome of *H. pylori* has been sequenced, it is still not clear how this organism, which resides in the stomach, causes ulceration in the duodenum. *H. pylori* eradication efforts may lead to a decrease in gastric cancer in high-risk populations particularly in individuals who have not developed chronic atrophic gastritis and gastric metaplasia.

The Bacterium The bacterium, initially named *Campylobacter pyloridis*, is a gram-negative microaerophilic rod found most commonly in the deeper portions of the mucous gel coating the gastric mucosa or between the mucous layer and the gastric epithelium. It may attach to gastric epithelium but under normal circumstances does not appear to invade cells. It is strategically designed to live within the aggressive environment of the stomach. It is S-shaped (~0.5–3 μm in size) and contains multiple sheathed flagella. Initially, *H. pylori* resides in the antrum but, over time, migrates toward the more proximal segments of the stomach. The organism is capable of transforming into a coccoid form, which represents a dormant state that may facilitate survival in adverse conditions. The genome of *H. pylori* (1.65 million base pairs) encodes ~1500 proteins. Among this multitude of proteins there are factors that are essential determinants of *H. pylori*–mediated pathogenesis and colonization such as the outer membrane protein (Hop proteins), urease, and the vacuolating cytotoxin (Vac A). Moreover, the majority of *H. pylori* strains contain a genomic fragment that encodes the cag pathogenicity island (cag-PAI). Several of the genes that make up cag-PAI encode components of a type IV secretion island that translocates Cag A into host cells. Once in the cell, Cag A activates a series of cellular events important in cell growth and cytokine production. *H. pylori* also has extensive genetic diversity that in turn enhances its ability to promote disease. The first step in infection by *H. pylori* is dependent on the bacteria's motility and its ability to produce urease. Urease produces ammonia from urea, an essential step in alkalinizing the surrounding pH. Additional bacterial factors include catalase, lipase, adhesins, platelet-activating factor, and pic B (induces cytokines). Multiple strains of *H. pylori* exist and are characterized by their ability to express several of these factors (Cag A, Vac A, etc.). It is possible that the different diseases related to *H. pylori* infection can be attributed to different strains of the organism with distinct pathogenic features.

Epidemiology The prevalence of *H. pylori* varies throughout the world and depends largely on the overall standard of living in the region. In developing parts of the world, 80% of the population may be infected by the age of 20, whereas the prevalence is 20–50% in industrialized countries. In contrast, in the United States this organism is rare in childhood. The overall prevalence of *H. pylori* in the United States is ~30%, with individuals born before 1950 having a higher rate of infection than those born later. About 10% of Americans <30 years of age are colonized with the bacteria. The rate of infection with *H. pylori* in industrialized countries has decreased substantially in recent decades. The steady increase in the prevalence of *H. pylori* noted with increasing age is due primarily to a cohort effect, reflecting higher transmission during a period in which the earlier cohorts were children. It has been calculated through mathematical models that improved sanitation during the latter half of the nineteenth century dramatically decreased transmission of *H. pylori.* Moreover, with the present rate of intervention, the organism will be ultimately eliminated from the United States. Two factors that predispose to higher colonization rates include poor socioeconomic status and less education. These factors, not race, are responsible for the rate of *H. pylori* infection in blacks and Hispanic Americans being double the rate seen in whites of comparable age. Other risk factors for *H. pylori* infection are (1) birth or residence in a developing country, (2) domestic crowding, (3) unsanitary living conditions, (4) unclean food or water, and (5) exposure to gastric contents of an infected individual.

Transmission of *H. pylori* occurs from person to person, following an oral-oral or fecal-oral route. The risk of *H. pylori* infection is declining in developing countries. The rate of infection in the United States has fallen by >50% when compared to 30 years ago.

Pathophysiology *H. pylori* infection is virtually always associated with a chronic active gastritis, but only 10–15% of infected individuals develop frank peptic ulceration. The basis for this difference is unknown, but is likely due to a combination of host and bacterial factors some of which are outlined below. Initial studies suggested that >90% of all DUs were associated with *H. pylori*, but *H. pylori* is present in only 30–60% of individuals with GUs and 50–70% of patients with DUs. The pathophysiology of ulcers not associated with *H. pylori* or NSAID ingestion (or the rare Zollinger-Ellison syndrome [ZES]) is becoming more relevant as the incidence of *H. pylori* is dropping, particularly in the Western world (see below).

The particular end result of *H. pylori* infection (gastritis, PUD, gastric MALT lymphoma, gastric cancer) is determined by a complex interplay between bacterial and host factors **(Fig. 317-6)**.

Bacterial factors: H. pylori is able to facilitate gastric residence, induce mucosal injury, and avoid host defense. Different strains of *H. pylori* produce different virulence factors including γ-glutamyl transpeptidase (GGT), cytotoxin-associated gene A (cagA) product, and virulence components vacuolating toxin (vacA), in addition to pathogen-associated molecular patterns (PAMPs) such as flagella and lipopolysaccharide (LPS). A specific region of the bacterial genome, the pathogenicity island (cag-PAI), encodes the virulence factors Cag A and pic B. Vac A also contributes to pathogenicity, although it is not encoded within the pathogenicity island. These virulence factors, in conjunction with additional bacterial constituents, can cause mucosal damage, in part through their ability to target the host immune cells.

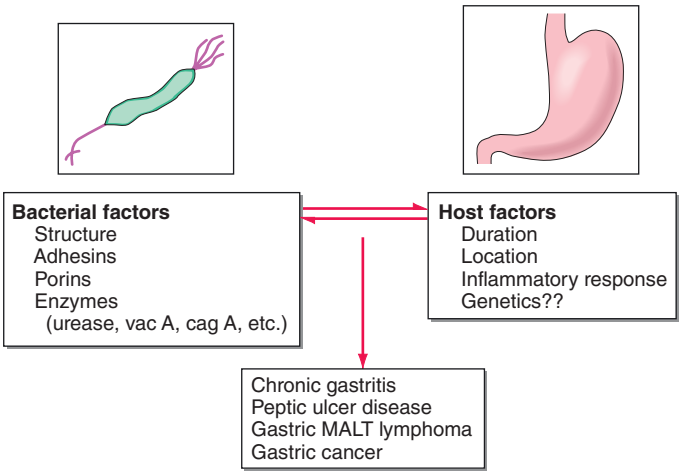


FIGURE 317-6 Outline of the bacterial and host factors important in determining *H. pylori*–induced gastrointestinal disease. MALT, mucosal-associated lymphoid tissue.

For example, Vac A targets human CD4 T cells, inhibiting their proliferation and in addition can disrupt normal function of B cells, CD8 T cells, macrophages, and mast cells. Multiple studies have demonstrated that *H. pylori* strains that are cag-PAI positive are associated with a higher risk of PUD, premalignant gastric lesions, and gastric cancer than are strains that lack the cag-PAI. In addition, *H. pylori* may directly inhibit parietal cell H^+,K^+-ATPase activity through a Cag A–dependent mechanism, leading in part to the low acid production observed after acute infection with the organism. Urease, which allows the bacteria to reside in the acidic stomach, generates NH_3, which can damage epithelial cells. The bacteria produce surface factors that are chemotactic for neutrophils and monocytes, which in turn contribute to epithelial cell injury (see below). *H. pylori* makes proteases and phospholipases that break down the glycoprotein lipid complex of the mucous gel, thus reducing the efficacy of this first line of mucosal defense. *H. pylori* expresses adhesins (OMPs like BabA), which facilitate attachment of the bacteria to gastric epithelial cells. Although LPS of gram-negative bacteria often plays an important role in the infection, *H. pylori* LPS has low immunologic activity compared to that of other organisms. It may promote a smoldering chronic inflammation.

Host factors: Studies in twins suggest that there may be genetic predisposition to acquire *H. pylori.* The inflammatory response to *H. pylori* includes recruitment of neutrophils, lymphocytes (T and B), macrophages, and plasma cells. The pathogen leads to local injury by binding to class II major histocompatibility complex (MHC) molecules expressed on gastric epithelial cells, leading to cell death (*apoptosis*). Moreover, bacterial strains that encode cag-PAI can introduce Cag A into the host cells, leading to further cell injury and activation of cellular pathways involved in cytokine production and repression of tumor-suppressor genes. Elevated concentrations of multiple cytokines are found in the gastric epithelium of *H. pylori*–infected individuals, including interleukin (IL) 1α/β, IL-2, IL-6, IL-8, tumor necrosis factor (TNF) α, and interferon (IFN) γ. *H. pylori* infection also leads to both a mucosal and a systemic humoral response, which does not lead to eradication of the bacteria but further compounds epithelial cell injury. Additional mechanisms by which *H. pylori* may cause epithelial cell injury include (1) activated neutrophil-mediated production of reactive oxygen or nitrogen species and enhanced epithelial cell turnover and (2) apoptosis related to interaction with T cells (T helper 1, or T_H1, cells) and IFN-γ. Finally, the human stomach is colonized by a host of commensal organisms that may affect the likelihood of *H. pylori*–infection and subsequent mucosal injury. Moreover, colonization of the stomach with *H. pylori* likely alters the composition of the gastric microbiota. The impact of the latter on gastric pathophysiology remains unknown. *H. pylori* also appears to regulate NO formation via different mechanisms that in turn may contribute to the organism's cytotoxic effects. Specifically, *H. pylori* derived factors, such as urease, or the bacterium itself, stimulate NO synthase (NOS2) expression in macrophages and in gastric epithelial cells leading to NO release and subsequent cytotoxic effect on surrounding cells. *H. pylori* also leads to the formation of 8-nitroguanine (8-NO2-Gua), which in conjunction with oncoprotein CagA, may contribute to the development of gastric cancer.

The reason for *H. pylori*–mediated duodenal ulceration remains unclear. Studies suggest that *H. pylori* associated with duodenal ulceration may be more virulent. In addition, certain specific bacterial factors such as the DU-promoting gene A (*dupA*), may be associated with the development of DUs. Another potential contributing factor is that gastric metaplasia in the duodenum of DU patients, which may be due to high acid exposure (see below), permits *H. pylori* to bind to it and produce local injury secondary to the host response. Another hypothesis is that *H. pylori* antral infection could lead to increased acid production, increased duodenal acid, and mucosal injury. Basal and stimulated (meal, gastrin-releasing peptide [GRP]) gastrin release are increased in *H. pylori*–infected individuals, and somatostatin-secreting D cells may be decreased. *H. pylori* infection might induce increased acid secretion through both direct and indirect actions of *H. pylori* and proinflammatory cytokines (IL-8, TNF, and IL-1) on G, D, and parietal cells **(Fig. 317-7)**. GUs, in contrast, are associated with *H. pylori*–induced

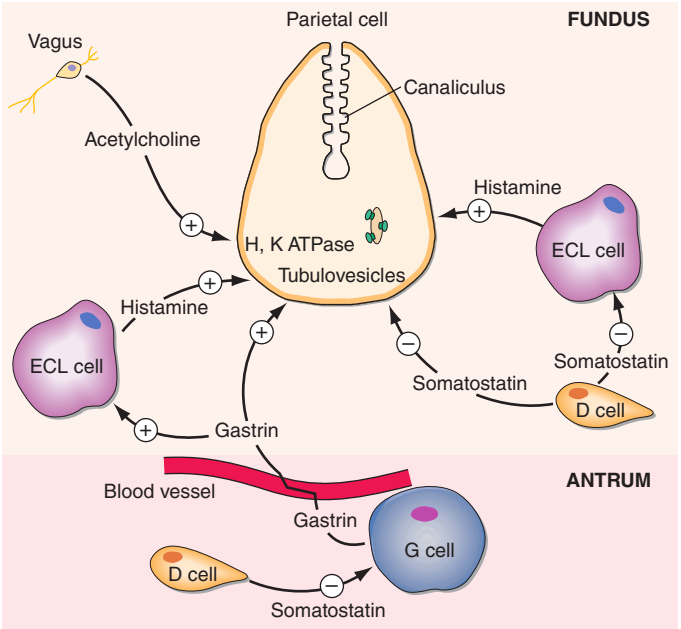


FIGURE 317-7 Summary of potential mechanisms by which *H. pylori* may lead to gastric secretory abnormalities. D, somatostatin cell; ECL, enterochromaffin-like cell; G, G cell. *(Adapted from J Calam et al: Gastroenterology 113:543, 1997.)*

pangastritis and normal or low gastric acid secretion. *H. pylori* infection has also been associated with decreased duodenal mucosal bicarbonate production. Data supporting and contradicting each of these interesting theories have been demonstrated. Thus, the mechanism by which *H. pylori* infection of the stomach leads to duodenal ulceration remains to be established.

In summary, the final effect of *H. pylori* on the GI tract is variable and determined by microbial and host factors. The type and distribution of gastritis correlate with the ultimate gastric and duodenal pathology observed. Specifically, the presence of antral-predominant gastritis is associated with DU formation; gastritis involving primarily the corpus predisposes to the development of GUs, gastric atrophy, and ultimately gastric carcinoma **(Fig. 317-8)**.

NSAID-INDUCED DISEASE • Epidemiology NSAIDs represent a group of the most commonly used medications in the world and the United States. It is estimated that 7 billion dollars per year are spent on NSAIDs

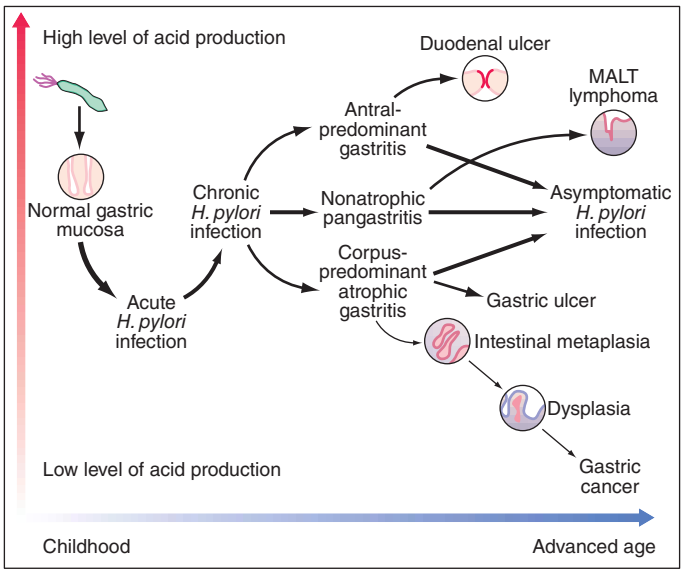


FIGURE 317-8 Natural history of *H. pylori* infection. MALT, mucosal-associated lymphoid tissue. *(Used with permission from S Suerbaum, P Michetti: N Engl J Med 347:1175, 2002.)*

world wide with more than 30 billion over-the-counter tablets and over 100 million prescriptions sold yearly in the United States alone. In fact, after the introduction of COX-2 inhibitors in the year 2000, the number of prescriptions written for NSAIDs was >111 million at a cost of $4.8 billion. Side effects and complications due to NSAIDs are considered the most common drug-related toxicities in the United States. The spectrum of NSAID-induced morbidity ranges from nausea and dyspepsia (prevalence reported as high as 50–60%) to a serious GI complication such as endoscopy-documented peptic ulceration (15–30% of individuals taking NSAIDs regularly) complicated by bleeding or perforation in as many as 1.5% of users per year. It is estimated that NSAID-induced GI bleeding accounts for 60,000–120,000 hospital admissions per year, and deaths related to NSAID-induced toxicity may be as high as 16,000 per year in the United States. Approximately 4–5% of patients develop symptomatic ulcers within 1 year. Unfortunately, dyspeptic symptoms do not correlate with NSAID-induced pathology. Over 80% of patients with serious NSAID-related complications did not have preceding dyspepsia. In view of the lack of warning signs, it is important to identify patients who are at increased risk for morbidity and mortality related to NSAID usage. Even 75 mg/d of aspirin may lead to serious GI ulceration; thus, no dose of NSAID is completely safe. In fact, the incidence of mucosal injury (ulcers and erosions) in patients taking low-dose aspirin (75–325 mg) has been estimated to range from as low as 8 to as high as 60%. It appears that *H. pylori* infection increases the risk of PUD-associated GI bleeding in chronic users of low-dose aspirin. Established risk factors include advanced age, history of ulcer, concomitant use of glucocorticoids, high-dose NSAIDs, multiple NSAIDs, concomitant use of anticoagulants, clopidogrel, and serious or multisystem disease. Possible risk factors include concomitant infection with *H. pylori*, cigarette smoking, and alcohol consumption.

Pathophysiology Prostaglandins play a critical role in maintaining gastroduodenal mucosal integrity and repair. It therefore follows that interruption of prostaglandin synthesis can impair mucosal defense and repair, thus facilitating mucosal injury via a systemic mechanism. Animal studies have demonstrated that neutrophil adherence to the gastric microcirculation plays an essential role in the initiation of NSAID-induced mucosal injury. A summary of the pathogenetic pathways by which systemically administered NSAIDs may lead to mucosal injury is shown in Fig. 317-9. Single nucleotide polymorphisms (SNPs) have been found in several genes, including those encoding certain subtypes of cytochrome P450 (see below), interleukin-1β (*IL-1β*), angiotensinogen (*AGT*), and an organic ion transporting polypeptide (*SLCO1B1*), but these findings need confirmation in larger scale studies.

Injury to the mucosa also occurs as a result of the topical encounter with NSAIDs leading to increased epithelial surface permeability. Aspirin and many NSAIDs are weak acids that remain in a nonionized lipophilic form when found within the acid environment of the stomach. Under these conditions, NSAIDs migrate across lipid membranes of epithelial cells, leading to cell injury once trapped intracellularly in an ionized form. Topical NSAIDs can also alter the surface mucous layer, permitting back diffusion of H^+ and pepsin, leading to further epithelial cell damage. Moreover, enteric-coated or buffered preparations are also associated with risk of peptic ulceration. NSAIDs can also lead to mucosal injury via production of additional pro-inflammatory mediators like TNF and leukotrienes through simultaneous activation of the lipoxygenase pathway.

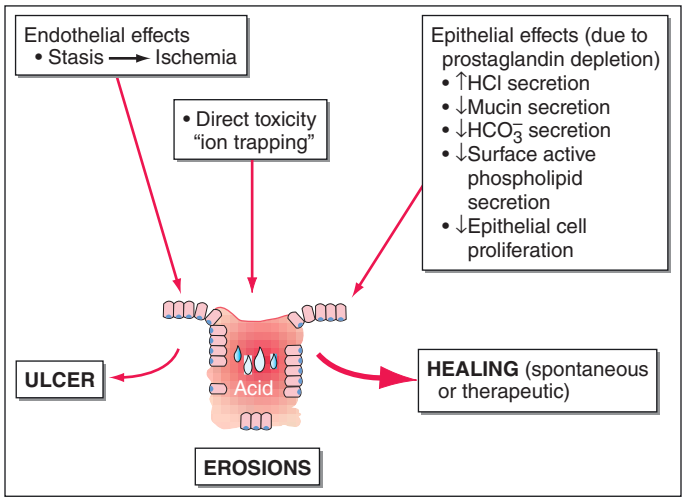


FIGURE 317-9 Mechanisms by which nonsteroidal anti-inflammatory drugs may induce mucosal injury. *(Adapted from J Scheiman et al: J Clin Outcomes Management 3:23, 1996. Copyright 2003 Turner White Communications, Inc., www.turner-white.com. Used with permission.)*

The interplay between *H. pylori* and NSAIDs in the pathogenesis of PUD is complex. Meta-analysis supports the conclusion that each of these aggressive factors is independent and synergistic risk factors for PUD and its complications such as GI bleeding. For example, eradication of *H. pylori* reduces the likelihood of GI complications in high-risk individuals to levels observed in individuals with average risk of NSAID-induced complications.

PATHOGENETIC FACTORS UNRELATED TO *H. PYLORI* AND NSAIDS IN ACID PEPTIC DISEASE Cigarette smoking has been implicated in the pathogenesis of PUD. Not only have smokers been found to have ulcers more frequently than do nonsmokers, but smoking appears to decrease healing rates, impair response to therapy, and increase ulcer-related complications such as perforation. The mechanism responsible for increased ulcer diathesis in smokers is unknown. Theories have included altered gastric emptying, decreased proximal duodenal bicarbonate production, increased risk for *H. pylori* infection, and cigarette-induced generation of noxious mucosal free radicals. Genetic predisposition may play a role in ulcer development. First-degree relatives of DU patients are three times as likely to develop an ulcer; however, the potential role of *H. pylori* infection in contacts is a major consideration. Increased frequencies of blood group O and of the nonsecretor status have also been implicated as genetic risk factors for peptic diathesis. However, *H. pylori* preferentially binds to group O antigens. Additional genetic factors have been postulated to predispose certain individuals to developing PUD and/or upper GI bleeding. Specifically, genes encoding the NSAID-metabolizing enzymes cytochrome P450 2C9 and 2C8 (CYP2C9 and CYP2C8) are potential susceptibility genes for NSAID-induced PUD, but unfortunately, the studies have not been consistent in demonstrating this association. In a United Kingdom study, the *CYP2C19*17* gain-of-function polymorphism was associated with PUD in a Caucasian cohort, irrespective of ulcer etiology. These findings need to be confirmed in broader studies. Psychological stress has been thought to contribute to PUD, but studies examining the role of psychological factors in its pathogenesis have generated conflicting results. Although PUD is associated with certain personality traits (neuroticism), these same traits are also present in individuals with nonulcer dyspepsia (NUD) and other functional and organic disorders.

Diet has also been thought to play a role in peptic diseases. Certain foods and beverages can cause dyspepsia, but no convincing studies indicate an association between ulcer formation and a specific diet. Specific chronic disorders have been shown to have a strong association with PUD: (1) advanced age, (2) chronic pulmonary disease, (3) chronic renal failure, (4) cirrhosis, (5) nephrolithiasis, (6) α_1-antitrypsin deficiency, and (7) systemic mastocytosis. Disorders with a possible association are (1) hyperparathyroidism, (2) coronary artery disease, (3) polycythemia vera, (4) chronic pancreatitis, (5) former alcohol use, (6) obesity, (7) African-American race, and (8) three or more doctor visits in a year.

Multiple factors play a role in the pathogenesis of PUD. The two predominant causes are *H. pylori* infection and NSAID ingestion. PUD not related to *H. pylori* or NSAIDs is increasing. Other less common causes of PUD are shown in Table 317-1. These etiologic agents should be considered as the incidence of *H. pylori* is decreasing. Independent of the inciting or injurious agent, peptic ulcers develop as a result of an imbalance between mucosal protection/repair and aggressive factors. Gastric acid plays an important role in mucosal injury.

CLINICAL FEATURES

History Abdominal pain is common to many GI disorders, including DU and GU, but has a poor predictive value for the presence of

TABLE 317-1 Causes of Ulcers Not Caused by *Helicobacter pylori* and NSAIDs

Pathogenesis of Non-Hp and Non-NSAID Ulcer Disease
Infection
Cytomegalovirus
Herpes simplex virus
Helicobacter heilmannii
Drug/Toxin
Bisphosphonates
Chemotherapy
Clopidogrel
Crack cocaine
Glucocorticoids (when combined with NSAIDs)
Mycophenolate mofetil
Potassium chloride
Miscellaneous
Basophilia in myeloproliferative disease
Duodenal obstruction (e.g., annular pancreas)
Infiltrating disease
Ischemia
Radiation therapy
Eosinophilic infiltration
Sarcoidosis
Crohn's disease
Idiopathic hypersecretory state

Abbreviations: Hp, *H. pylori*; NSAIDs, nonsteroidal anti-inflammatory drugs.

either DU or GU. Up to 87% of patients with NSAID-induced mucosal disease can present with a complication (bleeding, perforation, and obstruction) without antecedent symptoms. Despite this poor correlation, a careful history and physical examination are essential components of the approach to a patient suspected of having peptic ulcers.

Epigastric pain described as a burning or gnawing discomfort can be present in both DU and GU. The discomfort is also described as an ill-defined, aching sensation or as hunger pain. The typical pain pattern in DU occurs 90 min to 3 h after a meal and is frequently relieved by antacids or food. Pain that awakes the patient from sleep (between midnight and 3 A.M.) is the most discriminating symptom, with two-thirds of DU patients describing this complaint. Unfortunately, this symptom is also present in one-third of patients with NUD (see below). Elderly patients are less likely to have abdominal pain as a manifestation of PUD and may instead present with a complication such as ulcer bleeding or perforation. The pain pattern in GU patients may be different from that in DU patients, where discomfort may actually be precipitated by food. Nausea and weight loss occur more commonly in GU patients. Endoscopy detects ulcers in <30% of patients who have dyspepsia.

The mechanism for development of abdominal pain in ulcer patients is unknown. Several possible explanations include acid-induced activation of chemical receptors in the duodenum, enhanced duodenal sensitivity to bile acids and pepsin, or altered gastroduodenal motility.

Variation in the intensity or distribution of the abdominal pain, as well as the onset of associated symptoms such as nausea and/or vomiting, may be indicative of an ulcer complication. Dyspepsia that becomes constant, is no longer relieved by food or antacids, or radiates to the back may indicate a penetrating ulcer (pancreas). Sudden onset of severe, generalized abdominal pain may indicate perforation. Pain worsening with meals, nausea, and vomiting of undigested food suggest gastric outlet obstruction. Tarry stools or coffee-ground emesis indicate bleeding.

Physical Examination Epigastric tenderness is the most frequent finding in patients with GU or DU. Pain may be found to the right of the midline in 20% of patients. Unfortunately, the predictive value of this finding is low. Physical examination is critically important for discovering evidence of ulcer complication. Tachycardia and orthostasis suggest dehydration secondary to vomiting or active GI blood loss. A severely tender, board-like abdomen suggests a perforation. Presence of a succussion splash indicates retained fluid in the stomach, suggesting gastric outlet obstruction.

PUD-Related Complications • **GASTROINTESTINAL BLEEDING** GI bleeding is the most common complication observed in PUD. Bleeding is estimated to occur in 19.4–57 per 100,000 individuals in a general population or in ~15% of patients. Bleeding and complications of ulcer disease occur more often in individuals >60 years of age. The 30-day mortality rate is as high as 2.5–10%. The higher incidence in the elderly is likely due to the increased use of NSAIDs in this group. In addition, up to 80% of the mortality in PUD-related bleeding is due to nonbleeding causes such as multiorgan failure (24%), pulmonary complications (24%), and malignancy (34%).

Greater than 50% of patients with ulcer-related hemorrhage bleed without any preceding warning signs or symptoms.

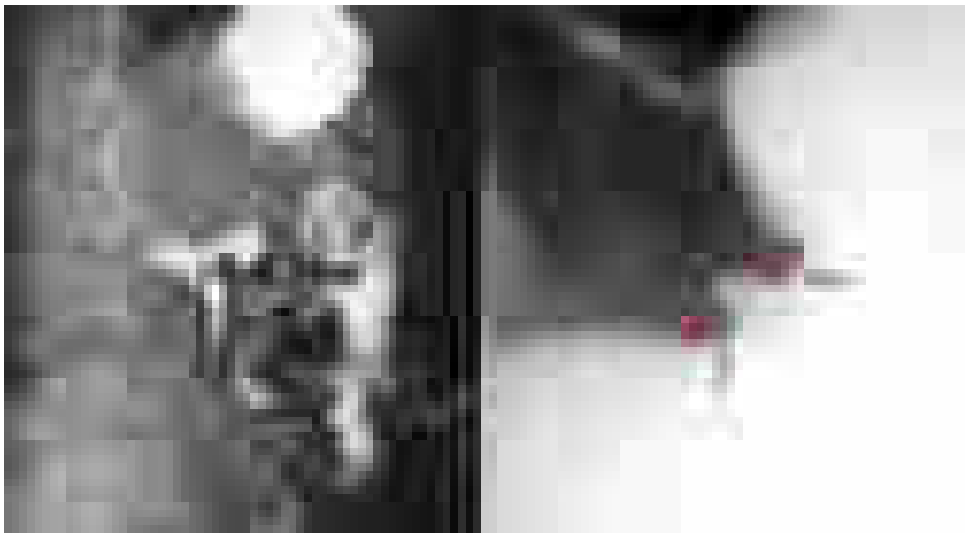
PERFORATION The second most common ulcer-related complication is perforation, being reported in as many as 6–7% of PUD patients with an estimated 30-day mortality of >20%. As in the case of bleeding, the incidence of perforation in the elderly appears to be increasing secondary to increased use of NSAIDs. Perforation of DUs has become less common in light of the increased rates of *H. pylori* eradication with NSAID induced GUs leading to perforation occurring more commonly. *Penetration* is a form of perforation in which the ulcer bed tunnels into an adjacent organ. DUs tend to penetrate posteriorly into the pancreas, leading to pancreatitis, whereas GUs tend to penetrate into the left hepatic lobe. Gastrocolic fistulas associated with GUs have also been described.

GASTRIC OUTLET OBSTRUCTION Gastric outlet obstruction is the least common ulcer-related complication, occurring in 1–2% of patients. A patient may have relative obstruction secondary to ulcer-related inflammation and edema in the peripyloric region. This process often resolves with ulcer healing. A fixed, mechanical obstruction secondary to scar formation in the peripyloric areas is also possible. The latter requires endoscopic (balloon dilation) or surgical intervention. Signs and symptoms relative to mechanical obstruction may develop insidiously. New onset of early satiety, nausea, vomiting, increase of postprandial abdominal pain, and weight loss should make gastric outlet obstruction a possible diagnosis.

Differential Diagnosis The list of GI and non-GI disorders that can mimic ulceration of the stomach or duodenum is quite extensive. The most commonly encountered diagnosis among patients seen for upper abdominal discomfort is functional dyspepsia (FD) or *essential dyspepsia* which refers to a group of heterogeneous disorders typified by upper abdominal pain without the presence of an ulcer. The symptoms can range from postprandial fullness and early satiety to epigastric burning pain. The dichotomy of this symptom complex has led to the identification of two subcategories of FD including postprandial distress syndrome (PDS) and epigastric pain syndrome (EPS). Dyspepsia has been reported to occur in up to 30% of the U.S. population. Up to 80% of patients seeking medical care for dyspepsia have a negative diagnostic evaluation. The etiology of FD is not established, but recent studies suggest that post-infectious states, certain foods and *H. pylori* infection may contribute to the pathogenesis of this common disorder.

Several additional disease processes that may present with "ulcer-like" symptoms include proximal GI tumors, gastroesophageal reflux, vascular disease, pancreaticobiliary disease (biliary colic, chronic pancreatitis), and gastroduodenal Crohn's disease.

Diagnostic Evaluation In view of the poor predictive value of abdominal pain for the presence of a gastroduodenal ulcer and the multiple disease processes that can mimic this disease, the clinician is often confronted with having to establish the presence of an ulcer. Documentation of an ulcer requires either a radiographic (barium study) or an endoscopic procedure. However, a large percentage of patients with symptoms suggestive of an ulcer have NUD; testing for *H. pylori* and antibiotic therapy (see below) is appropriate for individuals who



A B

FIGURE 317-10 Barium study demonstrating (***A***) a benign duodenal ulcer and (***B***) a benign gastric ulcer.

are otherwise healthy and <45 years of age, before embarking on a diagnostic evaluation **(Chap. 41)**.

Barium studies of the proximal GI tract are rarely used as a first test for documenting an ulcer. The sensitivity of older single-contrast barium meals for detecting a DU is as high as 80%, with a double-contrast study providing detection rates as high as 90%. Sensitivity for detection is decreased in small ulcers (<0.5 cm), with presence of previous scarring, or in postoperative patients. A DU appears as a well-demarcated crater, most often seen in the bulb **(Fig. 317-10*A*)**. A GU may represent benign or malignant disease. Typically, a benign GU also appears as a discrete crater with radiating mucosal folds originating from the ulcer margin **(Fig. 317-10*B*)**. Ulcers >3 cm in size or those associated with a mass are more often malignant. Unfortunately, up to 8% of GUs that appear to be benign by radiographic appearance are malignant by endoscopy or surgery. Radiographic studies that show a GU must be followed by endoscopy and biopsy.

Endoscopy provides the most sensitive and specific approach for examining the upper GI tract **(Fig. 317-11)**. In addition to permitting direct visualization of the mucosa, endoscopy facilitates photographic documentation of a mucosal defect and tissue biopsy to rule out malignancy (GU) or *H. pylori.* Endoscopic examination is particularly helpful in identifying lesions too small to detect by radiographic examination, for evaluation of atypical radiographic abnormalities, or to determine if an ulcer is a source of blood loss.

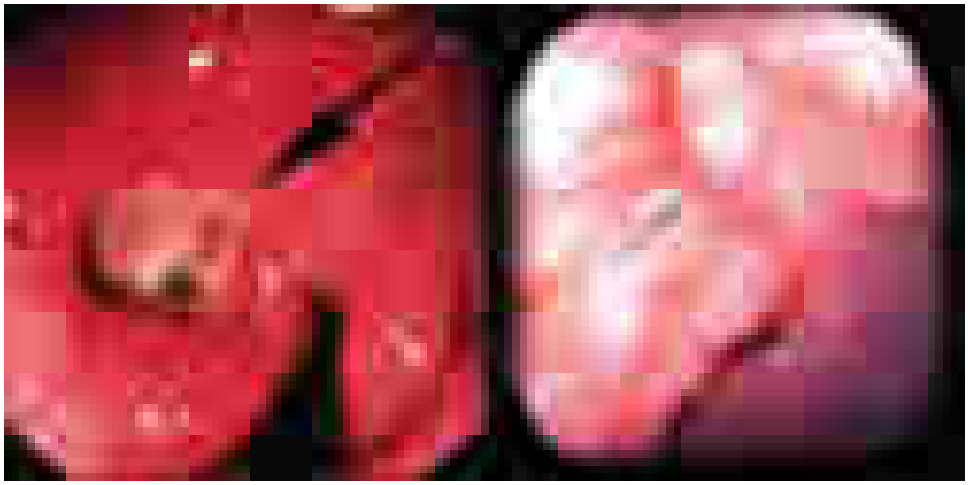
Although the methods for diagnosing *H. pylori* are outlined in **Chap. 158**, a brief summary will be included here **(Table 317-2)**. Several biopsy urease tests have been developed (PyloriTek, CLOtest, Hpfast, Pronto Dry) that have a sensitivity and specificity of >90–95%. Several noninvasive methods for detecting this organism have been developed. Three types of studies routinely used include serologic testing, the ^{13}C- or ^{14}C-urea breath test, and the fecal *H. pylori* (Hp) antigen test (monoclonal antibody test). A urinary Hp antigen test appears promising.

Occasionally, specialized testing such as serum gastrin and gastric acid analysis or sham feeding may be needed in individuals with complicated or refractory PUD (see "Zollinger-Ellison Syndrome [ZES]," below). Screening for aspirin or NSAIDs (blood or urine) may also be necessary in refractory *H. pylori*–negative PUD patients.

TREATMENT

Peptic Ulcer Disease

Before the discovery of *H. pylori*, the therapy of PUD was centered on the old dictum by Schwartz of "no acid, no ulcer." Although acid secretion is still important in the pathogenesis of PUD, eradication of *H. pylori* and therapy/prevention of NSAID-induced disease is the mainstay of treatment. A summary of commonly used drugs for treatment of acid peptic disorders is shown in **Table 317-3**.



A B

FIGURE 317-11 Endoscopy demonstrating (***A***) a benign duodenal ulcer and (***B***) a benign gastric ulcer.

TABLE 317-2 Tests for Detection of *Helicobacter pylori*

TEST	SENSITIVITY/ SPECIFICITY, %	COMMENTS
Invasive (Endoscopy/Biopsy Required)		
Rapid urease	80–95/95–100	Simple, false negative with recent use of PPIs, antibiotics, or bismuth compounds
Histology	80–90/>95	Requires pathology processing and staining; provides histologic information
Culture	—/—	Time-consuming, expensive, dependent on experience; allows determination of antibiotic susceptibility
Noninvasive		
Serology	>80/>90	Inexpensive, convenient; not useful for early follow-up
Urea breath test	>90/>90	Simple, rapid; useful for early follow-up; false negatives with recent therapy (see rapid urease test); exposure to low-dose radiation with ^{14}C test
Stool antigen	>90/>90	Inexpensive, convenient

Abbreviation: PPIs, proton pump inhibitors.

ACID-NEUTRALIZING/INHIBITORY DRUGS

Antacids Before we understood the important role of histamine in stimulating parietal cell activity, neutralization of secreted acid with antacids constituted the main form of therapy for peptic ulcers. They are now rarely, if ever, used as the primary therapeutic agent but instead are often used by patients for symptomatic relief of dyspepsia. The most commonly used agents are mixtures of aluminum hydroxide and magnesium hydroxide. Aluminum hydroxide can produce constipation and phosphate depletion; magnesium hydroxide may cause loose stools. Many of the commonly used antacids (e.g., Maalox, Mylanta) have a combination of both aluminum and magnesium hydroxide in order to avoid these side effects. The magnesium-containing preparation should not be used in chronic renal failure patients because of possible hypermagnesemia, and aluminum may cause chronic neurotoxicity in these patients.

Calcium carbonate and sodium bicarbonate are potent antacids with varying levels of potential problems. The long-term use of calcium carbonate (converts to calcium chloride in the stomach) can lead to milk-alkali syndrome (hypercalcemia, hyperphosphatemia with possible renal calcinosis and progression to renal insufficiency). Sodium bicarbonate may induce systemic alkalosis.

TABLE 317-3 Drugs Used in the Treatment of Peptic Ulcer Disease

DRUG TYPE/MECHANISM	EXAMPLES	DOSE
Acid-Suppressing Drugs		
Antacids	Mylanta, Maalox, Tums, Gaviscon	100–140 meq/L 1 and 3 h after meals and hs
H_2 receptor antagonists	Cimetidine	400 mg bid
	Ranitidine	300 mg hs
	Famotidine	40 mg hs
	Nizatidine	300 mg hs
Proton pump inhibitors	Omeprazole	20 mg/d
	Lansoprazole	30 mg/d
	Rabeprazole	20 mg/d
	Pantoprazole	40 mg/d
	Esomeprazole	20 mg/d
	Dexlansoprazole	30 mg/d
Mucosal Protective Agents		
Sucralfate	Sucralfate	1 g qid
Prostaglandin analogue	Misoprostol	200 μg qid
Bismuth-containing compounds	Bismuth subsalicylate (BSS)	See anti-*H. pylori* regimens (Table 317-4)

Abbreviation: hs, at bedtime (*hora somni*).

H_2 Receptor Antagonists Four of these agents are presently available (cimetidine, ranitidine, famotidine, and nizatidine), and their structures share homology with histamine. Although each has different potency, all will significantly inhibit basal and stimulated acid secretion to comparable levels when used at therapeutic doses. Moreover, similar ulcer-healing rates are achieved with each drug when used at the correct dosage. Presently, this class of drug is often used for treatment of active ulcers (4–6 weeks) in combination with antibiotics directed at eradicating *H. pylori* (see below).

Cimetidine was the first H_2 receptor antagonist used for the treatment of acid peptic disorders.

Cimetidine may have weak antiandrogenic side effects resulting in reversible gynecomastia and impotence, primarily in patients receiving high doses for prolonged periods of time (months to years). In view of cimetidine's ability to inhibit cytochrome P450, careful monitoring of drugs such as warfarin, phenytoin, and theophylline is indicated with long-term usage. Other rare reversible adverse effects reported with cimetidine include confusion and elevated levels of serum aminotransferases, creatinine, and serum prolactin. Ranitidine, famotidine, and nizatidine are more potent H_2 receptor antagonists than cimetidine. Each can be used once a day at bedtime for ulcer prevention, which was commonly done before the discovery of *H. pylori* and the development of proton pump inhibitors (PPIs). Patients may develop tolerance to H_2 blockers, a rare event with PPIs (see below). Comparable nighttime dosing regimens are cimetidine 800 mg, ranitidine 300 mg, famotidine 40 mg, and nizatidine 300 mg.

Additional rare, reversible systemic toxicities reported with H_2 receptor antagonists include pancytopenia, neutropenia, anemia, and thrombocytopenia, with a prevalence rate varying from 0.01 to 0.2%. Cimetidine and ranitidine (to a lesser extent) can bind to hepatic cytochrome P450; famotidine and nizatidine do not.

Proton Pump (H^+,K^+-ATPase) Inhibitors Omeprazole, esomeprazole, lansoprazole, rabeprazole, and pantoprazole are substituted benzimidazole derivatives that covalently bind and irreversibly inhibit H^+,K^+-ATPase. Esomeprazole is the S-enantiomer of omeprazole, which is a racemic mixture of both S- and R-optical isomers. The R-isomer of lansoprazole, dexlansoprazole, is the most recent PPI approved for clinical use. Its reported advantage is a dual delayed-release system, aimed at improving treatment of gastroesophageal reflux disease (GERD). These are the most potent acid inhibitory agents available. Omeprazole and lansoprazole are the PPIs that have been used for the longest time. Both are acid-labile and are administered as enteric-coated granules in a sustained-release capsule that dissolves within the small intestine at a pH of 6. Lansoprazole is available in an orally disintegrating tablet that can be taken with or without water, an advantage for individuals who have significant dysphagia. Absorption kinetics are similar to the capsule. In addition, a lansoprazole-naproxen combination preparation that has been made available is targeted at decreasing NSAID-related GI injury (see below). Omeprazole is available as nonenteric-coated granules mixed with sodium bicarbonate in a powder form that can be administered orally or via gastric tube. The sodium bicarbonate has two purposes: to protect the omeprazole from acid degradation and to promote rapid gastric alkalinization and subsequent proton pump activation, which facilitates rapid action of the PPI. Pantoprazole and rabeprazole are available as enteric-coated tablets. Pantoprazole is also available as a parenteral formulation for intravenous use. These agents are lipophilic compounds; upon entering the parietal cell, they are protonated and trapped within the acid environment of the tubulovesicular and canalicular system. These agents potently inhibit all phases of gastric acid secretion. Onset of action is rapid, with a maximum acid inhibitory effect between 2 and 6 h after administration and

duration of inhibition lasting up to 72–96 h. With repeated daily dosing, progressive acid inhibitory effects are observed, with basal and secretagogue-stimulated acid production being inhibited by >95% after 1 week of therapy. The half-life of PPIs is ~18 h; thus, it can take between 2 and 5 days for gastric acid secretion to return to normal levels once these drugs have been discontinued. Because the pumps need to be activated for these agents to be effective, their efficacy is maximized if they are administered before a meal (except for the immediate-release formulation of omeprazole) (e.g., in the morning before breakfast). Mild to moderate hypergastrinemia has been observed in patients taking these drugs. Carcinoid tumors developed in some animals given the drugs preclinically; however, extensive experience has failed to demonstrate gastric carcinoid tumor development in humans. Serum gastrin levels return to normal levels within 1–2 weeks after drug cessation. Rebound gastric acid hypersecretion has been described in *H. pylori*–negative individuals after discontinuation of PPIs. It occurs even after relatively short-term usage (2 months) and may last for up to 2 months after the PPI has been discontinued. The mechanism involves gastrin-induced hyperplasia and hypertrophy of histamine-secreting ECL cells. The clinical relevance of this observation is that individuals may have worsening symptoms of GERD or dyspepsia upon stopping the PPI. Gradual tapering of the PPI and switching to an H_2 receptor antagonist may prevent this from occurring. *H. pylori*–induced inflammation and concomitant decrease in acid production may explain why this does not occur in *H. pylori*–positive patients. IF production is also inhibited, but vitamin B_{12}-deficiency anemia is uncommon, probably because of the large stores of the vitamin. As with any agent that leads to significant hypochlorhydria, PPIs may interfere with absorption of drugs such as ketoconazole, ampicillin, iron, and digoxin. Hepatic cytochrome P450 can be inhibited by the earlier PPIs (omeprazole, lansoprazole). Rabeprazole, pantoprazole, and esomeprazole do not appear to interact significantly with drugs metabolized by the cytochrome P450 system. The overall clinical significance of this observation is not definitely established. Caution should be taken when using theophylline, warfarin, diazepam, atazanavir, and phenytoin concomitantly with PPIs.

The list of potential side effects with long-term PPI use has steadily grown over the years. These agents are commonly used since several formulations have become available as over the counter medications. Moreover, up to 70% of current prescriptions for long-term PPIs may be unwarranted. Interpretation of the multiple studies should take into consideration that the vast majority were retrospective observational studies in which confounding factors could not be accounted for entirely.

Long-term acid suppression, especially with PPIs, has been associated with a higher incidence of community-acquired pneumonia as well as community and hospital acquired *Clostridium difficile*–associated disease. A meta-analysis showed a 74% increased risk of *Clostridium difficile* infection and a 2.5-fold higher risk of reinfection as compared to non-users. In light of these concerns the FDA published a safety alert regarding the association between *Clostridium difficile* infection and PPI use. Although the risk of spontaneous bacterial peritonitis in cirrhotics was thought to be increased, the data here are less supportive. The impact of PPI-induced changes in the host microbiome is postulated to play a role in the increased risk of infection, but this theory needs to be confirmed. These observations require confirmation but should alert the practitioner to take caution when recommending these agents for long-term use, especially in elderly patients at risk for developing pneumonia or *Clostridium difficile* infection.

A population-based study revealed that long-term use of PPIs was associated with the development of hip fractures in older women. The absolute risk of fracture remained low despite an observed increase associated with the dose and duration of acid suppression. The mechanism for this observation is not clear, and this finding must be confirmed before making broad recommendations regarding the discontinuation of these agents in patients who benefit from them. Long-term use of PPIs has also been implicated in the development of iron, vitamin B12, and magnesium deficiency. A meta-analysis of nine observational studies found a 40% increase in hypomagnesemia in PPI users as compared to non-users. One approach to consider in patients needing to take PPI's long term, is to check a complete blood count looking for evidence of anemia due to iron or B12 deficiency, vitamin B12 level and a magnesium level after 1–2 years of PPI use but these recommendations are not evidence-based nor recommended by expert opinion. PPIs may exert a negative effect on the antiplatelet effect of clopidogrel. Although the evidence is mixed and inconclusive, a small increase in mortality and readmission rate for coronary events was seen in patients receiving a PPI while on clopidogrel in earlier studies. Subsequently, three meta-analyses reported an inverse correlation between clopidogrel and PPI use; therefore, the influence of this drug interaction on mortality is not clearly established. The mechanism involves the competition of the PPI and clopidogrel with the same cytochrome P450 (CYP2C19). Whether this is a class effect of PPIs is unclear; there appears to be at least a theoretical advantage of pantoprazole over the other PPIs, but this has not been confirmed. This drug interaction is particularly relevant in light of the common use of aspirin and clopidogrel for prevention of coronary events and the efficacy of PPIs in preventing GI bleeding in these patients. The FDA has made several recommendations while awaiting further evidence to clarify the impact of PPI therapy on clopidogrel use. Health care providers should continue to prescribe clopidogrel to patients who require it and should reevaluate the need for starting or continuing treatment with a PPI. From a practical standpoint, additional recommendations to consider include the following: Patients taking clopidogrel with aspirin, especially with other GI risk factors for bleeding, should receive GI protective therapy. Although high-dose H_2 blockers have been considered an option, these do not appear to be as effective as PPIs. If PPIs are to be given, some have recommended that there be a 12-h separation between administration of the PPI and clopidogrel to minimize competition of the two agents with the involved cytochrome P450. One option is to give the PPI 30 min before breakfast and the clopidogrel at bedtime. Insufficient data are available to firmly recommend one PPI over another. Additional concerning side effects with long-term PPI use include increased cardiac risks independent of clopidogrel use, dementia, acute and chronic kidney injury. Again, the data are often retrospective and confounding variables were not consistently eliminated thus making it difficult to develop definitive association between PPIs and the toxicities outlined. A summary of the side effects with the corresponding relative risks is shown in **Table 317-4**. Ultimately, heightened awareness of inappropriate long-term use of PPIs is paramount. Patients aged ≥65 years of age have a higher risk for some of the long-term side effects of PPIs highlighted above, in part due to the higher prevalence of concomitant chronic diseases. It is therefore essential to carefully select individuals, especially among the elderly, who need long-term PPI therapy and discontinue it in those individuals who do not need it.

Development of novel acid inhibitory agents continues in an attempt to primarily address the need for better agents to treat GERD. For example, modified H_2 blockers with greater potency and duration as well as novel PPIs with longer half-life and potency are under study. For example, tenatoprazole is a PPI containing an imidazopyridine ring instead of a benzimidazole ring, which promotes irreversible proton pump inhibition. This agent has a longer half-life than the other PPIs and may be beneficial for inhibiting nocturnal acid secretion, which has significant relevance in GERD. Additional PPIs with longer half-life and combined with other agents are being studied but the details are beyond the scope of this chapter. A second new class of agents is the potassium-competitive acid pump antagonists (P-CABs). These compounds inhibit gastric acid secretion via potassium competitive binding of the H^+,K^+-ATPase. Revaprazan and venoprazan are the first two agents approved for use in Korea and Japan, respectively.

TABLE 317-4 Evidence Supporting the Potential Adverse Effects of Proton Pump Inhibitor Drugs

ADVERSE EFFECT	ADJUSTED OR (95% CI)
Chronic kidney disease	1.50 (1.11–1.90)
Acute kidney disease	2.52 (2.27–2.79)
Acute interstitial nephritis	3.00 (1.47–6.14)
Hypomagnesemia	1.43 (1.08–1.88)
Clostridium difficile	1.74 (1.47–2.85)
Community-acquired pneumonia	1.34 (1.14–1.57)
Community-acquired pneumonia	1.05 (0.89–1.25)
Bone fracture	1.33 (1.15–1.54)

Abbreviation: OR, odds ratio.

Source: Adapted from AJ Schoenfeld, D Grady: Adverse effects associated with proton pump inhibitors. JAMA Intern Med 176:172, 2016.

CYTOPROTECTIVE AGENTS

Sucralfate Sucralfate is a complex sucrose salt in which the hydroxyl groups have been substituted by aluminum hydroxide and sulfate. This compound is insoluble in water and becomes a viscous paste within the stomach and duodenum, binding primarily to sites of active ulceration. Sucralfate may act by several mechanisms: serving as a physicochemical barrier, promoting a trophic action by binding growth factors such as EGF, enhancing prostaglandin synthesis, stimulating mucus and bicarbonate secretion, and enhancing mucosal defense and repair. Toxicity from this drug is rare, with constipation being most common (2–3%). It should be avoided in patients with chronic renal insufficiency to prevent aluminum-induced neurotoxicity. Hypophosphatemia and gastric bezoar formation have also been reported rarely. Standard dosing of sucralfate is 1 g qid.

Bismuth-Containing Preparations Sir William Osler considered bismuth-containing compounds the drug of choice for treating PUD. The resurgence in the use of these agents is due to their effect against *H. pylori.* Colloidal bismuth subcitrate (CBS) and bismuth subsalicylate (BSS, Pepto-Bismol) are the most widely used preparations. The mechanism by which these agents induce ulcer healing is unclear. Adverse effects with short-term use include black stools, constipation, and darkening of the tongue. Long-term use with high doses, especially with the avidly absorbed CBS, may lead to neurotoxicity. These compounds are commonly used as one of the agents in an anti-*H. pylori* regimen (see below).

Prostaglandin Analogues In view of their central role in maintaining mucosal integrity and repair, stable prostaglandin analogues were developed for the treatment of PUD. The mechanism by which this rapidly absorbed drug provides its therapeutic effect is through enhancement of mucosal defense and repair. The most common toxicity noted with this drug is diarrhea (10–30% incidence). Other major toxicities include uterine bleeding and contractions; misoprostol is contraindicated in women who may be pregnant, and women of childbearing age must be made clearly aware of this potential drug toxicity. The standard therapeutic dose is 200 µg qid.

Miscellaneous Drugs A number of drugs including anticholinergic agents and tricyclic antidepressants were used for treating acid peptic disorders, but in light of their toxicity and the development of potent antisecretory agents, these are rarely, if ever, used today.

THERAPY OF *H. PYLORI*

The physician's goal in treating PUD is to provide relief of symptoms (pain or dyspepsia), promote ulcer healing, and ultimately prevent ulcer recurrence and complications. The greatest influence of understanding the role of *H. pylori* in peptic disease has been the ability to prevent recurrence. Documented eradication of *H. pylori* in patients with PUD is associated with a dramatic decrease in ulcer recurrence to <10–20% as compared to 59% in GU patients and 67% in DU patients when the organism is not eliminated. Eradication of the organism may lead to diminished recurrent ulcer bleeding. The effect of its eradication on ulcer perforation is unclear.

Extensive effort has been made in determining who of the many individuals with *H. pylori* infection should be treated. The common conclusion arrived at by multiple consensus conferences around the world is that *H. pylori* should be eradicated in patients with documented PUD. This holds true independent of time of presentation (first episode or not), severity of symptoms, presence of confounding factors such as ingestion of NSAIDs, or whether the ulcer is in remission. Some have advocated treating patients with a history of documented PUD who are found to be *H. pylori*–positive by stool antigen or breath testing. Between 60 and 90% of patients with gastric MALT lymphoma experience complete remission of the tumor in response to *H. pylori* eradication. The Maastricht IV/Florence Consensus Report recommends a test-and-treat approach for patients with uninvestigated dyspepsia if the local incidence of *H. pylori* is >20%. The American College of Gastroenterology (ACG) clinical guidelines (Developed for North America) recommend that individuals aged <60 with uninvestigated dyspepsia should be tested and treated for *H. pylori*. In addition, recommendations from this consensus report and the ACG clinical guidelines include testing and offering eradication of *H. pylori* in patients who will be using NSAIDs (including low-dose aspirin) on a long-term basis, especially if there is a prior history of PUD. These individuals will require continued PPI treatment as well as eradication treatment, because eradication of the organism alone does not eliminate the risk of gastroduodenal ulcers in patients already receiving long-term NSAIDs. Treating patients with NUD to prevent gastric cancer or patients with GERD requiring long-term acid suppression remains controversial. Guidelines from the ACG suggest eradication of *H. pylori* in patients who have undergone resection of early gastric cancer. The Maastricht IV/Florence Consensus Report also evaluated *H. pylori* treatment in gastric cancer prevention and recommends that eradication should be considered in the following situations: first-degree relatives of family members with gastric cancer; patients with previous gastric neoplasm treated by endoscopic or subtotal resection; individuals with a risk of gastritis (severe pangastritis or body-predominant gastritis) or severe atrophy; patients with gastric acid inhibition for >1 year; individuals with strong environmental risk factors for gastric cancer (heavy smoking; high exposure to dust, coal, quartz, or cement; and/or work in quarries); and *H. pylori*–positive patients with a fear of gastric cancer. Finally the ACG clinical guidelines recommend testing and offering *H. pylori* eradication to patients with unexplained iron deficiency anemia and idiopathic thrombocytopenic purpura.

Multiple drugs have been evaluated in the therapy of *H. pylori.* No single agent is effective in eradicating the organism. Combination therapy for 14 days provides the greatest efficacy, although regimens based on sequential administration of antibiotics also appear promising (see below). A shorter administration course (7–10 days), although attractive, has not proved as successful as the 14-day regimens. The agents used with the greatest frequency include amoxicillin, metronidazole, tetracycline, clarithromycin, and bismuth compounds.

Suggested treatment regimens for *H. pylori* are outlined in **Table 317-5** Choice of a particular regimen will be influenced by several factors, including efficacy, patient tolerance, existing antibiotic resistance, prior antibiotic use and cost of the drugs. The aim for initial eradication rates should be 85–90%. Dual therapy (PPI plus amoxicillin, PPI plus clarithromycin, ranitidine bismuth citrate [Tritec] plus clarithromycin) is not recommended in view of studies demonstrating eradication rates of <80–85%. The combination of bismuth, metronidazole, and tetracycline was the first triple regimen found effective against *H. pylori.* The combination of two antibiotics plus either a PPI, H_2 blocker, or bismuth compound has comparable success rates. Addition of acid suppression assists in providing early symptom relief and enhances bacterial eradication.

Triple therapy, although effective, has several drawbacks, including the potential for poor patient compliance and drug-induced side effects. Compliance is being addressed by simplifying the regimens

TABLE 317-5 Recommended First-Line Therapies for *H. pylori* Infection

REGIMEN	DRUGS (DOSES)	DOSING FREQUENCY	DURATION (DAYS)	FDA APPROVAL
Clarithromycin triple	PPI (standard or double dose)	BID	14	Yes[a]
	Clarithromycin (500mg)			
	Amoxicillin (1 g) or Metronidazole (500 mg TID)			
Bismuth quadruple	PPI (standard dose)	BID	10–14	No[b]
	Bismuth subcitrate (120–300 mg) or Subsalicylate (300 mg)	QID		
	Tetracycline (500 mg)	QID		
	Metronidazole (250–500 mg)	QID (250) TID to QID (500)		
Concomitant	PPI (standard dose)	BID	10–14	No
	Clarithromycin (500 mg)			
	Amoxicillin (1 g)			
	Nitroimidazole (500 mg)[c]			
Sequential	PPI (standard dose)	BID	5–7	No
	PPI, Clarithromycin (500 mg) + Nitroimidazole (500 mg)[c]	BID	5–7	
Hybrid	PPI (standard dose) + Amox (1 g)	BID	7	No
	PPI, Amox, Clarithromycin (500mg), Nitroimidazole (500 mg)[c]	BID	7	
Levofloxacin triple	PPI (standard or double dose) + Amox (1 g)	BID	5–7	No
	Levofloxacin (500 mg)	QD		
	Amox (1 g)	BID		
Levofloxacin sequential	PPI (standard or double dose) + Amox (1 g)	BID	5–7	No
	PPI, Amox, Levofloxacin (500 mg QD), Nitroimidazole (500 mg)[c]	BID	5–7	
LOAD	Levofloxacin (250 mg)	QD	7–10	No
	PPI (double dose)	QD		
	Nitazoxanide (500 mg)	BID		
	Doxycycline (100 mg)	QD		

[a]Several PPI, clarithromycin, and amoxicillin combinations have achieved FDA approval. PPI, clarithromycin and metronidazole is not an FDA-approved treatment regimen. [b]PPI, bismuth, tetracycline, and metronidazole combined with a PPI for 10 days is an FDA-approved treatment regimen. [c] Metronidazole or tinidazole.

Abbreviations: BID, twice daily; FDA, Food and Drug Administration; PPI, proton pump inhibitor; TID, three times daily; QD, once daily; QID, four times daily.

Source: Adapted from WD Chey et al: ACG Clinical Guideline: Treatment of *Helicobacter pylori* Infection. Am J Gastroenterol 112:212, 2017.

so that patients can take the medications twice a day. Simpler (dual therapy) and shorter regimens (7 and 10 days) are not as effective as triple therapy for 14 days. Two anti-*H. pylori* regimens are available in prepackaged formulation: Prevpac (lansoprazole, clarithromycin, and amoxicillin) and Helidac (BSS, tetracycline, and metronidazole). The contents of the Prevpac are to be taken twice per day for 14 days, whereas Helidac constituents are taken four times per day with an antisecretory agent (PPI or H_2 blocker), also for at least 14 days. Clarithromycin-based triple therapy should be avoided in settings where *H. pylori* resistance to this agent exceeds 15%.

Side effects have been reported in up to 20–30% of patients on triple therapy. Bismuth may cause black stools, constipation, or darkening of the tongue. The most feared complication with amoxicillin is pseudomembranous colitis, but this occurs in <1–2% of patients. Amoxicillin can also lead to antibiotic-associated diarrhea, nausea, vomiting, skin rash, and allergic reaction. Concomitant use of probiotics may ameliorate some of the antibiotic side effects (see below). Tetracycline has been reported to cause rashes and, very rarely, hepatotoxicity and anaphylaxis.

One important concern with treating patients who may not need therapy is the potential for development of antibiotic-resistant strains. The incidence and type of antibiotic-resistant *H. pylori* strains vary worldwide. Strains resistant to metronidazole, clarithromycin, amoxicillin, and tetracycline have been described, with the latter two being uncommon. Antibiotic-resistant strains are the most common cause for treatment failure in compliant patients. Unfortunately, in vitro resistance does not predict outcome in patients. Culture and sensitivity testing of *H. pylori* is not performed routinely. Although resistance to metronidazole has been found in as many as 30% of isolates in North America and 80% in developing countries, triple therapy is effective in eradicating the organism in >50% of patients infected with a resistant strain. Clarithromycin resistance is seen in 13–16% of individuals in the United States, with resistance to amoxicillin being <1% and resistance to both metronidazole and clarithromycin in the 5% range. Resistance to tetracycline and rifabutin (see below) is reported to be <2% in the United States. In light of the paucity of *H. pylori* antibiotic real time resistance data, asking the patient about prior antibiotic exposure should be included in the decision-making and used as a surrogate for potential antibiotic resistance especially when it comes to prior macrolide use. Clarithromycin use should be excluded in patients with prior macrolide usage. An approach to antibiotic selection for *H. pylori* therapy has been recommended in the ACG clinical guidelines **(Fig. 317-12)**.

Failure of *H. pylori* eradication with triple therapy in a compliant patient is usually due to infection with a resistant organism. A series of salvage therapies for *H. pylori* are shown in **Table 317-6**. Quadruple therapy (Table 317-4), where clarithromycin is substituted for metronidazole (or vice versa), should be the next step. The combination of PPI, amoxicillin, and rifabutin for 10 days has also been used successfully (86% cure rate) in patients infected with resistant strains. Additional regimens considered for second-line therapy include levofloxacin-based triple therapy (levofloxacin, amoxicillin, PPI) for 10 days and furazolidone-based triple therapy (furazolidone, amoxicillin, PPI) for 14 days. Unfortunately, there is no universally accepted treatment regimen recommended for patients in whom two courses of antibiotics have failed. If eradication is still not achieved in a compliant patient, then culture and sensitivity of the organism should be considered. One challenge with this approach is that culture and sensitivity testing is cumbersome and not widely available, thus *H. pylori* resistance data within specific communities

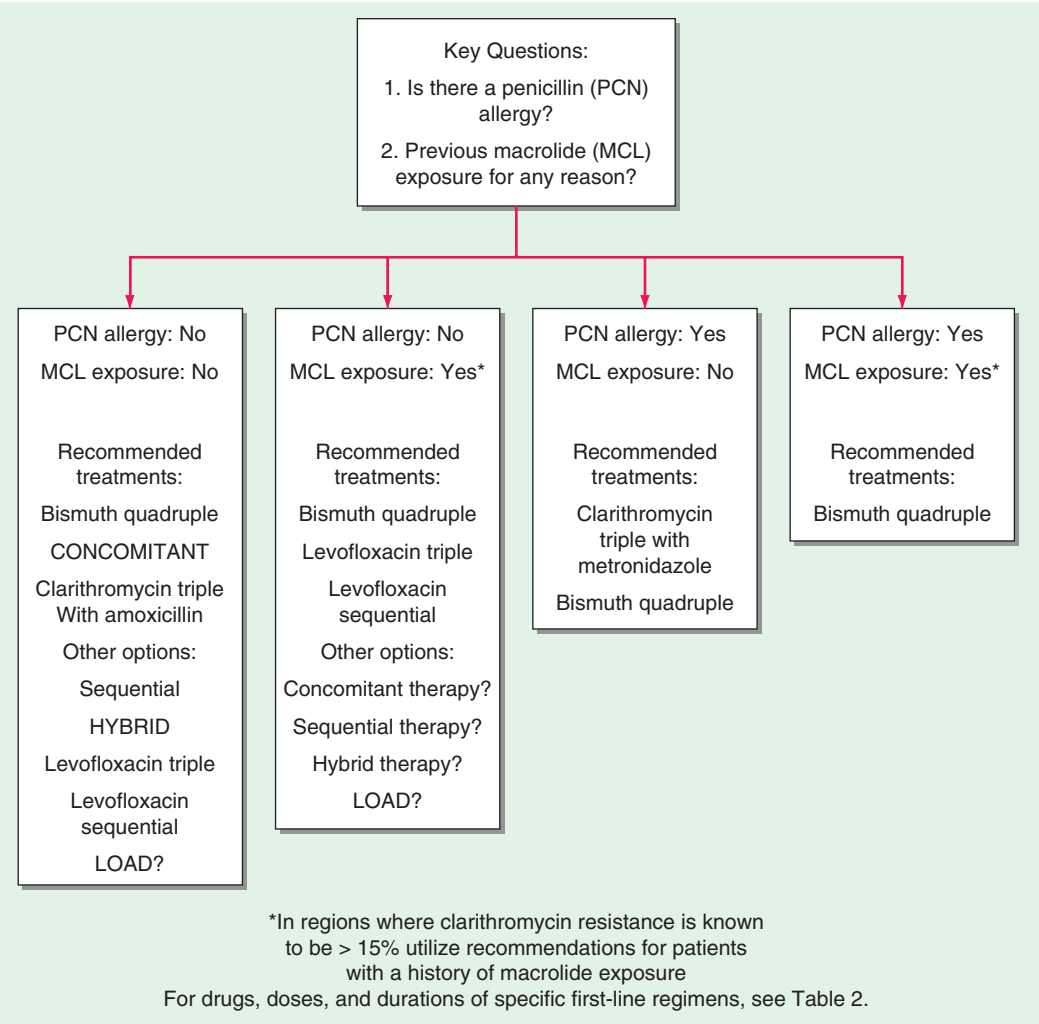


FIGURE 317-12 Approach to selecting antibiotics for patients with *H. pylori* infection. *(Adapted from WD Chey et al: ACG Clinical Guideline: Treatment of Helicobacter pylori Infection. Am J Gastroenterol 112:212, 2017.)*

TABLE 317-6 Salvage Therapies for *H. pylori* Infection

REGIMEN	DRUGS (DOSES)	DOSING FREQUENCY	DURATION (DAYS)	FDA APPROVAL
Bismuth quadruple	PPI (standard dose)	BID	14	No[a]
	Bismuth subcitrate (120–300 mg) or subsalicylate (300 mg)	QID		
	Tetracycline (500 mg)	QID		
	Metronidazole (500 mg)	TID or QID		
Levofloxacin triple	PPI (standard dose)	BID	14	No
	Levofloxacin (500 mg)	QD		
	Amox (1 g)	BID		
Concomitant	PPI (standard dose)	BID	10–14	No
	Clarithromycin (500 mg)	BID		
	Amoxicillin (1 g)	BID		
	Nitroimidazole (500 mg)	BID or TID		
Rifabutin triple	PPI (standard dose)	BID	10	No
	Rifabutin (300 mg)	QD		
	Amox (1 g)	BID		
High-dose dual	PPI (standard to double dose)	TID or QID	14	No
	Amox (1 g TID or 750 mg QID)	TID or QID		

[a]PPI, bismuth, tetracycline, and metronidazole prescribed separately is not an FDA-approved treatment regimen. However, Pylera, a combination product containing bismuth subcitrate, tetracycline, and metronidazole combined with a PPI for 10 days is an FDA-approved treatment regimen.

Abbreviations: BID, twice daily; FDA, Food and Drug Administration; PPI, proton pump inhibitor; TID, three times daily; QD, once daily; QID, four times daily.

Source: Adapted from WD Chey et al: ACG Clinical Guideline: Treatment of *Helicobacter pylori* Infection. Am J Gastroenterol 112:212–238, 2017.

are often not available. Non-culture-based approaches using molecular markers to determine potential resistance through stool testing are being developed but are not widely available. Additional factors that may lower eradication rates include the patient's country of origin (higher in Northeast Asia than other parts of Asia or Europe) and cigarette smoking. In addition, meta-analysis suggests that even the most effective regimens (quadruple therapy including PPI, bismuth, tetracycline, and metronidazole and triple therapy including PPI, clarithromycin, and amoxicillin) may have suboptimal eradication rates (<80%), thus, demonstrating the need for the development of more efficacious treatments.

In view of the observation that 15–25% of patients treated with first-line therapy may still remain infected with the organism, new approaches to treatment have been explored. One promising approach is sequential therapy. Regimens examined consist of 5 days of amoxicillin and a PPI, followed by an additional 5 days of PPI plus tinidazole and clarithromycin or levofloxacin. One promising regimen that has the benefit of being shorter in duration, easier to take, and less expensive is 5 days of concomitant therapy (PPI twice daily, amoxicillin 1 g twice daily, levofloxacin 500 mg twice daily, and tinidazole 500 mg twice daily). Initial studies have demonstrated eradication rates of >90% with good patient tolerance. Confirmation of these findings and applicability of this approach in the United States are needed, although some experts are recommending abandoning clarithromycin-based triple therapy in the United States for the concomitant therapy or the alternative sequential therapies highlighted above.

Innovative non–antibiotic-mediated approaches have been explored in an effort to improve eradication rates of *H. pylori*. Pretreatment of patients with *N*-acetylcysteine as a mucolytic agent to destroy the *H. pylori* biofilm and therefore impair antibiotic resistance has been examined, but more studies are needed to confirm the applicability of this approach. In vitro studies suggest that certain probiotics like *Lactobacillus* or its metabolites can inhibit *H. pylori.* Administration of probiotics has been attempted in several clinical studies in an effort to maximize antibiotic-mediated eradication with varying results. Overall, it appears that the use of certain probiotics, such as *Lactobacillus* spp., *Saccharomyces* spp., *Bifidobacterium* spp., and *Bacillus clausii*, did not alter eradication rates but importantly decreased antibiotic-associated side effects including nausea, dysgeusia, diarrhea, and abdominal discomfort/pain, resulting in enhanced tolerability of *H. pylori* therapies. Additional studies are needed to confirm the potential benefits of probiotics in this setting.

Reinfection after successful eradication of *H. pylori* is rare in the United States (<1% per year). If recurrent infection occurs within the first 6 months after completing therapy, the most likely explanation is recrudescence as opposed to reinfection.

THERAPY OF NSAID-RELATED GASTRIC OR DUODENAL INJURY

Medical intervention for NSAID-related mucosal injury includes treatment of an active ulcer and primary prevention of future injury. Recommendations for the treatment and primary prevention of NSAID-related mucosal injury are listed in **Table 317-7**. Ideally, the injurious agent should be stopped as the first step in the therapy of an active NSAID-induced ulcer. If that is possible, then treatment with one of the acid inhibitory agents (H_2 blockers, PPIs) is indicated. Cessation of NSAIDs is not always possible because of the patient's severe underlying disease. Only PPIs can heal GUs or DUs, independent of whether NSAIDs are discontinued.

The widespread use of NSAIDs has created some concern due to the increasing likelihood of GI and CV side effects associated with these agents. The approach to primary prevention has included avoiding the agent, using the lowest possible dose of the agent for the shortest period of time possible, using NSAIDs that are theoretically less injurious, using newer topical NSAID preparations, and/or using concomitant medical therapy to prevent NSAID-induced injury. Several nonselective NSAIDs that are associated with a lower likelihood of GI and CV toxicity include naproxen and ibuprofen, although the beneficial effect may be eliminated if higher dosages of the agents are used. Primary prevention of NSAID-induced ulceration can be accomplished by misoprostol (200 μg qid) or a PPI. High-dose H_2 blockers (famotidine, 40 mg bid) have also shown some promise in preventing endoscopically documented ulcers, although PPIs are superior. The highly selective COX-2 inhibitors, celecoxib and rofecoxib, are 100 times more selective inhibitors of COX-2 than standard NSAIDs, leading to gastric or duodenal mucosal injury that is comparable to placebo; their utilization led to an increase in CV events and withdrawal from the market. Additional caution was engendered when the CLASS study demonstrated that the advantage of celecoxib in preventing GI complications was offset when low-dose aspirin was used simultaneously. Therefore, gastric protection therapy is required in individuals taking COX-2 inhibitors and aspirin prophylaxis. Finally, much of the work demonstrating the benefit of COX-2 inhibitors and PPIs on GI injury has been performed in individuals of average risk; it is unclear if the same level of benefit will be achieved in high-risk patients. For example, concomitant use of warfarin and a COX-2 inhibitor was associated with rates of GI bleeding similar to those observed in patients taking nonselective NSAIDs. A combination of factors, including withdrawal of the majority of COX-2 inhibitors from the market, the observation that low-dose aspirin appears to diminish the beneficial effect of COX-2 selective inhibitors, and the growing use of aspirin for prophylaxis of CV events, have significantly altered the approach to gastric protective therapy during the use of NSAIDs. A set of guidelines for the approach to the use of NSAIDs was published by the ACG and is shown in **Table 317-8**. Individuals who are not at risk for CV events do not use aspirin and are without risk for GI complications can receive nonselective NSAIDs without gastric protection. In those without CV risk factors but with a high potential risk (prior GI bleeding or multiple GI risk factors) for NSAID-induced GI toxicity, cautious use of a selective COX-2 inhibitor and co-therapy with misoprostol or high-dose PPI are recommended. Individuals at moderate GI risk without cardiac risk factors can be treated with a COX-2 inhibitor alone or with a nonselective NSAID with misoprostol or a PPI. Individuals with CV risk factors, who

TABLE 317-7 Recommendations for Treatment of NSAID-Related Mucosal Injury

CLINICAL SETTING	RECOMMENDATION
Active ulcer	
NSAID discontinued	H_2 receptor antagonist or PPI
NSAID continued	PPI
Prophylactic therapy	Misoprostol PPI Selective COX-2 inhibitor
H. pylori infection	Eradication if active ulcer present or there is a past history of peptic ulcer disease

Abbreviations: COX-2, isoenzyme of cyclooxygenase; NSAID, nonsteroidal anti-inflammatory drug; PPI, proton pump inhibitor.

TABLE 317-8 Guide to NSAID Therapy

	NO/LOW NSAID GI RISK	NSAID GI RISK
No CV risk (no aspirin)	Traditional NSAID	Coxib *or* Traditional NSAID + PPI *or* misoprostol Consider non-NSAID therapy
CV risk (consider aspirin)	Traditional NSAID + PPI *or* misoprostol if GI risk warrants gastroprotection Consider non-NSAID therapy	A gastroprotective agent must be added if a traditional NSAID is prescribed Consider non-NSAID therapy

Abbreviations: CV, cardiovascular; GI, gastrointestinal; NSAID, nonsteroidal anti-inflammatory drug; PPI, proton pump inhibitor.

Source: Adapted from AM Fendrick: Am J Manag Care 10:740, 2004. Reproduced with permission of INTELLISPHERE, LLC via Copyright Clearance Center.

require low-dose aspirin and have low potential for NSAID-induced toxicity, should be considered for a non-NSAID agent or use of a traditional NSAID such as naproxen (lower CV side effects) in combination with gastric protection, if warranted. Finally, individuals with CV and GI risks who require aspirin must be considered for non-NSAID therapy, but if that is not an option, then gastric protection with any type of NSAID must be considered. Any patient, regardless of risk status, who is being considered for long-term traditional NSAID therapy, should also be considered for *H. pylori* testing and treatment if positive. Assuring the use of GI protective agents with NSAIDs is difficult, even in high-risk patients. This is in part due to under prescribing of the appropriate protective agent; other times the difficulty is related to patient compliance. The latter may be due to patients forgetting to take multiple pills or preferring not to take the extra pill, especially if they have no GI symptoms. Several NSAID gastroprotective-containing combination pills are now commercially available, including double-dose famotidine with ibuprofen, diclofenac with misoprostol, and naproxen with esomeprazole. Although initial studies suggested improved compliance and a cost advantage when taking these combination drugs, their clinical benefit over the use of separate pills has not been established. One additional concern with NSAID-induced GI complications is the relatively low rate of primary care provider compliance with established guidelines outlining preventative measures. An intervention including professional education, informatics to facilitate review, and financial incentives for practices to review patients' charts to assess appropriateness showed reduced rate of high-risk prescribing of antiplatelet medications and NSAIDs with a tendency towards improved clinical outcomes. Efforts continue toward developing safer NSAIDs, including topical NSAIDs, NSAID formulations that are rapidly absorbed (diclofenac potassium powder mixed with a buffering agent, Prosorb and SoluMatrix technology), NO–releasing NSAIDs, hydrogen sulfide–releasing NSAIDs, dual COX/5-LOX inhibitors, NSAID prodrugs, or agents that can effectively sequester unbound NSAIDs without interfering with their efficacy.

APPROACH AND THERAPY: SUMMARY

Controversy continues regarding the best approach to the patient who presents with dyspepsia **(Chap. 41)**. The discovery of *H. pylori* and its role in pathogenesis of ulcers has added a new variable to the equation. Previously, if a patient <50 years of age presented with dyspepsia and without alarming signs or symptoms suggestive of an ulcer complication or malignancy, an empirical therapeutic trial with acid suppression was commonly recommended. Although this approach is practiced by some today, an approach presently gaining approval for the treatment of patients with dyspepsia is outlined in **Fig. 317-13**. The referral to a gastroenterologist is for the potential need of endoscopy and subsequent evaluation and treatment if the endoscopy is negative.

Once an ulcer (GU or DU) is documented, the main issue at stake is whether *H. pylori* or an NSAID is involved. With *H. pylori* present, independent of the NSAID status, triple therapy is recommended for 14 days, followed by continued acid-suppressing drugs (H_2 receptor antagonist or PPIs) for a total of 4–6 weeks. *H. pylori* eradication should be documented 4 weeks after completing antibiotics. The test of choice for documenting eradication is the laboratory-based validated monoclonal stool antigen test or a urea breath test (UBT). The patient must be off antisecretory agents when being tested for eradication of *H. pylori* with UBT or stool antigen. Serologic testing is not useful for the purpose of documenting eradication because antibody titers fall slowly and often do not become undetectable. Some recommend that patients with complicated ulcer disease, or who are frail, should be treated with long-term acid suppression, thus making documentation of *H. pylori* eradication a moot point. In view of this discrepancy in practice, it would be best to discuss with the patient the different options available.

Several issues differentiate the approach to a GU versus a DU. GUs, especially of the body and fundus, have the potential of being malignant. Multiple biopsies of a GU should be taken initially; even

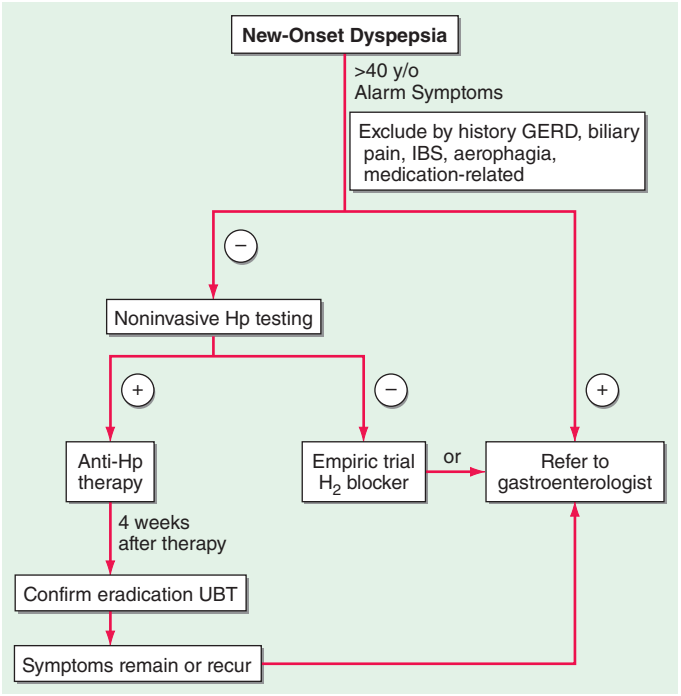


FIGURE 317-13 Overview of new-onset dyspepsia. GERD, gastroesophageal reflux disease; Hp, *Helicobacter pylori*; IBS, irritable bowel syndrome; UBT, urea breath test. *(Adapted from BS Anand, DY Graham: Endoscopy 31:215, 1999.)*

if these are negative for neoplasm, repeat endoscopy to document healing at 8–12 weeks should be performed, with biopsy if the ulcer is still present. About 70% of GUs eventually found to be malignant undergo significant (usually incomplete) healing. Repeat endoscopy is warranted in patients with DU if symptoms persist despite medical therapy or a complication is suspected.

The majority (>90%) of GUs and DUs heal with the conventional therapy outlined above. A GU that fails to heal after 12 weeks and a DU that does not heal after 8 weeks of therapy should be considered refractory. Once poor compliance and persistent *H. pylori* infection have been excluded, NSAID use, either inadvertent or surreptitious, must be excluded. In addition, cigarette smoking must be eliminated. For a GU, malignancy must be meticulously excluded. Next, consideration should be given to a gastric acid hypersecretory state such as ZES (see "Zollinger-Ellison Syndrome," below) or the idiopathic form, which can be excluded with gastric acid analysis. Although a subset of patients have gastric acid hypersecretion of unclear etiology as a contributing factor to refractory ulcers, ZES should be excluded with a fasting gastrin or secretin stimulation test (see below). More than 90% of refractory ulcers (either DUs or GUs) heal after 8 weeks of treatment with higher doses of PPI (omeprazole, 40 mg/d; lansoprazole 30–60 mg/d). This higher dose is also effective in maintaining remission. Surgical intervention may be a consideration at this point; however, other rare causes of refractory ulcers must be excluded before recommending surgery. Rare etiologies of refractory ulcers that may be diagnosed by gastric or duodenal biopsies include ischemia, Crohn's disease, amyloidosis, sarcoidosis, lymphoma, eosinophilic gastroenteritis, smoking crack cocaine or infection (cytomegalovirus [CMV], tuberculosis, or syphilis).

SURGICAL THERAPY

Surgical intervention in PUD can be viewed as being either elective, for treatment of medically refractory disease, or as urgent/emergent, for the treatment of an ulcer-related complication. The development of pharmacologic and endoscopic approaches for the treatment of peptic disease and its complications has led to a substantial decrease in the number of operations needed for this disorder with a drop of >90% for elective ulcer surgery over the last four decades. Refractory ulcers are an exceedingly rare occurrence. Surgery is more often required for treatment of an ulcer-related complication.

Hemorrhage is the most common ulcer-related complication, occurring in ~15–25% of patients. Bleeding may occur in any age group but is most often seen in older patients (sixth decade or beyond). The majority of patients stop bleeding spontaneously, but endoscopic therapy **(Chap. 315)** is necessary in some. Parenterally and orally administered PPIs also decrease ulcer rebleeding in patients who have undergone endoscopic therapy. Patients unresponsive or refractory to endoscopic intervention will require angiographic intervention or surgery (~5% of transfusion-requiring patients).

Free peritoneal perforation occurs in ~2–3% of DU patients with NSAID-induced GU perforations occurring more commonly. Sudden onset of severe abdominal pain with peritoneal signs and evidence of pneumoperitoneum on abdominal imaging is the classic presentation of a perforated viscous but this presentation occurs in only two-thirds of patients. The latter is especially true in elderly patients (>70 years old), obese individuals and in immunocompromised patients. It is important to keep in mind that as in the case of bleeding, up to 10% of these patients will not have antecedent ulcer symptoms. Delay in diagnosis clearly leads to higher mortality thus early suspicion and intervention with nasogastric suction, intravenous PPI, antibiotics and surgical consultation is essential. Concomitant bleeding may occur in up to 10% of patients with perforation, with mortality being increased substantially. Peptic ulcer can also penetrate into adjacent organs, especially with a posterior DU, which can penetrate into the pancreas, colon, liver, or biliary tree.

Pyloric channel ulcers or DUs can lead to gastric outlet obstruction in ~2–3% of patients. This can result from chronic scarring or from impaired motility due to inflammation and/or edema with pylorospasm. Patients may present with early satiety, nausea, vomiting of undigested food, and weight loss. Conservative management with nasogastric suction, intravenous hydration/nutrition, and antisecretory agents is indicated for 7–10 days with the hope that a functional obstruction will reverse. If a mechanical obstruction persists, endoscopic intervention with balloon dilation may be effective. Surgery should be considered if all else fails.

Specific Operations for Duodenal Ulcers Surgical treatment was originally designed to decrease gastric acid secretion. Operations most commonly performed include (1) vagotomy and drainage (by pyloroplasty, gastroduodenostomy, or gastrojejunostomy), (2) highly selective vagotomy (which does not require a drainage procedure), and (3) vagotomy with antrectomy. The specific procedure performed is dictated by the underlying circumstances: elective versus emergency, the degree and extent of duodenal ulceration, the etiology of the ulcer (*H. pylori*, NSAIDs, malignancy), and the expertise of the surgeon. Moreover, the trend has been toward a dramatic decrease in the need for surgery for treatment of refractory PUD, and when needed, minimally invasive and anatomy-preserving operations are preferred.

Vagotomy is a component of each of these procedures and is aimed at decreasing acid secretion through ablating cholinergic input to the stomach. Unfortunately, both truncal and selective vagotomy (preserves the celiac and hepatic branches) result in gastric atony despite successful reduction of both basal acid output (BAO; decreased by 85%) and maximal acid output (MAO; decreased by 50%). Drainage through pyloroplasty or gastroduodenostomy is required in an effort to compensate for the vagotomy-induced gastric motility disorder. This procedure has an intermediate complication rate and a 10% ulcer recurrence rate. To minimize gastric dysmotility, highly selective vagotomy (also known as parietal cell, super-selective, or proximal vagotomy) was developed. Only the vagal fibers innervating the portion of the stomach that contains parietal cells is transected, thus leaving fibers important for regulating gastric motility intact. Although this procedure leads to an immediate decrease in both BAO and stimulated acid output, acid secretion recovers over time. By the end of the first postoperative year, basal and stimulated acid output are ~30 and 50%, respectively, of preoperative levels. Ulcer recurrence rates are higher with highly selective vagotomy (≥10%), although the overall complication rates are the lowest of the three procedures.

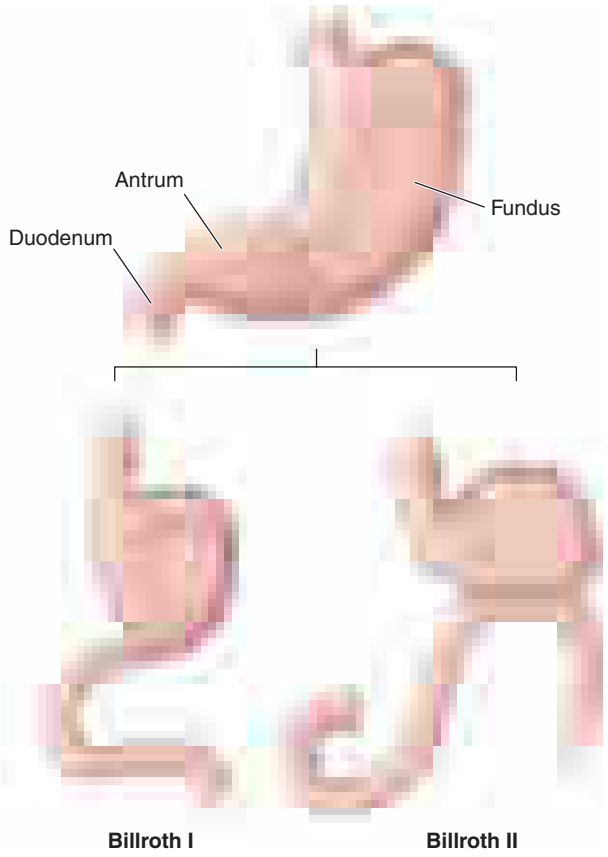


FIGURE 317-14 Schematic representation of Billroth I and II procedures.

The procedure that provides the lowest rates of ulcer recurrence (1%) but has the highest complication rate is vagotomy (truncal or selective) in combination with antrectomy. Antrectomy is aimed at eliminating an additional stimulant of gastric acid secretion, gastrin. Two principal types of reanastomoses are used after antrectomy: gastroduodenostomy (Billroth I) or gastrojejunostomy (Billroth II) **(Fig. 317-14)**. Although Billroth I is often preferred over II, severe duodenal inflammation or scarring may preclude its performance. Prospective, randomized studies confirm that partial gastrectomy followed by Roux-en-Y reconstruction leads to a significantly better clinical, endoscopic, and histologic outcome than Billroth II reconstruction.

Of these procedures, highly selective vagotomy may be the one of choice in the elective setting, except in situations where ulcer recurrence rates are high (prepyloric ulcers and those refractory to medical therapy). Selection of vagotomy and antrectomy may be more appropriate in these circumstances.

These procedures have been traditionally performed by standard laparotomy. The advent of laparoscopic surgery has led several surgical teams to successfully perform highly selective vagotomy, truncal vagotomy/pyloroplasty, and truncal vagotomy/antrectomy through this approach. An increase in the number of laparoscopic procedures for treatment of PUD has occurred. Laparoscopic repair of perforated peptic ulcers is safe, feasible for the experienced surgeon and is associated with decreased postoperative pain, although it does take longer than an open approach. Moreover, no difference between the two approaches is noted in postoperative complications or length of hospital stay.

Specific Operations for GUs The location and the presence of a concomitant DU dictate the operative procedure performed for a GU. Antrectomy (including the ulcer) with a Billroth I anastomosis is the treatment of choice for an antral ulcer. Vagotomy is performed only if a DU is present. Although ulcer excision with

vagotomy and drainage procedure has been proposed, the higher incidence of ulcer recurrence makes this a less desirable approach. Ulcers located near the esophagogastric junction may require a more radical approach, a subtotal gastrectomy with a Roux-en-Y esophagogastrojejunostomy (Csendes' procedure). A less aggressive approach, including antrectomy, intraoperative ulcer biopsy, and vagotomy (Kelling-Madlener procedure), may be indicated in fragile patients with a high GU. Ulcer recurrence approaches 30% with this procedure.

Surgery-Related Complications Complications seen after surgery for PUD are related primarily to the extent of the anatomic modification performed. Minimal alteration (highly selective vagotomy) is associated with higher rates of ulcer recurrence and less GI disturbance. More aggressive surgical procedures have a lower rate of ulcer recurrence but a greater incidence of GI dysfunction. Overall, morbidity and mortality related to these procedures are quite low. Morbidity associated with vagotomy and antrectomy or pyloroplasty is ≤5%, with mortality ~1%. Highly selective vagotomy has lower morbidity and mortality rates of 1 and 0.3%, respectively.

In addition to the potential early consequences of any intraabdominal procedure (bleeding, infection, thromboembolism), gastroparesis, duodenal stump leak, and efferent loop obstruction can be observed.

Recurrent Ulceration The risk of ulcer recurrence is directly related to the procedure performed. Ulcers that recur after partial gastric resection tend to develop at the anastomosis (stomal or marginal ulcer). Epigastric abdominal pain is the most frequent presenting complaint (>90%). Severity and duration of pain tend to be more progressive than observed with DUs before surgery.

Ulcers may recur for several reasons, including incomplete vagotomy, inadequate drainage, retained antrum, and, less likely, persistent or recurrent *H. pylori* infection. ZES should have been excluded preoperatively. Surreptitious use of NSAIDs is an important reason for recurrent ulcers after surgery, especially if the initial procedure was done for an NSAID-induced ulcer. Once *H. pylori* and NSAIDs have been excluded as etiologic factors, the question of incomplete vagotomy or retained gastric antrum should be explored. For the latter, fasting plasma gastrin levels should be determined. If elevated, retained antrum or ZES (see below) should be considered. Incomplete vagotomy can be ruled out by gastric acid analysis coupled with sham feeding. In this test, gastric acid output is measured while the patient sees, smells, and chews a meal (without swallowing). The cephalic phase of gastric secretion, which is mediated by the vagus, is being assessed with this study. An increase in gastric acid output in response to sham feeding is evidence that the vagus nerve is intact. A rise in serum pancreatic polypeptide >50% within 30 min of sham feeding is also suggestive of an intact vagus nerve.

Medical therapy with H_2 blockers will heal postoperative ulceration in 70–90% of patients. The efficacy of PPIs has not been fully assessed in this group, but one may anticipate greater rates of ulcer healing compared to those obtained with H_2 blockers. Repeat operation (complete vagotomy, partial gastrectomy) may be required in a small subgroup of patients who have not responded to aggressive medical management.

Afferent Loop Syndromes Although rarely seen today as a result of the decrease in the performance of Billroth II anastomosis, two types of afferent loop syndrome can occur in patients who have undergone this type of partial gastric resection. The more common of the two is bacterial overgrowth in the afferent limb secondary to stasis. Patients may experience postprandial abdominal pain, bloating, and diarrhea with concomitant malabsorption of fats and vitamin B_{12}. Cases refractory to antibiotics may require surgical revision of the loop. The less common afferent loop syndrome can present with severe abdominal pain and bloating that occur 20–60 min after meals. Pain is often followed by nausea and vomiting of bile-containing material. The pain and bloating may improve after emesis. The cause of this clinical picture is theorized to be incomplete drainage of bile and pancreatic secretions from an afferent loop that is partially obstructed. Cases refractory to dietary measures may need surgical revision or conversion of the Billroth II anastomosis to a Roux-en-Y gastrojejunostomy.

Dumping Syndrome Dumping syndrome consists of a series of vasomotor and GI signs and symptoms and occurs in patients who have undergone vagotomy and drainage (especially Billroth procedures). Two phases of dumping, early and late, can occur. Early dumping takes place 15–30 min after meals and consists of crampy abdominal discomfort, nausea, diarrhea, belching, tachycardia, palpitations, diaphoresis, light-headedness, and, rarely, syncope. These signs and symptoms arise from the rapid emptying of hyperosmolar gastric contents into the small intestine, resulting in a fluid shift into the gut lumen with plasma volume contraction and acute intestinal distention. Release of vasoactive GI hormones (vasoactive intestinal polypeptide, neurotensin, motilin) is also theorized to play a role in early dumping.

The late phase of dumping typically occurs 90 min to 3 h after meals. Vasomotor symptoms (light-headedness, diaphoresis, palpitations, tachycardia, and syncope) predominate during this phase. This component of dumping is thought to be secondary to hypoglycemia from excessive insulin release.

Dumping syndrome is most noticeable after meals rich in simple carbohydrates (especially sucrose) and high osmolarity. Ingestion of large amounts of fluids may also contribute. Up to 50% of postvagotomy and drainage patients will experience dumping syndrome to some degree early on. Signs and symptoms often improve with time, but a severe protracted picture can occur in up to 1% of patients.

Dietary modification is the cornerstone of therapy for patients with dumping syndrome. Small, multiple (six) meals devoid of simple carbohydrates coupled with elimination of liquids during meals is important. Antidiarrheals and anticholinergic agents are complementary to diet. Guar and pectin, which increase the viscosity of intraluminal contents, may be beneficial in more symptomatic individuals. Acarbose, an α-glucosidase inhibitor that delays digestion of ingested carbohydrates, has also been shown to be beneficial in the treatment of the late phases of dumping. The somatostatin analogue octreotide has been successful in diet-refractory cases. This drug is administered subcutaneously (50 μg tid), titrated according to clinical response. A long-acting depot formulation of octreotide can be administered once every 28 days and provides symptom relief comparable to the short-acting agent. In addition, patient weight gain and quality of life appear to be superior with the long-acting form.

Postvagotomy Diarrhea Up to 10% of patients may seek medical attention for the treatment of postvagotomy diarrhea. This complication is most commonly observed after truncal vagotomy, which is rarely performed today. Patients may complain of intermittent diarrhea that occurs typically 1–2 h after meals. Occasionally the symptoms may be severe and relentless. This is due to a motility disorder from interruption of the vagal fibers supplying the luminal gut. Other contributing factors may include decreased absorption of nutrients (see below), increased excretion of bile acids, and release of luminal factors that promote secretion. Diphenoxylate or loperamide is often useful in symptom control. The bile salt–binding agent cholestyramine may be helpful in severe cases. Surgical reversal of a 10-cm segment of jejunum may yield a substantial improvement in bowel frequency in a subset of patients.

Bile Reflux Gastropathy A subset of post–partial gastrectomy patients who present with abdominal pain, early satiety, nausea, and vomiting will have mucosal erythema of the gastric remnant as the only finding. Histologic examination of the gastric mucosa reveals minimal inflammation but the presence of epithelial cell injury. This clinical picture is categorized as bile or alkaline reflux gastropathy/gastritis. Although reflux of bile is implicated as the reason for this disorder, the mechanism is unknown. Prokinetic agents, cholestyramine, and sucralfate have been somewhat effective treatments. Severe refractory symptoms may require using either nuclear scanning with ^{99m}Tc-HIDA to document reflux or an alkaline challenge test, where 0.1 *N* NaOH is infused into the

stomach in an effort to reproduce the patient's symptoms. Surgical diversion of pancreaticobiliary secretions away from the gastric remnant with a Roux-en-Y gastrojejunostomy consisting of a long (50–60 cm) Roux limb has been used in severe cases. Bilious vomiting improves, but early satiety and bloating may persist in up to 50% of patients.

Maldigestion And Malabsorption Weight loss can be observed in up to 60% of patients after partial gastric resection. Patients can experience a 10% loss of body weight, which stabilizes 3 months postoperatively. A significant component of this weight reduction is due to decreased oral intake. However, mild steatorrhea can also develop. Reasons for maldigestion/malabsorption include decreased gastric acid production, rapid gastric emptying, decreased food dispersion in the stomach, reduced luminal bile concentration, reduced pancreatic secretory response to feeding, and rapid intestinal transit.

Decreased serum vitamin B_{12} levels can be observed after partial gastrectomy. This is usually not due to deficiency of intrinsic factor (IF), since a minimal amount of parietal cells (source of IF) are removed during antrectomy. Reduced vitamin B_{12} may be due to competition for the vitamin by bacterial overgrowth or inability to split the vitamin from its protein-bound source due to hypochlorhydria.

Iron-deficiency anemia may be a consequence of impaired absorption of dietary iron in patients with a Billroth II gastrojejunostomy. Absorption of iron salts is normal in these individuals; thus, a favorable response to oral iron supplementation can be anticipated. Folate deficiency with concomitant anemia can also develop in these patients. This deficiency may be secondary to decreased absorption or diminished oral intake.

Malabsorption of vitamin D and calcium resulting in osteoporosis and osteomalacia is common after partial gastrectomy and gastrojejunostomy (Billroth II). Osteomalacia can occur as a late complication in up to 25% of post–partial gastrectomy patients. Bone fractures occur twice as commonly in men after gastric surgery as in a control population. It may take years before x-ray findings demonstrate diminished bone density. Elevated alkaline phosphatase, reduced serum calcium, bone pain, and pathologic fractures may be seen in patients with osteomalacia. The high incidence of these abnormalities in this subgroup of patients justifies treating them with vitamin D and calcium supplementation indefinitely. Therapy is especially important in females. Copper deficiency has also been reported in patients undergoing surgeries that bypass the duodenum, where copper is primarily absorbed. Patients may present with a rare syndrome that includes ataxia, myelopathy, and peripheral neuropathy.

Gastric Adenocarcinoma The incidence of adenocarcinoma in the gastric stump is increased 15 years after resection. Some have reported a four- to fivefold increase in gastric cancer 20–25 years after resection. The pathogenesis is unclear but may involve alkaline reflux, bacterial proliferation, or hypochlorhydria. The role of endoscopic screening is not clear, and most guidelines do not support its use.

Additional Complications Reflux esophagitis and a higher incidence of gallstones and cholecystitis have been reported to patients undergoing subtotal gastrectomy. The latter is thought to be due to decreased gallbladder contractility associated with vagotomy and bypass of the duodenum, leading to decreased postprandial release of cholecystokinin.

RELATED CONDITIONS

ZOLLINGER–ELLISON SYNDROME

Severe peptic ulcer diathesis secondary to gastric acid hypersecretion due to unregulated gastrin release from a non-β cell often well-differentiated neuroendocrine tumor (gastrinoma) defines the components of ZES. Initially, ZES was typified by aggressive and refractory ulceration in which total gastrectomy provided the only chance for enhancing survival. Today it can be cured by surgical resection in up to 40% of patients.

Epidemiology The true incidence of ZES is unknown but estimates suggest that it varies from 0.1 to 1% of individuals presenting with PUD with 0.1–3 individuals per year having this rare diagnosis. Females are slightly more commonly affected than males, and the majority of patients are diagnosed between ages 30 and 50. Gastrinomas are classified into sporadic tumors (80%) and those associated with multiple endocrine neoplasia (MEN) type 1 (see below). The widespread availability and use of PPIs has led to a decreased patient referral for gastrinoma evaluation, delay in diagnosis, and an increase in false-positive diagnoses of ZES. In fact, diagnosis may be delayed for ≥6 years after symptoms consistent with ZES are displayed.

Pathophysiology Hypergastrinemia originating from an autonomous neoplasm is the driving force responsible for the clinical manifestations in ZES. Gastrin stimulates acid secretion through gastrin receptors on parietal cells and by inducing histamine release from ECL cells. Gastrin also has a trophic action on gastric epithelial cells. Longstanding hypergastrinemia leads to markedly increased gastric acid secretion through both parietal cell stimulation and increased parietal cell mass. The increased gastric acid output leads to peptic ulcer diathesis, erosive esophagitis, and diarrhea.

Tumor Distribution Although early studies suggested that the vast majority of gastrinomas occurred within the pancreas, a significant number of these lesions are extrapancreatic. Over 80% of these tumors are found within the hypothetical gastrinoma triangle (confluence of the cystic and common bile ducts superiorly, junction of the second and third portions of the duodenum inferiorly, and junction of the neck and body of the pancreas medially). Duodenal tumors constitute the most common nonpancreatic lesion; between 50 and 75% of gastrinomas are found here. Duodenal tumors are smaller, slower growing, and less likely to metastasize than pancreatic lesions. Less common extrapancreatic sites include stomach, bones, ovaries, heart, liver, and lymph nodes. More than 60% of tumors are considered malignant, with up to 30–50% of patients having multiple lesions or metastatic disease at presentation. Histologically, gastrin-producing cells appear well-differentiated, expressing markers typically found in endocrine neoplasms (chromogranin, neuron-specific enolase).

Clinical Manifestations Gastric acid hypersecretion is responsible for the signs and symptoms observed in patients with ZES. Peptic ulcer is the most common clinical manifestation, occurring in >90% of gastrinoma patients. Initial presentation and ulcer location (duodenal bulb) may be indistinguishable from common PUD. Clinical situations that should create suspicion of gastrinoma are ulcers in unusual locations (second part of the duodenum and beyond), ulcers refractory to standard medical therapy, ulcer recurrence after acid-reducing surgery, ulcers presenting with frank complications (bleeding, obstruction, and perforation), or ulcers in the absence of *H. pylori* or NSAID ingestion. Symptoms of esophageal origin are present in up to two-thirds of patients with ZES, with a spectrum ranging from mild esophagitis to frank ulceration with stricture and Barrett's mucosa.

Diarrhea, the next most common clinical manifestation, is found in up to 50% of patients. Although diarrhea often occurs concomitantly with acid peptic disease, it may also occur independent of an ulcer. Etiology of the diarrhea is multifactorial, resulting from marked volume overload to the small bowel, pancreatic enzyme inactivation by acid, and damage of the intestinal epithelial surface by acid. The epithelial damage can lead to a mild degree of maldigestion and malabsorption of nutrients. The diarrhea may also have a secretory component due to the direct stimulatory effect of gastrin on enterocytes or the co-secretion of additional hormones from the tumor such as vasoactive intestinal peptide.

Gastrinomas can develop in the presence of MEN 1 syndrome (Chaps. 80 and 381) in ~25% of patients. This autosomal dominant disorder involves primarily three organ sites: the parathyroid glands (80–90%), pancreas (40–80%), and pituitary gland (30–60%). The syndrome is caused by inactivating mutations of the *MEN1* tumor suppressor gene found on the long arm of chromosome 11q13. The gene encodes for Menin, which has an important role in DNA replication

and transcriptional regulation. A genetic diagnosis is obtained by sequencing of the *MEN1* gene, which can reveal mutations in 70–90% of typical MEN 1 cases. A family may have an unknown mutation, making a genetic diagnosis impossible, and therefore certain individuals will require a clinical diagnosis, which is determined by whether a patient has tumors in two of the three endocrine organs (parathyroid, pancreas/duodenum, or pituitary) or has a family history of MEN 1 and one of the endocrine organ tumors. In view of the stimulatory effect of calcium on gastric secretion, the hyperparathyroidism and hypercalcemia seen in MEN 1 patients may have a direct effect on ulcer disease. Resolution of hypercalcemia by parathyroidectomy reduces gastrin and gastric acid output in gastrinoma patients. An additional distinguishing feature in ZES patients with MEN 1 is the higher incidence of gastric carcinoid tumor development (as compared to patients with sporadic gastrinomas). ZES presents and is diagnosed earlier in MEN 1 patients, and they have a more indolent course as compared to patients with sporadic gastrinoma. Gastrinomas tend to be smaller, multiple, and located in the duodenal wall more often than is seen in patients with sporadic ZES. Establishing the diagnosis of MEN 1 is critical in order to provide genetic counseling to the patient and his or her family and also to determine the recommended surgical approach. Therefore, gastrinoma patients should be screened for MEN I performing a detailed family history and obtaining several serum markers including calcium, parathyroid, prolactin and pancreatic polypeptide levels.

Diagnosis Biochemical measurements of gastrin and acid secretion in patients suspected of ZES play an important role is establishing this rare diagnosis. Often, patients suspected of having ZES will be treated with a PPI in an effort to ameliorate symptoms and decrease the likelihood of possible acid-related complications. The presence of the PPI, which will lower acid secretion and potentially elevate fasting gastrin levels in normal individuals, will make the diagnostic approach in these individuals somewhat difficult. Significant morbidity related to peptic diathesis has been described when stopping PPIs in gastrinoma patients; therefore, a systematic approach in stopping these agents is warranted (see below). The first step in the evaluation of a patient suspected of having ZES is to obtain a fasting gastrin level. A list of clinical scenarios that should arouse suspicion regarding this diagnosis is shown in **Table 317-9**. Fasting gastrin levels obtained using a dependable assay are usually <150 pg/mL. A normal fasting gastrin, on two separate occasions, especially if the patient is on a PPI, virtually excludes this diagnosis. Virtually all gastrinoma patients will have a gastrin level >150–200 pg/mL. Measurement of fasting gastrin should be repeated to confirm the clinical suspicion. Some of the commercial biochemical assays used for measuring serum gastrin may be inaccurate. Variable specificity of the antibodies used have led to both false-positive and false-negative fasting gastrin levels, placing in jeopardy the ability to make an accurate diagnosis of ZES.

Multiple processes can lead to an elevated fasting gastrin level, the most frequent of which are gastric hypochlorhydria and achlorhydria, with or without pernicious anemia. Gastric acid induces feedback inhibition of gastrin release. A decrease in acid production will subsequently lead to failure of the feedback inhibitory pathway, resulting in net hypergastrinemia. Gastrin levels will thus be high in patients using antisecretory agents for the treatment of acid peptic disorders and dyspepsia. *H. pylori* infection can also cause hypergastrinemia. Additional causes of elevated gastrin include retained gastric antrum; G cell hyperplasia; gastric outlet obstruction; renal insufficiency; massive small-bowel obstruction; and conditions such as rheumatoid arthritis, vitiligo, diabetes mellitus, and pheochromocytoma. Although a fasting gastrin >10 times normal is highly suggestive of ZES, two-thirds of patients will have fasting gastrin levels that overlap with levels found in the more common disorders outlined above, especially if a PPI is being taken by the patient. The effect of the PPI on gastrin levels and acid secretion will linger several days after stopping the PPI; therefore, it should be stopped for a minimum of 7 days before testing. During this period, the patient should be placed on a histamine H_2 antagonist, such as famotidine, twice to three times per day. Although this type of agent has a short-term effect on gastrin and acid secretion, it needs to be stopped 24 h before repeating fasting gastrin levels or performing some of the tests highlighted below. The patient may take antacids for the final day, stopping them ~12 h before testing is performed. Heightened awareness of complications related to gastric acid hypersecretion during the period of PPI cessation is critical.

The next step at times needed for establishing a biochemical diagnosis of gastrinoma is to assess acid secretion. Nothing further needs to be done if decreased acid output in the absence of a PPI is observed. A pH can be measured on gastric fluid obtained either during endoscopy or through nasogastric aspiration; a pH <3 is suggestive of a gastrinoma, but a pH >3 is not helpful in excluding the diagnosis. In those situations where the pH is >3, formal gastric acid analysis should be performed if available. Normal BAO in nongastric surgery patients is typically <5 meq/h. A BAO >15 meq/h in the presence of hypergastrinemia is considered pathognomonic of ZES, but up to 12% of patients with common PUD may have elevated BAO to a lesser degree that can overlap with levels seen in ZES patients. In an effort to improve the sensitivity and specificity of gastric secretory studies, a BAO/ MAO ratio was established using pentagastrin infusion as a way to maximally stimulate acid production, with a BAO/MAO ratio >0.6 being highly suggestive of ZES. Pentagastrin is no longer available in the United States, making measurement of MAO virtually impossible. An endoscopic method for measuring gastric acid output has been developed but requires further validation.

Gastrin provocative tests have been developed in an effort to differentiate between the causes of hypergastrinemia and are especially helpful in patients with indeterminate acid secretory studies. The tests are the secretin stimulation test and the calcium infusion study. The most sensitive and specific gastrin provocative test for the diagnosis of gastrinoma is the secretin study. An increase in gastrin of ≥120 pg within 15 min of secretin injection has a sensitivity and specificity of >90% for ZES. PPI-induced hypochlorhydria or achlorhydria may lead to a false-positive secretin test; thus, this agent must be stopped for 1 week before testing.

The calcium infusion study is less sensitive and specific than the secretin test, which, coupled with it being a more cumbersome study with greater potential for adverse effects, relegates it to rare utilization in the cases where the patient's clinical characteristics are highly suggestive of ZES but the secretin stimulation is inconclusive.

Tumor Localization Once the biochemical diagnosis of gastrinoma has been confirmed, the tumor must be located. Multiple imaging studies have been used in an effort to enhance tumor localization **(Table 317-10)**. The broad range of sensitivity is due to the variable success rates achieved by the different investigative groups. Endoscopic ultrasound (EUS) permits imaging of the pancreas with a high degree of resolution (<5 mm). This modality is particularly helpful in excluding small neoplasms within the pancreas and in assessing the presence of surrounding lymph nodes and vascular involvement, but it is not very sensitive for finding duodenal lesions. Several types of endocrine tumors express cell-surface receptors for somatostatin, in particular the sub-type 2 (SSTR2). This permits the localization, staging, and

TABLE 317-9 When to Obtain a Fasting Serum Gastrin Level

Multiple ulcers
Ulcers in unusual locations; associated with severe esophagitis; resistant to therapy with frequent recurrences; in the absence of nonsteroidal anti-inflammatory drug ingestion or *H. pylori* infection
Ulcer patients awaiting surgery
Extensive family history for peptic ulcer disease
Postoperative ulcer recurrence
Basal hyperchlorhydria
Unexplained diarrhea or steatorrhea
Hypercalcemia
Family history of pancreatic islet, pituitary, or parathyroid tumor
Prominent gastric or duodenal folds

TABLE 317-10 Sensitivity of Imaging Studies in Zollinger-Ellison Syndrome

STUDY	SENSITIVITY, %	
	PRIMARY GASTRINOMA	METASTATIC GASTRINOMA
Ultrasound	21–28	14
CT scan	55–70	>85
Selective angiography	35–68	33–86
Portal venous sampling	70–90	N/A
SASI	55–78	41
MRI	55–70	>85
OctreoScan	67–86	80–100
EUS	80–100	N/A

Abbreviations: CT, computed tomography; EUS, endoscopic ultrasonography; MRI, magnetic resonance imaging; N/A, not applicable; OctreoScan, imaging with ^{111}In-pentreotide; SASI, selective arterial secretin injection.

prediction of therapeutic response to somatostatin analogues (see below) bygastrinomas. The original functional scinitigraphic tool developed measuring the uptake of the stable somatostatin analogue^{111} In-pentreotide (OctreoScan) has demonstrated sensitivity and specificity rates of >80%. More recently, PET-CT with ^{68}Ga-DOTATATE has been developed and is superior than Octreoscan for assessing tumor presence in patients with well-differentiated neuroendocrine tumors such as gastrinomas, with sensitivity and specificity of >90%, making it the functional imaging study of choice when available.

Up to 50% of patients have metastatic disease at diagnosis. Success in controlling gastric acid hypersecretion has shifted the emphasis of therapy toward providing a surgical cure. Detecting the primary tumor and excluding metastatic disease are critical in view of this paradigm shift. Once a biochemical diagnosis has been confirmed, the patient should first undergo an abdominal computed tomography (CT) scan, magnetic resonance imaging (MRI), or OctreoScan/PET-CT with ^{68}Ga-DOTATATE (depending on availability) to exclude metastatic disease. Once metastatic disease has been excluded, an experienced endocrine surgeon may opt for exploratory laparotomy with intraoperative ultrasound or transillumination. In other centers, careful examination of the peripancreatic area with EUS, accompanied by endoscopic exploration of the duodenum for primary tumors, will be performed before surgery. Selective arterial secretin injection may be a useful adjuvant for localizing tumors in a subset of patients. The extent of the diagnostic and surgical approach must be carefully balanced with the patient's overall physiologic condition and the natural history of a slow-growing gastrinoma.

TREATMENT

Zollinger-Ellison Syndrome

Treatment of functional endocrine tumors is directed at ameliorating the signs and symptoms related to hormone overproduction, curative resection of the neoplasm, and attempts to control tumor growth in metastatic disease.

PPIs are the treatment of choice and have decreased the need for total gastrectomy. Initial PPI doses tend to be higher than those used for treatment of GERD or PUD. The initial dose of omeprazole, lansoprazole, rabeprazole, or esomeprazole should be in the range of 60 mg in divided doses in a 24-h period. Dosing can be adjusted to achieve a BAO <10 meq/h (at the drug trough) in surgery-naive patients and to <5 meq/h in individuals who have previously undergone an acid-reducing operation. Although the somatostatin analogue has inhibitory effects on gastrin release from receptor-bearing tumors and inhibits gastric acid secretion to some extent, PPIs have the advantage of reducing parietal cell activity to a greater degree. Despite this, octreotide or lanreotide may be considered as adjunctive therapy to the PPI in patients with tumors that express somatostatin receptors and have peptic symptoms that are difficult to control with high-dose PPI.

The ultimate goal of surgery would be to provide a definitive cure. Improved understanding of tumor distribution has led to immediate cure rates as high as 33% with 10-year disease-free intervals as high as 95% in sporadic gastrinoma patients undergoing surgery. A positive outcome is highly dependent on the experience of the surgical team treating these rare tumors. Surgical therapy of gastrinoma patients with MEN 1 remains controversial because of the difficulty in rendering these patients disease-free with surgery. In contrast to the encouraging postoperative results observed in patients with sporadic disease, only 6% of MEN 1 patients are disease-free 5 years after an operation. Moreover, in contrast to patients with sporadic ZES, the clinical course of MEN 1 patients is benign and rarely leads to disease-related mortality, recommending that early surgery be deferred. Some groups suggest surgery only if a clearly identifiable, nonmetastatic lesion is documented by structural studies. Others advocate a more aggressive approach, where all patients free of hepatic metastasis are explored and all detected tumors in the duodenum are resected; this is followed by enucleation of lesions in the pancreatic head, with a distal pancreatectomy to follow. The outcome of the two approaches has not been clearly defined. Laparoscopic surgical interventions may provide attractive approaches in the future but currently seem to be of some limited benefit in patients with gastrinoma because a significant percentage of the tumors may be extrapancreatic and difficult to localize with a laparoscopic approach. Finally, patients selected for surgery should be individuals whose health status would lead them to tolerate a more aggressive operation and obtain the long-term benefits from such aggressive surgery, which are often witnessed after 10 years.

Therapy of metastatic endocrine tumors in general remains suboptimal; gastrinomas are no exception. In light of the observation that in many instances tumor growth is indolent and that many individuals with metastatic disease remain relatively stable for significant periods of time, many advocate not instituting systemic tumor-targeted therapy until evidence of tumor progression or refractory symptoms not controlled with PPIs are noted. Medical approaches, including biological therapy (IFN-α, long-acting somatostatin analogues, peptide receptor radionuclides), systemic chemotherapy (streptozotocin, 5-fluorouracil, and doxorubicin), and hepatic artery embolization, may lead to significant toxicity without a substantial improvement in overall survival. Use of temozolomide with capecitabine has demonstrated radiographic regression and progression-free survival in patients with well-differentiated NETs in the range of 70% and 18 months respectively. Systemic therapy with radiolabeled somatostatin analogues (Peptide Receptor Radiotherapy, PRRT) has been used in the therapy of metastatic neuroendocrine tumors and appears to be very promising in terms of radiographic, symptom, and progression-free survival, but additional studies are warranted. Several promising therapies are being explored, including radiofrequency ablation or cryoablation of liver lesions and use of agents that block the vascular endothelial growth receptor pathway (sunitinib) or the mammalian target of rapamycin **(Chap. 80)**.

Surgical approaches, including debulking surgery and liver transplantation for hepatic metastasis, have also produced limited benefit.

The overall 5- and 10-year survival rates for gastrinoma patients are 62–75% and 47–53%, respectively. Individuals with the entire tumor resected or those with a negative laparotomy have 5- and 10-year survival rates >90%. Patients with incompletely resected tumors have 5- and 10-year survival rates of 43 and 25%, respectively. Patients with hepatic metastasis have <20% survival at 5 years. Favorable prognostic indicators include primary duodenal wall tumors, isolated lymph node tumor, the presence of MEN 1, and undetectable tumor upon surgical exploration. Poor outcome is seen in patients with shorter disease duration; higher gastrin levels (>10,000 pg/mL); large pancreatic primary tumors (>3 cm); metastatic disease to lymph nodes, liver, and bone; and Cushing's syndrome. Rapid growth of hepatic metastases is also predictive of poor outcome.

STRESS-RELATED MUCOSAL INJURY

Patients suffering from shock, sepsis, massive burns, severe trauma, or head injury can develop acute erosive gastric mucosal changes or frank ulceration with bleeding. Classified as stress-induced gastritis or ulcers, injury is most commonly observed in the acid-producing (fundus and body) portions of the stomach. The most common presentation is GI bleeding, which is usually minimal but can occasionally be life-threatening. Respiratory failure requiring mechanical ventilation and underlying coagulopathy are risk factors for bleeding, which tends to occur 48–72 h after the acute injury or insult.

Histologically, stress injury does not contain inflammation or *H. pylori*; thus, "gastritis" is a misnomer. Although elevated gastric acid secretion may be noted in patients with stress ulceration after head trauma (Cushing's ulcer) and severe burns (Curling's ulcer), mucosal ischemia, breakdown of the normal protective barriers of the stomach, systemic release of cytokines, poor GI motility, and oxidative stress also play an important role in the pathogenesis. Acid must contribute to injury in view of the significant drop in bleeding noted when acid inhibitors are used as prophylaxis for stress gastritis.

Improvement in the general management of intensive care unit patients has led to a significant decrease in the incidence of GI bleeding due to stress ulceration. The estimated decrease in bleeding is from 20–30% to <5%. This improvement has led to some debate regarding the need for prophylactic therapy. The high mortality associated with stress-induced clinically important GI bleeding (>40%) and the limited benefit of medical (endoscopic, angiographic) and surgical therapy in a patient with hemodynamically compromising bleeding associated with stress ulcer/gastritis support the use of preventive measures in high-risk patients (mechanically ventilated, coagulopathy, multiorgan failure, or severe burns). Metaanalysis comparing H_2 blockers with PPIs for the prevention of stress-associated clinically important and overt GI bleeding demonstrates superiority of the latter without increasing the risk of nosocomial infections, increasing mortality, or prolonging intensive care unit length of stay. Therefore, PPIs are the treatment of choice for stress prophylaxis. Oral PPI is the best option if the patient can tolerate enteral administration. Pantoprazole is available as an intravenous formulation for individuals in whom enteral administration is not possible. If bleeding occurs despite these measures, endoscopy, intraarterial vasopressin, and embolization are options. If all else fails, then surgery should be considered. Although vagotomy and antrectomy may be used, the better approach would be a total gastrectomy, which has an exceedingly high mortality rate in this setting.

GASTRITIS

The term *gastritis* should be reserved for histologically documented inflammation of the gastric mucosa. Gastritis is not the mucosal erythema seen during endoscopy and is not interchangeable with "dyspepsia." The etiologic factors leading to gastritis are broad and heterogeneous. Gastritis has been classified based on time course (acute vs chronic), histologic features, and anatomic distribution or proposed pathogenic mechanism **(Table 317-11)**.

The correlation between the histologic findings of gastritis, the clinical picture of abdominal pain or dyspepsia, and endoscopic findings noted on gross inspection of the gastric mucosa is poor. Therefore, there is no typical clinical manifestation of gastritis.

Acute Gastritis The most common causes of acute gastritis are infectious. Acute infection with *H. pylori* induces gastritis. However, *H. pylori* acute gastritis has not been extensively studied. It is reported as presenting with sudden onset of epigastric pain, nausea, and vomiting, and limited mucosal histologic studies demonstrate a marked infiltrate of neutrophils with edema and hyperemia. If not treated, this picture will evolve into one of chronic gastritis. Hypochlorhydria lasting for up to 1 year may follow acute *H. pylori* infection.

Bacterial infection of the stomach or phlegmonous gastritis is a rare, potentially life-threatening disorder characterized by marked and diffuse acute inflammatory infiltrates of the entire gastric wall, at times accompanied by necrosis. Elderly individuals, alcoholics, and AIDS patients may be affected. Potential iatrogenic causes include polypectomy and mucosal injection with India ink. Organisms associated with this entity include streptococci, staphylococci, *Escherichia coli*, *Proteus*, and *Haemophilus* species. Failure of supportive measures and antibiotics may result in gastrectomy.

Other types of infectious gastritis may occur in immunocompromised individuals such as AIDS patients. Examples include herpetic (herpes simplex) or CMV gastritis. The histologic finding of intranuclear inclusions would be observed in the latter.

TABLE 317-11 Classification of Gastritis

I. Acute gastritis
  A. Acute *H. pylori* infection
  B. Other acute infectious gastritides
    1. Bacterial (other than *H. pylori)*
    2. *H. heilmannii*
    3. Phlegmonous
    4. Mycobacterial
    5. Syphilitic
    6. Viral
    7. Parasitic
    8. Fungal
II. Chronic atrophic gastritis
  A. Type A: Autoimmune, body-predominant
  B. Type B: *H. pylori*–related, antral-predominant
  C. Indeterminate
III. Uncommon forms of gastritis
  A. Lymphocytic
  B. Eosinophilic
  C. Crohn's disease
  D. Sarcoidosis
  E. Isolated granulomatous gastritis
  F. Russell body gastritis

Chronic Gastritis Chronic gastritis is identified histologically by an inflammatory cell infiltrate consisting primarily of lymphocytes and plasma cells, with very scant neutrophil involvement. Distribution of the inflammation may be patchy, initially involving superficial and glandular portions of the gastric mucosa. This picture may progress to more severe glandular destruction, with atrophy and metaplasia. Chronic gastritis has been classified according to histologic characteristics. These include superficial atrophic changes and gastric atrophy. The association of atrophic gastritis with the development of gastric cancer has led to the development of endoscopic and serologic markers of severity. Some of these include gross inspection and classification of mucosal abnormalities during standard endoscopy, magnification endoscopy, endoscopy with narrow band imaging and/or autofluorescence imaging, and measurement of several serum biomarkers including pepsinogen I and II levels, gastrin-17, and anti-*H. pylori* serologies. The clinical utility of these tools is currently being explored.

The early phase of chronic gastritis is *superficial gastritis.* The inflammatory changes are limited to the lamina propria of the surface mucosa, with edema and cellular infiltrates separating intact gastric glands. The next stage is *atrophic gastritis.* The inflammatory infiltrate extends deeper into the mucosa, with progressive distortion and destruction of the glands. The final stage of chronic gastritis is *gastric atrophy.* Glandular structures are lost, and there is a paucity of inflammatory infiltrates. Endoscopically, the mucosa may be substantially thin, permitting clear visualization of the underlying blood vessels.

Gastric glands may undergo morphologic transformation in chronic gastritis. Intestinal metaplasia denotes the conversion of gastric glands to a small intestinal phenotype with small-bowel mucosal glands containing goblet cells. The metaplastic changes may vary in distribution from patchy to fairly extensive gastric involvement. Intestinal metaplasia is an important predisposing factor for gastric cancer **(Chap. 76)**.

Chronic gastritis is also classified according to the predominant site of involvement. Type A refers to the body-predominant form (autoimmune), and type B is the antral-predominant form (*H. pylori*–related).

This classification is artificial in view of the difficulty in distinguishing between these two entities. The term *AB gastritis* has been used to refer to a mixed antral/body picture.

TYPE A GASTRITIS The less common of the two forms involves primarily the fundus and body, with antral sparing. Traditionally, this form of gastritis has been associated with pernicious anemia **(Chap. 95)** in the presence of circulating antibodies against parietal cells and IF; thus, it is also called *autoimmune gastritis. H. pylori* infection can lead to a similar distribution of gastritis. The characteristics of an autoimmune picture are not always present.

Antibodies to parietal cells have been detected in >90% of patients with pernicious anemia and in up to 50% of patients with type A gastritis. The parietal cell antibody is directed against H^+,K^+-ATPase. T cells are also implicated in the injury pattern of this form of gastritis. A subset of patients infected with *H. pylori* develop antibodies against H^+,K^+-ATPase, potentially leading to the atrophic gastritis pattern seen in some patients infected with this organism. The mechanism is thought to involve molecular mimicry between *H. pylori* LPS and H^+,K^+-ATPase.

Parietal cell antibodies and atrophic gastritis are observed in family members of patients with pernicious anemia. These antibodies are observed in up to 20% of individuals aged >60 and in ~20% of patients with vitiligo and Addison's disease. About one-half of patients with pernicious anemia have antibodies to thyroid antigens, and about 30% of patients with thyroid disease have circulating antiparietal cell antibodies. Anti-IF antibodies are more specific than parietal cell antibodies for type A gastritis, being present in ~40% of patients with pernicious anemia. Another parameter consistent with this form of gastritis being autoimmune in origin is the higher incidence of specific familial histocompatibility haplotypes such as HLA-B8 and HLA-DR3.

The parietal cell–containing gastric gland is preferentially targeted in this form of gastritis, and achlorhydria results. Parietal cells are the source of IF, the lack of which will lead to vitamin B_{12} deficiency and its sequelae (megaloblastic anemia, neurologic dysfunction).

Gastric acid plays an important role in feedback inhibition of gastrin release from G cells. Achlorhydria, coupled with relative sparing of the antral mucosa (site of G cells), leads to hypergastrinemia. Gastrin levels can be markedly elevated (>500 pg/mL) in patients with pernicious anemia. ECL cell hyperplasia with frank development of gastric carcinoid tumors may result from gastrin trophic effects. Hypergastrinemia and achlorhydria may also be seen in nonpernicious anemia–associated type A gastritis.

TYPE B GASTRITIS Type B, or antral-predominant, gastritis is the more common form of chronic gastritis. *H. pylori* infection is the cause of this entity. Although described as "antral-predominant," this is likely a misnomer in view of studies documenting the progression of the inflammatory process toward the body and fundus of infected individuals. The conversion to a pangastritis is time-dependent and estimated to require 15–20 years. This form of gastritis increases with age, being present in up to 100% of persons aged >70. Histology improves after *H. pylori* eradication. The number of *H. pylori* organisms decreases dramatically with progression to gastric atrophy, and the degree of inflammation correlates with the level of these organisms. Early on, with antral-predominant findings, the quantity of *H. pylori* is highest and a dense chronic inflammatory infiltrate of the lamina propria is noted, accompanied by epithelial cell infiltration with polymorphonuclear leukocytes **(Fig. 317-15)**.

Multifocal atrophic gastritis, gastric atrophy with subsequent metaplasia, has been observed in chronic *H. pylori*–induced gastritis. This may ultimately lead to development of gastric adenocarcinoma (Fig. 317-8; **Chap. 76**). *H. pylori* infection is now considered an independent risk factor for gastric cancer. Worldwide epidemiologic studies have documented a higher incidence of *H. pylori* infection in patients with adenocarcinoma of the stomach as compared to control subjects. Seropositivity for *H. pylori* is associated with a three- to sixfold increased risk of gastric cancer. This risk may be as high as ninefold after adjusting for the inaccuracy of serologic testing in the elderly. The mechanism by which *H. pylori* infection leads to cancer is unknown, but it appears to be related to the chronic inflammation induced by the organism. Eradication of *H. pylori* as a general preventative measure for gastric cancer is being evaluated but is not yet recommended.

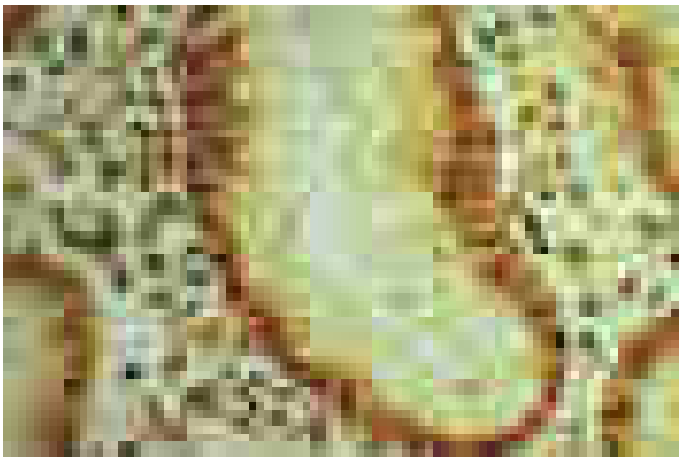


FIGURE 317-15 Chronic gastritis and *H. pylori* organisms. Steiner silver stain of superficial gastric mucosa showing abundant darkly stained microorganisms layered over the apical portion of the surface epithelium. Note that there is no tissue invasion.

Infection with *H. pylori* is also associated with development of a low-grade B cell lymphoma, gastric MALT lymphoma **(Chap. 104)**. The chronic T cell stimulation caused by the infection leads to production of cytokines that promote the B cell tumor. The tumor should be initially staged with a CT scan of the abdomen and EUS. Tumor growth remains dependent on the presence of *H. pylori*, and its eradication is often associated with complete regression of the tumor. The tumor may take more than a year to regress after treating the infection. Such patients should be followed by EUS every 2–3 months. If the tumor is stable or decreasing in size, no other therapy is necessary. If the tumor grows, it may have become a high-grade B cell lymphoma. When the tumor becomes a high-grade aggressive lymphoma histologically, it loses responsiveness to *H. pylori* eradication.

TREATMENT

Chronic Gastritis

Treatment in chronic gastritis is aimed at the sequelae and not the underlying inflammation. Patients with pernicious anemia will require parenteral vitamin B_{12} supplementation on a long-term basis. Eradication of *H. pylori* is often recommended even if PUD or a low-grade MALT lymphoma is not present. Expert opinion suggests that patients with atrophic gastritis complicated by intestinal metaplasia without dysplasia should undergo surveillance endoscopy every 3 years.

Miscellaneous Forms of Gastritis *Lymphocytic gastritis* is characterized histologically by intense infiltration of the surface epithelium with lymphocytes. The infiltrative process is primarily in the body of the stomach and consists of mature T cells and plasmacytes. The etiology of this form of chronic gastritis is unknown. It has been described in patients with celiac sprue, but whether there is a common factor associating these two entities is unknown. No specific symptoms suggest lymphocytic gastritis. A subgroup of patients have thickened folds noted on endoscopy. These folds are often capped by small nodules that contain a central depression or erosion; this form of the disease is called *varioliform gastritis. H. pylori* probably plays no significant role in lymphocytic gastritis. Therapy with glucocorticoids or sodium cromoglycate has obtained unclear results.

Marked eosinophilic infiltration involving any layer of the stomach (mucosa, muscularis propria, and serosa) is characteristic of *eosinophilic gastritis.* Affected individuals will often have circulating eosinophilia with clinical manifestation of systemic allergy. Involvement may range

from isolated gastric disease to diffuse eosinophilic gastroenteritis. Antral involvement predominates, with prominent edematous folds being observed on endoscopy. These prominent antral folds can lead to outlet obstruction. Patients can present with epigastric discomfort, nausea, and vomiting. Treatment with glucocorticoids has been successful.

Several systemic disorders may be associated with *granulomatous gastritis.* Gastric involvement has been observed in Crohn's disease. Involvement may range from granulomatous infiltrates noted only on gastric biopsies to frank ulceration and stricture formation. Gastric Crohn's disease usually occurs in the presence of small-intestinal disease. Several rare infectious processes can lead to granulomatous gastritis, including histoplasmosis, candidiasis, syphilis, and tuberculosis. Other unusual causes of this form of gastritis include sarcoidosis, idiopathic granulomatous gastritis, and eosinophilic granulomas involving the stomach. Establishing the specific etiologic agent in this form of gastritis can be difficult, at times requiring repeat endoscopy with biopsy and cytology. Occasionally, a surgically obtained full-thickness biopsy of the stomach may be required to exclude malignancy.

Russell body gastritis (RBG) is a mucosal lesion of unknown etiology that has a pseudotumoral endoscopic appearance. Histologically, it is defined by the presence of numerous plasma cells containing Russell bodies (RBs) that express kappa and lambda light chains. Only 10 cases have been reported, and 7 of these have been associated with *H. pylori* infection. The lesion can be confused with a neoplastic process, but it is benign in nature, and the natural history of the lesion is not known. There have been cases of resolution of the lesion when *H. pylori* was eradicated.

MÉNÉTRIER'S DISEASE

Ménétrier's disease (MD) is a very rare gastropathy characterized by large, tortuous mucosal folds. MD has an average age of onset of 40–60 years with a male predominance. The differential diagnosis of large gastric folds includes ZES, malignancy (lymphoma, infiltrating carcinoma), infectious etiologies (CMV, histoplasmosis, syphilis, tuberculosis), gastritis polyposa profunda, and infiltrative disorders such as sarcoidosis. MD is most commonly confused with large or multiple gastric polyps (prolonged PPI use) or familial polyposis syndromes. The mucosal folds in MD are often most prominent in the body and fundus, sparing the antrum. Histologically, massive foveolar hyperplasia (hyperplasia of surface and glandular mucous cells) and a marked reduction in oxyntic glands and parietal cells and chief cells are noted. This hyperplasia produces the prominent folds observed. The pits of the gastric glands elongate and may become extremely dilated and tortuous. Although the lamina propria may contain a mild chronic inflammatory infiltrate including eosinophils and plasma cells, MD is not considered a form of gastritis. The etiology of this unusual clinical picture in children is often CMV, but the etiology in adults is unknown. Overexpression of the growth factor TGF-α has been demonstrated in patients with MD. The overexpression of TGF-α in turn results in overstimulation of the epidermal growth factor receptor (EGFR) pathway and increased proliferation of mucus cells, resulting in the observed foveolar hyperplasia.

The clinical presentation in adults is usually insidious and progressive. Epigastric pain, nausea, vomiting, anorexia, peripheral edema, and weight loss are signs and symptoms in patients with MD. Occult GI bleeding may occur, but overt bleeding is unusual and, when present, is due to superficial mucosal erosions. In fact, bleeding is more often seen in one of the common mimics of MD, gastric polyposis. Twenty to 100% of patients (depending on time of presentation) develop a protein-losing gastropathy due to hypersecretion of gastric mucus accompanied by hypoalbuminemia and edema. Gastric acid secretion is usually reduced or absent because of the decreased parietal cells. Large gastric folds are readily detectable by either radiographic (barium meal) or endoscopic methods. Endoscopy with deep mucosal biopsy, preferably full thickness with a snare technique, is required to establish the diagnosis and exclude other entities that may present similarly. A nondiagnostic biopsy may lead to a surgically obtained full-thickness biopsy to exclude malignancy. Although MD is considered premalignant by some, the risk of neoplastic progression is not defined. Complete blood count, serum gastrin, serum albumin, CMV and *H. pylori* serology, and pH testing of gastric aspirate during endoscopy should be included as part of the initial evaluation of patients with large gastric folds.

TREATMENT

Ménétrier's Disease

Medical therapy with anticholinergic agents, prostaglandins, PPIs, prednisone, somatostatin analogues (octreotide) and H_2 receptor antagonists yields varying results. Ulcers should be treated with a standard approach. The discovery that MD is associated with overstimulation of the EGFR pathway has led to the successful use of the EGF inhibitory antibody, cetuximab, in these patients. Specifically, four of seven patients who completed a 1-month trial with this agent demonstrated near complete histologic remission and improvement in symptoms. Cetuximab is now considered the first-line treatment for MD, leaving total gastrectomy for severe disease with persistent and substantial protein loss despite therapy with this agent.

FURTHER READING

Basuroy R et al: Neuroendocrine tumors. Gastroenterol Clin N Am 45:487, 2016.

Chan JA, Kulke MH: Medical management of pancreatic neuroendocrine tumors: Current and future therapy. Surg Oncol Clin N Am 25:423, 2016.

Chey WD et al: ACG Clinical Guideline: Treatment of *Helicobacter pylori* Infection. Am J Gastroenterol 112:212, 2017.

Clancy TE: Surgical management of pancreatic neuroendocrine tumors. Hematol Oncol Clin N Am 30:103, 2016.

Davis JL, Ripley RT: Postgastrectomy syndromes and nutritional considerations following gastric surgery. Surg Clin N Am 97:277, 2017.

Dreischulte T: Safer prescribing—A trial of education, informatics, and financial incentives. N Engl J Med 374:1053, 2016.

Epelboym I, Mazeh H: Zollinger-Ellison Syndrome: Classical considerations and current controversies. Oncologist 19:44, 2014.

Graham DY: *Helicobacter pylori* update: Gastric cancer, reliable therapy, and possible benefits. Gastroenterology 148:719, 2015.

Hunt RH et al: The stomach in health and disease. Gut 64:1650, 2015.

Kim HU: Diagnostic and treatment approaches for refractory peptic ulcers. Clin Endosc 48:285, 2015.

Laine L: Upper gastrointestinal bleeding due to a peptic ulcer. N Engl J Med 374:2367, 2016.

McCarberg BH, Cryer B: Evolving therapeutic strategies to improve nonsteroidal anti-inflammatory drug safety. Am J Ther 22:167, 2015.

Melcarne L et al: Management of NSAID-associated peptic ulcer disease. Expert Rev Gastroenterol Hepatol 6:723, 2016.

Patel KA, Howden CW: Update on the diagnosis and management of *Helicobacter pylori* infection in adults. J Clin Gastroenterol 49:461, 2015.

Pelletier JP et al: Efficacy and safety of oral and analgesics in the management of osteoarthritis: Evidence from real-life setting trials and surveys. Semin Arthritis Rheum 45:S22, 2016.

Scarpignato C et al: Effective and safe proton pump inhibitor therapy in acid-related disease—A position paper addressing benefits and potential harms of acid suppression. BMC Med 14:179, 2016.

Schoenfeld AJ, Grady D: Adverse effects associate with proton pump inhibitors. JAMA Int Med 176:172, 2016.

Schubert ML: Gastric acid secretion. Curr Opin Gastroenterol 32:452, 2016.

Singh S et al: Diagnosis and management of gastrointestinal neuroendocrine tumors: An evidence-based Canadian consensus. Cancer Treat Rev 47:32, 2016.

Sugano K et al: Kyoto global consensus report on *Helicobacter pylori* gastritis. Gut 64:1353, 2015.

Talley NJ: Functional dyspepsia: New insights into pathogenesis and therapy. Korean J Intern Med 31:444, 2016.

Vinik AI, Chaya C: Clinical presentation and diagnosis of neuroendocrine tumors. Hematol Oncol Clin N Am 30:21, 2016.

318 Disorders of Absorption

Henry J. Binder

Disorders of absorption constitute a broad spectrum of conditions with multiple etiologies and varied clinical manifestations. Almost all of these clinical problems are associated with *diminished* intestinal absorption of one or more dietary nutrients and are often referred to as the *malabsorption syndrome.* This term is not ideal as it represents a pathophysiologic state, does *not* provide an etiologic explanation for the underlying problem, and should not be considered an adequate final diagnosis. The only clinical conditions in which absorption is *increased* are hemochromatosis and Wilson's disease, in which absorption of iron and copper, respectively, is elevated.

Most malabsorption syndromes are associated with *steatorrhea*, an increase in stool fat excretion to >7% of dietary fat intake. Some malabsorption disorders are not associated with steatorrhea: primary lactase deficiency, a congenital absence of the small-intestinal brush border disaccharidase enzyme lactase, is associated with lactose "malabsorption," and pernicious anemia is associated with a marked decrease in intestinal absorption of cobalamin (vitamin B_{12}) due to an absence of gastric parietal-cell intrinsic factor, which is required for cobalamin absorption.

Disorders of absorption must be included in the differential diagnosis of diarrhea **(Chap. 42)**. First, diarrhea is frequently associated with and/or is a consequence of the diminished absorption of one or more dietary nutrients. The diarrhea may be secondary either to the intestinal process that is responsible for the steatorrhea or to steatorrhea per se. Thus, celiac disease (see below) is associated with both extensive morphologic changes in the small-intestinal mucosa and reduced absorption of several dietary nutrients; in contrast, the diarrhea of steatorrhea is the result of the effect of nonabsorbed dietary fatty acids on intestinal (usually colonic) ion transport. For example, oleic and ricinoleic acids (a bacterially hydroxylated fatty acid that is also the active ingredient in castor oil, a widely used laxative) induce active colonic Cl ion secretion, most likely secondary to increasing intracellular Ca. In addition, diarrhea per se may result in mild steatorrhea (<11 g of fat excretion while on a 100-g fat diet). Second, most patients will indicate that they have diarrhea, not that they have fat malabsorption. Third, many intestinal disorders that have diarrhea as a prominent symptom (e.g., ulcerative colitis, traveler's diarrhea secondary to an enterotoxin produced by *Escherichia coli*) do not necessarily have diminished absorption of any dietary nutrient.

Diarrhea as a *symptom* (i.e., when the term is used by patients to describe their bowel movement pattern) may reflect a decrease in stool consistency, an increase in stool volume, an increase in number of bowel movements, or any combination of these three changes. In contrast, diarrhea as a *sign* is a quantitative increase in stool water or weight of >200–225 mL or g per 24 h when a Western-type diet is consumed. Individuals consuming a diet with higher-fiber content may normally have a stool weight of up to 400 g/24 h. Thus, the clinician must clarify what an individual patient means by diarrhea. Some 10% of patients referred to gastroenterologists for further evaluation of unexplained diarrhea do not have an increase in stool water when this variable is determined quantitatively. Such patients may have small, frequent, somewhat loose bowel movements with stool urgency that is indicative of proctitis, but do not have an increase in stool weight or volume. In addition, an occasional patient will describe their fecal incontinence as diarrhea due to social embarrassment.

It is also critical to establish whether a patient's diarrhea is secondary to diminished absorption of one or more dietary nutrients rather than being due to small- and/or large-intestinal fluid and electrolyte secretion. The former has often been termed *osmotic diarrhea*, while the latter has been referred to as *secretory diarrhea.* Unfortunately, both secretory and osmotic elements can be present simultaneously in the same disorder; thus, this distinction is not always precise. Nonetheless, two studies—determination of stool electrolytes and observation of the effect of a fast on stool output—can help make this distinction.

The demonstration of the effect of prolonged (>24 h) fasting on stool output can suggest that a *dietary nutrient* is responsible for the individual's diarrhea. Secretory diarrhea associated with enterotoxin-induced traveler's diarrhea would not be affected by prolonged fasting, as enterotoxin-induced stimulation of intestinal fluid and electrolyte secretion is not altered by eating. In contrast, diarrhea secondary to lactose malabsorption in primary lactase deficiency would undoubtedly cease during a prolonged fast. Thus, a substantial decrease in stool output by a fasting patient during quantitative stool collection lasting at least 24 h is presumptive evidence that the diarrhea is related to malabsorption of one or more dietary nutrients. The persistence of stool output during fasting indicates that the diarrhea is likely secretory and that its cause is *not* a dietary nutrient. Either a luminal (e.g., *E. coli* enterotoxin) or a circulating (e.g., vasoactive intestinal peptide) secretagogue could be responsible for unaltered persistence of a patient's diarrhea during a prolonged fast. The observed effects of fasting can be compared and correlated with stool electrolyte and osmolality determinations.

Measurement of stool electrolytes and osmolality requires comparison of Na^+ and K^+ concentrations in liquid stool with the osmolality of the stool in order to determine the presence or absence of a so-called stool osmotic gap. The following formula is used:

$$2 \times (\text{stool } [Na^+] + \text{stool } [K^+]) \leq \text{stool osmolality}$$

The cation concentrations are doubled to estimate stool anion concentrations. The presence of a significant osmotic gap suggests the presence in stool water of a substance (or substances) other than Na/K/anions, which is presumably responsible for the patient's diarrhea. Originally, stool osmolality was measured, but it is almost invariably greater than the required 290–300 mosmol/kg H_2O, reflecting bacterial degradation of nonabsorbed carbohydrate either immediately before defecation or in the stool jar while specimen awaits chemical analysis, even when the stool is refrigerated. As a result, the stool osmolality should be assumed to be 300 mosmol/kg H_2O. A low stool osmolality (<290 mosmol/kg H_2O) reflects the addition of either dilute urine or water, indicating either collection of urine and stool together or so-called factitious diarrhea, a form of Münchausen's syndrome. When the calculated difference in the formula above is >50, an osmotic gap exists; its presence suggests that the diarrhea is due to a nonabsorbed dietary nutrient—for example a fatty acid and/or a carbohydrate. When this difference is <25, it is presumed that a dietary nutrient is not responsible for the diarrhea. Since elements of both osmotic diarrhea (i.e., due to malabsorption of a dietary nutrient) and secretory diarrhea may be present, this distinction at times is less clear-cut at the bedside than when used as a teaching example. Ideally, the presence of an osmotic gap will be associated with a marked decrease in stool output during a prolonged fast, while an osmotic gap will likely be absent in an individual whose stool output is not reduced substantially during a period of fasting.

NUTRIENT DIGESTION AND ABSORPTION

The lengths of the small intestine and the colon are ~300 and ~80 cm, respectively. However, the effective functional surface area is ~600-fold greater than that of a hollow tube as a result of folds, villi (in the small intestine), and microvilli. The functional surface area of the small intestine is somewhat greater than that of a doubles tennis court. In addition to nutrient digestion and absorption, the intestinal epithelia have several other functions:

1. *Barrier and immune defense.* The intestine is exposed to a large number of potential antigens and enteric and invasive microorganisms, and it is extremely effective at preventing the entry of almost all of these agents. The intestinal mucosa also synthesizes and secretes secretory IgA.
2. *Fluid and electrolyte absorption and secretion.* The intestine absorbs ~7–8 L of fluid daily, a volume comprising dietary fluid intake (1–2 L/d) and salivary, gastric, pancreatic, biliary, and intestinal

fluid (6–7 L/d). Several stimuli, especially bacteria and bacterial enterotoxins, induce fluid and electrolyte secretion that may lead to diarrhea **(Chap. 128)**.

3. *Synthesis and secretion of several proteins.* The intestinal mucosa is a major site for the production of proteins, including apolipoproteins.
4. *Production of several bioactive amines and peptides.* The intestine is one of the largest endocrine organs in the body and produces several amines (e.g., 5-hydroxytryptophan) and peptides that serve as paracrine and hormonal mediators of intestinal function.

The small and large intestines are distinct anatomically (villi are present in the small intestine but are absent in the colon) and functionally (nutrient digestion and absorption take place in the small intestine but not in the colon). No precise anatomic characteristics separate duodenum, jejunum, and ileum, although certain nutrients are absorbed exclusively in specific areas of the small intestine. However, villous cells in the small intestine (surface epithelial cells in the colon) and crypt cells have distinct anatomic and functional characteristics. Intestinal epithelial cells are continuously renewed; new proliferating epithelial cells at the base of the crypt migrate over 48–72 h to the tip of the villus (or surface of the colon), where they exist as well-developed epithelial cells with digestive and absorptive function. This high rate of cell turnover explains the relatively rapid resolution of diarrhea and other digestive-tract side effects during chemotherapy as new cells not exposed to these toxic agents are produced. Equally important is the paradigm of separation of villous/surface cell and crypt cell functions. Digestive hydrolytic enzymes are present primarily in the brush border of villous epithelial cells. Absorptive and secretory functions are also separate: villous/surface cells are primarily, but not exclusively, the site for absorptive function, while secretory function is located in crypts of both the small and large intestines.

Nutrients, minerals, and vitamins are absorbed by one or more active-transport mechanisms. These mechanisms are energy dependent and are mediated by membrane transport proteins. These processes will result in the *net* movement of a substance against or in the absence of an electrochemical concentration gradient. Intestinal absorption of amino acids and monosaccharides (e.g., glucose) is also a specialized form of active transport—*secondary active transport.* The movement of actively transported nutrients against a concentration gradient is Na^+ dependent and is due to a Na^+ gradient across the apical membrane. The Na^+ gradient is maintained by Na^+,K^+-adenosine triphosphatase (ATPase), the so-called Na^+ pump located on the basolateral membrane, which extrudes Na^+ and maintains low intracellular [Na] as well as the Na^+ gradient across the apical membrane. As a result, active glucose absorption and glucose-stimulated Na^+ absorption require both the apical membrane transport protein SGLT1 and the basolateral Na^+,K^+-ATPase. In addition to requiring Na^+ for its absorption, glucose stimulates Na^+ and fluid absorption; this effect is the physiologic basis of oral rehydration therapy for the treatment of diarrhea **(Chap. 42)**. **The mechanisms of intestinal fluid and electrolyte absorption and secretion are discussed in Chap. 42.**

Although the intestinal epithelial cells are crucial mediators of absorption and of ion and water flow, the several cell types in the lamina propria (e.g., mast cells, macrophages, myofibroblasts) and the enteric nervous system interact with the epithelium to regulate mucosal cell function. Intestinal function results from the integrated responses and interactions of intestinal epithelial cells and intestinal muscle.

ENTEROHEPATIC CIRCULATION OF BILE ACIDS

Bile acids are not present in the diet but are synthesized in the liver by a series of enzymatic steps that also represent cholesterol catabolism. Indeed, interruption of the enterohepatic circulation of bile acids can reduce serum cholesterol levels by 10% before a new steady state is established. Bile acids are either primary or secondary. Primary bile acids are synthesized in the liver from cholesterol, and secondary bile acids are synthesized from primary bile acids in the intestine by colonic bacterial enzymes. The two primary bile acids in humans are cholic acid and chenodeoxycholic acid; the two most abundant secondary bile acids are deoxycholic acid and lithocholic acid. The liver synthesizes

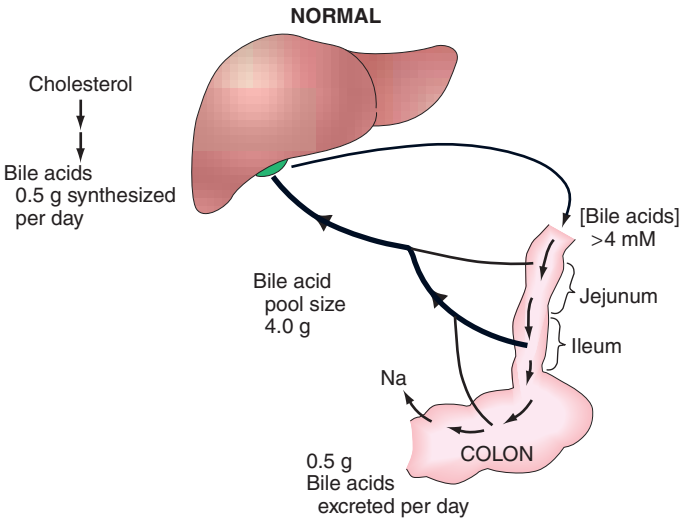


FIGURE 318-1 Schematic representation of the enterohepatic circulation of bile acids. Bile-acid synthesis is cholesterol catabolism and occurs in the liver. Bile acids are secreted in bile and are stored in the gallbladder between meals and at night. Food in the duodenum induces the release of cholecystokinin, a potent stimulus for gallbladder contraction resulting in bile-acid entry into the duodenum. Bile acids are primarily absorbed via an Na-dependent transport process that is located only in the ileum. A relatively small quantity of bile acids (~500 mg) is not absorbed in a 24-h period and is lost in stool. Fecal bile-acid losses are matched by bile-acid synthesis. The bile-acid pool (the total amount of bile acids in the body) is ~4 g and is circulated twice during each meal or six to eight times in a 24-h period.

~500 mg of bile acids daily; the bile acids are conjugated to either taurine or glycine (to form tauroconjugated and glycoconjugated bile acids, respectively) and are secreted into the duodenum in bile. The primary functions of bile acids are to (1) promote bile flow, (2) solubilize cholesterol and phospholipid in the gallbladder by mixed micelle formation, and (3) enhance dietary lipid digestion and absorption by forming mixed micelles in the proximal small intestine.

Bile acids are primarily absorbed by an active, Na^+-dependent process that takes place exclusively in the ileum; to a lesser extent, they are absorbed by non-carrier-mediated transport processes in the jejunum, ileum, and colon. Conjugated bile acids that enter the colon are deconjugated by colonic bacterial enzymes. The unconjugated bile acids are rapidly absorbed by nonionic diffusion. Colonic bacterial enzymes also dehydroxylate bile acids to secondary bile acids.

Bile acids absorbed from the intestine return to the liver via the portal vein and are then re-secreted **(Fig. 318-1)**. Bile-acid synthesis is largely autoregulated by 7α-hydroxylase, the initial enzyme in cholesterol degradation. A decrease in the quantity of bile acids returning to the liver from the intestine is associated with an increase in bile-acid synthesis/cholesterol catabolism (mediated by fibroblast growth factor [FGF19]), which helps keep the bile-acid pool size relatively constant. However, the capacity to increase bile-acid synthesis is limited to ~2- to 2.5-fold (see below). The bile-acid pool size is ~4 g. The pool is circulated via the enterohepatic circulation about twice during each meal, or six to eight times during a 24-h period. A relatively small quantity of bile acids is not absorbed and is excreted in stool daily; this fecal loss is matched by hepatic bile-acid synthesis.

Defects in any of the steps in enterohepatic circulation of bile acids can result in a decrease in the duodenal concentration of conjugated bile acids and consequently in the development of steatorrhea. Thus, steatorrhea can be caused by abnormalities in bile-acid synthesis and excretion, their physical state in the intestinal lumen, and reabsorption **(Table 318-1)**.

Synthesis Decreased bile-acid synthesis and steatorrhea have been demonstrated in chronic liver disease, but steatorrhea often is not a major component of illness in these patients.

Secretion Although bile-acid secretion may be reduced or absent in biliary obstruction, steatorrhea is rarely a significant medical

TABLE 318-1 Defects in Enterohepatic Circulation of Bile Acids

PROCESS	PATHOPHYSIOLOGIC DEFECT	DISEASE EXAMPLE
Synthesis	Decreased hepatic function	Cirrhosis
Biliary secretion	Altered canalicular function	Primary biliary cirrhosis
Maintenance of conjugated bile acids	Bacterial overgrowth	Jejunal diverticulosis
Reabsorption	Abnormal ileal function	Crohn's disease

problem in these patients. In contrast, primary biliary cirrhosis represents a defect in canalicular excretion of organic anions, including bile acids, and not infrequently is associated with steatorrhea and its consequences (e.g., chronic bone disease). Thus, the osteopenia/osteomalacia and other chronic bone abnormalities often present in patients with primary biliary cirrhosis and other cholestatic syndromes are secondary to steatorrhea that then leads to calcium and vitamin D malabsorption as well as to the effects of cholestasis (e.g., bile acids and inflammatory cytokines).

Maintenance of Conjugated Bile Acids In bacterial overgrowth syndromes associated with diarrhea, steatorrhea, and macrocytic anemia, a colonic type of bacterial flora is increased in the small intestine. Steatorrhea is primarily a result of the decrease in conjugated bile acids secondary to their deconjugation by colonic-type bacteria. Two complementary explanations account for the resulting impairment of micelle formation: (1) Unconjugated bile acids are rapidly absorbed in the jejunum by nonionic diffusion, and the result is a reduced concentration of duodenal bile acids; and (2) the critical micellar concentration (CMC) of unconjugated bile acids is higher than that of conjugated bile acids; therefore, unconjugated bile acids are less effective than conjugated bile acids in micelle formation.

Reabsorption Ileal dysfunction caused by either Crohn's disease or surgical resection results in a decrease in bile-acid reabsorption in the ileum and an *increase* in the delivery of bile acids to the large intestine. The resulting clinical consequences—diarrhea with or without steatorrhea—are determined by the *degree* of ileal dysfunction and the *response* of the enterohepatic circulation to bile-acid losses **(Table 318-2)**. Patients with limited ileal disease or resection often have diarrhea but not steatorrhea. The diarrhea, a result of stimulation of active Cl secretion by bile acids in the colon, has been called *bile-acid diarrhea* or *choleretic enteropathy* and responds promptly to cholestyramine, an anion-binding resin. Steatorrhea does not develop because hepatic synthesis of bile acids increases to compensate for the rate of fecal bile-acid losses, resulting in maintenance of both the bile-acid pool size and the intraduodenal concentrations of bile acids. In contrast, patients with greater degrees of ileal disease and/or resection often have diarrhea and steatorrhea that do not respond to cholestyramine. In this situation, ileal disease is also associated with increased volumes of bile acids entering the colon; however, hepatic synthesis can no longer increase sufficiently to maintain the bile-acid pool size. As a consequence, the intraduodenal concentration of bile acids is reduced to less than the CMC, and the result is impaired micelle formation and steatorrhea. This second situation is often called *fatty acid diarrhea.* Cholestyramine may not be effective (and may even exacerbate the diarrhea by further depleting the intraduodenal bile-acid concentration); however, a low-fat diet to reduce fatty acid entry into the colon can be effective. Two clinical features—the length of the ileal section removed and the degree of steatorrhea—can predict whether an individual patient will respond to cholestyramine. Unfortunately, these predictors are imperfect, and a therapeutic trial of cholestyramine is often necessary to establish whether an individual patient will benefit from cholestyramine. Table 318-2 contrasts the characteristics of bile-acid diarrhea (small ileal dysfunction) and fatty acid diarrhea (large ileal dysfunction).

TABLE 318-2 Comparison of Bile Acid and Fatty Acid Diarrhea

	BILE-ACID DIARRHEA	FATTY ACID DIARRHEA
Extent of ileal disease	Limited	Extensive
Ileal bile-acid absorption	Reduced	Reduced
Fecal bile-acid excretion	Increased	Increased
Fecal bile-acid loss compensated by hepatic synthesis	Yes	No
Bile-acid pool size	Normal	Reduced
Intraduodenal (bile acid)	Normal	Reduced
Steatorrhea	None or mild	>20 g
Response to cholestyramine	Yes	No
Response to low-fat diet	No	Yes

Bile-acid diarrhea can also occur in the absence of ileal inflammation and/or resection and is characterized by an abnormal ^{75}SeHCAT retention study and reduced ileal release of FGF19, a negative regulator of bile-acid synthesis, with a consequent increase in bile-acid synthesis and secretion that exceeds ileal bile-acid absorption. The diarrhea in these patients also responds to cholestyramine.

LIPIDS

Steatorrhea is caused by one or more defects in the digestion and absorption of dietary fat. The average intake of dietary fat in the United States is ~120–150 g/d, and fat absorption is linear to dietary fat intake. The total load of fat presented to the small intestine is considerably greater, as substantial amounts of lipid are secreted in bile each day (see "Enterohepatic Circulation of Bile Acids," above). Three types of fatty acids compose fats: long-chain fatty acids (LCFAs), medium-chain fatty acids (MCFAs), and short-chain fatty acids (SCFAs) **(Table 318-3)**. Dietary fat is exclusively composed of long-chain triglycerides (LCTs)—that is, glycerol that is bound via ester linkages to three LCFAs. While the majority of dietary LCFAs have carbon chain lengths of 16 or 18, all fatty acids of carbon chain length >12 are metabolized in the same manner; saturated and unsaturated fatty acids are handled identically.

Assimilation of dietary lipid requires three integrated processes: (1) an intraluminal, or digestive, phase; (2) a mucosal, or absorptive, phase; and (3) a delivery, or postabsorptive, phase. An abnormality at any site involved in these processes can cause steatorrhea **(Table 318-4)**. Therefore, it is essential that any patient with steatorrhea be evaluated

TABLE 318-3 Comparison of Different Types of Fatty Acids

	LONG-CHAIN	MEDIUM-CHAIN	SHORT-CHAIN
Carbon chain length	>12	8–12	<8
Present in diet	In large amounts	In small amounts	No
Origin	In diet as triglycerides	Only in small amounts in diet as triglycerides	Bacterial degradation in colon of nonabsorbed carbohydrate to fatty acids
Primary site of absorption	Small intestine	Small intestine	Colon
Requires pancreatic lipolysis	Yes	No	No
Requires micelle formation	Yes	No	No
Present in stool	Minimal	No	Substantial

TABLE 318-4 Defects in Lipid Digestion and Absorption in Steatorrhea

PHASE, PROCESS	PATHOPHYSIOLOGIC DEFECT	DISEASE EXAMPLE
Digestive		
Lipolysis formation	Decreased lipase secretion	Chronic pancreatitis
Micelle formation	Decreased intraduodenal bile acids	See Table 318-1
Absorptive		
Mucosal uptake and re-esterification	Mucosal dysfunction	Celiac disease
Postabsorptive		
Chylomicron formation	Absent β-lipoproteins	Abetalipoproteinemia
Delivery from intestine	Abnormal lymphatics	Intestinal lymphangiectasia

to identify the specific physiologic defect in overall lipid digestion/absorption, as therapy will be determined by the specific etiology.

The digestive phase has two components: *lipolysis* and *micelle formation.* Although dietary lipid is in the form of LCTs, the intestinal mucosa does not absorb triglycerides; they must first be hydrolyzed **(Fig. 318-2)**. The initial step in lipid digestion is the formation of emulsions of finely dispersed lipid, which is accomplished by mastication and gastric contractions. Lipolysis, the hydrolysis of triglycerides to free fatty acids, monoglycerides, and glycerol by lipase, is initiated in the stomach by lingual and gastric lipases that have a pH optimum of 4.5–6.0. About 20–30% of total lipolysis occurs in the stomach. Lipolysis is completed in the duodenum and jejunum by pancreatic lipase, which is inactivated by a pH <7.0. Pancreatic lipolysis is greatly enhanced by the presence of a second pancreatic enzyme, colipase, which facilitates the movement of lipase to the triglyceride.

Impaired lipolysis can lead to steatorrhea and can occur in the presence of pancreatic insufficiency due to chronic pancreatitis in adults or cystic fibrosis in children and adolescents. Normal lipolysis can be maintained by ~5% of maximal pancreatic lipase secretion; thus, steatorrhea is a late manifestation of these disorders. A reduction in intraduodenal pH can also result in altered lipolysis, as pancreatic lipase is inactivated at pH <7. Thus, ~15% of patients who have gastrinoma **(Chap. 317)**, with substantial increases in gastric acid secretion from ectopic production of gastrin (usually from an islet cell adenoma), have diarrhea, and some have steatorrhea believed to be secondary to acid inactivation of pancreatic lipase. Similarly, patients who have chronic pancreatitis (with reduced lipase secretion) often have a decrease in pancreatic bicarbonate secretion, which will also result in a lowering of intraduodenal pH and inactivation of endogenous pancreatic lipase or of therapeutically administered lipase.

Overlying the microvillus membrane of the small intestine is the so-called unstirred water layer, a relatively stagnant aqueous phase that must be traversed by the products of lipolysis that are primarily water insoluble. Water-soluble mixed micelles provide a mechanism by which the water-insoluble products of lipolysis can reach the luminal plasma membrane of villous epithelial cells—the site for lipid absorption. Mixed micelles are molecular aggregates composed of fatty acids, monoglycerides, phospholipids, cholesterol, and conjugated bile acids. These mixed micelles are formed when the concentration of conjugated bile acids is greater than its CMC, which differs among the several bile acids present in the small-intestinal lumen. Conjugated bile acids, synthesized in the liver and excreted into the duodenum in bile, are regulated by the enterohepatic circulation (see above). Steatorrhea can result from impaired movement of fatty acids across the unstirred aqueous fluid layer in two situations: (1) an increase in the relative thickness of the unstirred water layer that occurs in bacterial overgrowth syndromes (see below) secondary to functional stasis (e.g., scleroderma); and (2) a decrease in the *duodenal* concentration of conjugated bile acids below the CMC, resulting in impaired micelle formation. Thus, steatorrhea can be caused by one or more defects in the enterohepatic circulation of bile acids.

Uptake and re-esterification constitute the *absorptive phase* of lipid digestion/absorption. Although passive diffusion has been thought to be responsible, a carrier-mediated process may mediate fatty acid and monoglyceride uptake. Regardless of the uptake process, fatty acids and monoglycerides are re-esterified by a series of enzymatic steps in the endoplasmic reticulum to form triglycerides, in which lipid exits from the intestinal epithelial cell. Impaired lipid absorption as a result of mucosal inflammation (e.g., celiac disease) and/or intestinal resection can also lead to steatorrhea.

The re-esterified triglycerides require the formation of *chylomicrons* to permit their exit from the small-intestinal epithelial cell and their delivery to the liver via the lymphatics. Chylomicrons are composed of β-lipoprotein and contain triglycerides, cholesterol, cholesterol esters, and phospholipids and enter the lymphatics, not the portal vein. Defects in the *postabsorptive phase* of lipid digestion/absorption can also result in steatorrhea, but these disorders are uncommon. Abetalipoproteinemia, or acanthocytosis, is a rare disorder of impaired synthesis of β-lipoprotein associated with abnormal erythrocytes (acanthocytes), neurologic problems, and steatorrhea **(Chap. 400)**. Lipolysis, micelle formation, and lipid uptake are all normal in patients with abetalipoproteinemia, but the re-esterified triglyceride cannot exit the epithelial cell because of the failure to produce chylomicrons. Small-intestinal biopsy samples obtained from these rare patients in the postprandial state reveal lipid-laden small-intestinal epithelial cells that become perfectly normal in appearance after a 72- to 96-h fast. Similarly, abnormalities of intestinal lymphatics (e.g., intestinal lymphangiectasia) may also be associated with steatorrhea as well as protein loss (see below). Steatorrhea can result from defects at any of the several steps in lipid digestion/absorption.

The mechanism of lipid digestion/absorption outlined above is limited to *dietary* lipid, which is almost exclusively in the form of LCTs (Table 318-3). Medium-chain triglycerides (MCTs), composed of fatty acids with carbon chain lengths of 8–12, are present in large amounts in coconut oil and are used as a nutritional supplement. MCTs can be digested and absorbed by a pathway different from that involved in LCT digestion and absorption; at one time, MCTs held promise as an important treatment for steatorrhea of almost all etiologies. Unfortunately, they have been less therapeutically effective than expected because, for reasons that are not completely understood, their use often is not associated with an increase in body weight.

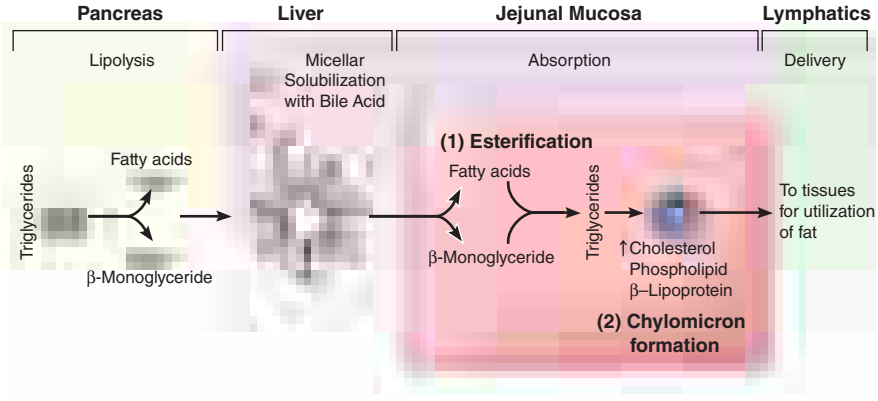


FIGURE 318-2 Schematic representation of lipid digestion and absorption. Dietary lipid is in the form of long-chain triglycerides. The overall process can be divided into (1) a digestive phase that includes both lipolysis and micelle formation requiring pancreatic lipase and conjugated bile acids, respectively, in the duodenum; (2) an absorptive phase for mucosal uptake and re-esterification; and (3) a postabsorptive phase that includes chylomicron formation and exit from the intestinal epithelial cell via lymphatics. *(Courtesy of John M. Dietschy, MD; with permission.)*

In contrast to LCTs, MCTs do not require pancreatic lipolysis as they can be absorbed intact by the intestinal epithelial cell. Further, micelle formation is not necessary for the absorption of MCTs (or MCFAs, if hydrolyzed by pancreatic lipase). MCTs are absorbed more efficiently than LCTs for the following reasons: (1) The rate of absorption is greater for MCTs than for LCFAs; (2) after absorption, MCFAs are not re-esterified; (3) after absorption, MCTs are hydrolyzed to MCFAs; (4) MCTs do not require chylomicron formation to exit intestinal epithelial cells; and (5) the route of MCT exit is via the portal vein and not via lymphatics. Thus, the absorption of MCTs is greater than that of LCTs in pancreatic insufficiency, conditions with reduced intraduodenal bile-acid concentrations, small-intestinal mucosal disease, abetalipoproteinemia, and intestinal lymphangiectasia.

SCFAs are not dietary lipids but are synthesized by colonic bacterial enzymes from nonabsorbed carbohydrate and are the anions present at the highest concentration in stool (80–130 m*M*). The SCFAs in stool are primarily acetate, propionate, and butyrate, whose carbon chain lengths are 2, 3, and 4, respectively. Butyrate is the primary nutrient for colonic epithelial cells, and its deficiency can be associated with one or more colitides. SCFAs conserve calories and carbohydrate: carbohydrates that are not completely absorbed in the small intestine will not be absorbed in the large intestine because of the absence of both disaccharidases and SGLT1, the transport protein that mediates monosaccharide absorption. In contrast, SCFAs are rapidly absorbed and stimulate colonic NaCl and fluid absorption. Most antibiotic-associated diarrhea not caused by *Clostridium difficile* is due to antibiotic suppression of the colonic microbiota, with a resulting decrease in SCFA production. As *C. difficile* accounts for only ~15–20% of all antibiotic-associated diarrhea, a relative decrease in colonic production of SCFA is likely the cause of most antibiotic-associated diarrhea.

The clinical manifestations of steatorrhea are a consequence both of the underlying disorder responsible for its development and of steatorrhea per se. Depending on the degree of steatorrhea and the level of dietary intake, significant fat malabsorption may lead to weight loss. Steatorrhea per se can be responsible for diarrhea; if the primary cause of the steatorrhea has not been identified, a low-fat diet can often ameliorate the diarrhea by decreasing fecal fat excretion. Steatorrhea is commonly associated with fat-soluble vitamin deficiency, which requires replacement with water-soluble preparations of these vitamins.

Disorders of absorption may also be associated with malabsorption of other dietary nutrients—most often carbohydrates—with or without a decrease in dietary lipid digestion and absorption. Therefore, knowledge of the mechanisms of digestion and absorption of carbohydrates, proteins, and other minerals and vitamins is useful in the evaluation of patients with altered intestinal nutrient absorption.

CARBOHYDRATES

Carbohydrates in the diet are present in the form of starch, disaccharides (sucrose and lactose), and glucose. Carbohydrates are absorbed only in the small intestine and only in the form of monosaccharides. Therefore, before their absorption, starch and disaccharides must first be digested by pancreatic amylase and intestinal brush border disaccharidases to monosaccharides. Monosaccharide absorption occurs by a Na-dependent process mediated by the brush border transport protein SGLT1.

Lactose malabsorption is the only clinically important disorder of carbohydrate absorption. Lactose, the disaccharide present in milk, requires digestion by brush border lactase to its two constituent monosaccharides, glucose and galactose. Lactase is present in almost all species in the postnatal period but then disappears throughout the animal kingdom, except in humans. Lactase activity persists in many individuals throughout life. Two different types of lactase deficiency exist—primary and secondary. In *primary lactase deficiency*, a genetically determined decrease or absence of lactase is noted, while all other aspects of both intestinal absorption and brush border enzymes are normal. In a number of nonwhite groups, primary lactase deficiency is common in adulthood. In fact, Northern European and North American whites are the only groups to maintain small-intestinal lactase activity throughout adult life. Table 318-5 presents the incidence of primary lactase deficiency in several ethnic groups. Lactase persistence in adults is an abnormality due to a defect in the regulation of its maturation. In contrast, *secondary lactase deficiency* occurs in association with small-intestinal mucosal disease, with abnormalities in both structure and function of other brush border enzymes and transport processes. Secondary lactase deficiency is often seen in celiac disease.

TABLE 318-5 Primary Lactase Deficiency in Adult Ethnic Groups

ETHNIC GROUP	PREVALENCE OF LACTASE DEFICIENCY, %
Northern European	5–15
Mediterranean	60–85
African black	85–100
African American	45–80
American white	10–25
Native American	50–95
Mexican American	40–75
Asian	90–100

Source: From FJ Simoons: Am J Dig Dis 23:963, 1978.

As lactose digestion is rate-limiting compared to glucose/galactose absorption, lactase deficiency is associated with significant lactose malabsorption. Some individuals with lactose malabsorption develop symptoms such as diarrhea, abdominal pain, cramps, and/or flatus. Most individuals with primary lactase deficiency do not have symptoms. Since lactose intolerance may be associated with symptoms suggestive of irritable bowel syndrome, persistence of such symptoms in an individual who exhibits lactose intolerance while on a strict lactose-free diet suggests that the person's symptoms were related to irritable bowel syndrome.

The development of symptoms of lactose intolerance is related to several factors:

1. *Amount of lactose in the diet.*
2. *Rate of gastric emptying.* Symptoms are more likely when gastric emptying is rapid than when it is slower. Therefore, skim milk is more likely to be associated with symptoms of lactose intolerance than whole milk, as the rate of gastric emptying after skim milk intake is more rapid. Similarly, diarrhea following subtotal gastrectomy is often a result of lactose intolerance, as gastric emptying is accelerated in patients with a gastrojejunostomy.
3. *Small-intestinal transit time.* Although the small and large intestines both contribute to the development of symptoms, many symptoms of lactase deficiency are related to the interaction of colonic bacteria and nonabsorbed lactose. More rapid small-intestinal transit makes symptoms more likely.
4. *Colonic compensation by production of SCFAs from nonabsorbed lactose.* Reduced levels of colonic microflora, which can follow antibiotic use, are associated with increased symptoms after lactose ingestion, especially in a lactase-deficient individual.

Glucose-galactose or monosaccharide malabsorption may also be associated with diarrhea and is due to a congenital absence of SGLT1. Diarrhea develops when individuals with this disorder ingest carbohydrates that contain actively transported monosaccharides (e.g., glucose, galactose) but not when they ingest monosaccharides that are not actively transported (e.g., fructose). Fructose is absorbed by the brush border transport protein GLUT 5, a facilitated diffusion process that is not Na-dependent and is distinct from SGLT1. In contrast, some individuals develop diarrhea as a result of the consumption of large quantities of sorbitol, a sugar used in diabetic candy; sorbitol is only minimally absorbed because of the absence of an intestinal absorptive transport mechanism for this sugar.

PROTEINS

Protein is present in food almost exclusively as polypeptides and requires extensive hydrolysis to di- and tripeptides and amino acids before absorption. Proteolysis occurs in both the stomach and the small intestine; it is mediated by pepsin, which is secreted as pepsinogen

by gastric chief cells, and by trypsinogen and other peptidases from pancreatic acinar cells. The proenzymes pepsinogen and trypsinogen must be activated to pepsin (by pepsin at a pH <5) and trypsin (by the intestinal brush border enzyme enterokinase and subsequently by trypsin), respectively. Proteins are absorbed by separate transport systems for di- and tripeptides and for different types of amino acids—for example neutral and dibasic. Alterations in either protein or amino acid digestion and absorption are rarely observed clinically, even in the presence of extensive small-intestinal mucosal inflammation. However, three rare genetic disorders involve protein digestion/absorption: (1) *Enterokinase deficiency* is due to an absence of the brush border enzyme that converts the proenzyme trypsinogen to trypsin and is associated with diarrhea, growth retardation, and hypoproteinemia; (2) *Hartnup's syndrome*, a defect in neutral amino acid transport, is characterized by a pellagra-like rash and neuropsychiatric symptoms; and (3) *cystinuria*, a defect in dibasic amino acid transport, is associated with renal calculi and chronic pancreatitis.

APPROACH TO THE PATIENT

Malabsorption

The clues provided by the history, symptoms, and initial preliminary observations will serve to limit extensive, ill-focused, and expensive laboratory and imaging studies. For example, a clinician evaluating a patient who has symptoms suggestive of malabsorption and has recently undergone extensive small-intestinal resection for mesenteric ischemia should direct the initial assessment almost exclusively to defining whether a short-bowel syndrome might explain the entire clinical picture. Similarly, the development of a pattern of bowel movements suggestive of steatorrhea in a patient with long-standing alcohol abuse and chronic pancreatitis should prompt an assessment of pancreatic exocrine function.

The classic picture of malabsorption is rarely seen today in most parts of the United States. As a consequence, diseases with malabsorption must be suspected in individuals who have less severe symptoms and signs and subtle evidence of the altered absorption of only a *single* nutrient rather than obvious evidence of the malabsorption of multiple nutrients.

Although diarrhea can be caused by changes in fluid and electrolyte movement in either the small or the large intestine, dietary nutrients are absorbed almost exclusively in the small intestine. Therefore, the demonstration of diminished absorption of a dietary nutrient provides unequivocal evidence for small-intestinal disease, although colonic dysfunction may also be present (e.g., Crohn's disease may involve both the small and large intestines). Dietary nutrient absorption may be segmental or diffuse along the small intestine and is site specific. Thus, for example, calcium, iron, and folic acid are exclusively absorbed by active-transport processes in the proximal small intestine, especially the duodenum; in contrast, the active-transport mechanisms for both cobalamin and bile acids are operative only in the ileum. Therefore, in an individual who years previously has had an intestinal resection, the details of which are not presently available, a presentation with evidence of calcium, folic acid, and/or iron malabsorption but without cobalamin deficiency makes it likely that the duodenum and proximal jejunum, but not the ileum, were resected.

Some nutrients—for example glucose, amino acids, and lipids—are absorbed throughout the small intestine, although their rate of absorption is greater in the proximal than in the distal segments. However, after segmental resection of the small intestine, the remaining segments undergo both morphologic and functional "adaptation" to enhance absorption. Such adaptation is secondary to the presence of luminal nutrients and hormonal stimuli and may not be complete in humans for several months after resection. Adaptation is critical for the survival of individuals who have undergone massive resection of the small intestine and/or colon.

Establishing the presence of steatorrhea and identifying its specific cause are often quite difficult. The "gold standard" remains a timed, quantitative stool-fat determination. From a practical standpoint, stool collections are invariably difficult and often incomplete, as nobody wants to handle stool. A qualitative test—Sudan III staining—has long been available to document an increase in stool fat. This test is rapid and inexpensive but, as a qualitative test, does not establish the degree of fat malabsorption and is best used as a preliminary screening study. Many of the blood, breath, and isotopic tests that have been developed (1) do not directly measure fat absorption; (2) exhibit excellent sensitivity when steatorrhea is obvious and severe but poor sensitivity when steatorrhea is mild (e.g., assays for stool chymotrypsin and elastase, which can potentially distinguish pancreatic from nonpancreatic etiologies of steatorrhea); or (3) have not survived the transition from the research laboratory to commercial application.

Nevertheless, routine laboratory studies (i.e., complete blood count, prothrombin time, serum protein determination, alkaline phosphatase) may suggest dietary nutrient depletion, especially deficiencies of iron, folate, cobalamin, and vitamins D and K. Additional studies include measurement of serum carotene, cholesterol, albumin, iron, folate, and cobalamin levels. The serum carotene level can also be reduced if the patient's dietary intake of leafy vegetables is poor.

If steatorrhea and/or altered absorption of other nutrients are suspected, history, clinical observations, and laboratory testing can help detect deficiency of a nutrient, especially of a fat-soluble vitamin (A, D, E, or K). Thus, evidence of metabolic bone disease with elevated alkaline phosphatase concentrations and/or reduced serum calcium levels suggests vitamin D malabsorption. A deficiency of vitamin K is suggested by an elevated prothrombin time in an individual without liver disease who is not taking anticoagulants. Macrocytic anemia leads to an evaluation for possible cobalamin or folic acid malabsorption. Iron-deficiency anemia in the absence of occult bleeding from the gastrointestinal tract in either a male patient or a nonmenstruating female patient requires an evaluation for iron malabsorption and the exclusion of celiac disease, as iron is absorbed exclusively in the proximal small intestine.

At times, however, a timed (72-h) quantitative stool collection, preferably while the patient is on a defined diet, must be undertaken in order to determine stool fat content and establish the diagnosis of steatorrhea. The presence of steatorrhea then requires further assessment to identify the pathophysiologic process(es) responsible for the defect in dietary lipid digestion/absorption (Table 318-4). Other studies include the D-xylose test, duodenal mucosal biopsy, small-intestinal radiologic examination, and tests of pancreatic exocrine function.

URINARY D-XYLOSE TEST

The urinary D-xylose test for carbohydrate absorption provides an assessment of proximal small-intestinal mucosal function. D-Xylose, a pentose, is absorbed almost exclusively in the proximal small intestine. The D-xylose test is usually performed by administering 25 g of D-xylose and collecting urine for 5 h. An abnormal test (excretion of <4.5 g) primarily reflects duodenal/jejunal mucosal disease. The D-xylose test can also be abnormal in patients with blind loop syndrome (as a consequence primarily of an abnormal intestinal mucosa) and, as a false-positive study, in patients with large collections of fluid in a third space (i.e., ascites, pleural fluid). The ease of obtaining a mucosal biopsy of the small intestine by endoscopy and the false-negative rate of the D-xylose test have led to its diminished use. When small-intestinal mucosal disease is suspected, a small-intestinal mucosal biopsy should be performed.

RADIOLOGIC EXAMINATION

Radiologic examination of the small intestine using barium contrast (small-bowel series or study) can provide important information in the evaluation of the patient with presumed or suspected malabsorption. This study is most often performed in conjunction with an examination of the esophagus, stomach, and duodenal bulb. Because insufficient barium is given to the patient to permit an

adequate examination of the small-intestinal mucosa, especially in the ileum, many gastrointestinal radiologists alter the procedure by performing either a small-bowel series in which a large amount of barium is given by mouth, without concurrent examination of the esophagus and stomach, or an enteroclysis study in which a large amount of barium is introduced into the duodenum via a fluoroscopically placed tube. In addition, many of the diagnostic features initially described by radiologists to denote the presence of small-intestinal disease (e.g., flocculation, segmentation) are rarely seen with current barium suspensions. Nonetheless, in skilled hands, barium contrast examination of the small intestine can yield important information. For example, with extensive mucosal disease, intestinal dilation can be seen as a dilution of barium from increased intestinal fluid secretion **(Fig. 318-3)**. A normal barium contrast study does *not* exclude the possibility of small-intestinal disease. However, a small-bowel series remains useful in the search for anatomic abnormalities, such as strictures and fistulas (as in Crohn's disease) or blind loop syndrome (e.g., multiple jejunal diverticula) and to define the extent of a previous surgical resection. Other imaging studies that assess the integrity of small-intestinal morphology are CT enterography and magnetic resonance enterography. Capsule endoscopy and double-balloon enteroscopy are other useful aids in the diagnostic assessment of small-intestinal pathology and most often are used to identify a small-intestinal bleeding site.

BIOPSY OF SMALL-INTESTINAL MUCOSA

A small-intestinal mucosal biopsy is essential in the evaluation of a patient with documented steatorrhea or chronic diarrhea (i.e., that lasting >3 weeks) **(Chap. 42)**. The ready availability of endoscopic equipment to examine the stomach and duodenum has led to its almost uniform use as the preferred method of obtaining histologic material from the proximal small-intestinal mucosa. The primary indications for a small-intestinal biopsy are evaluation of a patient (1) either with documented or suspected steatorrhea or with chronic diarrhea, and (2) with diffuse or focal abnormalities of the small intestine defined on a small-intestinal series. Lesions seen on small-bowel biopsy can be classified into three categories **(Table 318-6)**:

1. *Diffuse, specific lesions.* Relatively few diseases associated with altered nutrient absorption have specific histopathologic abnormalities on small-intestinal mucosal biopsy, and these diseases are uncommon. *Whipple's disease* is characterized by the presence of periodic acid–Schiff (PAS)–positive macrophages in the lamina propria; the bacilli that are also present may require electron microscopic examination for identification **(Fig. 318-4)**. *Abetalipoproteinemia* is characterized by a normal mucosal appearance except for the presence of mucosal absorptive cells that contain lipid postprandially and disappear after a prolonged period of either fat-free intake or fasting. *Immune globulin deficiency* is associated with a variety of histopathologic findings on small-intestinal mucosal biopsy. The characteristic feature is the absence of or substantial reduction in the number of plasma cells in the lamina propria; the mucosal architecture may be either perfectly normal or flat (i.e., villous atrophy). As patients with immune globulin deficiency are often infected with *Giardia lamblia, Giardia* trophozoites may also be seen in the biopsy.
2. *Patchy, specific lesions.* Several diseases feature an abnormal small-intestinal mucosa with a patchy distribution. As a result, biopsy samples obtained randomly or in the absence of endoscopically visualized abnormalities may not reveal diagnostic features. Intestinal *lymphoma* can at times be diagnosed on mucosal biopsy by the identification of malignant lymphoma

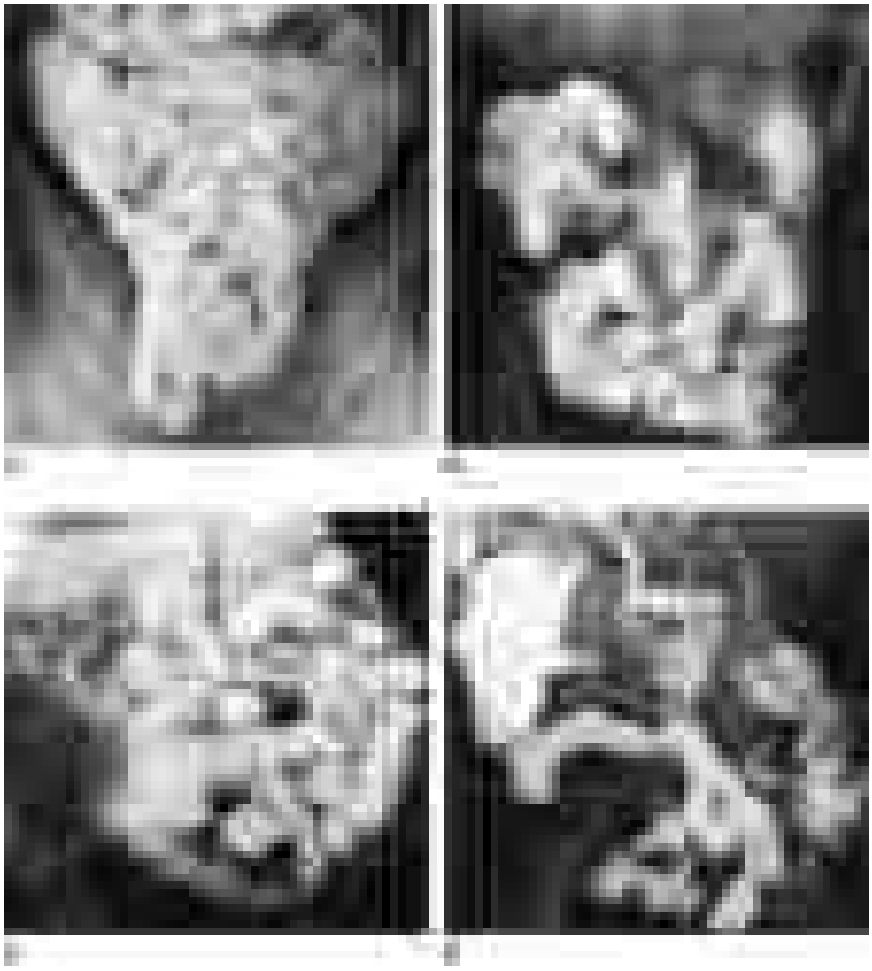


FIGURE 318-3 Barium contrast small-intestinal radiologic examinations. *A.* Normal individual. ***B.*** Celiac disease. ***C.*** Jejunal diverticulosis. ***D.*** Crohn's disease. *(Courtesy of Morton Burrell, MD, Yale University; with permission.)*

TABLE 318-6 Diseases That Can Be Diagnosed by Small-Intestinal Mucosal Biopsies

LESIONS	PATHOLOGIC FINDINGS
Diffuse, Specific	
Whipple's disease	Lamina propria includes macrophages containing material positive on periodic acid–Schiff staining
Agammaglobulinemia	No plasma cells; either normal or absent villi ("flat mucosa")
Abetalipoproteinemia	Normal villi; epithelial cells vacuolated with fat postprandially
Patchy, Specific	
Intestinal lymphoma	Malignant cells in lamina propria and submucosa
Intestinal lymphangiectasia	Dilated lymphatics; clubbed villi
Eosinophilic gastroenteritis	Eosinophil infiltration of lamina propria and mucosa
Amyloidosis	Amyloid deposits
Crohn's disease	Noncaseating granulomas
Infection by one or more microorganisms (see text)	Specific organisms
Mastocytosis	Mast cell infiltration of lamina propria
Diffuse, Nonspecific	
Celiac disease	Short or absent villi; mononuclear infiltrate; epithelial cell damage; hypertrophy of crypts
Tropical sprue	Similar to celiac disease
Bacterial overgrowth	Patchy damage to villi; lymphocyte infiltration
Folate deficiency	Short villi; decreased mitosis in crypts; megalocytosis
Vitamin B_{12} deficiency	Similar to folate deficiency
Radiation enteritis	Similar to folate deficiency
Zollinger-Ellison syndrome	Mucosal ulceration and erosion from acid
Protein-calorie malnutrition	Villous atrophy; secondary bacterial overgrowth
Drug-induced enteritis	Variable histology

cells in the lamina propria and submucosa **(Chap. 104)**. Dilated lymphatics in the submucosa and sometimes in the lamina propria indicate *lymphangiectasia* associated with hypoproteinemia secondary to protein loss into the intestine. *Eosinophilic gastroenteritis* comprises a heterogeneous group of disorders with a spectrum of presentations and symptoms, with an eosinophilic infiltrate of the lamina propria, and with or without peripheral eosinophilia. The patchy nature of the infiltrate and its presence in the submucosa often lead to an absence of histopathologic findings on mucosal biopsy. As the involvement of the duodenum in *Crohn's disease* is also submucosal and not necessarily continuous, mucosal biopsies are not the most direct approach to the diagnosis of duodenal Crohn's disease **(Chap. 319)**. Amyloid deposition can be identified by Congo Red staining in some patients with *amyloidosis* involving the duodenum **(Chap. 108)**.

3. *Diffuse, nonspecific lesions*. *Celiac disease* presents with a characteristic mucosal appearance on duodenal/proximal jejunal mucosal biopsy that is *not* diagnostic of the disease. The diagnosis of celiac disease is established by clinical, histologic, and immunologic responses to a gluten-free diet. *Tropical sprue* (see below) is associated with histologic findings similar to those of celiac disease after a tropical or subtropical exposure but does not respond to gluten restriction; most often symptoms improve with antibiotics and folate administration.

Several microorganisms can be identified in small-intestinal biopsy samples, establishing a correct diagnosis. At times, the biopsy is performed specifically to diagnose infection (e.g., Whipple's disease or giardiasis). In most other instances, the infection is detected incidentally during the workup for diarrhea or other abdominal symptoms. Many of these infections occur in immunocompromised patients with diarrhea; the etiologic agents include *Cryptosporidium*, *Isospora belli*, microsporidia, *Cyclospora*, *Toxoplasma*, cytomegalovirus, adenovirus, *Mycobacterium avium-intracellulare*, and *G. lamblia*. In immunocompromised patients, when *Candida*, *Aspergillus*, *Cryptococcus*, or *Histoplasma* organisms are seen on duodenal biopsy, their presence generally reflects systemic infection. Apart from Whipple's disease and infections in the immunocompromised host, small-bowel biopsy is seldom used as the primary mode of diagnosis of infection. Even giardiasis is more easily diagnosed by stool antigen studies and/or duodenal aspiration than by duodenal biopsy.

Patients with steatorrhea require assessment of *pancreatic exocrine function*, which is often abnormal in chronic pancreatitis. The secretin test that collects pancreatic secretions by duodenal intubation following intravenous administration of secretin is the only test that directly measures pancreatic exocrine function but is available only at a few specialized centers. Endoscopic approaches (endoscopic retrograde cholangiopancreatography, endoscopic ultrasound) provide an excellent assessment of pancreatic duct anatomy but do *not* assess exocrine function **(Chap. 340)**.

Table 318-7 summarizes the results of the D-xylose test, the Schilling test, and small-intestinal mucosal biopsy in patients with steatorrhea of various etiologies.

SPECIFIC DISEASE ENTITIES

CELIAC DISEASE

Celiac disease is a common cause of malabsorption of one or more nutrients. Although celiac disease was originally considered largely a disease of white individuals, especially persons of European descent, recent observations have established that it is a common disease with protean manifestations, a worldwide distribution, and an estimated incidence in the United States that is as high as 1 in 113 people. Its incidence has increased over the past 50 years. Celiac disease has had several other names, including nontropical sprue, celiac sprue, adult celiac disease, and gluten-sensitive enteropathy. The etiology of celiac disease is not known, but environmental, immunologic, and genetic factors are important. Celiac disease is considered an "iceberg" disease. A small number of individuals have classic symptoms and manifestations related to nutrient malabsorption along with a varied natural history; the onset of symptoms can occur at all points from the first year of life through the eighth decade. A much larger number of individuals have "atypical celiac disease," with manifestations that are not obviously related to intestinal malabsorption (e.g., anemia, osteopenia, infertility, and neurologic symptoms). Finally, an even larger number of persons have "silent celiac disease"; they are essentially asymptomatic despite abnormal small-intestinal histopathology and serologies (see below).

The hallmark of celiac disease is an abnormal small-intestinal biopsy (Fig. 318-4) and the response of the condition (including symptoms and histologic changes on small-intestinal biopsy) to the elimination of gluten from the diet. The histologic changes have a proximal-to-distal intestinal distribution of severity, which probably reflects the exposure of the intestinal mucosa to varied amounts of dietary gluten. The symptoms do not necessarily correlate with histologic changes, especially as many newly diagnosed patients with celiac disease may be asymptomatic or only minimally symptomatic (often with no gastrointestinal symptoms).

The symptoms of celiac disease may appear with the introduction of cereals into an infant's diet, although spontaneous remissions often occur during the second decade of life that may be either permanent or followed by the reappearance of symptoms over several years. Alternatively, the symptoms of celiac disease may first become evident at almost any age throughout adulthood. In many patients, frequent spontaneous remissions and exacerbations occur. The symptoms range from significant malabsorption of multiple nutrients, with diarrhea, steatorrhea, weight loss, and the consequences of nutrient depletion (i.e., anemia and metabolic bone disease), to the total absence of gastrointestinal symptoms despite evidence of the depletion of a single

FIGURE 318-4 Small-intestinal mucosal biopsies. ***A.*** Normal individual. ***B.*** Untreated celiac disease. ***C.*** Treated celiac disease. ***D.*** Intestinal lymphangiectasia. ***E.*** Whipple's disease. ***F.*** Lymphoma. ***G.*** Giardiasis. *(Courtesy of Marie Robert, MD, Yale University; with permission.)*

nutrient (e.g., iron or folate deficiency, osteomalacia, edema from protein loss). Asymptomatic relatives of patients with celiac disease have been identified as having this disease either by small-intestinal biopsy or by serologic studies (e.g., antiendomysial antibodies, tissue transglutaminase [tTG], deamidated gliadin peptide). The availability of these "celiac serologies" has led to a substantial increase in the frequency of diagnosis of celiac disease, and the diagnosis is now being made primarily in patients without "classic" symptoms but with atypical and subclinical presentations.

Etiology The etiology of celiac disease is not known, but environmental, immunologic, and genetic factors all appear to contribute to the disease. One *environmental* factor is the clear association of the disease with gliadin, a component of gluten that is present in wheat, barley, and rye. In addition to the role of gluten restriction in treatment, the instillation of gluten into the normal-appearing rectum and the distal ileum of patients with celiac disease results in morphologic changes within hours.

An *immunologic* component in the pathogenesis of celiac disease is critical and involves both adaptive and innate immune responses. Serum antibodies—IgA antigliadin, antiendomysial, and anti-tTG antibodies—are present, but it is not known whether such antibodies are primary or secondary to the tissue damage. The presence of antiendomysial antibody is 90–95% sensitive and 90–95% specific; the antigen recognized by antiendomysial antibody is tTG, which deaminates gliadin, which is presented to HLA-DQ2 or HLA-DQ8 (see below). Antibody studies are frequently used to identify patients with

TABLE 318-7 Results of Diagnostic Studies in Steatorrhea of Various Etiologies

	D-XYLOSE TEST	SCHILLING TEST	DUODENAL MUCOSAL BIOPSY
Chronic pancreatitis	Normal	50% abnormal; if abnormal, normal with pancreatic enzyme treatment	Normal
Bacterial overgrowth syndromes	Normal or only modestly abnormal	Often abnormal; if abnormal, normal after antibiotic treatment	Usually normal
Ileal disease	Normal	Abnormal	Normal
Celiac disease	Decreased	Normal	Abnormal: probably "flat"
Intestinal lymphangiectasia	Normal	Normal	Abnormal: "dilated lymphatics"

celiac disease; patients with these antibodies should undergo duodenal biopsy. This autoantibody has not been linked to a pathogenetic mechanism (or mechanisms) responsible for celiac disease. Nonetheless, this antibody is useful in establishing the true prevalence of celiac disease in the general population. A 4-week course of treatment with prednisolone induces a remission in a patient with celiac disease who continues to eat gluten and converts the "flat" abnormal duodenal biopsy to a more normal-appearing one. In addition, gliadin peptides interact with gliadin-specific T cells that mediate tissue injury and induce the release of one or more cytokines (e.g., interferon γ) that cause tissue injury.

Genetic factor(s) are also involved in celiac disease. The incidence of symptomatic celiac disease varies widely in different population groups (high among whites, low among blacks and Asians) and is 10% among first-degree relatives of celiac disease patients. However, serologic studies provide clear evidence that celiac disease is present worldwide. Furthermore, all patients with celiac disease express the HLA-DQ2 or HLA-DQ8 allele, although only a minority of people expressing DQ2/DQ8 have celiac disease. Absence of DQ2/DQ8 excludes the diagnosis of celiac disease.

Diagnosis A small-intestinal biopsy is required to establish a diagnosis of celiac disease (Fig. 318-4). A biopsy should be performed when patients have symptoms and laboratory findings suggestive of nutrient malabsorption and/or deficiency as well as a positive tTG antibody test. Since the presentation of celiac disease is often subtle, without overt evidence of malabsorption or nutrient deficiency, a relatively low threshold for biopsy performance is important. It is more prudent to perform a biopsy than another test of intestinal absorption that can never completely exclude or establish this diagnosis.

The diagnosis of celiac disease requires the detection of characteristic histologic changes on small-intestinal biopsy together with a prompt clinical and histologic response after the institution of a gluten-free diet. If IgA antiendomysial or tTG antibodies have been detected in serologic studies, they too should disappear after a gluten-free diet is started. With the increase in the number of patients diagnosed with celiac disease (mostly by serologic studies), the spectrum of histologic changes seen on duodenal biopsy has increased and includes findings that are not as severe as the classic changes shown in Fig. 318-4. The classic changes seen on duodenal/jejunal biopsy are restricted to the mucosa and include (1) an increase in the number of intraepithelial lymphocytes; (2) absence or a reduced height of villi, which causes a flat appearance with increased crypt cell proliferation resulting in crypt hyperplasia and loss of villous structure, with consequent villous, but not mucosal, atrophy; (3) a cuboidal appearance and nuclei that are no longer oriented basally in surface epithelial cells; and (4) increased numbers of lymphocytes and plasma cells in the lamina propria (Fig. 318-4*B*). Although these features are characteristic of celiac disease, they are not diagnostic because a similar appearance can develop eosinophilic enteritis, and milk-protein intolerance in children and occasionally in lymphoma, bacterial overgrowth, Crohn's disease, and gastrinoma with acid hypersecretion. However, a characteristic histologic appearance that reverts toward normal after the initiation of a gluten-free diet establishes the diagnosis of celiac disease (Fig. 318-4*C*). Readministration of gluten, with or without an additional small-intestinal biopsy, is not necessary.

A number of patients exhibit *nonceliac gluten sensitivity*; that is, they have gastrointestinal symptoms that respond to gluten restriction but do not have celiac disease. The basis for such gluten sensitivity is not known.

Failure to Respond to Gluten Restriction The most common cause of persistent symptoms in a patient who fulfills all the criteria for the diagnosis of celiac disease is *continued intake of gluten*. Gluten is ubiquitous, and a significant effort must be made to exclude all gluten from the diet. Use of rice flour in place of wheat flour is very helpful, and several support groups provide important aid to patients with celiac disease and to their families. More than 90% of patients who have the characteristic findings of celiac disease respond to complete dietary gluten restriction. The remainder constitute a heterogeneous group (whose condition is often called *refractory celiac disease* or *refractory sprue*) that includes some patients who (1) respond to restriction of other dietary protein (e.g., soy); (2) respond to glucocorticoid treatment; (3) are "temporary" (i.e., whose clinical and morphologic findings disappear after several months or years); or (4) fail to respond to all measures and have a fatal outcome, with or without documented complications of celiac disease, such as the development of intestinal T cell lymphoma or autoimmune enteropathy.

Therapeutic approaches that do not include a gluten-free diet are being developed and include the use of peptidases to inactivate toxic gliadin peptides and of small molecules to block toxic peptide uptake across intestinal tight junctions.

Mechanism of Diarrhea The diarrhea in celiac disease has several pathogenetic mechanisms. Diarrhea may be secondary to (1) steatorrhea, which is primarily a result of changes in jejunal mucosal function; (2) secondary lactase deficiency, a consequence of changes in jejunal brush border enzymatic function; (3) bile-acid malabsorption resulting in bile-acid–induced fluid secretion in the colon (in cases with more extensive disease involving the ileum); and (4) endogenous fluid secretion resulting from crypt hyperplasia. Celiac disease patients with more severe involvement may improve temporarily with *dietary lactose and fat restriction* while awaiting the full effects of total gluten restriction, which constitutes primary therapy.

Associated Diseases Celiac disease is associated with dermatitis herpetiformis (DH), but this association has not been explained. Patients with DH have characteristic papulovesicular lesions that respond to dapsone. Almost all patients with DH have histologic changes in the small intestine consistent with celiac disease, although usually much milder and less diffuse in distribution. Most patients with DH have mild or no gastrointestinal symptoms. In contrast, relatively few patients with celiac disease have DH.

Celiac disease is also associated with diabetes mellitus type 1, IgA deficiency, Down's syndrome, and Turner's syndrome. The clinical importance of the association with diabetes is that, although severe watery diarrhea without evidence of malabsorption is most often diagnosed as "diabetic diarrhea" **(Chap. 396)**, assay of antiendomysial antibodies and/or a small-intestinal biopsy must be considered to exclude celiac disease.

Complications The most important complication of celiac disease is the development of cancer. The incidences of both gastrointestinal and nongastrointestinal neoplasms as well as intestinal lymphoma are elevated among patients with celiac disease. For unexplained reasons, the frequency of lymphoma in patients with celiac disease is higher in Ireland and the United Kingdom than in the United States. The possibility of lymphoma must be considered whenever a patient with celiac disease who has previously done well on a gluten-free diet is no longer responsive to gluten restriction or a patient who presents with clinical and histologic features consistent with celiac disease does not respond to a gluten-free diet. Other complications of celiac disease include the development of intestinal ulceration independent of lymphoma and so-called refractory sprue (see above) and collagenous sprue. In *collagenous sprue*, a layer of collagen-like material is present beneath the basement membrane; patients with collagenous sprue generally do not respond to a gluten-free diet and often have a poor prognosis.

TROPICAL SPRUE

Tropical sprue is a poorly understood syndrome that affects both expatriates and natives in certain but not all tropical areas and is manifested by chronic diarrhea, steatorrhea, weight loss, and nutritional deficiencies, including those of both folate and cobalamin. This disease affects 5–10% of the population in some tropical areas.

Chronic diarrhea in a tropical environment is most often caused by infectious agents, including *G. lamblia*, *Yersinia enterocolitica*, *C. difficile*, *Cryptosporidium parvum*, and *Cyclospora cayetanensis*. Tropical sprue should not be entertained as a possible diagnosis until the presence of cysts and trophozoites has been excluded in three stool samples. **Chronic**

infections of the gastrointestinal tract and diarrhea in patients with or without AIDS are discussed in Chaps. 128, 129, and 197.

The small-intestinal mucosa of individuals living in tropical areas is not identical to that of individuals who reside in temperate climates. In residents of tropical areas, biopsies reveal a mild alteration of villous architecture with a modest increase in mononuclear cells in the lamina propria, which on occasion can be as severe as that seen in celiac disease. These changes are observed both in native residents and in expatriates living in tropical regions and are usually associated with mild decreases in absorptive function, but they revert to "normal" when an individual moves or returns to a temperate area. Some have suggested that the changes seen in tropical enteropathy and in tropical sprue represent different ends of the spectrum of a single entity, but convincing evidence to support this concept is lacking.

In the past few years the term "environmental enteropathy" has been introduced as the diagnosis of many patients (especially infants and children) who had previously been diagnosed as tropical sprue. Such patients have physical and/or mental shunting. However, exact delineation of this newly designated entity is lacking.

Etiology Because tropical sprue responds to antibiotics, the consensus is that it may be caused by one or more infectious agents. Nonetheless, the etiology and pathogenesis of tropical sprue are uncertain. First, its occurrence is not evenly distributed in all tropical areas; rather, it is found in specific locations, including southern India, the Philippines, and several Caribbean islands (e.g., Puerto Rico, Haiti), but is rarely observed in Africa, Jamaica, or Southeast Asia. Second, an occasional individual does not develop symptoms of tropical sprue until long after having left an endemic area. For this reason, celiac disease (often referred to as celiac sprue) was originally called *nontropical sprue* to distinguish it from tropical sprue. Third, multiple microorganisms have been identified in jejunal aspirates, with relatively little consistency among studies. *Klebsiella pneumoniae*, *Enterobacter cloacae*, and *E. coli* have been implicated in some studies of tropical sprue, while other studies have favored a role for a toxin produced by one or more of these bacteria. Fourth, the incidence of tropical sprue appears to have decreased substantially during the past two or three decades, perhaps in relation to improved sanitation in many tropical countries during this time. Some have speculated that the reduced occurrence is attributable to the wider use of antibiotics in acute diarrhea, especially in travelers to tropical areas from temperate countries. Fifth, the role of folic acid deficiency in the pathogenesis of tropical sprue requires clarification. Folic acid is absorbed exclusively in the duodenum and proximal jejunum, and most patients with tropical sprue have evidence of folate malabsorption and depletion. Although folate deficiency can cause changes in small-intestinal mucosa that are corrected by folate replacement, several earlier studies reporting that tropical sprue could be cured by folic acid did not provide an explanation for the "insult" that was initially responsible for folate malabsorption.

The clinical pattern of tropical sprue varies in different areas of the world (e.g., India vs Puerto Rico). Not infrequently, individuals in southern India initially report the occurrence of acute enteritis before the development of steatorrhea and malabsorption. In contrast, in Puerto Rico, a more insidious onset of symptoms and a more dramatic response to antibiotics are seen than in some other locations. Tropical sprue in different areas of the world may not be the same disease, and similar clinical entities may have different etiologies.

Diagnosis The diagnosis of tropical sprue is best based on an abnormal small-intestinal mucosal biopsy in an individual with chronic diarrhea and evidence of malabsorption who is either residing or has recently lived in a tropical country. The small-intestinal biopsy in tropical sprue does not reveal pathognomonic features but resembles, and can often be indistinguishable from, that seen in celiac disease (Fig. 318-4). The biopsy sample in tropical sprue has less villous architectural alteration and more mononuclear cell infiltrate in the lamina propria. In contrast to those of celiac disease, the histologic features of tropical sprue manifest with a similar degree of severity throughout the small intestine, and a gluten-free diet does not result in either clinical or histologic improvement in tropical sprue.

TREATMENT

Tropical Sprue

Broad-spectrum antibiotics and folic acid are most often curative, especially if the patient leaves the tropical area and does not return. Tetracycline should be used for up to 6 months and may be associated with improvement within 1–2 weeks. Folic acid alone induces hematologic remission as well as improvement in appetite, weight gain, and some morphologic changes in small-intestinal biopsy. Because of marked folate deficiency, folic acid is most often given together with antibiotics.

SHORT-BOWEL SYNDROME

Short-bowel syndrome is a descriptive term for the myriad clinical problems that follow resection of various lengths of the small intestine or, on rare occasions, are congenital (e.g., microvillous inclusion disease). The factors that determine both the type and degree of symptoms include (1) the specific segment (jejunum vs ileum) resected, (2) the length of the resected segment, (3) the integrity of the ileocecal valve, (4) whether any large intestine has also been removed, (5) residual disease in the remaining small and/or large intestine (e.g., Crohn's disease, mesenteric artery disease), and (6) the degree of adaptation in the remaining intestine. Short-bowel syndrome can occur in persons of any age, from neonates to the elderly.

Three different situations in adults mandate intestinal resection: (1) mesenteric vascular disease, including atherosclerosis, thrombotic phenomena, and vasculitides; (2) primary mucosal and submucosal disease (e.g., Crohn's disease); and (3) operations without preexisting small-intestinal disease (e.g., after trauma).

After resection of the small intestine, the residual intestine undergoes adaptation of both structure and function that may last for up to 6–12 months. Continued intake of dietary nutrients and calories is required to stimulate adaptation via direct contact with the intestinal mucosa, the release of one or more intestinal hormones, and pancreatic and biliary secretions. Thus, enteral nutrition with calorie administration must be maintained, especially in the early postoperative period, even if an extensive intestinal resection requiring parenteral nutrition (PN) has been performed. The subsequent ability of such patients to absorb nutrients will not be known for several months, until adaptation is complete.

Multiple factors besides the absence of intestinal mucosa (required for lipid, fluid, and electrolyte absorption) contribute to diarrhea and steatorrhea in these patients. Removal of the ileum, and especially the ileocecal valve, is often associated with more severe diarrhea than jejunal resection. Without part or all of the ileum, diarrhea can be caused by an increase in bile acids entering the colon; these acids stimulate colonic fluid and electrolyte secretion. Absence of the ileocecal valve is also associated with a decrease in intestinal transit time and bacterial overgrowth from the colon. The presence of the colon (or a major portion) is associated with substantially less diarrhea and a lower likelihood of *intestinal failure* (an inability to maintain nutrition without parenteral support) as a result of fermentation of nonabsorbed carbohydrates to SCFAs. The latter are absorbed in the colon and stimulate Na and water absorption, improving overall fluid balance. Lactose intolerance as a result of the removal of lactase-containing mucosa as well as gastric hypersecretion may also contribute to the diarrhea.

In addition to diarrhea and/or steatorrhea, a range of nonintestinal symptoms is observed in some patients. The frequency of renal calcium oxalate calculi increases significantly in patients with a small-intestinal resection and an intact colon; this greater frequency is due to an increase in oxalate absorption by the large intestine, with subsequent *enteric hyperoxaluria*. Two possible mechanisms for the increase in oxalate absorption in the colon have been suggested: (1) increased bile acids and fatty acids that augment colonic mucosal permeability, resulting in enhanced oxalate absorption; and (2) increased fatty acids that bind calcium, resulting in an enhanced amount of soluble oxalate that is then absorbed. Since oxalate is high in relatively few foods (e.g., spinach, rhubarb, tea), dietary restrictions alone do not constitute

adequate treatment. Cholestyramine (an anion-binding resin) and calcium have proved useful in reducing hyperoxaluria. Similarly, an increase in cholesterol gallstones is related to a decrease in the bile-acid pool size, which results in the generation of cholesterol supersaturation in gallbladder bile. Gastric hypersecretion of acid occurs in many patients after large resections of the small intestine. The etiology is unclear but may be related to either reduced hormonal inhibition of acid secretion or increased gastrin levels due to reduced small-intestinal catabolism of circulating gastrin. The resulting gastric acid secretion may be an important factor contributing to diarrhea and steatorrhea. A reduced pH in the duodenum can inactivate pancreatic lipase and/or precipitate duodenal bile acids, thereby increasing steatorrhea, and an increase in gastric secretion can create a volume overload relative to the reduced small-intestinal absorptive capacity. Inhibition of gastric acid secretion with proton pump inhibitors can help reduce diarrhea and steatorrhea.

TREATMENT

Short-Bowel Syndrome

Treatment of short-bowel syndrome depends on the severity of symptoms and on whether the individual is able to maintain caloric and electrolyte balance with oral intake alone. Initial treatment includes judicious use of opiates (including codeine) to reduce stool output and establish an effective diet. If the colon is in situ, the initial diet should be low in fat and high in carbohydrate in order to minimize diarrhea from fatty acid stimulation of colonic fluid secretion. MCTs (see Table 318-3), a low-lactose diet, and various soluble fiber–containing diets should also be tried. In the absence of an ileocecal valve, possible bacterial overgrowth must be considered and treated. If gastric acid hypersecretion is contributing to diarrhea and steatorrhea, a proton pump inhibitor may be helpful. Usually none of these therapeutic approaches provides an instant solution, but each can contribute to the reduction of disabling diarrhea.

The patient's vitamin and mineral status must also be monitored; replacement therapy should be initiated if indicated. Fat-soluble vitamins, folate, cobalamin, calcium, iron, magnesium, and zinc are the most critical factors to monitor on a regular basis. If these approaches are not successful, home PN is an established therapy that can be maintained for many years. Small-intestinal transplantation is becoming established as a possible approach for individuals with extensive intestinal resection who cannot be maintained without PN—that is, those with intestinal failure. A recombinant analogue of glucagon-like peptide 2 (GLP-2; teduglutide) is approved for use in patients with PN-dependent short-bowel syndrome on the basis of its ability to increase intestinal growth, improve absorption, and reduce requirement for PN.

BACTERIAL OVERGROWTH SYNDROMES

Bacterial overgrowth syndromes comprise a group of disorders with diarrhea, steatorrhea, and macrocytic anemia whose common feature is the proliferation of colonic-type bacteria within the small intestine. This bacterial proliferation is due to stasis caused by impaired peristalsis (*functional stasis*), changes in intestinal anatomy (*anatomic stasis*), or direct communication between the small and large intestine. These conditions have also been referred to as *stagnant bowel syndrome* or *blind loop syndrome.*

Pathogenesis The manifestations of bacterial overgrowth syndromes are a direct consequence of the presence of increased amounts of a colonic-type bacterial flora, such as *E. coli* or *Bacteroides*, in the small intestine. *Macrocytic anemia* is due to cobalamin—not folate—deficiency. Most bacteria require cobalamin for growth, and increased concentrations of bacteria use up the relatively small amounts of dietary cobalamin. *Steatorrhea* is due to impaired micelle formation as a consequence of a reduced intraduodenal concentration of conjugated bile acids and the presence of unconjugated bile acids. Certain bacteria, including *Bacteroides*, deconjugate conjugated bile acids to unconjugated bile acids. Unconjugated bile acids are absorbed more rapidly than conjugated bile acids; as a result, the intraduodenal concentration of bile acids is reduced. In addition, the CMC of unconjugated bile acids is higher than that of conjugated bile acids, and the result is a decrease in micelle formation. *Diarrhea* is due, at least in part, to steatorrhea, when it is present. However, some patients manifest diarrhea *without* steatorrhea, and it is assumed that the colonic-type bacteria in these patients are producing one or more bacterial enterotoxins that are responsible for fluid secretion and diarrhea.

Etiology The etiology of these different disorders is bacterial proliferation in the small-intestinal lumen secondary to anatomic or functional stasis or to a communication between the relatively sterile small intestine and the colon, with its high levels of aerobic and anaerobic bacteria. Several examples of *anatomic* stasis have been identified: (1) one or more diverticula (both duodenal and jejunal) (Fig. 318-3*C*); (2) fistulas and strictures related to Crohn's disease (Fig. 318-3*D*); (3) a proximal duodenal afferent loop following subtotal gastrectomy and gastrojejunostomy; (4) a bypass of the intestine (e.g., a jejunoileal bypass for obesity); and (5) dilation at the site of a previous intestinal anastomosis. These anatomic derangements are often associated with the presence of a segment (or segments) of intestine out of continuity of propagated peristalsis, with consequent stasis and bacterial proliferation. Bacterial overgrowth syndromes can also occur in the *absence* of an anatomic blind loop when *functional* stasis is present. Impaired peristalsis and bacterial overgrowth in the absence of a blind loop occur in scleroderma, where motility abnormalities exist in both the esophagus and the small intestine **(Chap. 353)**. Functional stasis and bacterial overgrowth can also develop in association with diabetes mellitus and in the small intestine when a direct connection exists between the small and large intestines, including an ileocolonic resection, or occasionally after an enterocolic anastomosis that permits entry of bacteria into the small intestine as a result of bypassing the ileocecal valve.

Diagnosis The diagnosis may be suspected from the combination of a low serum cobalamin level and an elevated serum folate level, as enteric bacteria frequently produce folate compounds that are absorbed in the duodenum. Ideally, the bacterial overgrowth syndromes are diagnosed by the demonstration of increased levels of aerobic and/or anaerobic colonic-type bacteria in a jejunal aspirate obtained by intubation. However, this specialized test is rarely available. Breath hydrogen testing with administration of lactulose (a nondigestible disaccharide) has also been used to detect bacterial overgrowth. Often, the diagnosis is suspected clinically and confirmed by the response to treatment.

TREATMENT

Bacterial Overgrowth Syndromes

Primary treatment should be directed, if at all possible, to the surgical correction of an anatomic blind loop. In the absence of functional stasis, it is important to define the anatomic relationships responsible for stasis and bacterial overgrowth. For example, bacterial overgrowth secondary to strictures, one or more diverticula, or a proximal afferent loop can potentially be cured by surgical correction of the anatomic state. In contrast, the functional stasis of scleroderma or certain anatomic stasis states (e.g., multiple jejunal diverticula) cannot be corrected surgically, and these conditions should be treated with broad-spectrum antibiotics. Tetracycline used to be the initial drug of choice; because of increasing resistance, however, other antibiotics, such as metronidazole, amoxicillin/clavulanic acid, rifaximin and cephalosporins, have been employed. The antibiotic should be given for ~3 weeks or until symptoms remit. Although the natural history of these conditions is chronic, antibiotics should not be given continuously. Symptoms usually remit within 2–3 weeks of initial antibiotic therapy. Treatment need not be repeated until symptoms recur. For frequent recurrences, several treatment strategies exist, but the use of antibiotics for 1 week per month, whether or not symptoms are present, is often most effective.

Unfortunately, therapy for bacterial overgrowth syndromes is largely empirical, with an absence of clinical trials on which to base rational decisions regarding antibiotic choice, treatment duration, and/or the best approach to therapy for recurrences. Bacterial overgrowth may also occur as a component of another chronic disease, such as Crohn's disease, radiation enteritis, or short-bowel syndrome. Treatment of the bacterial overgrowth in these settings will not cure the underlying problem but may be very important in ameliorating a subset of clinical problems that are related to bacterial overgrowth.

WHIPPLE'S DISEASE

Whipple's disease is a chronic multisystemic disease associated with diarrhea, steatorrhea, weight loss, arthralgia, and central nervous system (CNS) and cardiac problems; it is caused by the bacterium *Tropheryma whipplei.* Until the identification of *T. whipplei* by polymerase chain reaction, the hallmark of Whipple's disease had been the presence of PAS-positive macrophages in the small intestine (Fig. 318-4*E*) and other organs with evidence of disease.

Etiology *T. whipplei*, a small (50–500 nm) gram-positive bacillus in the group Actinobacteria, has low virulence but high infectivity. Symptoms of Whipple's disease are relatively minimal compared to the bacterial burden in multiple tissues.

Clinical Presentation The onset of Whipple's disease is insidious and is characterized by diarrhea, steatorrhea, abdominal pain, weight loss, migratory large-joint arthropathy, and fever as well as ophthalmologic and CNS symptoms. Dementia is a relatively late symptom and an extremely poor prognostic sign, especially in patients who experience relapse after the induction of a remission with antibiotics. For unexplained reasons, the disease occurs primarily in middle-aged white men. The steatorrhea in these patients is generally believed to be secondary to both small-intestinal mucosal injury and lymphatic obstruction due to the increased number of PAS-positive macrophages in the lamina propria of the small intestine.

Diagnosis The diagnosis of Whipple's disease is suggested by a multisystemic disease in a patient with diarrhea and steatorrhea. Tissue biopsy of the small intestine and/or other organs that may be involved (e.g., liver, lymph nodes, heart, eyes, CNS, or synovial membranes), given the patient's symptoms, is the primary approach. The presence of PAS-positive macrophages containing the characteristic small bacilli is suggestive of this diagnosis. However, *T. whipplei*–containing macrophages can be confused with PAS-positive macrophages containing *M. avium* complex, which may be a cause of diarrhea in AIDS. The presence of the *T. whipplei* bacillus outside of macrophages is a more important indicator of active disease than is their presence within the macrophages. *T. whipplei* has now been successfully grown in culture.

TREATMENT

Whipple's Disease

The treatment for Whipple's disease is prolonged use of antibiotics. The current regimen of choice is ceftriaxone or meropenem for 2 weeks followed by oral TMP-SMX (160/800 mg) twice a day for 1 year. PAS-positive macrophages can persist after successful treatment, and the presence of bacilli outside of macrophages is indicative of persistent infection or an early sign of recurrence. Recurrence of disease activity, especially with dementia, is an extremely poor prognostic sign and requires an antibiotic that crosses the blood-brain barrier. If trimethoprim-sulfamethoxazole is not tolerated, chloramphenicol is an appropriate second choice.

PROTEIN-LOSING ENTEROPATHY

Protein-losing enteropathy is not a specific disease but rather a group of gastrointestinal and nongastrointestinal disorders with hypoproteinemia and edema in the absence of either proteinuria or defects in protein synthesis (e.g., chronic liver disease). These diseases are characterized by excess protein loss into the gastrointestinal tract. Normally, ~10% of total protein catabolism occurs via the gastrointestinal tract. Evidence of increased protein loss into the gastrointestinal tract is found in >65 different diseases, which can be classified into three groups: (1) mucosal ulceration, such that the protein loss primarily represents exudation across damaged mucosa (e.g., ulcerative colitis, gastrointestinal carcinomas, and peptic ulcer); (2) nonulcerated mucosa, but with evidence of mucosal damage so that the protein loss represents loss across epithelia with altered permeability (e.g., celiac disease and Ménétrier's disease in the small intestine and stomach, respectively); and (3) lymphatic dysfunction, representing either primary lymphatic disease or lymphatic disease secondary to partial lymphatic obstruction that may occur as a result of enlarged lymph nodes or cardiac disease.

Diagnosis The diagnosis of protein-losing enteropathy is suggested by peripheral edema and low serum albumin and globulin levels in the absence of renal and hepatic disease. An individual with protein-losing enteropathy only rarely has selective loss of *only* albumin or *only* globulins. Therefore, marked reduction of serum albumin with normal serum globulins should not initiate an evaluation for protein-losing enteropathy but should suggest renal and/or hepatic disease. Likewise, reduced serum globulins with normal serum albumin levels are more likely a result of reduced globulin synthesis rather than enhanced globulin loss into the intestine. An increase in protein loss into the gastrointestinal tract has been documented by the administration of one of several radiolabeled proteins and its quantitation in stool during a 24- or 48-h period. Unfortunately, none of these radiolabeled proteins is available for routine clinical use. α_1-Antitrypsin, a protein that accounts for ~4% of total serum proteins and is resistant to proteolysis, can be used to detect enhanced rates of serum protein loss into the intestinal tract but cannot be used to assess gastric protein loss because of its degradation in an acid milieu. α_1-Antitrypsin clearance is measured by determining stool volume as well as both stool and plasma α_1-antitrypsin concentrations. In addition to the loss of protein via abnormal and distended lymphatics, peripheral lymphocytes may be lost via lymphatics, with consequent relative lymphopenia. Thus, lymphopenia in a patient with hypoproteinemia indicates increased loss of protein into the gastrointestinal tract.

Patients with increased protein loss into the gastrointestinal tract from lymphatic obstruction often have steatorrhea and diarrhea. The steatorrhea is a result of altered lymphatic flow as lipid-containing chylomicrons exit from intestinal epithelial cells via intestinal lymphatics (Table 318-4; Fig. 318-4). In the absence of mechanical or anatomic lymphatic obstruction, intrinsic intestinal lymphatic dysfunction—with or without lymphatic dysfunction in the peripheral extremities—has been designated *intestinal lymphangiectasia*. Similarly, ~50% of individuals with intrinsic peripheral lymphatic disease (Milroy's disease) also have intestinal lymphangiectasia and hypoproteinemia. Other than steatorrhea and enhanced protein loss into the gastrointestinal tract, all other aspects of intestinal absorptive function are normal in intestinal lymphangiectasia.

Other Causes Patients who appear to have idiopathic protein-losing enteropathy without evidence of gastrointestinal disease should be examined for cardiac disease—especially right-sided valvular disease and chronic pericarditis **(Chaps. 263 and 265)**. On occasion, hypoproteinemia can be the only presenting manifestation in these two types of heart disease. Ménétrier's disease (also called *hypertrophic gastropathy*) is an uncommon entity that involves the body and fundus of the stomach and is characterized by large gastric folds, reduced gastric acid secretion, and, at times, enhanced protein loss into the stomach.

TREATMENT

Protein-Losing Enteropathy

As excess protein loss into the gastrointestinal tract is most often secondary to a specific disease, treatment should be directed primarily to the underlying disease process and not to the hypoproteinemia.

For example, if significant hypoproteinemia with resulting peripheral edema is secondary to celiac disease or ulcerative colitis, a gluten-free diet and mesalamine, respectively, would be the initial therapy. When enhanced protein loss is secondary to lymphatic obstruction, it is critical to establish the nature of this obstruction.

TABLE 318-8 Classification of Malabsorption Syndromes

- Inadequate digestion
  - Postgastrectomy[a]
  - Deficiency or inactivation of pancreatic lipase
    - Exocrine pancreatic insufficiency
      - Chronic pancreatitis
      - Pancreatic carcinoma
      - Cystic fibrosis
      - Pancreatic insufficiency—congenital or acquired
  - Gastrinoma—acid inactivation of lipase
  - Drugs—orlistat
- Reduced intraduodenal bile-acid concentration/impaired micelle formation
  - Liver disease
    - Parenchymal liver disease
    - Cholestatic liver disease
  - Bacterial overgrowth in small intestine:

Anatomic stasis	Functional stasis
Afferent loop	Diabetes[a]
Stasis/blind	Scleroderma[a]
Loop/strictures/fistulae	Intestinal pseudo-obstruction

  - Interrupted enterohepatic circulation of bile salts

Ileal resection
Crohn's disease

  - Drugs (binding or precipitating bile salts)—neomycin, cholestyramine, calcium carbonate
  - Impaired mucosal absorption/mucosal loss or defect
    - Intestinal resection or bypass[a]
    - Inflammation, infiltration, or infection:

Crohn's disease[a]	Celiac disease
Amyloidosis	Collagenous sprue
Scleroderma[a]	Whipple's disease[a]
Lymphoma[a]	Radiation enteritis[a]
Eosinophilic enteritis	Folate and vitamin B_{12} deficiency
Mastocytosis	Infections—giardiasis
Tropical sprue	Graft versus host disease

  - Genetic disorders
    - Disaccharidase deficiency
    - Agammaglobulinemia
    - Abetalipoproteinemia
    - Hartnup's disease
    - Cystinuria
- Impaired nutrient delivery to and/or from intestine:

Lymphatic obstruction	Circulatory disorders
Lymphoma[a]	Congestive heart failure
Lymphangiectasia	Constrictive pericarditis
	Mesenteric artery atherosclerosis
	Vasculitis

- Endocrine and metabolic disorders
  - Diabetes[a]
  - Hypoparathyroidism
  - Adrenal insufficiency
  - Hyperthyroidism
  - Carcinoid syndrome

[a]Malabsorption caused by more than one mechanism.

TABLE 318-9 Pathophysiology of Clinical Manifestations of Malabsorption Disorders

SYMPTOM OR SIGN	MECHANISM
Weight loss/malnutrition	Anorexia, malabsorption of nutrients
Diarrhea	Impaired absorption or secretion of water and electrolytes; colonic fluid secretion secondary to unabsorbed dihydroxy bile acids and fatty acids
Flatus	Bacterial fermentation of unabsorbed carbohydrate
Glossitis, cheilosis, stomatitis	Deficiency of iron, vitamin B_{12}, folate, and vitamin A
Abdominal pain	Bowel distention or inflammation, pancreatitis
Bone pain	Calcium, vitamin D malabsorption, protein deficiency, osteoporosis
Tetany, paresthesia	Calcium and magnesium malabsorption
Weakness	Anemia, electrolyte depletion (particularly K^+)
Azotemia, hypotension	Fluid and electrolyte depletion
Amenorrhea, decreased libido	Protein depletion, decreased calories, secondary hypopituitarism
Anemia	Impaired absorption of iron, folate, vitamin B_{12}
Bleeding	Vitamin K malabsorption, hypoprothrombinemia
Night blindness/ xerophthalmia	Vitamin A malabsorption
Peripheral neuropathy	Vitamin B_{12} and thiamine deficiency
Dermatitis	Deficiency of vitamin A, zinc, and essential fatty acid

Identification of mesenteric nodes or lymphoma may be possible by imaging studies. Similarly, it is important to exclude cardiac disease as a cause of protein-losing enteropathy, either by echosonography or, on occasion, by a right-heart catheterization.

The increased protein loss that occurs in intestinal lymphangiectasia is a result of distended lymphatics associated with lipid malabsorption. The hypoproteinemia is treated with a low-fat diet and the administration of MCTs (Table 318-3), which do not exit from the intestinal epithelial cells via lymphatics but are delivered to the body via the portal vein.

SUMMARY

The many conditions that can produce malabsorption are classified by their pathophysiology in Table 318-8. The pathophysiology of the various clinical manifestations of malabsorption is summarized in Table 318-9.

FURTHER READING

Binder HJ: Role of colonic short chain fatty acid transport in diarrhea. Annu Rev Physiol 72:297, 2010.

Carroll RE et al: Management and complications of short bowel syndrome: An updated review. Curr Gastroenterol Rep 18:40, 2016.

Desnues B, Al Moussawi K, Fenollar F: New insights into Whipple's disease and *Tropheryma whipplei* infections. Microbes Infect 12:1102, 2010.

Ghoshal UC et al: Tropical sprue in 2014: The new face of an old disease. Curr Gastroenterol Rep 16:391, 2014.

Hill ID et al: NASPGHAN Clinical Report on the diagnosis and treatment of gluten-related disorders. J Pediatr Gastroenerol Nutr 63:156, 2016.

Kelly CP et al: Advances in diagnosis and management of celiac disease. Gastroenterology 148:1175, 2015.

Korpe PS, Petri WA: Environmental enteropathy: critical implications of a poorly studied condition. Trends Mol Med 18:328, 2012.

Rezaeie A, Pimentel M, Rao SS: How to test and treat small intestinal bacterial overgrowth: An evidence-based approach. Curr Gastroenterol Rep 18:8, 2016.

319 Inflammatory Bowel Disease

Sonia Friedman, Richard S. Blumberg

Inflammatory bowel disease (IBD) is an immune-mediated chronic intestinal condition. Ulcerative colitis (UC) and Crohn's disease (CD) are the two major types of IBD.

GLOBAL CONSIDERATIONS: EPIDEMIOLOGY

The highest incidence rates of CD and UC have been reported in northern Europe, the United Kingdom, and North America. Countries in the Pacific, including New Zealand and Australia, which share many possible environmental risk factors and similar genetic background as northwest Europe and North America, have high incidence rates of IBD.

The highest reported incidence rates are in Canada (19.2 per 1000,000 for UC and 20.2 for CD), Northern Europe (24.3 per 100,000 for UC in Iceland and 10.6 per 100,000 for CD in the United Kingdom), and Australia (17.4 per 100,000 for UC and 29.3 per 100,000 for CD). Prevalence is highest in Europe (505 per 100,000 for UC in Norway and 322 per 100,000 for CD in Italy) and Canada (248 per 100,000 for UC and 319 per 100,000 for CD) **(Table 319-1)**. Based on these estimates ~0.6% of the Canadian population has IBD.

In countries that are becoming more Westernized, including China, South Korea, India, Lebanon, Iran, Thailand, and countries in the French West Indies and North Africa, IBD appears to be emerging, emphasizing the importance of environmental factors in disease pathogenesis. In Japan, the prevalence of CD has risen rapidly from 2.9 cases per 100,000 in 1986 to 13.5 per 100,000 in 1998, whereas in South Korea, the prevalence of UC has quadrupled from 7.6 per 100,000 in 1997 to 30.9 per 100,000 in 2005. In Hong Kong, the prevalence of UC almost tripled from 2.3 in 1997 to 6.3 per 100,000 over a 9-year period. In Singapore, the prevalence of CD increased from 1.3 in 1990 to 7.2 per 100,000 in 2004. In China the number of cases of UC has increased by fourfold between 1981–1990 and 1991–2000.

Increasing immigration to Western societies also has an impact on the incidence and prevalence of IBD. The prevalence of UC among southern Asians who immigrated to the United Kingdom (UK) was higher in comparison to the European UK population (17 cases per 100,000 persons vs 7 per 100,000). Spanish patients who emigrated within Europe, but not those who immigrated to Latin America, developed IBD more frequently than controls. Individuals who have immigrated to Westernized countries and then returned to their country of birth also continue to demonstrate an increased risk of developing IBD.

TABLE 319-1 Epidemiology of IBD

	ULCERATIVE COLITIS	CROHN'S DISEASE
Incidence (North America) per person-years	0–19.2 per 100,000	0–20.2 per 100,000
Age of onset	Second to fourth decades and seventh to ninth decades	Second to fourth decades and seventh to ninth decades
Ethnicity	Jewish > non-Jewish white > African American > Hispanic > Asian	
Female/male ratio	0.51–1.58	0.34–1.65
Smoking	May prevent disease (odds ratio 0.58)	May cause disease (odds ratio 1.76)
Oral contraceptives	No increased risk	Hazard ratio 2.82
Appendectomy	Protective (risk reduction 13–26%)	Not protective
Monozygotic twins	6–18% concordance	38–58% concordance
Dizygotic twins	0–2% concordance	4% concordance
Antibiotic use in the first year of life	2.9× the risk of developing childhood IBD	

Abbreviation: IBD, inflammatory bowel disease.

Peak incidence of UC and CD is in the second to fourth decades, with 78% of CD studies and 51% of UC studies reporting the highest incidence among those age 20–29 years old. A second modest rise in incidence occurs between the seventh and ninth decades of life. The female-to-male ratio ranges from 0.51 to 1.58 for UC studies and 0.34 to 1.65 for CD studies, suggesting that the diagnosis of IBD is not gender-specific. Pediatric IBD (patients <18 years of age) comprises ~20–25% of all IBD patients and about 5% of all IBD patients are <10 years of age. Very early onset IBD (VEOIBD) has been defined as IBD that occurs in children <6 years of age and infantile IBD in children <2 years of age. VEOIBD and infantile IBD mainly affect the colon, are resistant to standard medications, and patients often have a strong family history of IBD, with at least one first-degree related affected. Twenty-five percent of patients have an underlying immunodeficiency. In some cases, infantile IBD or VEOIBD can be caused by a number of rare, single genetic mutations.

The greatest incidence of IBD is among white and Jewish people, but the incidence of IBD in Hispanic and Asian people is increasing, as noted above. Urban areas have a higher prevalence of IBD than rural areas, and high socioeconomic classes have a higher prevalence than lower socioeconomic classes.

Epidemiologic studies have identified a number of potential environmental factors that are associated with disease risk **(Fig. 319-1)**. In Caucasian populations, smoking is an important risk factor in IBD with opposite effects on UC (odds ratio [OR] 0.58) and CD (OR 1.76), whereas in other ethnic groups with different genetic susceptibility, smoking may play a lesser role. There is a protective effect of previous appendectomy with confirmed appendicitis (risk reduction of 13–26%), particularly at a young age, on the development of UC across different geographical regions and populations. There is a modest association with the development of CD but this may be due to diagnostic bias. Oral contraceptive use is associated with the risk of CD with a reported hazard ratio as high as 2.82 among current users and 1.39 among past users. The association between oral contraceptive use and UC is limited to women with a history of smoking. There is an association between antibiotic use and the development of childhood IBD with children who received one or more dispensations of antibiotics during the first year of life having a 2.9-fold increase in the risk of developing IBD during childhood. Breastfeeding may also protect against the development of IBD. Infectious gastroenteritis with pathogens (e.g., *Salmonella*, *Shigella*, *Campylobacter* spp., *Clostridium difficile*) increases IBD risk by two- to threefold. Diets high in animal protein, sugars, sweets, oils, fish and shellfish, and dietary fat, especially ω-6 fatty acids, and low in ω-3 fatty acids have been implicated in increasing the risk of IBD. A protective effect of vitamin D on the risk of CD has been reported.

IBD is a familial disease in 5–10% of patients **(Fig. 319-2)**. Some of these patients may exhibit early-onset disease during the first decade of life and, in CD, a concordance of anatomic site and clinical type within families. In the remainder of patients, IBD is observed in the absence of a family history (i.e., sporadic disease). If a patient has IBD, the lifetime risk that a first-degree relative will be affected is ~10%. If two parents have IBD, each child has a 36% chance of being affected. In twin studies, 38–58% of monozygotic twins are concordant for CD and 6–18% are concordant for UC, whereas 4% of dizygotic twins are concordant for CD and 0–2% are concordant for UC in Swedish and Danish cohorts. The risks of developing IBD are higher in first-degree relatives of Jewish versus non-Jewish patients: 7.8% versus 5.2% for CD and 4.5% versus 1.6% for UC.

GLOBAL CONSIDERATIONS: IBD PHENOTYPES

There are racial differences in IBD location and behavior that may reflect underlying genetic variations and have important implications for diagnosis and management of disease. African Americans and Hispanics tend to have an ileocolonic CD distribution. Data from East Asia have observed that ileocolonic CD is the most

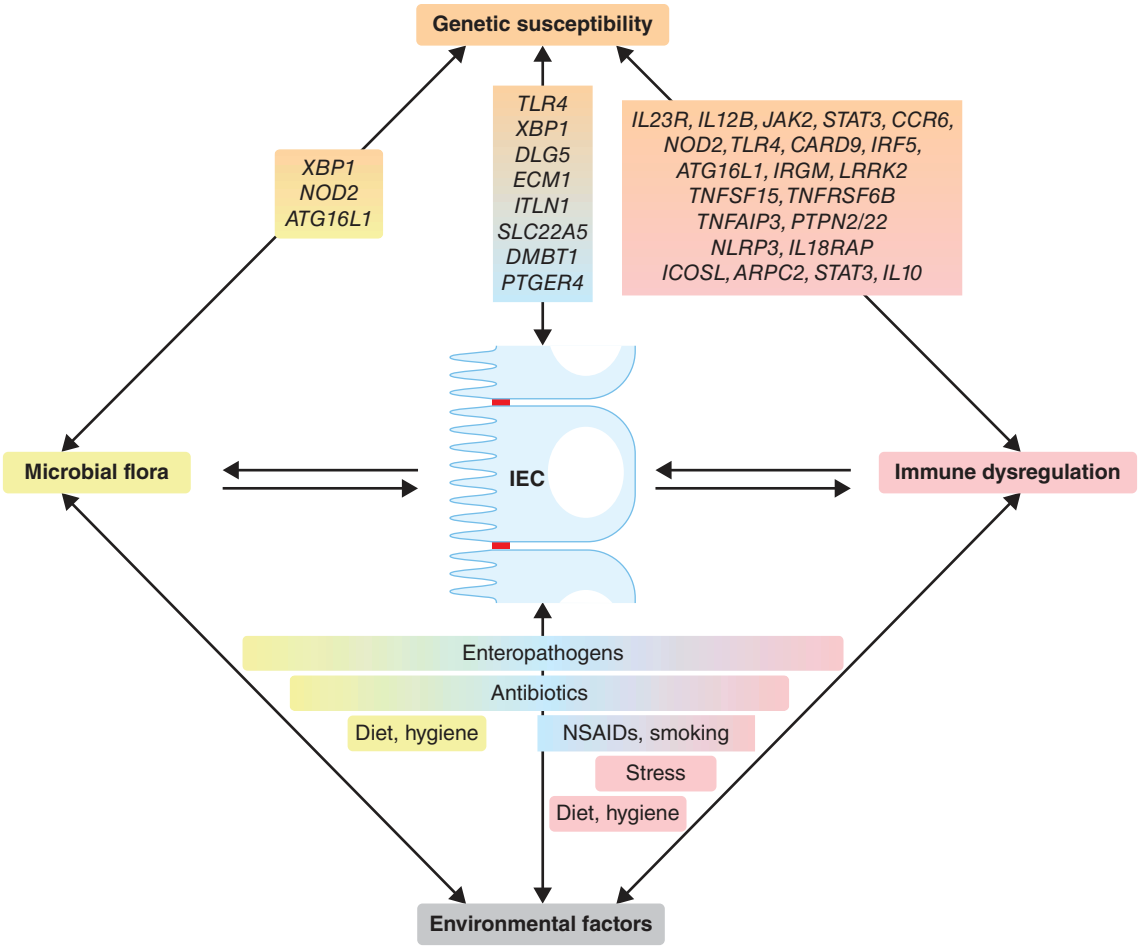


FIGURE 319-1 Pathogenesis of inflammatory bowel disease (IBD). In IBD, the tridirectional relationship between the commensal flora (microbiota), intestinal epithelial cells (IECs), and mucosal immune system is dysregulated, leading to chronic inflammation. Each of these three factors is affected by genetic and environmental factors that determine risk for the disease. NSAIDs, nonsteroidal anti-inflammatory drugs. *(Adapted from A Kaser et al: Annu Rev Immunol 28:573, 2010.)*

common CD phenotype (50.5–71%) and perianal disease is more common in East Asian patients (30.3–58.8%) than Caucasians (25.1–29.6%). Pancolonic disease is more common than left-sided colitis or proctitis among African Americans, Hispanics, and Asian patients with UC. Older Asian patients with UC (age >60) tend to have a more aggressive disease course. Among African Americans, joint involvement is the predominant extra intestinal manifestation (EIM) reported and ranges from 15.7 to 29.6%. Ocular involvement is also common in African Americans and ranges from 7.1 to 13%. Dermatologic manifestations are the most common EIM reported in Hispanics (10–13%). There are few data on all aspects of disease in Hispanics, on the incidence and prevalence of IBD in African Americans, and in Asians with IBD outside Asia. These ethnic variations implicate the importance of different genetic and/or environmental factors in the pathogenesis of this disorder.

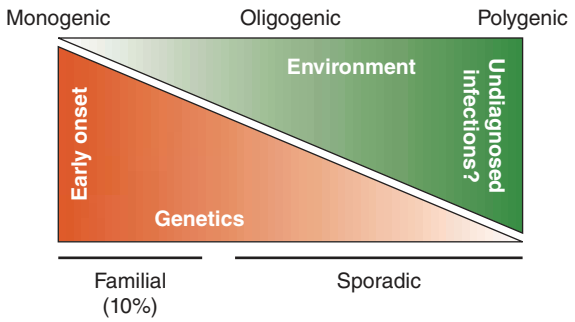


FIGURE 319-2 A model for the syndromic nature of inflammatory bowel disease. Genetic and environmental factors variably influence the development and phenotypic manifestations of IBD. At the one extreme, IBD is exemplified as a simple Mendelian disorder as observed in "early-onset IBD" due to single gene defects such as *IL10*, *IL10RA*, and *IL10RB*; and at the other extreme, it may be exemplified by as yet to be described emerging infectious diseases. *(Adapted from A Kaser et al: Dig Dis 28:395, 2010.)*

ETIOLOGY AND PATHOGENESIS

Under physiologic conditions, homeostasis normally exists between the commensal microbiota, epithelial cells that line the interior of the intestines (intestinal epithelial cells [IECs]) and immune cells within the tissues (Fig. 319-1). A consensus hypothesis is that each of these three major host compartments that function together as an integrated "supraorganism" (microbiota, IECs, and immune cells) are affected by specific environmental (e.g., smoking, antibiotics, enteropathogens) and genetic factors that, in a susceptible host, cumulatively and interactively disrupt homeostasis during the course of one's life, which in so doing culminates in a chronic state of dysregulated inflammation; that is IBD. Although chronic activation of the mucosal immune system may represent an appropriate response to an infectious agent, a search for such an agent has thus far been unrewarding in IBD. As such, IBD is currently considered an inappropriate immune response to the endogenous (autochthonous) commensal microbiota within the intestines, with or without some component of autoimmunity. Importantly, the normal, uninflamed intestines contain a large number of immune cells that are in a unique state of activation, in which the gut is restrained from full immunologic responses to the commensal microbiota and dietary antigens by very powerful regulatory pathways that function within the immune system (e.g., T-regulatory cells that express the FoxP3 transcription factor and suppress inflammation). During the course of infections or other environmental stimuli in the normal host, full activation of the lymphoid tissues in the intestines occurs but is

rapidly superseded by dampening of the immune response and tissue repair. In IBD such processes may not be regulated normally.

GENETIC CONSIDERATIONS

The genetic underpinning of IBD is known from its occurrence in the context of several genetic syndromes and the development of severe, refractory IBD in early life in the setting of single gene defects that affect the immune system (Table 319-2). These include mutations in genes encoding, for example, interleukin-10 (IL-10), the IL-10 receptor (IL-10R), cytotoxic T-lymphocyte associated protein-4 (*CTLA4*), neutrophil cytosolic factor 2 protein (*NCF2*), X-linked inhibitor of apoptosis protein (*XIAP*), lipopolysaccharide responsive and beige-like anchor protein (*LRBA*), or tetratricopeptide repeat domain 7A protein (*TTC7*), among many other genes that are involved in host-commensal interactions. In addition, IBD has a familial origin in at least 10% of afflicted individuals (Fig. 319-2). In the majority of patients, IBD is considered to be a polygenic disorder that gives rise to multiple clinical subgroups within UC and CD. A variety of genetic approaches including candidate gene studies, linkage analysis, and genome-wide association studies (GWAS) that focus on the identification of disease-associated, single-nucleotide polymorphisms (SNPs) within the human genome and, more recently, whole-genome sequencing have elucidated many of the genetic factors that affect risk for these diseases. GWAS have, to date, identified ~200 genetic loci with two-thirds of these loci observed to be associated with both disease phenotypes with the remainder being specific for either CD or UC (Table 319-3). These genetic similarities account for the overlapping immunopathogenesis and consequently epidemiologic observations of both diseases in the same families and similarities in response to therapies. Because the specific causal variants for each identified gene or locus are mostly unknown as most risk loci are contained with regulatory regions of the associated genes, it is not clear whether the similarities in the genetic risk factors associated with CD and UC that are observed are shared at structural or functional levels. The risk conferred by each identified gene or locus is unequal and generally small, such that only ~20% of

TABLE 319-2 Primary Genetic Disorders Associated with IBD

NAME	GENETIC ASSOCIATION	PHENOTYPE
Turner's syndrome	Loss of part or all of X chromosome	Associated with UC and colonic CD
Hermansky-Pudlak	Autosomal recessive chromosome 10q23	Granulomatous colitis, oculocutaneous albinism, platelet dysfunction, pulmonary fibrosis
Wiskott-Aldrich syndrome (WAS)	X-linked recessive disorder, loss of WAS protein function	Colitis, immunodeficiency, severely dysfunctional platelets, and thrombocytopenia
Glycogen storage disease	Deficiency of the glucose-6-phosphate transport protein type B1	Granulomatous colitis, presents in infancy with hypoglycemia, growth failure, hepatomegaly, and neutropenia
Immune dysregulation polyendocrinopathy, enteropathy X-linked (IPEX)	Loss of FoxP3 transcription factor and T regulatory cell function	UC-like autoimmune enteropathy, with endocrinopathy (neonatal type 1 diabetes or thyroiditis), dermatitis
Early-onset IBD	Deficient IL-10 and IL-10 receptor function	Severe, refractory IBD in early life

Abbreviations: CD, Crohn's disease; IBD, inflammatory bowel disease; IL, interleukin; UC, ulcerative colitis.

TABLE 319-3 Some Genetic Loci Associated with Crohn's Disease and/or Ulcerative Colitis

CHROMOSOME	PUTATIVE GENE	GENE NAME	PROTEIN FUNCTION	CD	UC
Unfolded Protein Response, Autophagy and Metabolism					
2q37	*ATG16L1*	ATG16 autophagy related 16-like 1	Autophagy	+	
5q31	*SLC22A5*	Solute carrier family 22, member 5	β carnitine transporter	+	
5q33	*IRGM*	Immunity-related GTPase family, M	Autophagy	+	
7p21	*AGR2*	Anterior gradient 2	Unfolded protein response	+	+
12q12	*LRRK2*	Leucine-rich repeat kinase 2	Autophagy	+	
13q14	*C13orf1*	FAMIN/ LACC1	Immunometabolic regulator	+	
17q21	*ORMDL3*	Orosomucoid related member 1-like 3	Unfolded protein response and lipid synthesis	+	+
22q12	*XBP1*	X-box binding protein 1	Unfolded protein response	+	+
Innate Immunity					
1q23	*ITLN1*	Intelectin 1	Bacterial binding	+	
16q12	*NOD2*	Nucleotide-binding oligomerization domain containing 2	Bacterial sensing and autophagy activation	+	
Adaptive Immunity					
1p31	*IL23R*	Interleukin 23 receptor	Th17 cell stimulation	+	+
1q32	*IL10*	Interleukin 10	Treg-associated cytokine		+
5q33	*IL12B*	Interleukin 12B	IL-12 p40 chain of IL-12/IL-23	+	+
18p11	*PTPN2*	Protein tyrosine phosphatase, nonreceptor type 2	T cell regulation	+	
Inflammation					
3p21	*MST1*	Macrophage stimulating 1	Macrophage activation	+	+
5p13	*PTGER4*	Prostaglandin E receptor 4	PGE_2 receptor	+	+
6q23	*TNFAIP3*	Tumor necrosis factor, alpha-induced protein 3 (A20)	Toll-like receptor regulation	+	
6q27	*CCR6*	Chemokine (C-C motif) receptor 6	Dendritic cell migration	+	
9p24	*JAK2*	Janus kinase 2	IL-6R and IL-23R signaling	+	+
17q21	*STAT3*	Signal transducer and activator of transcription 3	IL-6R, IL-23R, and IL-10R signaling	+	+

Abbreviations: CD, Crohn's disease; GTPase, guanosine triphosphatase; IL, interleukin; PGE_2, prostaglandin E_2; UC, ulcerative colitis.

Source: Adapted from A Kaser et al: Ann Rev Immunol 28:573, 2010; B Khor et al: Nature 474:307, 2011; and L Jostins et al: Nature 491:119, 2012.

the genetic variance is considered to be explained by the current genetic information. Further, many of the genetic risk factors identified are also observed to be associated with risk for other immune-mediated diseases, suggesting that related immunogenetic pathways are involved in the pathogenesis of multiple different disorders accounting for the common responsiveness to similar types of biologic therapies (e.g., anti–tumor necrosis factor [TNF] therapies) and possibly the simultaneous occurrence of these disorders. The diseases and the genetic risk factors that are shared with IBD include rheumatoid arthritis (*TNFAIP3*), psoriasis (*IL23R*, *IL12B*), ankylosing spondylitis (*IL23R*), type 1 diabetes mellitus (*IL10*, *PTPN2*), asthma (*ORMDL3*), and systemic lupus erythematosus (*TNFAIP3*, *IL10*) among others.

The genetic factors defined to date that are recognized to mediate risk for IBD have highlighted the importance of several common mechanisms of disease (Table 319-3). These include the following: those genes that are associated with fundamental cell biologic processes such as endoplasmic reticulum (ER) and metabolic stress (e.g., *XBP1*, *ORMDL3*, *OCTN*), which serve to regulate the secretory activity of cells involved in responses to the commensal microbiota such as Paneth and goblet cells and the manner in which intestinal cells respond to the metabolic products of bacteria; those associated with innate immunity and autophagy (e.g., *NOD2*, *ATG16L1*, *IRGM, JAK2, STAT3, C13orf31*) that function in innate immune cells (both parenchymal and hematopoietic) to respond to and effectively clear bacteria, mycobacteria, and viruses; those that are associated with the regulation of adaptive immunity (e.g., *IL23R, IL12B*, *IL10*, *PTPN2*), which regulate the balance between inflammatory and anti-inflammatory (regulatory) cytokines; and, finally, those that are involved in the development and resolution of inflammation (e.g., *MST1*, *CCR6, TNFAIP3, PTGER4*) and ultimately leukocyte recruitment and inflammatory mediator production. Some of these loci are associated with specific subtypes of disease such as the association between *NOD2* polymorphisms and fibrostenosing CD or *ATG16L1* and fistulizing disease, especially within the ileum. However, the clinical utility of these genetic risk factors for the diagnosis or determination of prognosis and therapeutic responses remains to be defined.

COMMENSAL MICROBIOTA AND IBD

The endogenous commensal microbiota within the intestines plays a central role in the pathogenesis of IBD. Humans are born sterile and acquire their commensal microbiota initially from the mother during egress through the birth canal and subsequently from environmental sources. A stable configuration of up to 1000 species of bacteria that achieves a biomass of ~10^{12} colony-forming units per gram of feces is achieved by 3 years of age, which likely persists into adult life, with each individual human possessing a unique combination of species. In addition, the intestines contain other microbial life forms including archae, viruses, and protists. The microbiota is thus considered as a critical and sustaining component of the human organism. The establishment and maintenance of the intestinal microbiota composition and function is under the control of host (e.g., immune and epithelial responses), environmental (e.g., diet and antibiotics), and likely genetic (e.g., *NOD2*) factors (Fig. 319-1). In turn, the microbiota, through its structural components and metabolic activity, has major influences on the epithelial and immune function of the host, which, through epigenetic effects, may have durable consequences. During early life when the commensal microbiota is being established, these microbial effects on the host may be particularly important in determining later life risk for IBD. Specific components of the microbiota can promote or protect from disease. The commensal microbiota in patients with both UC and CD is demonstrably different from nonafflicted individuals, a state of dysbiosis, suggesting the presence of microorganisms that drive disease (e.g., Proteobacteria such as enteroinvasive and adherent *Escherichia coli*) and to which the immune response is directed and/or the loss of microorganisms that hinder inflammation (e.g., Firmicutes such as *Faecalibacterium prausnitzii*). Many of the changes in the commensal microbiota occur as a consequence of the inflammation. In addition, agents that alter the intestinal microbiota such as metronidazole, ciprofloxacin, and elemental diets, may improve CD. CD may also respond to fecal diversion, demonstrating the ability of luminal contents to exacerbate disease.

DEFECTIVE IMMUNE REGULATION IN IBD

The mucosal immune system does not normally elicit an inflammatory immune response to luminal contents due to oral (mucosal) tolerance. Administration of soluble antigens orally, rather than subcutaneously or intramuscularly, leads to antigen-specific control of the response and the host's ability to tolerate the antigen. Multiple mechanisms are involved in the induction of oral tolerance and include deletion or anergy of antigen-reactive T cells or induction of CD4+ T cells that suppress gut inflammation (e.g., T-regulatory cells expressing the FoxP3 transcription factor) that secrete anti-inflammatory cytokines such as IL-10, IL-35, and transforming growth factor β (TGF-β). Oral tolerance may be responsible for the lack of immune responsiveness to dietary antigens and the commensal microbiota in the intestinal lumen. In IBD this suppression of inflammation is altered, leading to uncontrolled inflammation. The mechanisms of this regulated immune suppression are incompletely known.

Gene knockout ($^{-/-}$) or transgenic (Tg) mouse models of IBD, which include those that are directed at genes demonstrated to be associated with risk for the human disease, have revealed that deleting specific cytokines (e.g., IL-2, IL-10, TGF-β) or their receptors, deleting molecules associated with T-cell antigen recognition (e.g., T-cell antigen receptors), or interfering with IEC barrier function and the regulation of responses to commensal bacteria (e.g., XBP1, mucus glycoproteins, or nuclear factor-κB [NF-κB]) leads to spontaneous colitis or enteritis. In the majority of circumstances, intestinal inflammation in these animal models requires the presence of the commensal microbiota. However, in some cases, activation of certain elements of the intestinal immune system may be exacerbated by the absence of bacteria resulting in severe colitis emphasizing the presence of protective properties that are also contained within the commensal microbiota. Thus, a variety of specific alterations in either the microbiota or host can lead to uncontrolled immune activation and inflammation directed at the intestines in mice. How these relate to human IBD remains to be defined, but they are consistent with inappropriate responses of the genetically susceptible host to the commensal microbiota.

THE INFLAMMATORY CASCADE IN IBD

In both UC and CD, inflammation thus likely emerges from the genetic predisposition of the host in the context of yet-to-be defined environmental factors. Once initiated in IBD by abnormal innate immune sensing of bacteria by parenchymal cells (e.g., IECs) and hematopoietic cells (e.g., dendritic cells), the immune inflammatory response is perpetuated by T-cell activation when coupled together with inadequate regulatory pathways. A sequential cascade of inflammatory mediators extends the response making each step a potential target for therapy. Inflammatory cytokines from innate immune cells such as IL-1, IL-6, and TNF have diverse effects on tissues. They promote fibrogenesis, collagen production, activation of tissue metalloproteinases, and the production of other inflammatory mediators; they also activate the coagulation cascade in local blood vessels (e.g., increased production of von Willebrand's factor). These cytokines are normally produced in response to infection but are usually turned off or inhibited at the appropriate time to limit tissue damage. In IBD their activity is not regulated, resulting in an imbalance between the pro-inflammatory and anti-inflammatory mediators. Some cytokines activate other inflammatory cells (macrophages and B cells), and others act indirectly to recruit other lymphocytes, inflammatory leukocytes, and mononuclear cells from the bloodstream into the gut through interactions between homing receptors on leukocytes (e.g., α4β7 integrin) and addressins on vascular endothelium (e.g., MadCAM1). CD4+ T-helper (T_H) cells that promote inflammation are of three major types, all of which may be associated with colitis in animal models and perhaps humans: T_H1 cells (secrete interferon [IFN] γ), T_H2 cells (secrete IL-4, IL-5, IL-13), and T_H17 cells (secrete IL-17, IL-21). T_H17 cells may however also provide protective functions. Innate immune-like cells that lack T-cell receptors are also present in intestines, polarize to the same functional fates and may similarly participate in IBD. T_H1 cells induce transmural granulomatous inflammation that resembles CD; T_H2 cells,

and related natural killer T cells that secrete IL-4, IL-5, and IL-13, induce superficial mucosal inflammation resembling UC in animal models; and T_H17 cells may be responsible for neutrophilic recruitment. However, neutralization of the cytokines produced by these cells, such as IFN-γ or IL-17, has yet to show efficacy in therapeutic trials. Each of these T-cell subsets cross-regulate each other. The T_H1 cytokine pathway is initiated by IL-12, a key cytokine in the pathogenesis of experimental models of mucosal inflammation. IL-4 and IL-23, together with IL-6 and TGF-β, induce T_H2 and T_H17 cells, respectively, and IL-23 inhibits the suppressive function of regulatory T cells. Activated macrophages secrete TNF and IL-6.

These characteristics of the immune response in IBD explain the beneficial therapeutic effects of antibodies to block pro-inflammatory cytokines or the signaling by their receptors (e.g., anti-TNF, anti-IL-12, anti-IL-23, anti-IL-6, or Janus kinase [JAK] inhibitors) or molecules associated with leukocyte recruitment (e.g., anti-α4β7). They also highlight the potential usefulness of cytokines that inhibit inflammation and promote regulatory T cells or promote intestinal barrier (e.g., IL-10) in the treatment of IBD. Therapies such as the 5-aminosalicylic acid (5-ASA) compounds and glucocorticoids are also potent inhibitors of these inflammatory mediators through inhibition of transcription factors such as NF-κB that regulate their expression.

PATHOLOGY

ULCERATIVE COLITIS: MACROSCOPIC FEATURES

UC is a mucosal disease that usually involves the rectum and extends proximally to involve all or part of the colon. About 40–50% of patients have disease limited to the rectum and rectosigmoid, 30–40% have disease extending beyond the sigmoid but not involving the whole colon, and 20% have a total colitis. Proximal spread occurs in continuity without areas of uninvolved mucosa. When the whole colon is involved, the inflammation extends 2–3 cm into the terminal ileum in 10–20% of patients. The endoscopic changes of *backwash ileitis* are superficial and mild and are of little clinical significance. Although variations in macroscopic activity may suggest skip areas, biopsies from normal-appearing mucosa are usually abnormal. Thus, it is important to obtain multiple biopsies from apparently uninvolved mucosa, whether proximal or distal, during endoscopy. One caveat is that effective medical therapy can change the appearance of the mucosa such that either skip areas or the entire colon can be microscopically normal.

With mild inflammation, the mucosa is erythematous and has a fine granular surface that resembles sandpaper. In more severe disease, the mucosa is hemorrhagic, edematous, and ulcerated **(Fig. 319-3)**. In long-standing disease, inflammatory polyps (pseudopolyps) may be present as a result of epithelial regeneration. The mucosa may appear normal in remission, but in patients with many years of disease it appears atrophic and featureless, and the entire colon becomes narrowed and shortened. Patients with fulminant disease can develop a toxic colitis or megacolon where the bowel wall thins and the mucosa is severely ulcerated; this may lead to perforation.

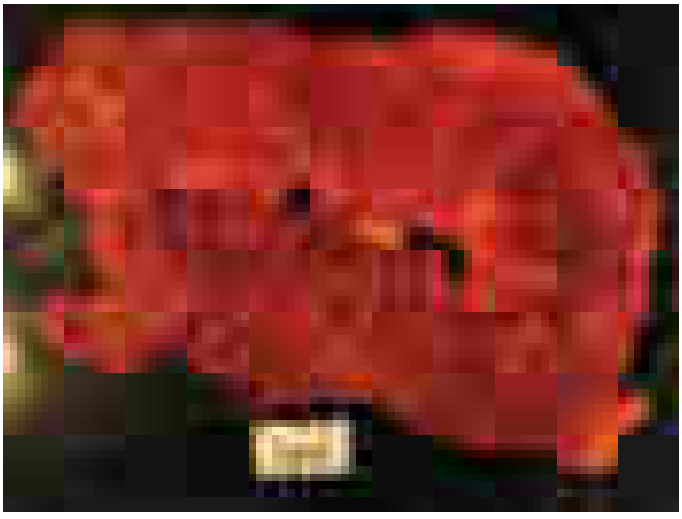


FIGURE 319-3 Ulcerative colitis. Diffuse (nonsegmental) mucosal disease, with broad areas of ulceration. The bowel wall is not thickened, and there is no cobblestoning. *(Courtesy of Dr. R. Odze, Division of Gastrointestinal Pathology, Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts; with permission.)*

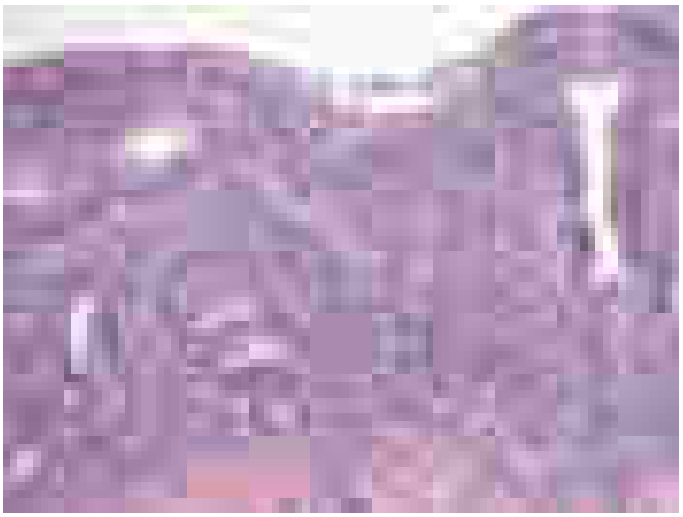


FIGURE 319-4 Medium-power view of colonic mucosa in ulcerative colitis showing diffuse mixed inflammation, basal lymphoplasmacytosis, crypt atrophy and irregularity, and superficial erosion. These features are typical of chronic active ulcerative colitis. *(Courtesy of Dr. R. Odze, Division of Gastrointestinal Pathology, Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts; with permission.)*

ULCERATIVE COLITIS: MICROSCOPIC FEATURES

Histologic findings correlate well with the endoscopic appearance and clinical course of UC. The process is limited to the mucosa and superficial submucosa, with deeper layers unaffected except in fulminant disease. In UC, two major histologic features suggest chronicity and help distinguish it from infectious or acute self-limited colitis. First, the crypt architecture of the colon is distorted; crypts may be bifid and reduced in number, often with a gap between the crypt bases and the muscularis mucosae. Second, some patients have basal plasma cells and multiple basal lymphoid aggregates. Mucosal vascular congestion, with edema and focal hemorrhage, and an inflammatory cell infiltrate of neutrophils, lymphocytes, plasma cells, and macrophages may be present. The neutrophils invade the epithelium, usually in the crypts, giving rise to cryptitis and, ultimately, to crypt abscesses **(Fig. 319-4)**. Ileal changes in patients with backwash ileitis include villous atrophy and crypt regeneration with increased inflammation, increased neutrophil and mononuclear inflammation in the lamina propria, and patchy cryptitis and crypt abscesses.

CROHN'S DISEASE: MACROSCOPIC FEATURES

CD can affect any part of the gastrointestinal (GI) tract from the mouth to the anus. Some 30–40% of patients have small bowel disease alone, 40–55% have disease involving both the small and large intestines, and 15–25% have colitis alone. In the 75% of patients with small intestinal disease, the terminal ileum is involved in 90%. Unlike UC, which almost always involves the rectum, the rectum is often spared in CD. CD is segmental with skip areas in the midst of diseased intestine **(Fig. 319-5)**. Perirectal fistulas, fissures, abscesses, and anal stenosis are present in one-third of patients with CD, particularly those with colonic involvement. Rarely, CD may also involve the liver and the pancreas.

Unlike UC, CD is a transmural process. Endoscopically, aphthous or small superficial ulcerations characterize mild disease; in more active disease, stellate ulcerations fuse longitudinally and transversely to demarcate islands of mucosa that frequently are histologically normal. This

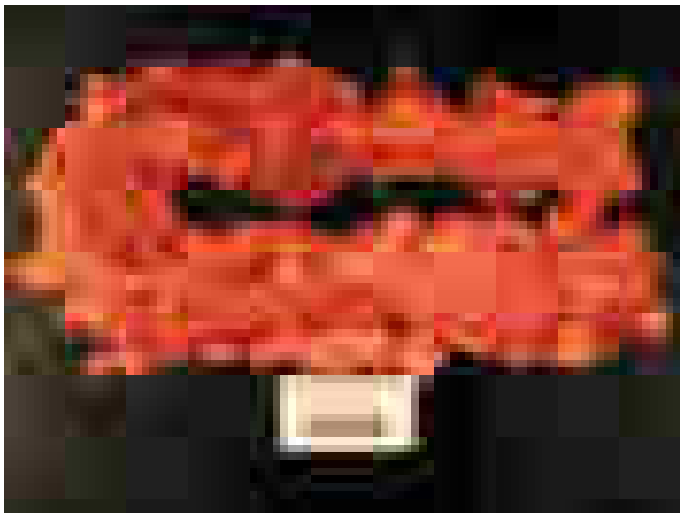


FIGURE 319-5 Crohn's disease of the colon showing thickening of the wall, with stenosis, linear serpiginous ulcers and cobblestoning of the mucosa. *(Courtesy of Dr. R Odze, Division of Gastrointestinal Pathology, Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts; with permission.)*

"cobblestone" appearance is characteristic of CD, both endoscopically and by barium radiography. As in UC, pseudopolyps can form in CD.

Active CD is characterized by focal inflammation and formation of fistula tracts, which resolve by fibrosis and stricturing of the bowel. The bowel wall thickens and becomes narrowed and fibrotic, leading to chronic, recurrent bowel obstructions. Projections of thickened mesentery encase the bowel ("creeping fat"), and serosal and mesenteric inflammation promotes adhesions and fistula formation.

CROHN'S DISEASE: MICROSCOPIC FEATURES

The earliest lesions are aphthoid ulcerations and focal crypt abscesses with loose aggregations of macrophages, which form noncaseating granulomas in all layers of the bowel wall **(Fig. 319-6)**. Granulomas can be seen in lymph nodes, mesentery, peritoneum, liver, and pancreas. Granulomas are a characteristic feature of CD. They are less commonly found on mucosal biopsies than on surgical resection specimens. Other histologic features of CD include submucosal or subserosal lymphoid aggregates, particularly away from areas of ulceration, gross and microscopic skip areas, and transmural inflammation that is accompanied by fissures that penetrate deeply into the bowel wall and sometimes form fistulous tracts or local abscesses.

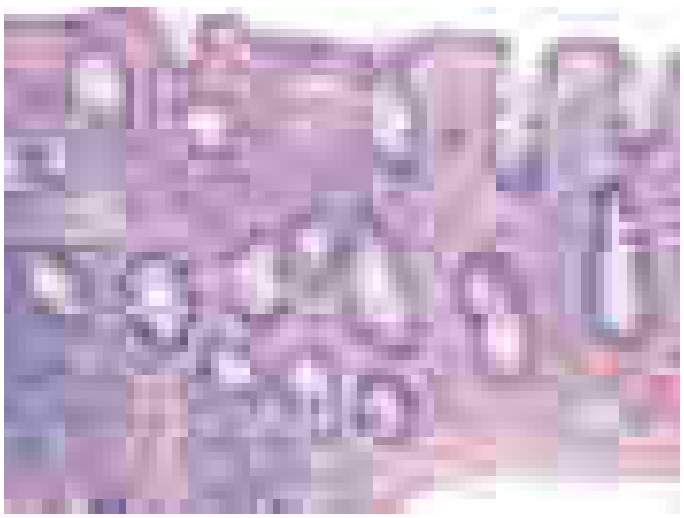


FIGURE 319-6 Medium-power view of Crohn's colitis showing mixed acute and chronic inflammation, crypt atrophy, and multiple small epithelioid granulomas in the mucosa. *(Courtesy of Dr. R Odze, Division of Gastrointestinal Pathology, Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts; with permission.)*

CLINICAL PRESENTATION

ULCERATIVE COLITIS

Signs and Symptoms The major symptoms of UC are diarrhea, rectal bleeding, tenesmus, passage of mucus, and crampy abdominal pain. The severity of symptoms correlates with the extent of disease. Although UC can present acutely, symptoms usually have been present for weeks to months. Occasionally, diarrhea and bleeding are so intermittent and mild that the patient does not seek medical attention.

Patients with proctitis usually pass fresh blood or blood-stained mucus, either mixed with stool or streaked onto the surface of a normal or hard stool. They also have tenesmus, or urgency with a feeling of incomplete evacuation, but rarely have abdominal pain. With proctitis or proctosigmoiditis, proximal transit slows, which may account for the constipation commonly seen in patients with distal disease.

When the disease extends beyond the rectum, blood is usually mixed with stool or grossly bloody diarrhea may be noted. Colonic motility is altered by inflammation with rapid transit through the inflamed intestine. When the disease is severe, patients pass a liquid stool containing blood, pus, and fecal matter. Diarrhea is often nocturnal and/or postprandial. Although severe pain is not a prominent symptom, some patients with active disease may experience lower abdominal discomfort or mild central abdominal cramping. Severe cramping and abdominal pain can occur with severe attacks of the disease. Other symptoms in moderate to severe disease include anorexia, nausea, vomiting, fever, and weight loss.

Physical signs of proctitis include a tender anal canal and blood on rectal examination. With more extensive disease, patients have tenderness to palpation directly over the colon. Patients with a toxic colitis have severe pain and bleeding, and those with megacolon have hepatic tympany. Both may have signs of peritonitis if a perforation has occurred. The classification of disease activity is shown in **Table 319-4**.

Laboratory, Endoscopic, and Radiographic Features Active disease can be associated with a rise in acute-phase reactants (C-reactive protein [CRP]), platelet count, and erythrocyte sedimentation rate (ESR), and a decrease in hemoglobin. Fecal lactoferrin, a glycoprotein present in activated neutrophils, is a highly sensitive and specific marker for detecting intestinal inflammation. Fecal calprotectin is present in neutrophils and monocytes and levels correlate well with histologic inflammation, predict relapses, and detect pouchitis. Both fecal lactoferrin and calprotectin are becoming an integral part of IBD management and are used frequently to rule out active inflammation versus symptoms of irritable bowel or bacterial overgrowth. In severely ill patients, the serum albumin level will fall rather quickly. Leukocytosis may be present but is not a specific indicator of disease

TABLE 319-4 Ulcerative Colitis: Disease Presentation

	MILD	MODERATE	SEVERE
Bowel movements	<4 per day	4–6 per day	>6 per day
Blood in stool	Small	Moderate	Severe
Fever	None	<37.5°C mean (<99.5°F)	>37.5°C mean (>99.5°F)
Tachycardia	None	<90 mean pulse	>90 mean pulse
Anemia	Mild	>75% of a normal hemoglobin	≤75% of a normal hemoglobin
Sedimentation rate	<30 mm		>30 mm
Endoscopic appearance	Erythema, decreased vascular pattern, fine granularity	Marked erythema, coarse granularity, absent vascular markings, contact bleeding, no ulcerations	Spontaneous bleeding, ulcerations

FIGURE 319-7 Colonoscopy with acute ulcerative colitis: severe colon inflammation with erythema, friability, and exudates. *(Courtesy of Dr. M. Hamilton, Gastroenterology Division, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts; with permission.)*

activity. Proctitis or proctosigmoiditis rarely causes a rise in CRP. Diagnosis relies on the patient's history; clinical symptoms; negative stool examination for bacteria, *C. difficile* toxin, and ova and parasites; sigmoidoscopic appearance **(see Fig. 315-4*A*)**; and histology of rectal or colonic biopsy specimens.

Sigmoidoscopy is used to assess disease activity and is usually performed before treatment. If the patient is not having an acute flare, colonoscopy is used to assess disease extent and activity **(Fig. 319-7)**. Endoscopically mild disease is characterized by erythema, decreased vascular pattern, and mild friability. Moderate disease is characterized by marked erythema, absent vascular pattern, friability and erosions, and severe disease by spontaneous bleeding and ulcerations. Histologic features change more slowly than clinical features but can also be used to grade disease activity.

Complications Only 15% of patients with UC present initially with catastrophic illness. Massive hemorrhage occurs with severe attacks of disease in 1% of patients, and treatment for the disease usually stops the bleeding. However, if a patient requires 6–8 units of blood within 24–48 h, colectomy is indicated. *Toxic megacolon* is defined as a transverse or right colon with a diameter of >6 cm, with loss of haustration in patients with severe attacks of UC. It occurs in about 5% of attacks and can be triggered by electrolyte abnormalities and narcotics. About 50% of acute dilations will resolve with medical therapy alone, but urgent colectomy is required for those that do not improve. Perforation is the most dangerous of the local complications, and the physical signs of peritonitis may not be obvious, especially if the patient is receiving glucocorticoids. Although perforation is rare, the mortality rate for perforation complicating a toxic megacolon is about 15%. In addition, patients can develop a toxic colitis and such severe ulcerations that the bowel may perforate without first dilating.

Strictures occur in 5–10% of patients and are always a concern in UC because of the possibility of underlying neoplasia. Although benign strictures can form from the inflammation and fibrosis of UC, strictures that are impassable with the colonoscope should be presumed malignant until proven otherwise. A stricture that prevents passage of the colonoscope is an indication for surgery. UC patients occasionally develop anal fissures, perianal abscesses, or hemorrhoids, but the occurrence of extensive perianal lesions should suggest CD.

CROHN'S DISEASE

Signs and Symptoms Although CD usually presents as acute or chronic bowel inflammation, the inflammatory process evolves toward one of two patterns of disease: a fibrostenotic obstructing pattern or a penetrating fistulous pattern, each with different treatments and prognoses. The site of disease influences the clinical manifestations.

ILEOCOLITIS Because the most common site of inflammation is the terminal ileum, the usual presentation of ileocolitis is a chronic history of recurrent episodes of right lower quadrant pain and diarrhea. Sometimes the initial presentation mimics acute appendicitis with pronounced right lower quadrant pain, a palpable mass, fever, and leukocytosis. Pain is usually colicky; it precedes and is relieved by defecation. A low-grade fever is usually noted. High-spiking fever suggests intraabdominal abscess formation. Weight loss is common—typically 10–20% of body weight—and develops as a consequence of diarrhea, anorexia, and fear of eating.

An inflammatory mass may be palpated in the right lower quadrant of the abdomen. The mass is composed of inflamed bowel, induration of the mesentery, and enlarged abdominal lymph nodes. Extension of the mass can cause obstruction of the right ureter or bladder inflammation, manifested by dysuria and fever. The ''string sign'' on barium studies results from a severely narrowed loop of bowel, which makes the lumen resemble a frayed cotton string. It is caused by incomplete filling of the lumen as the result of edema, irritability, and spasms associated with inflammation and ulcerations. The sign may be seen in both nonstenotic and stenotic phases of the disease.

Bowel obstruction may take several forms. In the early stages of disease, bowel wall edema and spasm produce intermittent obstructive manifestations and increasing symptoms of postprandial pain. Over several years, persistent inflammation gradually progresses to fibrostenotic narrowing and stricture. Diarrhea will decrease and be replaced by chronic bowel obstruction. Acute episodes of obstruction occur as well, precipitated by bowel inflammation and spasm or sometimes by impaction of undigested food or medication. These episodes usually resolve with intravenous fluids and gastric decompression.

Severe inflammation of the ileocecal region may lead to localized wall thinning, with microperforation and fistula formation to the adjacent bowel, the skin, or the urinary bladder, or to an abscess cavity in the mesentery. Enterovesical fistulas typically present as dysuria or recurrent bladder infections or, less commonly, as pneumaturia or fecaluria. Enterocutaneous fistulas follow tissue planes of least resistance, usually draining through abdominal surgical scars. Enterovaginal fistulas are rare and present as dyspareunia or as a feculent or foul-smelling, often painful vaginal discharge. They are unlikely to develop without a prior hysterectomy.

JEJUNOILEITIS Extensive inflammatory disease is associated with a loss of digestive and absorptive surface, resulting in malabsorption and steatorrhea. Nutritional deficiencies can also result from poor intake and enteric losses of protein and other nutrients. Intestinal malabsorption can cause anemia, hypoalbuminemia, hypocalcemia, hypomagnesemia, coagulopathy, and hyperoxaluria with nephrolithiasis in patients with an intact colon. Many patients need to take oral and often intravenous iron. Vertebral fractures are caused by a combination of vitamin D deficiency, hypocalcemia, and prolonged glucocorticoid use. Pellagra from niacin deficiency can occur in extensive small-bowel disease, and malabsorption of vitamin B_{12} can lead to megaloblastic anemia and neurologic symptoms. Other important nutrients to measure and replete if low are folate and vitamins A, E, and K. Levels of minerals such as zinc, selenium, copper, and magnesium are often low in patients with extensive small-bowel inflammation or resections, and these should be repleted as well. Most patients should take a daily multivitamin, calcium, and vitamin D supplements.

Diarrhea is characteristic of active disease; its causes include (1) bacterial overgrowth in obstructive stasis or fistulization, (2) bile-acid malabsorption due to a diseased or resected terminal ileum, and (3) intestinal inflammation with decreased water absorption and increased secretion of electrolytes.

COLITIS AND PERIANAL DISEASE Patients with colitis present with low-grade fevers, malaise, diarrhea, crampy abdominal pain, and sometimes hematochezia. Gross bleeding is not as common as in UC and appears in about one-half of patients with exclusively colonic disease. Only 1–2% exhibit massive bleeding. Pain is caused by passage of fecal

material through narrowed and inflamed segments of the large bowel. Decreased rectal compliance is another cause for diarrhea in Crohn's colitis patients. Toxic megacolon is rare but may be seen with severe inflammation and short duration disease.

Stricturing can occur in the colon in 4–16% of patients and produce symptoms of bowel obstruction. If the endoscopist is unable to traverse a stricture in Crohn's colitis, surgical resection should be considered, especially if the patient has symptoms of chronic obstruction. Colonic disease may fistulize into the stomach or duodenum, causing feculent vomiting, or to the proximal or mid-small bowel, causing malabsorption by "short circuiting" and bacterial overgrowth. Ten percent of women with Crohn's colitis will develop a rectovaginal fistula.

Perianal disease affects about one-third of patients with Crohn's colitis and is manifested by incontinence, large hemorrhoidal tags, anal strictures, anorectal fistulae, and perirectal abscesses. Not all patients with perianal fistula will have endoscopic evidence of colonic inflammation.

GASTRODUODENAL DISEASE Symptoms and signs of upper GI tract disease include nausea, vomiting, and epigastric pain. Patients usually have an *Helicobacter pylori*–negative gastritis. The second portion of the duodenum is more commonly involved than the bulb. Fistulas involving the stomach or duodenum arise from the small or large bowel and do not necessarily signify the presence of upper GI tract involvement. Patients with advanced gastroduodenal CD may develop a chronic gastric outlet obstruction. About 30% of children diagnosed with CD have esophagogastroduodenal involvement.

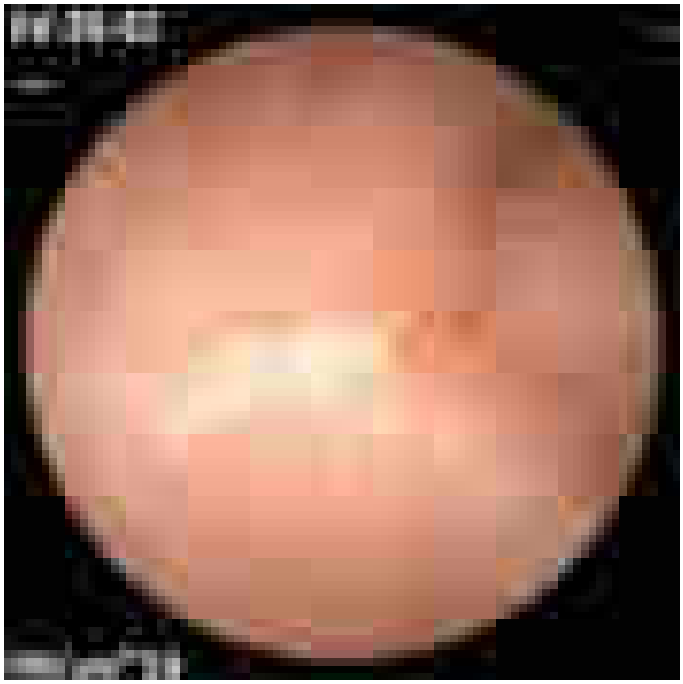


FIGURE 319-8 Wireless capsule endoscopy image in a patient with Crohn's disease of the ileum shows ulcerations and narrowing of the intestinal lumen. *(Courtesy of Dr. S. Reddy, Gastroenterology Division, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts; with permission.)*

Laboratory, Endoscopic, and Radiographic Features

Laboratory abnormalities include elevated ESR and CRP. In more severe disease, findings include hypoalbuminemia, anemia, and leukocytosis. Fecal calprotectin and lactoferrin levels have been used to distinguish IBD from irritable bowel syndrome (IBS), assess whether CD is active, and to detect postoperative recurrence of CD.

Endoscopic features of CD include rectal sparing, aphthous ulcerations, fistulas, and skip lesions. Colonoscopy allows examination and biopsy of mass lesions or strictures and biopsy of the terminal ileum. Upper endoscopy is useful in diagnosing gastroduodenal involvement in patients with upper tract symptoms. Ileal or colonic strictures may be dilated with balloons introduced through the colonoscope. Strictures ≤4 cm in length and those at anastomotic sites respond better to endoscopic dilation. The perforation rate is as high as 10%. Most endoscopists dilate only fibrotic strictures and not those associated with active inflammation. Wireless capsule endoscopy (WCE) allows direct visualization of the entire small-bowel mucosa **(Fig. 319-8)**. The diagnostic yield of detecting lesions suggestive of active CD is higher with WCE than CT or magnetic resonance (MR) enterography or small-bowel series. WCE cannot be used in the setting of a small-bowel stricture. Capsule retention occurs in <1% of patients with suspected CD, but retention rates of 4–6% are seen in patients with established CD. It is helpful to give the patient with CD a patency capsule, which is made of barium and starts to dissolve 30 h after ingestion. An abdominal x-ray can be taken at around 30 h after ingestion to see if the capsule is still present in the small bowel, which would indicate a stricture.

In CD, early radiographic findings in the small bowel include thickened folds and aphthous ulcerations. "Cobblestoning" from longitudinal and transverse ulcerations most frequently involves the small bowel. In more advanced disease, strictures, fistulas, inflammatory masses, and abscesses may be detected. The earliest macroscopic findings of colonic CD are aphthous ulcers. These small ulcers are often multiple and separated by normal intervening mucosa. As the disease progresses, aphthous ulcers become enlarged, deeper, and occasionally connected to one another, forming longitudinal stellate, serpiginous, and linear ulcers **(see Fig. 315-4*B*)**.

The transmural inflammation of CD leads to decreased luminal diameter and limited distensibility. As ulcers progress deeper, they can lead to fistula formation. The segmental nature of CD results in wide gaps of normal or dilated bowel between involved segments.

Although CT enterography (CTE), MR enterography (MRE), and small-bowel follow-through (SBFT) have been shown to be equally accurate in the identification of active small-bowel inflammation, CTE and MRE have been shown to be superior to SBFT in the detection of extraluminal complications, including fistulas, sinus tracts, and abscesses. MRI is thought to offer superior soft tissue contrast and has the added advantage of avoiding radiation exposure changes **(Figs. 319-9 and 319-10)**. The lack of ionizing radiation is particularly appealing in younger patients and when monitoring response to therapy where serial images will be obtained. Ultrasound is becoming increasingly more popular, especially in Europe, for measuring CD extent and activity. Pelvic MRI is superior to CT for demonstrating pelvic lesions such as ischiorectal abscesses and perianal fistulae **(Fig. 319-11)**.

Complications Because CD is a transmural process, serosal adhesions develop that provide direct pathways for fistula formation and reduce the incidence of free perforation. Perforation occurs in 1–2% of patients, usually in the ileum but occasionally in the jejunum or as a complication of toxic megacolon. The peritonitis of free perforation, especially colonic, may be fatal. Intraabdominal and pelvic abscesses occur in 10–30% of patients with CD at some time in the course of their illness. CT-guided percutaneous drainage of the abscess is standard therapy. Despite adequate drainage, most patients need resection of the offending bowel segment. Percutaneous drainage has an especially high failure rate in abdominal wall abscesses. Systemic glucocorticoid therapy increases the risk of intraabdominal and pelvic abscesses in CD patients who have never had an operation. Other complications include intestinal obstruction in 40%, massive hemorrhage, malabsorption, and severe perianal disease.

Serologic Markers Patients with CD show a wide variation in the way they present and progress over time. Some patients present with mild disease activity and do well with generally safe and mild medications, but many others exhibit more severe disease and can develop serious complications that will require surgery. Current and developing biologic therapies can help halt progression of disease and give patients with moderate to severe CD a better quality of life. There are potential risks of biologic therapies such as infection and malignancy, and it would be optimal to determine by genetic or serologic markers at the time of diagnosis which patients will require more aggressive medical therapy. This same argument holds true for UC patients as well.

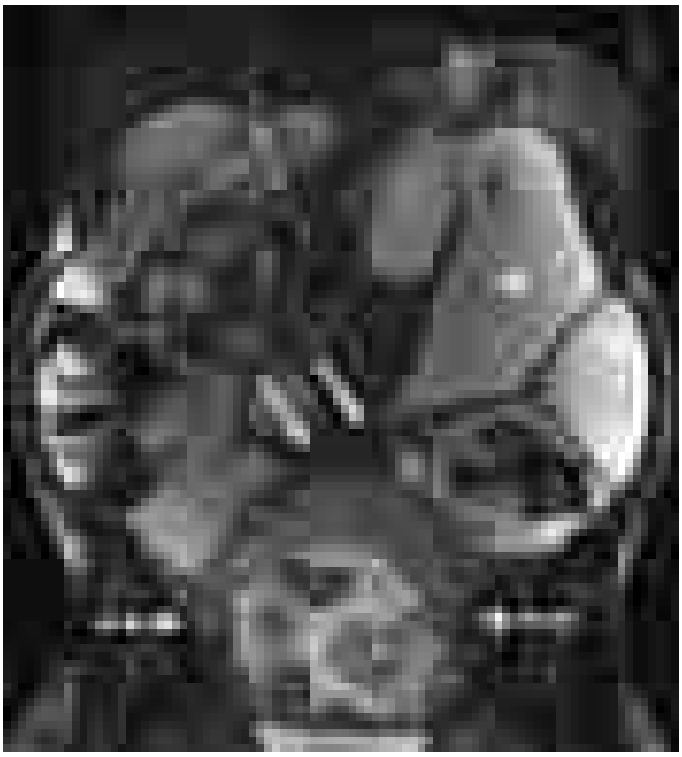


FIGURE 319-9 A coronal magnetic resonance image was obtained using a half Fourier single-shot T2-weighted acquisition with fat saturation in a 27-year-old pregnant (23 weeks' gestation) woman. The patient had Crohn's disease and was maintained on 6-mercaptopurine and prednisone. She presented with abdominal pain, distension, vomiting, and small-bowel obstruction. The image reveals a 7- to 10-cm long stricture at the terminal ileum (*white arrows*) causing obstruction and significant dilatation of the proximal small bowel (*white asterisk*). A fetus is seen in the uterus (*dashed white arrows*). *(Courtesy of Drs. J. F. B. Chick and P. B. Shyn, Abdominal Imaging and Intervention, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; with permission.)*

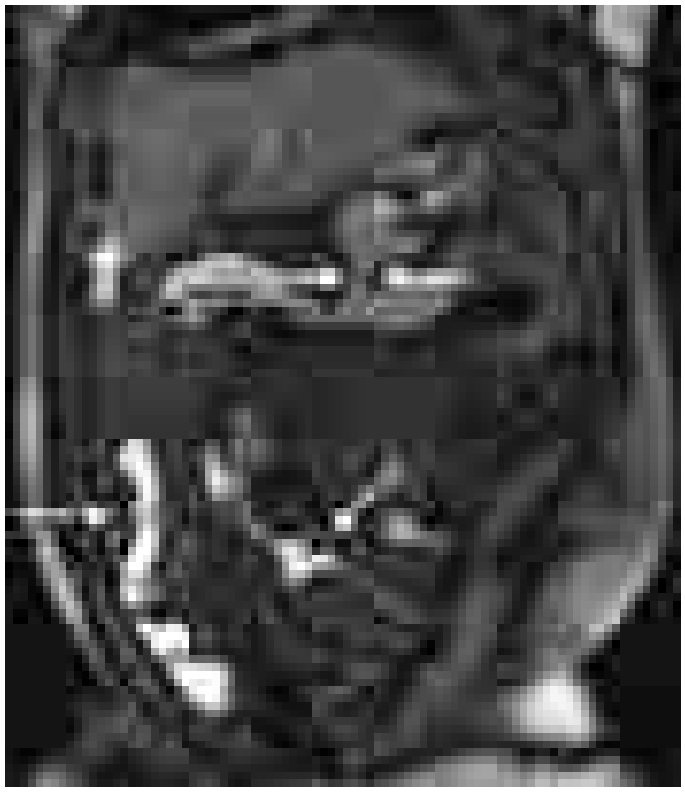


FIGURE 319-10 A coronal balanced, steady-state, free precession, T2-weighted image with fat saturation was obtained in a 32-year-old man with Crohn's disease and prior episodes of bowel obstruction, fistulas, and abscesses. He was being treated with 6-mercaptopurine and presented with abdominal distention and diarrhea. The image demonstrates a new gastrocolic fistula (*solid white arrows*). Multifocal involvement of the small bowel and terminal ileum is also present (*dashed white arrows*). *(Courtesy of Drs. J. F. B. Chick and P. B. Shyn, Abdominal Imaging and Intervention, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; with permission.)*

For success in diagnosing IBD and in differentiating between CD and UC, the efficacy of these serologic tests depends on the prevalence of IBD in a specific population. Increased titers of anti-Saccharomyces cerevisiae antibodies (ASCAs) have been associated with CD, whereas increased levels of perinuclear antineutrophil cytoplasmic antibodies (pANCA) are more commonly seen in patients with UC. However, when evaluated in a meta-analysis of 60 studies, the sensitivity and specificity of a ASCA+/pANCA- pattern for identification of CD was 55% and 93% respectively. In addition to ASCA, multiple other antibodies to bacterial proteins (Omp-C and I2), flagellin (CBir1) and bacterial carbohydrates have been studied and associated with CD, including laminaribioside (ALCA), chitobioside (ACCA) and mannobioside (SMCA). These serologic markers tend to have low sensitivity and specificity though due to elevation in levels cause by other autoimmune diseases, infections and inflammation outside the GI tract.

Clinical factors described at diagnosis are more helpful than serologies at predicting the natural history of CD. The initial requirements for glucocorticoid use, an age at diagnosis below 40 years and the presence of perianal disease at diagnosis, have been shown to be independently associated with subsequent disabling CD after 5 years. Except in special circumstances (such as before consideration of an ileoanal pouch anastomosis [IPAA] in a patient with indeterminate colitis), serologic markers have only minimal clinical utility.

DIFFERENTIAL DIAGNOSIS OF UC AND CD

UC and CD have similar features to many other diseases. In the absence of a key diagnostic test, a combination of features is used **(Table 319-5)**. Once a diagnosis of IBD is made, distinguishing between UC and CD is impossible initially in up to 15% of cases. These are termed *indeterminate colitis.* Fortunately, in most cases, the true nature of the underlying colitis becomes evident later in the course of the patient's disease. Approximately 5% (range 1–20%) of colon resection specimens are difficult to classify as either UC or CD because they exhibit overlapping histologic features.

INFECTIOUS DISEASES

Infections of the small intestines and colon can mimic CD or UC. They may be bacterial, fungal, viral, or protozoal in origin **(Table 319-6)**. *Campylobacter* colitis can mimic the endoscopic appearance of severe UC and can cause a relapse of established UC. *Salmonella* can cause watery or bloody diarrhea, nausea, and vomiting. Shigellosis causes watery diarrhea, abdominal pain, and fever followed by rectal tenesmus and by the passage of blood and mucus per rectum. All three are usually self-limited, but 1% of patients infected with *Salmonella* become asymptomatic carriers. *Yersinia enterocolitica* infection occurs mainly in the terminal ileum and causes mucosal ulceration, neutrophil invasion, and thickening of the ileal wall. Other bacterial infections that may mimic IBD include *C. difficile*, which presents with watery diarrhea, tenesmus, nausea, and vomiting; and *E. coli*, three categories of which can cause colitis. These are enterohemorrhagic, enteroinvasive, and enteroadherent *E. coli*, all of which can cause bloody diarrhea and abdominal tenderness. Diagnosis of bacterial colitis is made by sending stool specimens for bacterial culture and *C. difficile* toxin analysis. Gonorrhea, *Chlamydia*, and syphilis can also cause proctitis.

GI involvement with mycobacterial infection occurs primarily in the immunosuppressed patient but may occur in patients with normal immunity. Distal ileal and cecal involvement predominates, and patients present with symptoms of small-bowel obstruction and a tender abdominal mass. The diagnosis is made most directly by colonoscopy with

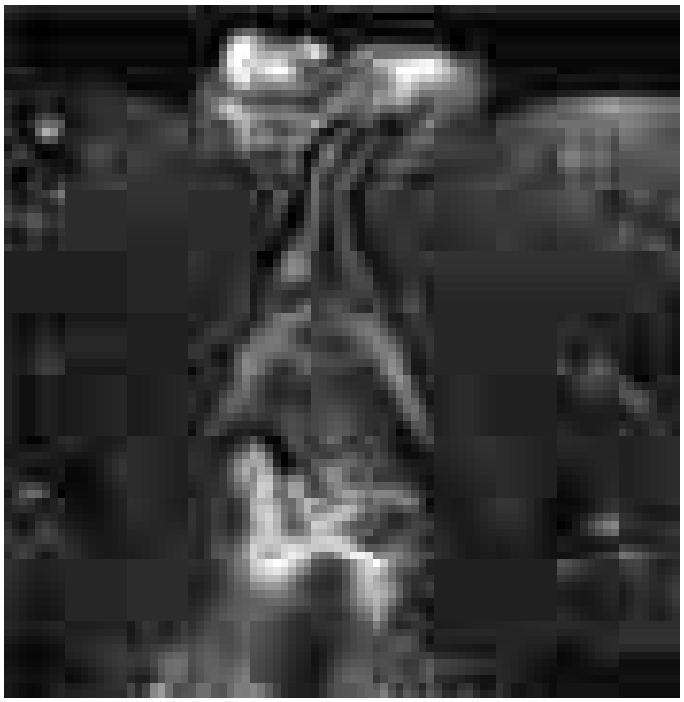


FIGURE 319-11 Axial T2-weighted fat-saturated image obtained in a 39-year-old male with Crohn's disease shows a defect in the internal sphincter at the 6:00 position of the mid anal canal (*open white arrow*) communicating with a 1.1-cm intersphincteric collection (*black arrow*). Wide defect in the external sphincter at the 7:00 position *(solid white arrow)* leads to a moderate sized perianal abscess in the ischioanal fossa (*asterisk*). *(Courtesy of Drs. J.S. Quon and P.B. Shyn, Abdominal Imaging and Intervention, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; with permission.)*

TABLE 319-6 Diseases That Mimic IBD

Infectious Etiologies		
Bacterial	**Mycobacterial**	**Viral**
Salmonella	Tuberculosis	Cytomegalovirus
Shigella	*Mycobacterium avium*	Herpes simplex
Toxigenic *Escherichia coli*		HIV
Campylobacter	**Parasitic**	**Fungal**
Yersinia	Amebiasis	Histoplasmosis
Clostridium difficile	*Isospora*	*Candida*
Gonorrhea	*Trichuris trichiura*	*Aspergillus*
Chlamydia trachomatis	Hookworm	
	Strongyloides	

Noninfectious Etiologies		
Inflammatory	**Neoplastic**	**Drugs and Chemicals**
Appendicitis	Lymphoma	NSAIDs
Diverticulitis	Metastatic	Phosphosoda
Diversion colitis	Carcinoma	Cathartic colon
Collagenous/ lymphocytic colitis	Carcinoma of the ileum	Gold
Ischemic colitis	Carcinoid	Oral contraceptives
Radiation colitis/ enteritis	Familial polyposis	Cocaine
Solitary rectal ulcer syndrome		Ipilimumab
Eosinophilic gastroenteritis		Mycophenolate mofetil
Neutropenic colitis		
Behçet's syndrome		
Graft-versus-host disease		

Abbreviation: IBD, inflammatory bowel disease; NSAIDs, nonsteroidal anti-inflammatory drugs.

TABLE 319-5 Different Clinical, Endoscopic, and Radiographic Features

	ULCERATIVE COLITIS	CROHN'S DISEASE
Clinical		
Gross blood in stool	Yes	Occasionally
Mucus	Yes	Occasionally
Systemic symptoms	Occasionally	Frequently
Pain	Occasionally	Frequently
Abdominal mass	Rarely	Yes
Significant perineal disease	No	Frequently
Fistulas	No	Yes
Small intestinal obstruction	No	Frequently
Colonic obstruction	Rarely	Frequently
Response to antibiotics	No	Yes
Recurrence after surgery	No	Yes
Endoscopic		
Rectal sparing	Rarely	Frequently
Continuous disease	Yes	Occasionally
"Cobblestoning"	No	Yes
Granuloma on biopsy	No	Occasionally
Radiographic		
Small bowel significantly abnormal	No	Yes
Abnormal terminal ileum	No	Yes
Segmental colitis	No	Yes
Asymmetric colitis	No	Yes
Stricture	Occasionally	Frequently

biopsy and culture. *Mycobacterium avium-intracellulare* complex infection occurs in advanced stages of HIV infection and in other immunocompromised states; it usually manifests as a systemic infection with diarrhea, abdominal pain, weight loss, fever, and malabsorption. Diagnosis is established by acid-fast smear and culture of mucosal biopsies.

Although most of the patients with viral colitis are immunosuppressed, cytomegalovirus (CMV) and herpes simplex proctitis may occur in immunocompetent individuals. CMV occurs most commonly in the esophagus, colon, and rectum but may also involve the small intestine. Symptoms include abdominal pain, bloody diarrhea, fever, and weight loss. With severe disease, necrosis and perforation can occur. Diagnosis is made by identification of characteristic intranuclear inclusions in mucosal cells on biopsy. Herpes simplex infection of the GI tract is limited to the oropharynx, anorectum, and perianal areas. Symptoms include anorectal pain, tenesmus, constipation, inguinal adenopathy, difficulty with urinary voiding, and sacral paresthesias. Diagnosis is made by rectal biopsy with identification of characteristic cellular inclusions and viral culture. HIV itself can cause diarrhea, nausea, vomiting, and anorexia. Small intestinal biopsies show partial villous atrophy; small bowel bacterial overgrowth and fat malabsorption may also be noted.

Protozoan parasites include *Isospora belli*, which can cause a self-limited infection in healthy hosts but causes a chronic profuse, watery diarrhea, and weight loss in AIDS patients. *Entamoeba histolytica* or related species infect about 10% of the world's population; symptoms include abdominal pain, tenesmus, frequent loose stools containing blood and mucus, and abdominal tenderness. Colonoscopy reveals focal punctate ulcers with normal intervening mucosa; diagnosis is made by biopsy or serum amebic antibodies. Fulminant amebic colitis is rare but has a mortality rate of >50%.

Other parasitic infections that may mimic IBD include hookworm (*Necator americanus*), whipworm (*Trichuris trichiura*), and *Strongyloides*

stercoralis. In severely immunocompromised patients, *Candida* or *Aspergillus* can be identified in the submucosa. Disseminated histoplasmosis can involve the ileocecal area.

NONINFECTIOUS DISEASES

Diverticulitis can be confused with CD clinically and radiographically. Both diseases cause fever, abdominal pain, tender abdominal mass, leukocytosis, elevated ESR, partial obstruction, and fistulas. Perianal disease or ileitis on small-bowel series favors the diagnosis of CD. Significant endoscopic mucosal abnormalities are more likely in CD than in diverticulitis. Endoscopic or clinical recurrence following segmental resection favors CD. Diverticular-associated colitis is similar to CD, but mucosal abnormalities are limited to the sigmoid and descending colon.

Ischemic colitis is commonly confused with IBD. The ischemic process can be chronic and diffuse, as in UC, or segmental, as in CD. Colonic inflammation due to ischemia may resolve quickly or may persist and result in transmural scarring and stricture formation. Ischemic bowel disease should be considered in the elderly following abdominal aortic aneurysm repair or when a patient has a hypercoagulable state or a severe cardiac or peripheral vascular disorder. Patients usually present with sudden onset of left lower quadrant pain, urgency to defecate, and the passage of bright red blood per rectum. Endoscopic examination often demonstrates a normal-appearing rectum and a sharp transition to an area of inflammation in the descending colon and splenic flexure.

The effects of radiotherapy on the GI tract can be difficult to distinguish from IBD. Acute symptoms can occur within 1–2 weeks of starting radiotherapy. When the rectum and sigmoid are irradiated, patients develop bloody, mucoid diarrhea and tenesmus, as in distal UC. With small-bowel involvement, diarrhea is common. Late symptoms include malabsorption and weight loss. Stricturing with obstruction and bacterial overgrowth may occur. Fistulas can penetrate the bladder, vagina, or abdominal wall. Flexible sigmoidoscopy reveals mucosal granularity, friability, numerous telangiectasias, and occasionally discrete ulcerations. Biopsy can be diagnostic.

Solitary rectal ulcer syndrome is uncommon and can be confused with IBD. It occurs in persons of all ages and may be caused by impaired evacuation and failure of relaxation of the puborectalis muscle. Single or multiple ulcerations may arise from anal sphincter overactivity, higher intrarectal pressures during defecation, and digital removal of stool. Patients complain of constipation with straining and pass blood and mucus per rectum. Other symptoms include abdominal pain, diarrhea, tenesmus, and perineal pain. Ulceration, that may be as large as 5 cm in diameter, is usually observed anteriorly or anterolaterally 3–15 cm from the anal verge. Biopsies can be diagnostic.

Several types of colitis are associated with nonsteroidal anti-inflammatory drugs (NSAIDs), including de novo colitis, reactivation of IBD, and proctitis caused by use of suppositories. Most patients with NSAID-related colitis present with diarrhea and abdominal pain, and complications include stricture, bleeding, obstruction, perforation, and fistulization. Withdrawal of these agents is crucial, and in cases of reactivated IBD, standard therapies are indicated.

There are complications of two common drugs used in a hospital setting that mimic IBD. The first is ipilimumab, a drug that targets cytotoxic T lymphocyte antigen 4 (CTLA-4) and reverses T cell inhibition and is used to treat metastatic melanoma. Ipilimumab can cause an autoimmune colitis that is commonly associated with diarrhea: patients with diarrhea of grade 3 or greater and those who have colitis on colonoscopy often require glucocorticoid or infliximab therapy. The second is mycophenolate mofetil (MMF), an immunosuppressive agent that is anti-proliferative and commonly used to prevent post-transplant rejection. The colitis associated with MMF is common and can occur in more than one-third of patients taking the drug. Treatment is dose reduction or cessation of the drug. There have been case reports of entanercept (TNF receptor-Fc fusion protein) associated with de novo CD and UC.

THE ATYPICAL COLITIDES

Two atypical colitides—collagenous colitis and lymphocytic colitis—have completely normal endoscopic appearances. Collagenous colitis has two main histologic components: increased subepithelial collagen deposition and colitis with increased intraepithelial lymphocytes. The female to male ratio is 9:1, and most patients present in the sixth or seventh decades of life. The main symptom is chronic watery diarrhea. Treatments range from sulfasalazine or mesalamine and diphenoxylate/atropine (Lomotil) to bismuth to budesonide to prednisone or azathioprine/6-mercaptopurine for refractory disease. Risk factors include smoking, use of NSAIDs, proton pump inhibitors, or beta blockers; and a history of autoimmune disease.

Lymphocytic colitis has features similar to collagenous colitis, including age at onset and clinical presentation, but it has almost equal incidence in men and women and no subepithelial collagen deposition on pathologic section. However, intraepithelial lymphocytes are increased. Use of sertraline (but not beta blockers) is an additional risk factor. The frequency of celiac disease is increased in lymphocytic colitis and ranges from 9 to 27%. Celiac disease should be excluded in all patients with lymphocytic colitis, particularly if diarrhea does not respond to conventional therapy. Treatment is similar to that of collagenous colitis with the exception of a gluten-free diet for those who have celiac disease.

Diversion colitis is an inflammatory process that arises in segments of the large intestine that are excluded from the fecal stream. It usually occurs in patients with ileostomy or colostomy when a mucus fistula or a Hartmann's pouch has been created. Clinically, patients have mucus or bloody discharge from the rectum. Erythema, granularity, friability, and, in more severe cases, ulceration can be seen on endoscopy. Histopathology shows areas of active inflammation with foci of cryptitis and crypt abscesses. Crypt architecture is normal, which differentiates it from UC. It may be impossible to distinguish from CD. Short-chain fatty acid enemas may help in diversion colitis, but the definitive therapy is surgical reanastomosis.

EXTRAINTESTINAL MANIFESTATIONS

Up to one-third of IBD patients have at least one extraintestinal disease manifestation.

DERMATOLOGIC

Erythema nodosum (EN) occurs in up to 15% of CD patients and 10% of UC patients. Attacks usually correlate with bowel activity; skin lesions develop after the onset of bowel symptoms, and patients frequently have concomitant active peripheral arthritis. The lesions of EN are hot, red, tender nodules measuring 1–5 cm in diameter and are found on the anterior surface of the lower legs, ankles, calves, thighs, and arms. Therapy is directed toward the underlying bowel disease.

Pyoderma gangrenosum (PG) is seen in 1–12% of UC patients and less commonly in Crohn's colitis. Although it usually presents after the diagnosis of IBD, PG may occur years before the onset of bowel symptoms, run a course independent of the bowel disease, respond poorly to colectomy, and even develop years after proctocolectomy. It is usually associated with severe disease. Lesions are commonly found on the dorsal surface of the feet and legs but may occur on the arms, chest, stoma, and even the face. PG usually begins as a pustule and then spreads concentrically to rapidly undermine healthy skin. Lesions then ulcerate, with violaceous edges surrounded by a margin of erythema. Centrally, they contain necrotic tissue with blood and exudates. Lesions may be single or multiple and grow as large as 30 cm. They are sometimes very difficult to treat and often require IV antibiotics, IV glucocorticoids, dapsone, azathioprine, thalidomide, IV cyclosporine (CSA), infliximab or adalimumab.

Other dermatologic manifestations include pyoderma vegetans, which occurs in intertriginous areas; pyostomatitis vegetans, which involves the mucous membranes; Sweet syndrome, a neutrophilic dermatosis; and metastatic CD, a rare disorder defined by cutaneous granuloma formation. Psoriasis affects 5–10% of patients with IBD and is unrelated to bowel activity consistent with the potential shared immunogenetic basis of these diseases. Perianal skin tags are found in 75–80% of patients with CD, especially those with colon involvement. Oral mucosal lesions, seen often in CD and rarely in UC, include aphthous stomatitis and "cobblestone" lesions of the buccal mucosa.

RHEUMATOLOGIC

Peripheral arthritis develops in 15–20% of IBD patients, is more common in CD, and worsens with exacerbations of bowel activity. It is asymmetric, polyarticular, and migratory and most often affects large joints of the upper and lower extremities. Treatment is directed at reducing bowel inflammation. In severe UC, colectomy frequently cures the arthritis.

Ankylosing spondylitis (AS) occurs in about 10% of IBD patients and is more common in CD than UC. About two-thirds of IBD patients with AS express the HLA-B27 antigen. The AS activity is not related to bowel activity and does not remit with glucocorticoids or colectomy. It most often affects the spine and pelvis, producing symptoms of diffuse low-back pain, buttock pain, and morning stiffness. The course is continuous and progressive, leading to permanent skeletal damage and deformity. Anti-TNF therapy reduces spinal inflammation and improves functional status and quality of life.

Sacroiliitis is symmetric, occurs equally in UC and CD, is often asymptomatic, does not correlate with bowel activity, and does not always progress to AS. Other rheumatic manifestations include hypertrophic osteoarthropathy, pelvic/femoral osteomyelitis, and relapsing polychondritis.

OCULAR

The incidence of ocular complications in IBD patients is 1–10%. The most common are conjunctivitis, anterior uveitis/iritis, and episcleritis. Uveitis is associated with both UC and Crohn's colitis, may be found during periods of remission, and may develop in patients following bowel resection. Symptoms include ocular pain, photophobia, blurred vision, and headache. Prompt intervention, sometimes with systemic glucocorticoids, is required to prevent scarring and visual impairment. Episcleritis is a benign disorder that presents with symptoms of mild ocular burning. It occurs in 3–4% of IBD patients, more commonly in Crohn's colitis, and is treated with topical glucocorticoids.

HEPATOBILIARY

Hepatic steatosis is detectable in about one-half of the abnormal liver biopsies from patients with CD and UC; patients usually present with hepatomegaly. Fatty liver usually results from a combination of chronic debilitating illness, malnutrition, and glucocorticoid therapy. Cholelithiasis occurs in 10–35% of CD patients with ileitis or ileal resection. Gallstone formation is caused by malabsorption of bile acids, resulting in depletion of the bile salt pool and the secretion of lithogenic bile.

Primary sclerosing cholangitis (PSC) is a disorder characterized by both intrahepatic and extrahepatic bile duct inflammation and fibrosis, frequently leading to biliary cirrhosis and hepatic failure; ~5% of patients with UC have PSC, but 50–75% of patients with PSC have IBD. PSC occurs less often in patients with CD. Although it can be recognized after the diagnosis of IBD, PSC can be detected earlier or even years after proctocolectomy. Consistent with this, the immunogenetic basis for PSC appears to be overlapping but distinct from UC based on GWAS, although both IBD and PSC are commonly pANCA positive. Most patients have no symptoms at the time of diagnosis; when symptoms are present, they consist of fatigue, jaundice, abdominal pain, fever, anorexia, and malaise. The traditional gold standard diagnostic test is endoscopic retrograde cholangiopancreatography (ERCP), but magnetic resonance cholangiopancreatography (MRCP) is sensitive, specific and safer. MRCP is reasonable as an initial diagnostic test in children and adults and can visualize irregularities, multifocal strictures, and dilatations of all levels of the biliary tree. In patients with PSC, both ERCP and MRCP demonstrate multiple bile duct strictures alternating with relatively normal segments.

Gallbladder polyps in patients with PSC have a high incidence of malignancy and cholecystectomy is recommended, even if a mass lesion is less than 1 cm in diameter. Gallbladder surveillance with ultrasound should be performed annually. Endoscopic stenting may be palliative for cholestasis secondary to bile duct obstruction. Patients with symptomatic disease develop cirrhosis and liver failure over 5–10 years and eventually require liver transplantation. PSC patients have a 10–15% lifetime risk of developing cholangiocarcinoma and then cannot be transplanted. Patients with IBD and PSC are at increased risk of colon cancer and should be surveyed yearly by colonoscopy and biopsy.

In addition, cholangiography is normal in a small percentage of patients who have a variant of PSC known as *small duct primary sclerosing cholangitis*. This variant (sometimes referred to as "pericholangitis") is probably a form of PSC involving small-caliber bile ducts. It has similar biochemical and histologic features to classic PSC. It appears to have a significantly better prognosis than classic PSC, although it may evolve into classic PSC. Granulomatous hepatitis and hepatic amyloidosis are much rarer extraintestinal manifestations of IBD.

UROLOGIC

The most frequent genitourinary complications are calculi, ureteral obstruction, and ileal bladder fistulas. The highest frequency of nephrolithiasis (10–20%) occurs in patients with CD following small bowel resection. Calcium oxalate stones develop secondary to hyperoxaluria, which results from increased absorption of dietary oxalate. Normally, dietary calcium combines with luminal oxalate to form insoluble calcium oxalate, which is eliminated in the stool. In patients with ileal dysfunction, however, nonabsorbed fatty acids bind calcium and leave oxalate unbound. The unbound oxalate is then delivered to the colon, where it is readily absorbed, especially in the presence of inflammation.

METABOLIC BONE DISORDERS

Low bone mass occurs in 14–42% of IBD patients. The risk is increased by glucocorticoids, CSA, methotrexate (MTX), and total parenteral nutrition (TPN). Malabsorption and inflammation mediated by IL-1, IL-6, TNF, and other inflammatory mediators also contribute to low bone density. An increased incidence of hip, spine, wrist, and rib fractures has been noted: 36% in CD and 45% in UC. The absolute risk of an osteoporotic fracture is about 1% per person per year. Fracture rates, particularly in the spine and hip, are highest among the elderly (age >60). One study noted an OR of 1.72 for vertebral fracture and an OR of 1.59 for hip fracture. The disease severity predicted the risk of a fracture. Only 13% of IBD patients who had a fracture were on any kind of antifracture treatment. Up to 20% of bone mass can be lost per year with chronic glucocorticoid use. The effect is dosage-dependent. Budesonide may also suppress the pituitary-adrenal axis and thus carries a risk of causing osteoporosis.

Osteonecrosis is characterized by death of osteocytes and adipocytes and eventual bone collapse. The pain is aggravated by motion and swelling of the joints. It affects the hips more often than knees and shoulders, and in one series, 4.3% of patients developed osteonecrosis within 6 months of starting glucocorticoids. Diagnosis is made by bone scan or MRI, and treatment consists of pain control, cord decompression, osteotomy, and joint replacement.

THROMBOEMBOLIC DISORDERS

Patients with IBD have an increased risk of both venous and arterial thrombosis even if the disease is not active. Factors responsible for the hypercoagulable state have included abnormalities of the platelet-endothelial interaction, hyperhomocysteinemia, alterations in the coagulation cascade, impaired fibrinolysis, involvement of tissue factor-bearing microvesicles, disruption of the normal coagulation system by autoantibodies, and a genetic predisposition. A spectrum of vasculitides involving small, medium, and large vessels has also been observed.

OTHER DISORDERS

More common cardiopulmonary manifestations include endocarditis, myocarditis, pleuropericarditis, and interstitial lung disease. A secondary or reactive amyloidosis can occur in patients with long-standing IBD, especially in patients with CD. Amyloid material is deposited systemically and can cause diarrhea, constipation, and renal failure. The renal disease can be successfully treated with colchicine. Pancreatitis is a rare extraintestinal manifestation of IBD and results from duodenal fistulas; ampullary CD; gallstones; PSC; drugs such as 6-mercaptopurine, azathioprine, or, very rarely, 5-ASA agents; autoimmune pancreatitis; and primary CD of the pancreas.

Inflammatory Bowel Disease

5-ASA AGENTS

These agents are effective at inducing and maintaining remission in UC. They may have a limited role in inducing remission in CD but no clear role in maintenance of CD. Newer sulfa-free aminosalicylate preparations deliver increased amounts of the pharmacologically active ingredient of sulfasalazine (5-ASA, mesalamine) to the site of active bowel disease while limiting systemic toxicity. Peroxisome proliferator activated receptor γ (PPAR-γ) may mediate 5-ASA therapeutic action by decreasing nuclear localization of NF-κB. Sulfa-free aminosalicylate formulations include alternative azo-bonded carriers, 5-ASA dimers, and delayed-release and controlled-release preparations. Each has the same efficacy as sulfasalazine when equimolar concentrations are used.

Sulfasalazine's molecular structure provides a convenient delivery system to the colon by allowing the intact molecule to pass through the small intestine after only partial absorption and to be broken down in the colon by bacterial azo reductases that cleave the azo bond linking the sulfa and 5-ASA moieties. Sulfasalazine is effective treatment for mild to moderate UC and is occasionally used in Crohn's colitis, but its high rate of side effects limits its use. Although sulfasalazine is more effective at higher doses, at 6 or 8 g/d up to 30% of patients experience allergic reactions or intolerable side effects such as headache, anorexia, nausea, and vomiting that are attributable to the sulfapyridine moiety. Hypersensitivity reactions, independent of sulfapyridine levels, include rash, fever, hepatitis, agranulocytosis, hypersensitivity pneumonitis, pancreatitis, worsening of colitis, and reversible sperm abnormalities. Sulfasalazine can also impair folate absorption, and patients should be given folic acid supplements.

Balsalazide contains an azo bond binding mesalamine to the carrier molecule 4-aminobenzoyl-β-alanine; it is effective in the colon.

Delzicol and Asacol HD (high dose) are enteric-coated forms of mesalamine with the 5-ASA being released at pH >7. They disintegrate with complete breakup of the tablet occurring in many different parts of the gut ranging from the small intestine to the splenic flexure; they have increased gastric residence when taken with a meal. Asacol has been discontinued and replaced with Delzicol, which lacks dibutyl phthalate (DBP), an inactive ingredient in Asacol's enteric coating. DBP has been associated with adverse effects on the male reproductive system in animals at very high doses.

Lialda is a once-a-day formulation of mesalamine (Multi-Matrix System [MMX]) designed to release mesalamine in the colon. The MMX technology incorporates mesalamine into a lipophilic matrix within a hydrophilic matrix encapsulated in a polymer resistant to degradation at a low pH (<7) to delay release throughout the colon. The safety profile appears to be comparable to other 5-ASA formulations.

Apriso is a formulation containing encapsulated mesalamine granules that delivers mesalamine to the terminal ileum and colon via a proprietary extended-release mechanism (Intellicor). The outer coating of this agent (Eudragit L) dissolves at a pH >6. In addition, there is a polymer matrix core that aids in sustained release throughout the colon. Because Lialda and Apriso are given once daily, an anticipated benefit is improved compliance compared with two to four daily doses required for other mesalamine preparations.

Pentasa is another mesalamine formulation that uses an ethylcellulose coating to allow water absorption into small beads containing the mesalamine. Water dissolves the 5-ASA, which then diffuses out of the bead into the lumen. Disintegration of the capsule occurs in the stomach. The microspheres then disperse throughout the entire GI tract from the small intestine through the distal colon in both fasted and fed conditions.

Salofalk® Granu-Stix, an unencapsulated version of mesalamine, has been in use in Europe for induction and maintenance of remission for several years.

Appropriate doses of the 5-ASA compounds are shown in **Table 319-7**. Some 50–75% of patients with mild to moderate UC improve when treated with 5-ASA doses equivalent to 2 g/d of mesalamine; the dose response continues up to at least 4.8 g/d.

More common side effects of the 5-ASA medications include headaches, nausea, hair loss, and abdominal pain. Rare side effects of the 5-ASA medications include renal impairment, hematuria, pancreatitis, and paradoxical worsening of colitis. Renal function tests and urinalysis should be checked yearly.

Topical *Rowasa* enemas are composed of mesalamine and are effective in mild-to-moderate distal UC. Combination therapy with mesalamine in both oral and enema form is more effective than either treatment alone for both distal and extensive UC.

Canasa suppositories composed of mesalamine are effective in treating proctitis.

GLUCOCORTICOIDS

The majority of patients with moderate to severe UC benefit from oral or parenteral glucocorticoids. Prednisone is usually started at doses of 40–60 mg/d for active UC that is unresponsive to 5-ASA therapy. Parenteral glucocorticoids may be administered as hydrocortisone, 300 mg/d, or methylprednisolone, 40–60 mg/d. A new glucocorticoid for UC, budesonide (Uceris), is released entirely in the colon and has minimal to no glucocorticoid side effects. The dose is 9 mg/d for 8 weeks, and no taper is required. Topically applied

TABLE 319-7 Oral 5-ASA Preparations

PREPARATION	FORMULATION	DELIVERY	DOSING PER DAY
Azo-Bond			
Sulfasalazine (500 mg) (Azulfidine)	Sulfapyridine-5-ASA	Colon	3–6 g (acute) 2–4 g (maintenance)
Balsalazide (750 mg) (Colazal)	Aminobenzoyl-alanine–5-ASA	Colon	6.75–9 g
Delayed-Release			
Mesalamine (400, 800 mg) (Delzicol, Asacol HD)	Eudragit S (pH 7)	Distal ileum-colon	2.4–4.8 g (acute) 1.6–4.8 g (maintenance)
Mesalamine (1.2 g) (Lialda)	MMX mesalamine (SPD476)	Ileum-colon	2.4–4.8 g
Controlled-Release			
Mesalamine (250, 500, 1000 mg) (Pentasa)	Ethylcellulose microgranules	Stomach-colon	2–4 g (acute) 1.5–4 g (maintenance)
Delayed- and Extended-Release			
Mesalamine (0.375 g) (Apriso)	Intellicor extended-release mechanism	Ileum-colon	1.5 g (maintenance)

glucocorticoids are also beneficial for distal colitis and may serve as an adjunct in those who have rectal involvement plus more proximal disease. Hydrocortisone enemas or foam may control active disease, although they have no proven role as maintenance therapy. These glucocorticoids are significantly absorbed from the rectum and can lead to adrenal suppression with prolonged administration. Topical 5-ASA therapy is more effective than topical steroid therapy in the treatment of distal UC.

Glucocorticoids are also effective for treatment of moderate to severe CD and induce a 60–70% remission rate compared to a 30% placebo response. The systemic effects of standard glucocorticoid formulations have led to the development of more potent formulations that are less well-absorbed and have increased first-pass metabolism. Controlled ileal-release budesonide has been nearly equal to prednisone for ileocolonic CD with fewer glucocorticoid side effects. Budesonide is used for 2–3 months at a dose of 9 mg/d, and then tapered. Glucocorticoids play no role in maintenance therapy in either UC or CD. Once clinical remission has been induced, they should be tapered according to the clinical activity, normally at a rate of no more than 5 mg/week. They can usually be tapered to 20 mg/d within 4–5 weeks but often take several months to be discontinued altogether. The side effects are numerous, including fluid retention, abdominal striae, fat redistribution, hyperglycemia, subcapsular cataracts, osteonecrosis, osteoporosis, myopathy, emotional disturbances, and withdrawal symptoms. Most of these side effects, aside from osteonecrosis, are related to the dose and duration of therapy.

ANTIBIOTICS

Antibiotics have no role in the treatment of active or quiescent UC. However, pouchitis, which occurs in about 30–50% of UC patients after colectomy and IPAA, usually responds to treatment with metronidazole and/or ciprofloxacin.

Metronidazole is effective in active inflammatory, fistulizing, and perianal CD and may prevent recurrence after ileal resection. The most effective dose is 15–20 mg/kg per day in three divided doses; it is usually continued for several months. Common side effects include nausea, metallic taste, and disulfiram-like reaction. Peripheral neuropathy can occur with prolonged administration (several months) and on rare occasions is permanent despite discontinuation. *Ciprofloxacin* (500 mg bid) is also beneficial for inflammatory, perianal, and fistulizing CD but has been associated with tendinitis and tendon rupture. Both ciprofloxacin and metronidazole antibiotics can be used only for short period of time due to side effects.

AZATHIOPRINE AND 6-MERCAPTOPURINE

Azathioprine and 6-mercaptopurine (6-MP) are purine analogues used concomitantly with biologic therapy or, less often, as the sole immunosuppressants. Azathioprine is rapidly absorbed and converted to 6-MP, which is then metabolized to the active end product, thioinosinic acid, an inhibitor of purine ribonucleotide synthesis and cell proliferation. Efficacy can be seen as early as 3–4 weeks but can take up to 4–6 months. Adherence can be monitored by measuring the levels of 6-thioguanine and 6-methyl-mercaptopurine, end products of 6-MP metabolism. The doses used range from 2–3 mg/kg per day for azathioprine and 1–1.5 mg/kg per day for 6-MP.

Although azathioprine and 6-MP are usually well tolerated, pancreatitis occurs in 3–4% of patients, typically presents within the first few weeks of therapy, and is completely reversible when the drug is stopped. Other side effects include nausea, fever, rash, and hepatitis. Bone marrow suppression (particularly leukopenia) is dose-related and often delayed, necessitating regular monitoring of the complete blood cell count (CBC). Additionally, 1 in 300 individuals lacks thiopurine methyltransferase, the enzyme responsible for drug metabolism to inactive end-products (6-methylmercaptopurine); an additional 11% of the population are heterozygotes with intermediate enzyme activity. Both are at increased risk of toxicity because of increased accumulation of active 6-thioguanine metabolites. Although 6-thioguanine and 6-methylmercaptopurine levels can be followed to determine correct drug dosing and reduce toxicity, weight-based dosing is an acceptable alternative. CBCs and liver function tests should be monitored frequently regardless of dosing strategy.

METHOTREXATE

MTX inhibits dihydrofolate reductase, resulting in impaired DNA synthesis. Additional anti-inflammatory properties may be related to decrease in the production of IL-1. It is used most often concomitantly with biologic therapy to decrease antibody formation and improve disease response. Intramuscular (IM) or subcutaneous (SC) doses range from 15 to 25 mg/week. Potential toxicities include leukopenia and hepatic fibrosis, necessitating periodic evaluation of CBCs and liver enzymes. The role of liver biopsy in patients on long-term MTX is uncertain but is probably limited to those with increased liver enzymes. Hypersensitivity pneumonitis is a rare but serious complication of therapy.

CYCLOSPORINE

CSA is a lipophilic peptide with inhibitory effects on both the cellular and humoral immune systems. CSA blocks the production of IL-2 by T helper lymphocytes. CSA binds to cyclophilin, and this complex inhibits calcineurin, a cytoplasmic phosphatase enzyme involved in the activation of T cells. CSA also indirectly inhibits B cell function by blocking helper T cells. CSA has a more rapid onset of action than 6-MP and azathioprine.

CSA is most effective when given at 2–4 mg/kg per day IV in severe UC that is refractory to IV glucocorticoids, with 82% of patients responding. CSA can be an alternative to colectomy. The long-term success of oral CSA is not as dramatic, but if patients are started on 6-MP or azathioprine at the time of hospital discharge, remission can be maintained. For the 2 mg/kg dose, levels as measured by monoclonal radioimmunoassay or by the high-performance liquid chromatography assay should be maintained between 150 and 350 ng/mL.

CSA may cause significant toxicity; renal function should be monitored frequently. Hypertension, gingival hyperplasia, hypertrichosis, paresthesias, tremors, headaches, and electrolyte abnormalities are common side effects. Creatinine elevation calls for dose reduction or discontinuation. Seizures may also complicate therapy, especially if the patient is hypomagnesemic or if serum cholesterol levels are <3.1 mmol/L (<120 mg/dL). Opportunistic infections, most notably *Pneumocystis carinii* pneumonia, may occur with combination immunosuppressive treatment; prophylaxis should be given. Major adverse events occurred in 15% of patients in one large study, including nephrotoxicity not responding to dose adjustment, serious infections, seizures, anaphylaxis, and death of two patients. This high incidence suggests that vigorous monitoring by experienced clinicians at tertiary care centers may be required. To compare IV CSA versus infliximab, a large trial was conducted in Europe by the GETAID group. The results indicated identical 7-day response rates between CSA 2 mg/kg (with doses adjusted for levels of 150–250 ng/mL) and infliximab 5 mg/kg, with both groups achieving response rates of 85%. Serious infections occurred in 5 of 55 CSA patients and 4 of 56 infliximab patients. Response rates were similar in the two groups at day 98 among patients treated with oral CSA versus infliximab at the usual induction dose and maintenance dose regimen (40 and 46%, respectively). In light of data showing equal efficacy of CSA and infliximab in severe UC, more physicians are relying on infliximab rather than CSA in these patients.

TACROLIMUS

Tacrolimus is a macrolide antibiotic with immunomodulatory properties similar to CSA. It is 100 times as potent as CSA and is not dependent on bile or mucosal integrity for absorption. These pharmacologic properties enable tacrolimus to have good oral absorption despite proximal small bowel Crohn's involvement. It has shown efficacy in children with refractory IBD and in adults with extensive

involvement of the small bowel. It is also effective in adults with glucocorticoid-dependent or refractory UC and CD as well as refractory fistulizing CD.

BIOLOGIC THERAPIES

Biologic therapy is now commonly given as an initial therapy for patients with moderate to severe CD and UC. Patients who respond to biologic therapies enjoy an improvement in clinical symptoms; a better quality of life; less disability, fatigue, and depression; and fewer surgeries and hospitalizations.

Anti-TNF Therapies The first biologic therapy approved for moderate to severely active CD and also UC was *infliximab*, a chimeric IgG1 antibody against TNF-α,. Of active CD patients refractory to glucocorticoids, 6-MP, or 5-ASA, 65% will respond to IV infliximab (5 mg/kg); one-third will enter complete remission. The ACCENT I (A Crohn's Disease Clinical Trial Evaluating Infliximab in a New Long-Term Treatment Regimen) study showed that of the patients who experience an initial response, 40% will maintain remission for at least 1 year with repeated infusions of infliximab every 8 weeks.

Infliximab is also effective in CD patients with refractory perianal and enterocutaneous fistulas, with the ACCENT II trial showing a 68% response rate (50% reduction in fistula drainage) and a 50% complete remission rate. Reinfusion, typically every 8 weeks, is necessary to continue therapeutic benefits in many patients.

The SONIC (Study of Biologic and Immunomodulator-Naive Patients with Crohn's Disease) trial compared infliximab plus azathioprine, infliximab alone, and azathioprine alone in immunomodulator- and biologic-naive patients with moderate to severe CD. At 1 year, the infliximab plus azathioprine group had a glucocorticoid-free remission rate of 46% compared with 35% for infliximab alone and 24% for azathioprine alone. There was also complete mucosal healing at week 26 with the combined approach relative to either infliximab or azathioprine alone (44 vs 30 vs 17%). The adverse events were equal between groups.

Two large trials of infliximab in moderate to severe UC also showed efficacy with a response rate of 37–49%, with about one-fifth of patients maintaining remission after 54 weeks. Dosing for UC and CD are identical, with induction dosing at 0, 2, and 6 weeks and every 8 weeks thereafter. There is a similar study to SONIC in patients with moderate to severe UC. After 16 weeks of therapy, UC patients taking azathioprine plus infliximab had a glucocorticoid-free remission rate of 40% compared to 24% and 22% of those on azathioprine and infliximab alone, respectively. This is even further evidence for "top-down" or more aggressive therapy for both moderate to severe CD and UC.

Adalimumab is a recombinant human monoclonal IgG1 antibody containing only human peptide sequences and is injected subcutaneously. Adalimumab binds TNF and neutralizes its function by blocking the interaction between TNF and its cell-surface receptor. Therefore, it seems to have a similar mechanism of action to infliximab but with less immunogenicity. Adalimumab is approved for treatment of moderate to severe CD and UC. CHARM (Crohn's Trial of the Fully Human Adalimumab for Remission Maintenance) is an adalimumab maintenance study in patients who responded to adalimumab induction therapy. About 50% of the patients in this trial were previously treated with infliximab. Remission rates ranged from 42 to 48% of infliximab-naïve patients at 1 year compared with remission rates of 31–34% in patients who had previously received infliximab. Another trial showed a remission rate of 21% at 4 weeks in patients who had initially responded to and then failed infliximab. UC results are similar with a sustained remission rate at one year of 22% (12.4% placebo) among anti–TNF-naïve patients and a sustained remission rate at 1 year of 10.2% (3% placebo) among patients who had previously received anti-TNF agents. In clinical practice, the remission rate in both CD and UC patients taking adalimumab increases with a dose increase to 40 mg weekly instead of every other week.

Certolizumab pegol is a pegylated form of an anti-TNF Fab portion of an antibody administered SC once monthly. SC certolizumab pegol was effective for induction of clinical response in patients with active inflammatory CD. In the PRECISE II (Pegylated Antibody Fragment Evaluation in Crohn's Disease) trial of maintenance therapy with certolizumab in patients who responded to certolizumab induction, the results were similar to the CHARM trial. At week 26, the subgroup of patients who were infliximab naïve had a response of 69% as compared to 44% in patients who had previously received infliximab.

Golimumab is another fully human IgG1 antibody against TNF-α and is currently approved for the treatment of moderately to severely active UC. Like adalimumab and certolizumab, golimumab is injected SC.

Side Effects of Anti-TNF Therapies • **Development of Antibodies** The development of antibodies to infliximab is associated with an increased risk of infusion reactions and a decreased response to treatment. Current practice does not include giving on-demand or episodic infusions in contrast to periodic (every 8 week) infusions because patients are most likely to develop antibodies. Anti-infliximab antibodies are generally present when the quality of response or the response duration to infliximab infusion decreases. There are commercial assays for both infliximab and adalimumab antibodies and trough levels to determine optimal dosing. If a patient has high anti-infliximab antibodies and a low trough level of infliximab, it is best to switch to another anti-TNF therapy. Most acute infusion reactions and serum sickness can be managed with glucocorticoids and antihistamines. Some reactions can be serious and would necessitate a change in therapy, especially if a patient has anti-infliximab antibodies. It is now common practice to add an immunomodulator such as azathioprine, 6-mercaptopurine or MTX to anti-TNF therapy in order to prevent antibody formation.

Non-Hodgkin's Lymphoma (NHL) The baseline risk of NHL in CD patients is 2:10,000, which is slightly higher than in the general population. Azathioprine and/or 6-MP therapy increases the risk to about 4:10,000. The highest risk for thiopurine-associated NHL is in patients over 65 years old actively using thiopurines (yearly incidence rate per 1000 patient years of 5.41), with a moderate risk in those between the ages of 50 and 65 (incidence rate of 2.58 compared to an incidence rate of 0.37 in patients <50 years old). It is difficult to assess whether anti-TNF medications are associated with lymphoma because most patients are also receiving thiopurines. After adjustment for co-treatments, no excess risk of lymphoma was found in a recent adequately powered Danish study of a cohort of IBD patients exposed to anti-TNF medications.

Hepatosplenic T-Cell Lymphoma (HSTCL) HSTCL is a nearly universally fatal lymphoma in patients with or without CD. In patients with CD, events reported to the Food and Drug Administration Adverse Event Reporting System (FDA AERS) and search of PubMed and Embase published case reports demonstrate a total of 37 unique cases. Eighty-six percent of the patients were male, with a median age of 26 years. Patients had CD for a mean of 10 years before the diagnosis of HSTCL. Thirty-six cases had used either 6-MP or azathioprine, and 28 cases had used infliximab. Of these 28 cases, 27 had also used 6-MP or azathioprine. The other case had a history of both infliximab and adalimumab exposure.

Skin Lesions New-onset psoriasiform skin lesions develop in nearly 5% of IBD patients treated with anti-TNF therapy. Most often, these can be treated topically, and occasionally, anti-TNF therapy must be decreased, switched, or stopped. Patients with IBD may have a slight unexplained intrinsic higher risk of developing melanoma. The risk of melanoma is increased almost twofold with anti-TNF and not thiopurine use. The risk of nonmelanoma skin cancer is increased with thiopurines and biologics, especially with 1 year of follow-up or greater. Patients on these medications should have a skin check at least once a year.

Infections All of the anti-TNF drugs are associated with an increased risk of infections, particularly reactivation of latent tuberculosis and opportunistic fungal infections including disseminated histoplasmosis and coccidioidomycosis. It is recommended that

patients have a purified protein derivative (PPD) or a QuantiFERON-TB gold test as well as a chest x-ray before initiation of anti-TNF therapy. Patients >65 have a higher rate of infections and death on infliximab or adalimumab than those <65 years of age.

Other Acute liver injury due to reactivation of hepatitis B virus and to autoimmune effects and cholestasis has been reported. Rarely, infliximab and the other anti-TNF drugs have been associated with optic neuritis, seizures, new onset or exacerbation of clinical symptoms, and radiographic evidence of central nervous system demyelinating disorders, including multiple sclerosis. They may exacerbate symptoms in patients with New York Heart Association functional class III/IV heart failure.

ANTI-INTEGRINS

Integrins are expressed on the cell surface of leukocytes and serve as mediators of leukocyte adhesion to vascular endothelium. α4-Integrin along with its β1 or β7 subunit interact with endothelial ligands termed adhesion molecules. Interaction between α4β7 and mucosal addressin cellular adhesion molecule (MAdCAM-1) is important in lymphocyte trafficking to gut mucosa.

Natalizumab is a recombinant humanized IgG4 antibody against α4-integrin that has been shown to be effective in induction and maintenance of patients with CD. It has been approved since February 2008 for the treatment of patients with CD refractory or intolerant to anti-TNF therapy. The rates of response and remission at 3 months are about 60 and 40%, respectively, with a sustained remission rate of about 40% at 36 weeks.

Natalizumab is no longer widely used for CD due to the risk of progressive multifocal leukoencephalopathy (PML). The most important risk factor for development of PML is exposure to the John Cunningham (JC) polyomavirus, seen in 50–55% of the adult population. The other two risk factors for development of PML are longer duration of treatment, especially beyond 2 years, and prior treatment with an immunosuppressant medication. Patients with all three risk factors have an estimated risk of 11:1000.

The FDA approved a commercial enzyme-linked immunosorbent assay (ELISA) kit to assay anti-JC viral antibodies (Stratify JCV Antibody ELISA; Focus Diagnostics, Cypress, CA) in early 2012. The test is 99% accurate in stratifying risk of PML. It is recommended that all patients be tested prior to initiating natalizumab therapy. JC virus serologies are then measured every 6 months because 1–2% of patients will seroconvert yearly. Natalizumab is administered IV, 300 mg every 4 weeks. Labeling requirements mandate that it should not be used in combination with any immunosuppressant medications.

Vedolizumab, another leukocyte trafficking inhibitor, is a monoclonal antibody directed against α4β7 integrin specifically and has the ability to convey gut-selective immunosuppression. Vedolizumab is indicated for CD and UC patients who have had an inadequate response or lost response to, or were intolerant of a TNF blocker or immunomodulator; or had an inadequate response or were intolerant to, or demonstrated dependence on, glucocorticoids. It is also an option for patients who are JC antibody positive since unlike natalizumab it inhibits adhesion of a discrete gut-homing subset of T lymphocytes to MAdCAM-1, but not to vascular adhesion molecule-1. Vedolizumab decreases GI inflammation without inhibiting systemic immune responses or affecting T-cell trafficking to the central nervous system. Vedolizumab is given intravenously every 8 weeks after 3 induction doses at 0, 2, and 6 weeks. In the GEMINI I trial (A Phase 3, Randomized, Placebo-Controlled, Blinded, Multicenter Study of the Induction and Maintenance of Clinical Response and Remission by MLN002 in Patients With Moderate to Severe Ulcerative Colitis), 42% of the UC patients treated every 8 weeks and 45% of those treated every 4 weeks were in clinical remission at week 52 compared with 16% placebo. In the GEMINI II trial, the clinical remission rates for CD patients treated with vedolizumab were 36 to 39%, compared with 22% placebo at 52 weeks.

Ustekinumab, a fully human IgG1 monoclonal antibody, blocks the biologic activity of IL-12 and IL-23 through their common p40 subunit by inhibiting the interaction of these cytokines with their receptors on T cells, natural killer cells, and antigen presenting cells. It has recently been FDA-approved for use in Crohn's patients who have failed or were intolerant to immunomodulator or corticosteroid therapy but who never failed treatment with anti-TNF therapy or who failed or were intolerant to treatment with one or more anti-TNF medications. The result for the highest 6 mg/kg IV induction dose and subsequent 90 mg every 8 week dose in one major clinical trial was 41.7% remission rate versus 27.4% placebo at 22 weeks in Crohn's patients failing anti-TNF therapy.

THERAPIES IN CLINICAL DEVELOPMENT

Tofacitinib is an oral inhibitor of Janus kinases 1, 3, and, to a lesser extent, 2. It is expected to block signaling involving common gamma chain–containing cytokines including IL-2, IL-4, IL-7, IL-9, IL-15, and IL-21. These cytokines are integral to lymphocyte activation, function, and proliferation. It is effective in moderate to severe UC in clinical trials.

Biosimilars The FDA defines a biosimilar drug as a "biological product that is highly similar to the reference product not withstanding minor differences in clinically inactive components." There are no clinically meaningful differences between the biological product and the reference product in terms of the safety, purity, and potency of the product. The infliximab biosimilar CT-P13 is approved and available for use in almost 70 countries and many other biosimilars to infliximab and adalimumab are currently being manufactured. Biosimilars may be approved without randomized-controlled trials. The FDA examines quality considerations such as the expression system, manufacturing process, assessment of physiochemical properties, functional activities, receptor binding and immunochemical properties, measurement of impurities, stability under multiple stress conditions and effect of product formulation and shipping, Since biosimilars will likely cost about a third of the reference drug, they will likely be widely used in the near future in the United States.

Ozanimod is an oral agonist of the sphingosine-1-phosphate receptor subtypes 1 and 5 that causes peripheral lymphocyte sequestration, potentially decreasing the number of activated lymphocytes circulating to the GI tract. In a phase 2 trial of ozanimod in 197 patients with moderate to severe UC, at 32 weeks, 21% of patients who received 1 mg of ozanimod versus 6% achieved clinical remission. Phase 3 trials are now in progress.

NUTRITIONAL THERAPIES

Dietary antigens may stimulate the mucosal immune response. Patients with active CD respond to bowel rest, along with TPN. Bowel rest and TPN are as effective as glucocorticoids at inducing remission of active CD but are not effective as maintenance therapy. Enteral nutrition in the form of elemental or peptide-based preparations is also as effective as glucocorticoids or TPN, but these diets are not palatable. Enteral diets may provide the small intestine with nutrients vital to cell growth and do not have the complications of TPN. In contrast to CD, dietary intervention does not reduce inflammation in UC. Standard medical management of UC and CD is shown in **Fig. 319-12**.

SURGICAL THERAPY

Ulcerative Colitis Nearly one-half of patients with extensive chronic UC undergo surgery within the first 10 years of their illness. The indications for surgery are listed in **Table 319-8**. Morbidity is about 20% for elective, 30% for urgent, and 40% for emergency proctocolectomy. The risks are primarily hemorrhage, contamination and sepsis, and neural injury. The operation of choice is an IPAA.

Because UC is a mucosal disease, the rectal mucosa can be dissected and removed down to the dentate line of the anus or about 2 cm proximal to this landmark. The ileum is fashioned into a pouch that serves as a neorectum. This ileal pouch is then sutured circumferentially to the anus in an end-to-end fashion. If performed carefully, this operation preserves the anal sphincter

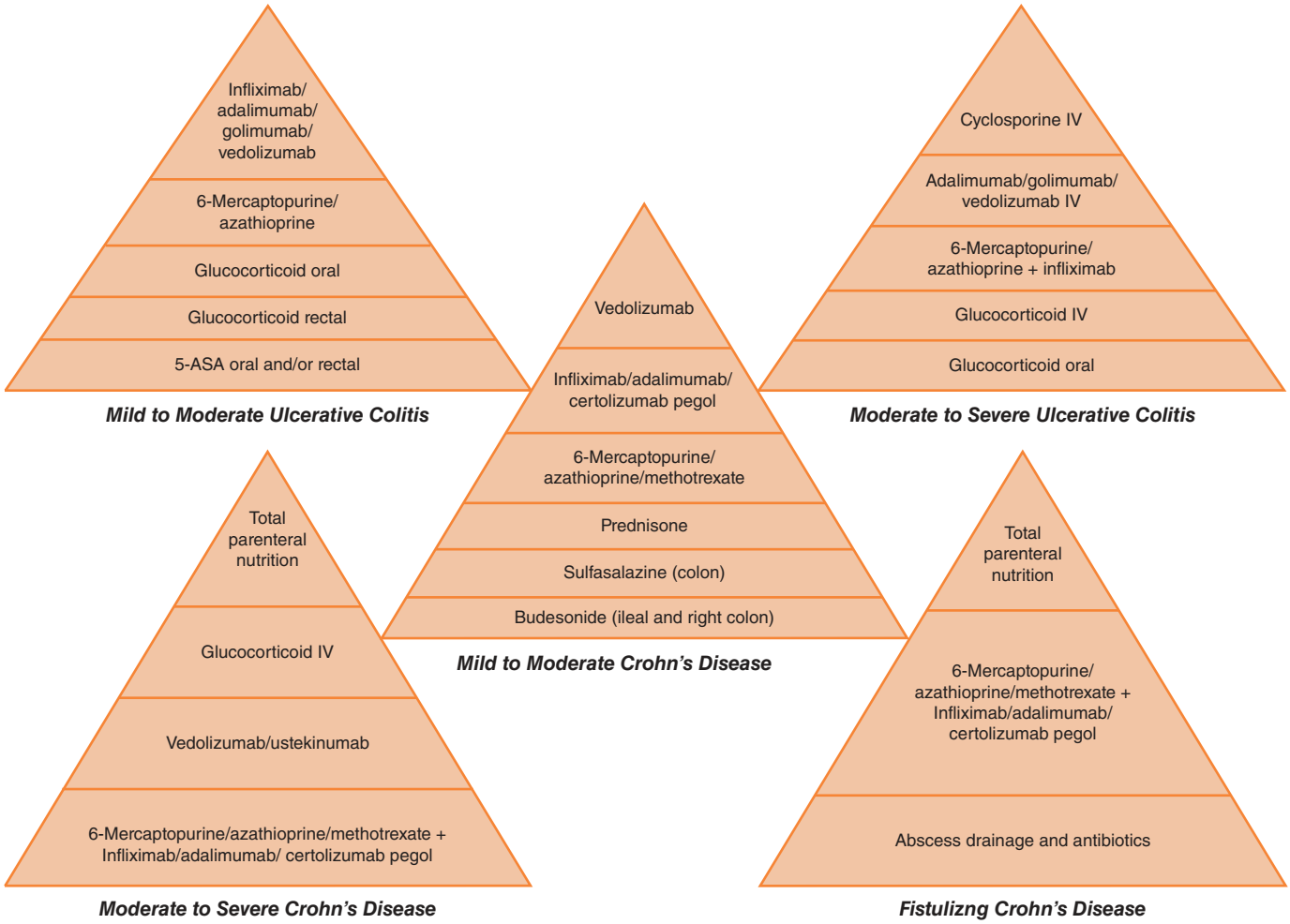


FIGURE 319-12 Medical management of inflammatory bowel disease. 5-ASA, 5-aminosalicylic acid; CD, Crohn's disease; UC, ulcerative colitis.

and maintains continence. The overall operative morbidity is 10%, with the major complication being bowel obstruction. Pouch failure necessitating conversion to permanent ileostomy occurs in 5–10% of patients. Some inflamed rectal mucosa is usually left behind, and thus endoscopic surveillance is necessary. Primary dysplasia of the ileal mucosa of the pouch has occurred rarely.

Patients with IPAA usually have about 6–10 bowel movements a day. On validated quality-of-life indices, they report better performance in sports and sexual activities than ileostomy patients.

TABLE 319-8 Indications for Surgery

ULCERATIVE COLITIS	CROHN'S DISEASE
Intractable disease	Small Intestine
Fulminant disease	Stricture and obstruction
Toxic megacolon	unresponsive to medical therapy
Colonic perforation	Massive hemorrhage
Massive colonic hemorrhage	Refractory fistula
Extracolonic disease	Abscess
Colonic obstruction	Colon and rectum
Colon cancer prophylaxis	Intractable disease
Colon dysplasia or cancer	Fulminant disease
	Perianal disease unresponsive to medical therapy
	Refractory fistula
	Colonic obstruction
	Cancer prophylaxis
	Colon dysplasia or cancer

The most frequent complication of IPAA is pouchitis in about 30–50% of patients with UC. This syndrome consists of increased stool frequency, watery stools, cramping, urgency, nocturnal leakage of stool, arthralgias, malaise, and fever. Pouch biopsies may distinguish true pouchitis from underlying CD. Although pouchitis usually responds to antibiotics, 3–5% of patients remain refractory and may require glucocorticoids, immunomodulators, biologics or even pouch removal. A highly concentrated probiotic preparation with four strains of *Lactobacillus*, three strains of *Bifidobacterium*, and one strain of *Streptococcus salivarius* may prevent the recurrence of pouchitis when taken daily.

Crohn's Disease Most patients with CD require at least one operation in their lifetime. The need for surgery is related to duration of disease and the site of involvement. Patients with small-bowel disease have an 80% chance of requiring surgery. Those with colitis alone have a 50% chance. Surgery is an option only when medical treatment has failed or complications dictate its necessity. The indications for surgery are shown in Table 319-8.

Small Intestinal Disease Because CD is chronic and recurrent, with no clear surgical cure, as little intestine as possible is resected. Current surgical alternatives for treatment of obstructing CD include resection of the diseased segment and strictureplasty. Surgical resection of the diseased segment is the most frequently performed operation, and in most cases, primary anastomosis can be done to restore continuity. If much of the small bowel has already been resected and the strictures are short, with intervening areas of normal mucosa, strictureplasties should be done to avoid a functionally insufficient length of bowel. The strictured area of intestine is incised longitudinally and the incision sutured transversely, thus widening the

narrowed area. Complications of strictureplasty include prolonged ileus, hemorrhage, fistula, abscess, leak, and restricture.

There is evidence that mesalamine, nitroimidazole antibiotics, 6-MP/azathioprine, infliximab, and adalimumab are all superior to placebo for the prevention of postoperative recurrence of CD. Mesalamine is the least effective, and the side effects of the nitroimidazole antibiotics limit their use. Risk factors for early recurrence of disease include cigarette smoking, penetrating disease (internal fistulas, abscesses, or other evidence of penetration through the wall of the bowel), early recurrence since a previous surgery, multiple surgeries, and a young age at the time of the first surgery. Aggressive postoperative treatment with 6-MP/azathioprine, infliximab, or adalimumab should be considered for this group of patients. It is also recommended to evaluate for endoscopic recurrence of CD via a colonoscopy, if possible, 6 months after surgery.

Colorectal Disease A greater percentage of patients with Crohn's colitis require surgery for intractability, fulminant disease, and anorectal disease. Several alternatives are available, ranging from the use of a temporary loop ileostomy to resection of segments of diseased colon or even the entire colon and rectum. For patients with segmental involvement, segmental colon resection with primary anastomosis can be performed. In 20–25% of patients with extensive colitis, the rectum is spared sufficiently to consider rectal preservation. Most surgeons believe that an IPAA is contraindicated in CD due to the high incidence of pouch failure. A diverting colostomy may help heal severe perianal disease or rectovaginal fistulas, but disease almost always recurs with reanastomosis. These patients often require a total proctocolectomy and ileostomy.

IBD AND PREGNANCY

Patients with quiescent UC and CD have normal fertility rates; the fallopian tubes can be scarred by the inflammatory process of CD, especially on the right side because of the proximity of the terminal ileum. In addition, perirectal, perineal, and rectovaginal abscesses and fistulae can result in dyspareunia. Infertility in men can be caused by sulfasalazine but reverses when treatment is stopped. In women who have an IPAA, most studies show that the fertility rate is reduced to about 50–80% of normal. This is due to scarring or occlusion of the fallopian tubes secondary to pelvic inflammation.

In mild or quiescent UC and CD, fetal outcome is nearly normal. The courses of CD and UC during pregnancy mostly correlate with disease activity at the time of conception. Patients should be in remission for 6 months before conceiving. Most CD patients can deliver vaginally, but cesarean delivery may be the preferred route of delivery for patients with anorectal and perirectal abscesses and fistulas to reduce the likelihood of fistulas developing or extending into the episiotomy scar. Unless they desire multiple children, UC patients with an IPAA should consider a cesarean delivery due to an increased risk of future fecal incontinence.

Sulfasalazine, Lialda, Apriso, Delzicol, balsalazide and now Asacol HD since the DBP has been removed from the capsule are safe for use in pregnancy and nursing with the caveat that additional folate supplementation must be given with sulfasalazine. Topical 5-ASA agents are safe during pregnancy and nursing. Glucocorticoids are generally safe for use during pregnancy and are indicated for patients with moderate to severe disease activity. The amount of glucocorticoids received by the nursing infant is minimal. The safest antibiotics to use for CD in pregnancy for short periods of time (weeks, not months) are ampicillin and cephalosporins. Metronidazole can be used in the second or third trimester. Ciprofloxacin causes cartilage lesions in immature animals and should be avoided because of the absence of data on its effects on growth and development in humans.

6-MP and azathioprine pose minimal or no risk during pregnancy, but experience is limited. If the patient cannot be weaned from the drug or has an exacerbation that requires 6-MP/azathioprine during pregnancy, she should continue the drug with informed consent. Breast milk has been shown to contain negligible levels of 6-MP/azathioprine when measured in a limited number of patients.

Little data exist on CSA in pregnancy. In a small number of patients with severe IBD treated with IV CSA during pregnancy, 80% of pregnancies were successfully completed without development of renal toxicity or congenital malformations. However, because of the lack of data, CSA should probably be avoided unless the patient would otherwise require surgery.

MTX is contraindicated in pregnancy and nursing. In a large prospective and multiple retrospective studies, no increased risk of stillbirths, miscarriages, or spontaneous abortions was seen with infliximab, adalimumab, or certolizumab. Infliximab and adalimumab are IgG1 antibodies and are actively transported across the placenta in the late second and third trimester. Infants can have serum levels of infliximab and adalimumab up to 12 months of age, and live vaccines should be avoided during this time. Certolizumab crosses the placenta by passive diffusion, and infant serum and cord blood levels are minimal. The anti-TNF drugs are relatively safe in nursing. Miniscule levels of infliximab, adalimumab, and certolizumab have been reported in breast milk. These levels are of no clinical significance. It is recommended that drugs should not be switched during pregnancy unless necessitated by the medical condition of the IBD. Vedolizumab and natalizumab appear safe during pregnancy although the data are limited. There are very little data available on ustekinemab use during pregnancy.

Surgery in UC should be performed only for emergency indications, including severe hemorrhage, perforation, and megacolon refractory to medical therapy. Total colectomy and ileostomy carry a 50% risk of postoperative spontaneous abortion. Fetal mortality is also high in CD requiring surgery. Patients with IPAAs have increased nighttime stool frequency during pregnancy that resolves postpartum. Transient small-bowel obstruction or ileus has been noted in up to 8% of patients with ileostomies.

CANCER IN IBD

ULCERATIVE COLITIS

Patients with long-standing UC are at increased risk for developing colonic epithelial dysplasia and carcinoma **(Fig. 319-13)**.

The risk of neoplasia in chronic UC increases with duration and extent of disease. In contrast to the relatively high risk in one large meta-analysis (2% after 10 years, 8% after 20 years, and 18% after 30 years of disease), a decrease in the risk of colorectal cancer has been noted over time potentially due to better control of inflammation, better colonoscopic surveillance, more frequent colectomies and use of 5-ASA chemoprophylaxis. The rates of colon cancer are still about 1.5 to 2 × higher than in the general population, and colonoscopic surveillance is the standard of care.

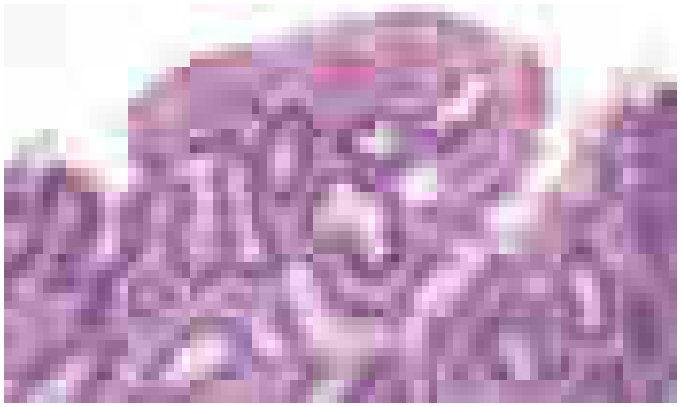


FIGURE 319-13 Medium-power view of low-grade dysplasia in a patient with chronic ulcerative colitis. Low-grade dysplastic crypts are interspersed among regenerating crypts. *(Courtesy of Dr. R. Odze, Division of Gastrointestinal Pathology, Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts; with permission.)*

Annual or biennial colonoscopy with multiple biopsies is recommended for patients with >8–10 years of extensive colitis (greater than one-third of the colon involved) or 12–15 years of proctosigmoiditis (less than one-third but more than just the rectum) and has been widely used to screen and survey for subsequent dysplasia and carcinoma. Risk factors for cancer in UC include long-duration disease, extensive disease, family history of colon cancer, PSC, a colon stricture, and the presence of post inflammatory pseudopolyps on colonoscopy.

CROHN'S DISEASE

Risk factors for developing cancer in Crohn's colitis are long-duration and extensive disease, bypassed colon segments, colon strictures, PSC, and family history of colon cancer. The cancer risks in CD and UC are probably equivalent for similar extent and duration of disease. In the CESAME study, a prospective observational cohort of IBD patients in France, the standardized incidence ratios of colorectal cancer were 2.2 for all IBD patients (95% confidence interval [CI], 1.5–3.0; $p < 0.001$) and 7.0 for patients with long-standing extensive colitis (both Crohn's and UC) (95% CI, 4.4–10.5; $p < 0.001$). Thus, the same endoscopic surveillance strategy used for UC is recommended for patients with chronic Crohn's colitis. A pediatric colonoscope can be used to pass narrow strictures in CD patients, but surgery should be considered in symptomatic patients with impassable strictures.

MANAGEMENT OF DYSPLASIA AND CANCER

Dysplasia can be flat or polypoid. If flat high-grade dysplasia is encountered on colonoscopic surveillance, the usual treatment is colectomy for UC and either colectomy or segmental resection for CD. If flat low-grade dysplasia is found (Fig. 319-13), most investigators recommend immediate colectomy. Adenomas may occur coincidently in UC and CD patients with chronic colitis and can be removed endoscopically provided that biopsies of the surrounding mucosa are free of dysplasia. High-definition and high-magnification colonoscopes and dye sprays have increased the rate of dysplasia detection.

IBD patients are also at greater risk for other malignancies. Patients with CD may have an increased risk of NHL, leukemia, and myelodysplastic syndromes. Severe, chronic, complicated perianal disease in CD patients may be associated with an increased risk of cancer in the lower rectum and anal canal (squamous cell cancers). Although the absolute risk of small-bowel adenocarcinoma in CD is low (2.2% at 25 years in one study), patients with long-standing, extensive, small-bowel disease should consider screening.

FURTHER READING

Beaugerie L, Itzkowitz SH: Cancers complicating inflammatory bowel disease. N Engl J Med 372:15, 2014.

Bonovas S: Biologic therapies and risk of infection and malignancy in patients with inflammatory bowel disease: a systemic review and network meta-analysis. Clin Gastroenetrol Hepatol 14:1385, 2016.

Columbel JF et al: The safety of vedolizumab for ulcerative colitis and Crohn's disease. Gut 66:839, 2017.

Ha C, Kornbluth A: A critical review of biosimilars in IBD: The confluence of biologic drug development, regulatory requirements, clinical outcomes, and big business. Inflamm Bowel Dis 22:2513, 2016.

Julsgaard M: Concentrations of adalimumab and infliximab in mothers and newborns and effects on infection. Gastroenterology 151:110, 2016.

Ng SC: Geographical variability and environmental risk factors in inflammatory bowel disease. Gut 62:630, 2013.

Panes J, Jairath V, Levesque BG: Advances in use of endoscopy, radiology and biomarkers to monitor inflammatory bowel disease. Gastroenterology 152:362, 2017.

Regueiro M: Infliximab reduces endoscopic, but not clinical, recurrence of Crohn's disease after ileocolonic resection. Gastroenterology 150:1568, 2016.

Sandborn WJ: Ustekinumab induction and maintenance therapy in refractory Crohn's disease. N Engl J Med 367:12, 2012.

Sandborn WJ: Tofactitinib as induction and maintenance therapy for ulcerative colitis. N Engl J Med 376:1723, 2017.

320 Irritable Bowel Syndrome

Chung Owyang

Irritable bowel syndrome (IBS) is a functional bowel disorder characterized by abdominal pain or discomfort and altered bowel habits in the absence of detectable structural abnormalities. No clear diagnostic markers exist for IBS; thus the diagnosis of the disorder is based on clinical presentation. In 2016, the Rome III criteria for the diagnosis of IBS were updated to Rome IV **(Table 320-1)**. Throughout the world, about 10–20% of adults and adolescents have symptoms consistent with IBS, and most studies show a female predominance. IBS symptoms tend to come and go over time and often overlap with other functional disorders such as fibromyalgia, headache, backache, and genitourinary symptoms. Severity of symptoms varies and can significantly impair quality of life, resulting in high health care costs. Advances in basic, mechanistic, and clinical investigations have improved our understanding of this disorder and its physiologic and psychosocial determinants. Altered gastrointestinal (GI) motility, visceral hyperalgesia, disturbance of brain–gut interaction, abnormal central processing, autonomic and hormonal events, genetic and environmental factors, and psychosocial disturbances are variably involved, depending on the individual. This progress may result in improved methods of treatment.

CLINICAL FEATURES

IBS is a disorder that affects all ages, although most patients have their first symptoms before age 45. Older individuals have a lower reporting frequency. Women are diagnosed with IBS two to three times as often as men and make up 80% of the population with severe IBS. As indicated in Table 320-1, pain is a key symptom for the diagnosis of IBS. This symptom should be associated with defecation and/or have their onset associated with a change in frequency or form of stool. In comparison to Rome III, the Rome IV criteria is more stringent, requiring abdominal pain to occur at a minimum of once a week and eliminates "discomfort" as one of the criteria. Painless diarrhea or constipation does not fulfill the diagnostic criteria to be classified as IBS. Supportive symptoms that are not part of the diagnostic criteria include defecation straining, urgency or a feeling of incomplete bowel movement, passing mucus, and bloating.

Abdominal Pain According to the current IBS diagnostic criteria, abdominal pain is a prerequisite clinical feature of IBS. Abdominal pain in IBS is highly variable in intensity and location. It is frequently episodic and crampy, but it may be superimposed on a background of constant ache. Pain may be mild enough to be ignored or it may interfere with daily activities. Despite this, malnutrition due to inadequate caloric intake is exceedingly rare with IBS. Sleep deprivation is also unusual because abdominal pain is almost uniformly present only during waking hours. However, patients with severe IBS frequently wake repeatedly during the night; thus, nocturnal pain is a poor discriminating factor between organic and functional bowel disease. Pain is often exacerbated by eating or emotional stress and improved by passage of flatus or stools. In addition, female patients with IBS commonly report worsening symptoms during the premenstrual and menstrual phases.

TABLE 320-1 Rome IV Diagnostic Criteria for Irritable Bowel Syndrome[a]

Recurrent abdominal pain, on average, at least 1 day per week in the last 3 months, associated with ≥2 of the following criteria:

1. Related to defecation
2. Associated with a change in frequency of stool
3. Associated with a change in form (appearance) of stool

[a]Criteria fulfilled for the last 3 months with symptom onset at least 6 months prior to diagnosis.

Altered Bowel Habits Alteration in bowel habits is the most consistent clinical feature in IBS. The most common pattern is constipation alternating with diarrhea, usually with one of these symptoms predominating. At first, constipation may be episodic, but eventually it becomes continuous and increasingly intractable to treatment with laxatives. Stools are usually hard with narrowed caliber, possibly reflecting excessive dehydration caused by prolonged colonic retention and spasm. Most patients also experience a sense of incomplete evacuation, thus leading to repeated attempts at defecation in a short time span. Patients whose predominant symptom is constipation may have weeks or months of constipation interrupted with brief periods of diarrhea. In other patients, diarrhea may be the predominant symptom. Diarrhea resulting from IBS usually consists of small volumes of loose stools. Most patients have stool volumes of <200 mL. Nocturnal diarrhea does not occur in IBS. Diarrhea may be aggravated by emotional stress or eating. Stool may be accompanied by passage of large amounts of mucus. Bleeding is not a feature of IBS unless hemorrhoids are present, and malabsorption or weight loss does not occur.

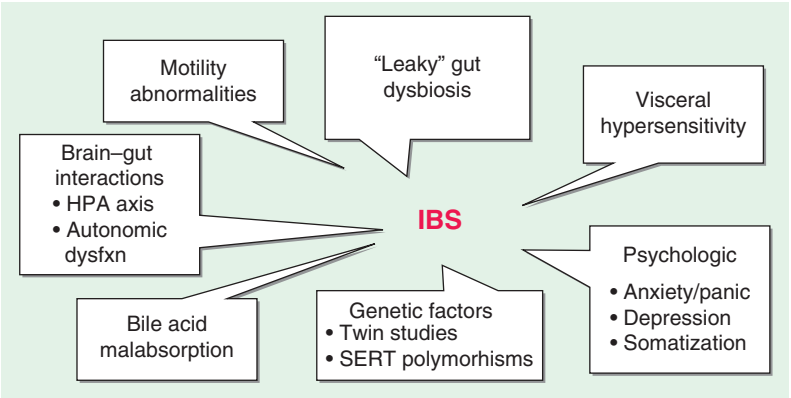


FIGURE 320-1 Pathophysiology of IBS. The cause of IBS is likely to be multifactorial. Patients often show evidence of visceral hypersensitivity and motility abnormalities. Many IBS patients have increased anxiety and/or depression and their symptoms are often exacerbated by mental or physical stress suggesting abnormal brain–gut interaction. Genetic studies suggest a few IBS patients may have genetic abnormalities affecting the serotonin transport system in the enteric nerves. Up to 30% of IBS patients may have bile acid malabsorption. Gut dysbiosis and impaired mucosa permeability also have been reported in many IBS patients. This may lead to subclinical mucosa inflammation.

Bowel pattern subtypes are highly unstable. In a patient population with ~33% prevalence rates of IBS-diarrhea predominant (IBS-D), IBS-constipation predominant (IBS-C), and IBS-mixed (IBS-M) forms, 75% of patients change subtypes and 29% switch between IBS-C and IBS-D over 1 year. The heterogeneity and variable natural history of bowel habits in IBS increase the difficulty of conducting pathophysiology studies and clinical trials.

Gas and Flatulence Patients with IBS frequently complain of abdominal distention and increased belching or flatulence, all of which they attribute to increased gas. Although some patients with these symptoms actually may have a larger amount of gas, quantitative measurements reveal that most patients who complain of increased gas generate no more than a normal amount of intestinal gas. Most IBS patients have impaired transit and tolerance of intestinal gas loads. In addition, patients with IBS tend to reflux gas from the distal to the more proximal intestine, which may explain the belching. Some patients with bloating may also experience visible distention with increase in abdominal girth. Both symptoms are more common among female patients and in those with higher overall Somatic Symptom Checklist scores.

Upper GI Symptoms Between 25 and 50% of patients with IBS complain of dyspepsia, heartburn, nausea, and vomiting. This suggests that other areas of the gut apart from the colon may be involved. Prolonged ambulant recordings of small-bowel motility in patients with IBS show a high incidence of abnormalities in the small bowel during the diurnal (waking) period; nocturnal motor patterns are not different from those of healthy controls. The overlap between dyspepsia and IBS is great. The prevalence of IBS is higher among patients with dyspepsia (31.7%) than among those who reported no symptoms of dyspepsia (7.9%). Conversely, among patients with IBS, 55.6% reported symptoms of dyspepsia. In addition, the functional abdominal symptoms can change over time. Those with predominant dyspepsia or IBS can flux between the two. Although the prevalence of functional GI disorders is stable over time, the turnover in symptom status is high. Many episodes of symptom disappearance are due to subjects changing symptoms rather than total symptom resolution. Thus it is conceivable that functional dyspepsia and IBS are two manifestations of a single, more extensive digestive system disorder. Furthermore, IBS symptoms are prevalent in noncardiac chest pain patients, suggesting overlap with other functional gut disorders.

PATHOPHYSIOLOGY

The pathogenesis of IBS is poorly understood, although roles of abnormal gut motor and sensory activity, central neural dysfunction, psychological disturbances, mucosal inflammation, stress, and luminal factors such as bile acid malabsorption and gut dysbiosis have been proposed (Fig. 320-1).

GI Motor Abnormalities Studies of colonic myoelectrical and motor activity under unstimulated conditions have not shown consistent abnormalities in IBS. In contrast, colonic motor abnormalities are more prominent under stimulated conditions in IBS. IBS patients may exhibit increased rectosigmoid motor activity for up to 3 h after eating. Similarly, inflation of rectal balloons both in IBS-D and IBS-C patients leads to marked and prolonged distention-evoked contractile activity. Recordings from the transverse, descending, and sigmoid colon showed that the motility index and peak amplitude of high-amplitude propagating contractions (HAPCs) in diarrhea-prone IBS patients were greatly increased compared to those in healthy subjects and were associated with rapid colonic transit and accompanied by abdominal pain.

Visceral Hypersensitivity As with studies of motor activity, IBS patients frequently exhibit exaggerated sensory responses to visceral stimulation. The frequency of perceptions of food intolerance is at least twofold more common than in the general population. Postprandial pain has been temporally related to entry of the food bolus into the cecum in 74% of patients. On the other hand, prolonged fasting in IBS patients is often associated with significant improvement in symptoms. Rectal balloon inflation produces nonpainful and painful sensations at lower volumes in IBS patients than in healthy controls without altering rectal tension, suggestive of visceral afferent dysfunction in IBS. Similar studies show gastric and esophageal hypersensitivity in patients with nonulcer dyspepsia and noncardiac chest pain, raising the possibility that these conditions have a similar pathophysiologic basis. Lipids lower the thresholds for the first sensation of gas, discomfort, and pain in IBS patients. Hence, postprandial symptoms in IBS patients may be explained in part by a nutrient-dependent exaggerated sensory component of the gastrocolonic response. In contrast to enhanced gut sensitivity, IBS patients do not exhibit heightened sensitivity elsewhere in the body. Thus, the afferent pathway disturbances in IBS appear to be selective for visceral innervation with sparing of somatic pathways. The mechanisms responsible for visceral hypersensitivity are still under investigation. It has been proposed that these exaggerated responses may be due to (1) increased end-organ sensitivity with recruitment of "silent" nociceptors; (2) spinal hyperexcitability with activation of nitric oxide and possibly other neurotransmitters; (3) endogenous (cortical and brainstem) modulation of caudad nociceptive transmission; and (4) over time, the possible development of long-term hyperalgesia due to development of neuroplasticity, resulting in permanent or semipermanent changes in neural responses to chronic or recurrent visceral stimulation.

Central Neural Dysregulation The role of central nervous system (CNS) factors in the pathogenesis of IBS is strongly suggested by the clinical association of emotional disorders and stress with symptom

exacerbation and the therapeutic response to therapies that act on cerebral cortical sites. Functional brain imaging studies such as magnetic resonance imaging (MRI) have shown that in response to distal colonic stimulation, the mid-cingulate cortex—a brain region concerned with attention processes and response selection—shows greater activation in IBS patients. Modulation of this region is associated with changes in the subjective unpleasantness of pain. In addition, IBS patients also show preferential activation of the prefrontal lobe, which contains a vigilance network within the brain that increases alertness. These may represent a form of cerebral dysfunction leading to the increased perception of visceral pain.

Abnormal Psychological Features Abnormal psychiatric features are recorded in up to 80% of IBS patients, especially in referral centers; however, no single psychiatric diagnosis predominates. Most of these patients demonstrated exaggerated symptoms in response to visceral distention, and this abnormality persists even after exclusion of psychological factors.

Psychological factors influence pain thresholds in IBS patients, as stress alters sensory thresholds. An association between prior sexual or physical abuse and development of IBS has been reported. Clinical studies suggest that IBS has a strong developmental component which involves interactions of genetic and epigenetic factors early in life. These may modulate brain networks related to emotional arousal and/or central autonomic control, salience and somatosensory integration. Abuse is associated with greater pain reporting, psychological distress, and poor health outcome. Brain functional MRI studies show greater activation of the posterior and middle dorsal cingulate cortex, which is implicated in affect processing in IBS patients with a past history of sexual abuse.

Thus, patients with IBS frequently demonstrate increased motor reactivity of the colon and small bowel to a variety of stimuli and altered visceral sensation associated with lowered sensation thresholds. These may result from CNS–enteric nervous system dysregulation.

Postinfectious IBS IBS may be induced by GI infection. In an investigation of 544 patients with confirmed bacterial gastroenteritis, one-quarter developed IBS subsequently. Conversely, about a third of IBS patients experienced an acute "gastroenteritis-like" illness at the onset of their chronic IBS symptomatology. This group of "postinfective" IBS occurs more commonly in females and affects younger rather than older patients. Risk factors for developing postinfectious IBS include, in order of importance, prolonged duration of initial illness, toxicity of infecting bacterial strain, smoking, mucosal markers of inflammation, female sex, depression, hypochondriasis, and adverse life events in the preceding 3 months. Age older than 60 years might protect against postinfectious IBS, whereas treatment with antibiotics has been associated with increased risk. The microbes involved in the initial infection are *Campylobacter, Salmonella*, and *Shigella*. Those patients with *Campylobacter* infection who are toxin-positive are more likely to develop postinfective IBS. Increased rectal mucosal enteroendocrine cells, T lymphocytes, and increased gut permeability are acute changes following *Campylobacter* enteritis that could persist for more than a year and may contribute to postinfective IBS.

Immune Activation and Mucosal Inflammation Some patients with IBS display persistent signs of low-grade mucosal inflammation with activated lymphocytes, mast cells, and enhanced expression of proinflammatory cytokines. Other studies also indicate that peripheral blood mononuclear cells (PBMCs) from IBS patients show abnormal release of proinflammatory cytokines such as IL6, IL1β, and TNF. These abnormalities may contribute to abnormal epithelial secretion and visceral hypersensitivity. There is increasing evidence that some members of the superfamily of transient receptor potential (TRP) cation channels such as TRPV1 (vanilloid) channels are central to the initiation and persistence of visceral hypersensitivity. Mucosal inflammation can lead to increased expression of TRPV1 in the enteric nervous system. Enhanced expression of TRPV1 channels in the sensory neurons of the gut has been observed in IBS, and such expression appears to correlate with visceral hypersensitivity and abdominal pain. Interestingly, clinical studies have also shown increased intestinal permeability in patients with IBS-D. Psychological stress and anxiety can increase the release of proinflammatory cytokine, and this in turn may alter intestinal permeability. A clinical study shows 39% of IBS-D patients had increased intestinal permeability as measured by the lactulose/mannitol ratio. These IBS patients also demonstrated higher Functional Bowel Disorder Severity Index (FBDSI) score and increased hypersensitivity to visceral nociceptive pain stimuli. This provides a functional link between psychological stress, immune activation, and symptom generation in patients with IBS.

Altered Gut Flora A high prevalence of small intestinal bacterial overgrowth in IBS patients has been noted based on positive lactulose hydrogen breath test. This finding, however, has been challenged by a number of other studies that found no increased incidence of bacterial overgrowth based on jejunal aspirate culture. Abnormal H_2 breath test can occur because of small-bowel rapid transit and may lead to erroneous interpretation. Hence, the role of testing for small intestinal bacterial overgrowth in IBS patients remains unclear.

Studies using culture-independent approaches such as 16S rRNA gene-based analysis found significant differences between the molecular profile of the fecal microbiota of IBS patients and that of healthy subjects. Twenty-two studies, comprising 827 subjects, reported significant changes in the microbial communities of healthy individuals versus patients with different subtypes of IBS. Despite a lack of consensus on the exact microbial differences between IBS patients and controls, in general IBS patients had decreased proportions of the genera *Bifidobacterium* and *Lactobacillus* and increased ratios of Firmicutes:Bacteroidetes. It has been speculated that these changes may be related to stress and diet. A temporary reduction in lactobacilli has been reported in animal models of early-life stress. On the other hand, Firmicutes is the dominant phylum in adults consuming a diet high in animal fat and protein. It is conceivable that gut dysbiosis acting in concert with genetic susceptibility and environmental insults may alter mucosal permeability and increase antigen presentation to the immune cells in the lamina propria. This may result in mast cell activation and altered enteric neuronal and smooth muscle function causing IBS symptoms. In addition, release of cytokines and chemokines from mucosal inflammation may generate extra GI symptoms such as chronic fatigue, muscle pain, and anxiety **(Fig. 320-2)**.

Abnormal Serotonin Pathways The serotonin (5-HT)-containing enterochromaffin cells in the colon are increased in a subset of IBS-D patients compared to healthy individuals or patients with ulcerative colitis. Furthermore, postprandial plasma 5-HT levels were significantly higher in this group of patients compared to healthy controls. Tryptophan hydroxylase 1 (TPH1) is the rate-limiting enzyme in enterochromaffin cell 5-HT biosynthesis, functional TPH1 polymorphism has been shown to be associated with IBS habit subtypes. Because serotonin plays an important role in the regulation of GI motility and visceral perception, the increased release of serotonin may contribute to the postprandial symptoms of these patients and provides a rationale for the use of serotonin antagonists in the treatment of this disorder.

APPROACH TO THE PATIENT

Irritable Bowel Syndrome

Because IBS is a disorder for which no pathognomonic abnormalities have been identified, its diagnosis relies on recognition of positive clinical features and elimination of other organic diseases. Symptom-based criteria have been developed for the purpose of differentiating patients with IBS from those with organic diseases. These include the Manning, Rome I, Rome II, Rome III, and Rome IV criteria. Rome IV criteria for the diagnosis of IBS were published in 2016 (Table 320-1) and defined IBS on the basis of abdominal pain and altered bowel habits that occur with sufficient frequency in affected patients. A careful history and physical examination are frequently helpful in establishing the diagnosis. Clinical features suggestive of

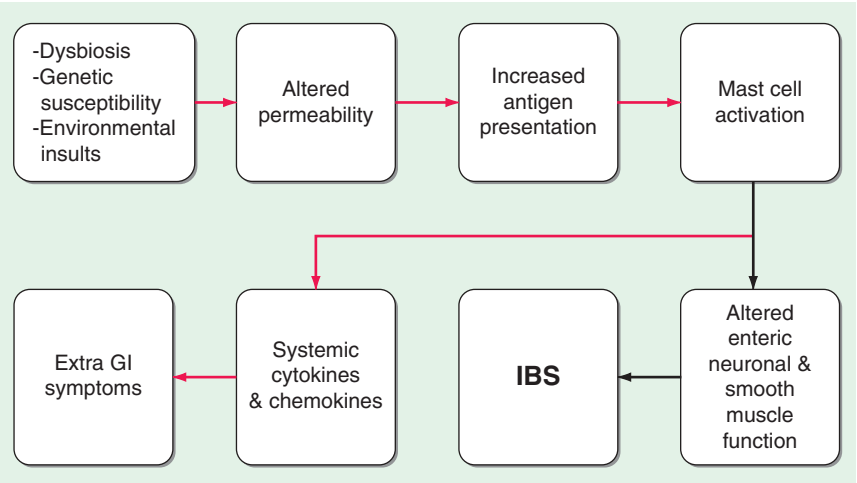


FIGURE 320-2 Gut dysbiosis and IBS. Gut dysbiosis acting in concert with genetic and environmental factors may alter intestinal permeability, increase antigen presentation resulting in mast cell activation. Products of mast cell degranulation may alter neuronal and smooth muscle function causing IBS symptoms. The cytokines and chemokines generated from mucosal inflammation may cause symptoms such as fibromyalgia, chronic fatigue, and mood changes. (*Adapted from NJ Talley, AA Fodor: Gastroenterology 141: 1555, 2011.*)

IBS include the following: recurrence of lower abdominal pain with altered bowel habits over a period of time without progressive deterioration, onset of symptoms during periods of stress or emotional upset, absence of other systemic symptoms such as fever and weight loss, and small-volume stool without any evidence of blood.

On the other hand, the appearance of the disorder for the first time in old age, progressive course from time of onset, persistent diarrhea after a 48-h fast, and presence of nocturnal diarrhea or steatorrheal stools argue against the diagnosis of IBS.

Because the major symptoms of IBS—abdominal pain, abdominal bloating, and alteration in bowel habits—are common complaints of many GI organic disorders, the list of differential diagnoses is a long one. The quality, location, and timing of pain may be helpful to suggest specific disorders. Pain due to IBS that occurs in the epigastric or periumbilical area must be differentiated from biliary tract disease, peptic ulcer disorders, intestinal ischemia, and carcinoma of the stomach and pancreas. If pain occurs mainly in the lower abdomen, the possibility of diverticular disease of the colon, inflammatory bowel disease (including ulcerative colitis and Crohn's disease), and carcinoma of the colon must be considered. Postprandial pain accompanied by bloating, nausea, and vomiting suggests gastroparesis or partial intestinal obstruction. Intestinal infestation with *Giardia lamblia* or other parasites may cause similar symptoms. When diarrhea is the major complaint, the possibility of lactase deficiency, laxative abuse, malabsorption, celiac sprue, hyperthyroidism, inflammatory bowel disease, and infectious diarrhea must be ruled out. On the other hand, constipation may be a side effect of many different drugs, such as anticholinergic, antihypertensive, and antidepressant medications. Endocrinopathies such as hypothyroidism and hypoparathyroidism must also be considered in the differential diagnosis of constipation, particularly if other systemic signs or symptoms of these endocrinopathies are present. In addition, acute intermittent porphyria and lead poisoning may present in a fashion similar to IBS, with painful constipation as the major complaint. These possibilities are suspected on the basis of their clinical presentations and are confirmed by appropriate serum and urine tests.

Few tests are required for patients who have typical IBS symptoms and no alarm features. Unnecessary investigations may be costly and even harmful. The American Gastroenterological Association has delineated factors to be considered when determining the aggressiveness of the diagnostic evaluation. These include the duration of symptoms, the change in symptoms over time, the age and sex of the patient, the referral status of the patient, prior diagnostic studies, a family history of colorectal malignancy, and the degree of psychosocial dysfunction. Thus, a younger individual with mild symptoms requires a minimal diagnostic evaluation, while an older person or an individual with rapidly progressive symptoms should undergo a more thorough exclusion of organic disease. Most patients should have a complete blood count and sigmoidoscopic examination; in addition, stool specimens should be examined for ova and parasites in those who have diarrhea. In patients with persistent diarrhea not responding to simple antidiarrheal agents, a sigmoid colon biopsy should be performed to rule out microscopic colitis. In those age >40 years, an air-contrast barium enema or colonoscopy should also be performed. If the main symptoms are diarrhea and increased gas, the possibility of lactase deficiency should be ruled out with a hydrogen breath test or with evaluation after a 3-week lactose-free diet. Excessive gas with bloating also raises the possibility of small bowel bacteria overgrowth and should be ruled out with a glucose hydrogen breath test. Some patients with IBS-D may have undiagnosed celiac sprue. Because the symptoms of celiac sprue respond to a gluten-free diet, testing for celiac sprue in IBS may prevent years of morbidity and attendant expense. Decision-analysis studies show that serology testing for celiac sprue in patients with IBS-D has an acceptable cost when the prevalence of celiac sprue is >1% and is the dominant strategy when the prevalence is >8%. In patients with concurrent symptoms of dyspepsia, upper GI radiographs or esophagogastroduodenoscopy may be advisable. In patients with postprandial right upper quadrant pain, an ultrasonogram of the gallbladder should be obtained. Laboratory features that argue against IBS include evidence of anemia, elevated sedimentation rate, presence of leukocytes or blood in stool, and stool volume >200–300 mL/d. These findings would necessitate other diagnostic considerations.

TREATMENT

Irritable Bowel Syndrome

Patient Counseling and Dietary Alterations Reassurance and careful explanation of the functional nature of the disorder and of how to avoid obvious food precipitants are important first steps in patient counseling and dietary change. Occasionally, a meticulous dietary history may reveal substances (such as coffee, disaccharides, legumes, and cabbage) that aggravate symptoms. Excessive fructose and artificial sweeteners, such as sorbitol or mannitol, may cause diarrhea, bloating, cramping, or flatulence. As a therapeutic trial, patients should be encouraged to eliminate any foodstuffs that appear to produce symptoms. However patients should avoid

TABLE 320-2 Some Common Food Sources of FODMAPs

FOOD TYPE	FREE FRUCTOSE	LACTOSE	FRUCTANS	GALACTO-OLIGOSACCHARIDES	POLYOLS
Fruits	Apple, cherry, mango, pear, watermelon		Peach, persimmon, watermelon		Apple, apricot, pear, avocado, blackberries, cherry, nectarine, plum, prune
Vegetables	Asparagus, artichokes, sugar snap peas		Artichokes, beetroot, Brussels sprout, chicory, fennel, garlic, leek, onion, peas		Cauliflower, mushroom, snow peas
Grains and cereals			Wheat, rye, barley		
Nuts and seeds			Pistachios		
Milk and milk products		Milk, yogurt, ice cream, custard, soft cheeses			
Legumes			Legumes, lentils, chickpeas	Legumes, chickpeas, lentils	
Other	Honey, high-fructose corn syrup		Chicory drinks		
Food additives			Inulin, FOS		Sorbitol, mannitol, maltitol, xylitol, isomalt

Abbreviations: FODMAPs, fermentable oligosaccharides, disaccharides, monosaccharides, and polyols; FOS, fructo-oligosaccharides.

Source: Adapted from PR Gibson et al: Am J Gastroenterol 107:657, 2012.

nutritionally depleted diets. A diet low in fermentable oligosaccharides, disaccharides, monosaccharides, and polyols (FODMAPs) **(Table 320-2)** has been shown to be helpful in IBS patients (see Low FODMAP Diet).

Stool-Bulking Agents High-fiber diets and bulking agents, such as bran or hydrophilic colloid, are frequently used in treating IBS. The water-holding action of fibers may contribute to increased stool bulk because of the ability of fiber to increase fecal output of bacteria. Fiber also speeds up colonic transit in most persons. In diarrhea-prone patients, whole-colonic transit is faster than average; however, dietary fiber can delay transit. Furthermore, because of their hydrophilic properties, stool-bulking agents bind water and thus prevent both excessive hydration and dehydration of stool. The latter observation may explain the clinical experience that a high-fiber diet relieves diarrhea in some IBS patients. Fiber supplementation with psyllium has been shown to reduce perception of rectal distention, indicating that fiber may have a positive effect on visceral afferent function.

The beneficial effects of dietary fiber on colonic physiology suggest that dietary fiber should be an effective treatment for IBS patients, but controlled trials of dietary fiber have produced variable results. This is not surprising since IBS is a heterogeneous disorder, with some patients being constipated and other having predominant diarrhea. Most investigations report increases in stool weight, decreases in colonic transit times, and improvement in constipation. Others have noted benefits in patients with alternating diarrhea and constipation, pain, and bloating. However, most studies observe no responses in patients with diarrhea- or pain-predominant IBS. It is possible that different fiber preparations may have dissimilar effects on selected symptoms in IBS. A cross-over comparison of different fiber preparations found that psyllium produced greater improvements in stool pattern and abdominal pain than bran. Furthermore, psyllium preparations tend to produce less bloating and distention. Despite the equivocal data regarding efficacy, most gastroenterologists consider stool-bulking agents worth trying in patients with IBS-C. Fiber should be started at a nominal dose and slowly titrated up as tolerated over the course of several weeks to a targeted dose of 20–30 g of total dietary and supplementary fiber per day. Even when used judiciously, fiber can exacerbate bloating, flatulence, constipation, and diarrhea.

Antispasmodics Clinicians have observed that anticholinergic drugs may provide temporary relief for symptoms such as painful cramps related to intestinal spasm. Although controlled clinical trials have produced mixed results, evidence generally supports beneficial effects of anticholinergic drugs for pain. A meta-analysis of 26 double-blind clinical trials of antispasmodic agents in IBS reported better global improvement (62%) and abdominal pain reductions (64%) compared to placebo (35 and 45%, respectively), suggesting efficacy in some patients. The drugs are most effective when prescribed in anticipation of predictable pain. Physiologic studies demonstrate that anticholinergic drugs inhibit the gastrocolic reflex; hence, postprandial pain is best managed by giving antispasmodics 30 min before meals so that effective blood levels are achieved shortly before the anticipated onset of pain. Most anticholinergics contain natural belladonna alkaloids, which may cause xerostomia, urinary hesitancy and retention, blurred vision, and drowsiness. They should be used in the elderly with caution. Some physicians prefer to use synthetic anticholinergics such as dicyclomine that have less effect on mucous membrane secretions and produce fewer undesirable side effects.

Antidiarrheal Agents Peripherally acting opiate-based agents are the initial therapy of choice for IBS-D. Physiologic studies demonstrate increases in segmenting colonic contractions, delays in fecal transit, increases in anal pressures, and reductions in rectal perception with these drugs. When diarrhea is severe, especially in the painless diarrhea variant of IBS, small doses of loperamide, 2–4 mg every 4–6 h up to a maximum of 12 g/d, can be prescribed. These agents are less addictive than paregoric, codeine, or tincture of opium. In general, the intestines do not become tolerant of the antidiarrheal effect of opiates, and increasing doses are not required to maintain antidiarrheal potency. These agents are most useful if taken before anticipated stressful events that are known to cause diarrhea. However, not infrequently, a high dose of loperamide may cause cramping because of increases in segmenting colonic contractions. Another antidiarrheal agent that may be used in IBS patients is the bile acid binder cholestyramine resin as up to 30% of IBS-D patients may have bile acid malabsorption.

Antidepressant Drugs In addition to their mood-elevating effects, antidepressant medications have several physiologic effects that suggest they may be beneficial in IBS. In IBS-D patients, the tricyclic antidepressant imipramine slows jejunal migrating motor complex transit propagation and delays orocecal and whole-gut transit, indicative of a motor inhibitory effect. Some studies also suggest that tricyclic agents may alter visceral afferent neural function.

A number of studies indicate that tricyclic antidepressants may be effective in some IBS patients. In a 2-month study of desipramine, abdominal pain improved in 86% of patients compared to 59% given placebo. Another study of desipramine in 28 IBS patients showed

improvement in stool frequency, diarrhea, pain, and depression. When stratified according to the predominant symptoms, improvements were observed in IBS-D patients, with no improvement being noted in IBS-C patients. The beneficial effects of the tricyclic compounds in the treatment of IBS appear to be independent of their effects on depression. The therapeutic benefits for the bowel symptoms occur faster and at a lower dosage. The efficacy of antidepressant agents in other chemical classes in the management of IBS is less well evaluated. In contrast to tricyclic agents, the selective serotonin reuptake inhibitor (SSRI) paroxetine accelerates orocecal transit, raising the possibility that this drug class may be useful in IBS-C patients. The SSRI citalopram blunts perception of rectal distention and reduces the magnitude of the gastrocolonic response in healthy volunteers. A small placebo-controlled study of citalopram in IBS patients reported reductions in pain. However, these findings could not be confirmed in another randomized controlled trial that showed that citalopram at 20 mg/d for 4 weeks was not superior to placebo in treating nondepressed IBS patients. Hence, the efficacy of SSRIs in the treatment of IBS needs further confirmation.

Antiflatulence Therapy The management of excessive gas is seldom satisfactory, except when there is obvious aerophagia or disaccharidase deficiency. Patients should be advised to eat slowly and not chew gum or drink carbonated beverages. Bloating may decrease if an associated gut syndrome such as IBS or constipation is improved. If bloating is accompanied by diarrhea and worsens after ingesting dairy products, fresh fruits, vegetables, or juices, further investigation or a dietary exclusion trial may be worthwhile. Avoiding flatogenic foods, exercising, losing excess weight, and taking activated charcoal are safe but unproven remedies. A low FODMAP diet has been shown to be quite effective to reduce gas and bloating (see Low FODMAP Diet). Data regarding the use of surfactants such as simethicone are conflicting. Antibiotics may help in a subgroup of IBS patients with predominant symptoms of bloating. Beano, an over-the-counter oral β-glycosidase solution, may reduce rectal passage of gas without decreasing bloating and pain. Pancreatic enzymes reduce bloating, gas, and fullness during and after high-calorie, high-fat meal ingestion.

Modulation of Gut Flora Because altered colonic flora (gut dysbiosis) may contribute to the pathogenesis of IBS, this has led to great interest in using antibiotics, prebiotics, and probiotics to treat IBS **(Fig. 320-3)**.

Antibiotics Antibiotic treatment benefits a subset of IBS patients. In a double-blind, randomized, placebo-controlled study, neomycin dosed at 500 mg twice daily for 10 days was more effective than placebo at improving symptom scores among IBS patients. The nonabsorbed oral antibiotic rifaximin is the most thoroughly studies antibiotic of the treatment of IBS. In a double-blind, placebo-controlled study, patients receiving rifaximin at a dose of 550 mg two times daily for 2 weeks experienced substantial improvement of global IBS symptoms over placebo. Rifaximin is the only antibiotic with demonstrated sustained benefit beyond therapy cessation in IBS patients. The drug has a favorable safety and tolerability profile compared with systemic antibiotics. A systemic review and meta-analysis of five studies of IBS patients found that rifaximin is more effective than placebo for global symptoms and bloating (odds ratio 1.57) with a number needed to treat (NNT) of 10.2. The modest therapeutic gain was similar to that yielded by other current available therapies for IBS. However, currently there are still insufficient data to recommend routine use of this antibiotic in the treatment of IBS.

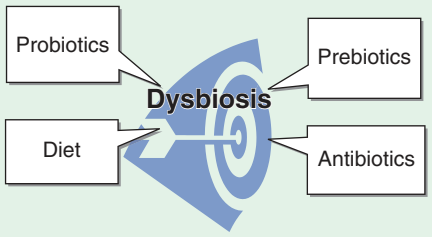


FIGURE 320-3 Gut dysbiosis: a potential treatment target. Prebiotics, probiotics, and low FODMAP diet may be used to modulate gut flora and treat IBS.

Prebiotics These are nondigestible food ingredients that stimulate growth and/or activity of bacteria in the GI tract. There have been four randomized trials to examine the effects of prebiotics. Three of the four studies reported that prebiotics worsened or did not improve IBS symptoms. This is not surprising given the adverse effects of high carbohydrate diet on IBS symptoms.

Probiotics These are defined as live microorganisms that when administered in adequate amounts confer a health benefit on the host. A meta-analysis of 10 probiotic studies in IBS patients found significant relief of pain and bloating with the use of *Bifidobacterium breve, B longum,* and *Lactobacillus acidophilus* species compared to placebo. However, there was no change in stool frequency or consistency. Large-scale studies of well-phenotyped IBS patients are needed to establish the efficacy of these probiotics.

Low FODMAP Diet A diet rich in FODMAP (fermentable oligosaccharides, disaccharides, monosaccharides, and polyols) often triggers symptoms in IBS patients. FODMAPs are poorly absorbed by the small intestine and fermented by bacteria in the colon to produce gas and osmotically active carbohydrates **(Fig. 320-4)**. At the same time, on entering the colon, FODMAPs may serve as nutrient for the colonic bacteria and promote the growth of gram negative commensal bacteria which may induce epithelial damage and subclinical mucosa inflammation. Fructose and fructans induce IBS symptoms in a dose-dependent manner. In contrast, a low FODMAP diet reduces IBS symptoms. A randomized controlled trial showed a 4-week low FODMAP diet improved symptoms in 68% of IBS patients compared with 23% on a habitual diet. Low FOMAP diets appeared to be superior to national guidelines for IBS management. These observations were confirmed by a double blind controlled study of 30 IBS patients and 18 healthy controls, in which a low FODMAP diet significantly reduced bloating, pain, passage of gas, and diarrhea. A double blind randomized control trial involving 92 IBS patients showed >50% of patients on the low FODMAP diet had major improvement of their abdominal pain compared

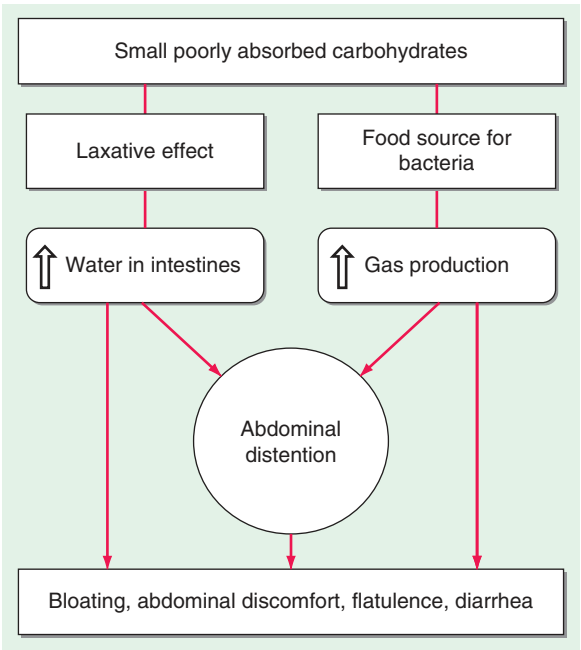


FIGURE 320-4 Pathogenesis of FODMAP-related symptoms. FODMAPs are poorly absorbed by the small intestine and fermented by gut bacteria to produce gas and osmotically active carbohydrates. These events act in concert to cause bloating, flatulence, and diarrhea. FODMAP may also serve as nutrients for colonic bacteria which may induce mucosa inflammation. *(Figure created using data from http://www.nutritiontoyou.com/wp-content/uploads/2014/06/IBS-symptoms.png.)*

with 20% of the control group. These observations demonstrate the impressive efficacy of low FODMAP diet for many IBS patients, and if confirmed may justify the recommendation of a low FODMAP diet as first-line treatment for IBS patients.

Serotonin Receptor Agonist and Antagonists Serotonin receptor antagonists have been evaluated as therapies for IBS-D. Serotonin acting on 5-HT_3 receptors enhances the sensitivity of afferent neurons projecting from the gut. In humans, a 5-HT_3 receptor antagonist such as alosetron reduces perception of painful visceral stimulation in IBS. It also induces rectal relaxation, increases rectal compliance, and delays colonic transit. Meta-analysis of 14 randomized controlled trials of alosetron or cilansetron showed that these antagonists are more effective than placebo in achieving global improvement in IBS symptoms and relief of abdominal pain and discomfort. These agents are more likely to cause constipation in IBS patients with diarrhea alternating with constipation. Also, 0.2% of patients using 5-HT_3 antagonists developed ischemic colitis versus none in the control group. In postrelease surveillance, 84 cases of ischemic colitis were observed, including 44 cases that required surgery and 4 deaths. As a consequence, the medication was voluntarily withdrawn by the manufacturer in 2000. Alosetron has been reintroduced under a new risk-management program where patients have to sign a patient-physician agreement. This has significantly limited its usage.

Novel 5-HT_4 receptor agonists such as tegaserod exhibit prokinetic activity by stimulating peristalsis. In IBS patients with constipation, tegaserod accelerated intestinal and ascending colon transit. Clinical trials involving >4000 IBS-C patients reported reductions in discomfort and improvements in constipation and bloating, compared to placebo. Diarrhea is the major side effect. However, tegaserod has been withdrawn from the market; a meta-analysis revealed an increase in serious cardiovascular events.

Chloride Channel Activators Lubiprostone is a bicyclic fatty acid that stimulates chloride channels in the apical membrane of intestinal epithelial cells. Chloride secretion induces passive movement of sodium and water into the bowel lumen and improves bowel function. Oral lubiprostone was effective in the treatment of patients with constipation-predominant IBS in large phase II and phase III randomized, double-blinded, placebo-controlled multicenter trials. Responses were significantly greater in patients receiving lubiprostone 8 μg twice daily for 3 months than in those receiving placebo. In general, the drug was quite well tolerated. The major side effects are nausea and diarrhea. Lubiprostone is a new class of compounds for treatment of chronic constipation with or without IBS.

Guanylate Cyclase-C Agonist Linaclotide is a minimally absorbed 14-amino-acid peptide guanylate cyclase-C (GC-C) agonist that binds to and activates GC-C on the luminal surface of intestinal epithelium. Activation of GC-C results in generation of cyclic guanosine monophosphate (cGMP), which triggers secretion of fluid, sodium, and bicarbonate. In animal models, linaclotide accelerates GI transit and reduces visceral nociception. The analgesic action of linaclotide appears to be mediated by cGMP acting on afferent pain fibers innervating the GI tract. A phase III, double-blind, controlled trial showed that linaclotide, 290 μg given once daily, significantly improved abdominal pain, bloating, and spontaneous bowel movement. The only significant side effect was diarrhea, which occurred in 4.5% of the patients. The drug has been approved for treatment of constipation in IBS-C patients.

TABLE 320-3 Spectrum of Severity in IBS

	MILD	MODERATE	SEVERE
Clinical Features			
Prevalence	70%	25%	5%
Correlations with gut physiology	+++	++	+
Symptoms constant	0	+	+++
Psychosocial difficulties	0	+	+++
Health care issues	+	++	+++
Practice type	Primary	Specialty	Referral

TABLE 320-4 Possible Drugs for a Dominant Symptom in IBS

SYMPTOM	DRUG	DOSE
Diarrhea	Loperamide	2–4 mg when necessary/ maximum 12 g/d
	Cholestyramine resin	4 g with meals
	Alosetron[a]	0.5–1 mg bid (for severe IBS, women)
Constipation	Psyllium husk	3–4 g bid with meals, then adjust
	Methylcellulose	2 g bid with meals, then adjust
	Calcium polycarbophil	1 g qd to qid
	Lactulose syrup	10–20 g bid
	70% sorbitol	15 mL bid
	Polyethylene glycol 3350	17 g in 250 mL water qd
	Lubiprostone (Amitiza)	24 mg bid
	Magnesium hydroxide	30–60 mL qd
	Linaclotide	290 μg qd
Abdominal pain	Smooth-muscle relaxant	qd to qid ac
	Tricyclic antidepressants	Start 25–50 mg hs, then adjust
	Selective serotonin reuptake inhibitors	Begin small dose, increase as needed
Gas and bloating	Low FODMAP diet	
	Probiotics	qd
	Rifaximin	550 mg bid

[a]Available only in the United States.

Abbreviation: FODMAP, fermentable oligosaccharides, disaccharides, monosaccharides, and polyols.

Source: Adapted from GF Longstreth et al: Gastroenterology 130:1480, 2006.

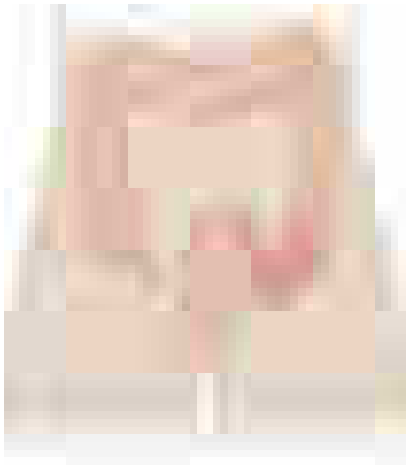
Summary The treatment strategy of IBS depends on the severity of the disorder **(Table 320-3)**. Most IBS patients have mild symptoms. They are usually cared for in primary care practices, have little or no psychosocial difficulties, and do not seek health care often. Treatment usually involves education, reassurance, and dietary/lifestyle changes. A smaller portion have moderate symptoms that are usually intermittent and correlate with altered gut physiology, e.g., worsened with eating or stress and relieved by defecation. For IBS-D patients, treatments include gut-acting pharmacologic agents such as antispasmodics, antidiarrheals, bile acid binders, and the newer gut serotonin modulators **(Table 320-4)**. In IBS-C patients, increased fiber intake and the use of osmotic agents such as polyethylene glycol may achieve satisfactory results. For patients with more severe constipation, a chloride channel opener (lubiprostone) or GC-C agonist (linaclotide) may be considered. For IBS patients with predominant gas and bloating, a low-FODMAP diet may provide significant relief. Some patients may benefit from probiotics and rifaximin treatment. A small proportion of IBS patients have severe and refractory symptoms, are usually seen in referral centers, and frequently have constant pain and psychosocial difficulties. This group of patients is best managed with antidepressants and other psychological treatments (Table 320-4). Clinical trials demonstrating success of low FODMAP diet in improving IBS symptoms and quality of life provide strong evidence supporting the use of this dietary approach in the treatment of IBS. These observations, if confirmed, may lead to the use of low FODMAP diet as the first line of treatment of IBS patients with moderate to severe symptoms.

FURTHER READING

BARBARA G et al: The immune system in irritable bowel syndrome. J Neurogastroenterol Motil 17:349, 2011.

DROSSMAN DA: Functional gastrointestinal disorders: History, pathophysiology, clinical features, and Rome IV. Gastroenterology 150:1262, 2016.

Eswaran SL et al: A randomized controlled trial comparing the low FODMAP diet vs. modified NICE guidelines in US adults with IBS-D. Am J Gastroenterol 111:1824, 2016.
Mayer EA et al: Brain-gut microbiome interactions and functional bowel disorders. Gastroenterology 146:1500; 2014.
Mayer EA et al: Towards a systems view of IBS. Nar Rev Gastroenterol Hepatol 12:592, 2015.
Natividad JM, Verdu EF: Modulation of intestinal barrier by intestinal microbiota: Pathological and therapeutic implications. Pharmacological Research 69:42, 2013.
Simren M et al: Intestinal microbiota in functional bowel disorders: A Rome foundation report. Gut 62:159, 2013.
Talley NJ, Fodor AA: Bugs, stool, and the irritable bowel syndrome: Too much is as bad as too little? Gastroenterology 141:1555, 2011.



321 Diverticular Disease and Common Anorectal Disorders

Rizwan Ahmed, Susan L. Gearhart

DIVERTICULAR DISEASE

Incidence and Epidemiology In the United States, diverticulosis affects 60% of the population aged >60 and up to 30% of individuals with diverticular disease will experience recurrent symptoms. Diverticular disease has become the fifth most costly gastrointestinal disorder in the United States and is the leading indication for elective colon resection. The incidence of diverticular disease is on the rise. Fortunately, only 20% of patients with diverticulosis develop diverticular disease and 4% require hospitalization. Previously overlooked, the majority of patients with diverticular disease report a lower health-related quality of life and more depression as compared to matched controls, thus adding to health care costs. Formerly, diverticular disease was confined to developed countries; however, with the adoption of westernized diets in underdeveloped countries, diverticulosis is on the rise across the globe. Immigrants to the United States develop diverticular disease at the same rate as U.S. natives. Although the prevalence among females and males is similar, males tend to present at a younger age. The mean age at presentation is now shifting to affect younger populations.

Anatomy and Pathophysiology Two types of diverticula occur in the intestine: true and false (or pseudo diverticula). A true diverticulum is a saclike herniation of the entire bowel wall, whereas a pseudo diverticulum involves only a protrusion of the mucosa and submucosa through the muscularis propria of the colon **(Fig. 321-1)**. The type of diverticulum most commonly affecting the colon is the pseudo diverticulum. Diverticula commonly affect the left and sigmoid colon; the rectum is always spared. However, in Asian populations, 70% of diverticula are seen in the right colon and cecum as well. Yamanda et al. found right-side colonic diverticulosis in 22% of Japanese patients undergoing colonoscopy. *Diverticulitis* is inflammation of a diverticulum. Previous understanding of the pathogenesis of diverticulosis attributed a low-fiber diet as the sole culprit, and onset of diverticulitis would occur acutely when these diverticula become obstructed. However, evidence now suggests that the pathogenesis is more complex and multifactorial. The diverticula occur at the point where the nutrient artery, or *vasa recti*, penetrates through the muscularis propria, resulting in a break in the integrity of the colonic wall. This anatomic restriction may be a result of the relative high-pressure zone within the muscular sigmoid colon. Thus, higher-amplitude contractions combined with constipated, high-fat-content stool within the sigmoid lumen in an area of weakness in the colonic wall results in the creation of these diverticula. Consequently, the vasa recti is either compressed or eroded, leading to either perforation or bleeding. Chronic low-grade inflammation is thought to play a key role in neuronal degeneration leading to dysmotility and high intraluminal pressure. As a consequence, pockets or outpouchings develop in the colonic wall where it is weakest. Furthermore, better understanding of the gut microbiota suggests that dysbiosis is an important aspect of disease.

FIGURE 321-1 Gross and microscopic view of sigmoid diverticular disease. *Arrows* mark an inflamed diverticulum with the diverticular wall made up only of mucosa.

Presentation, Evaluation, and Management of Diverticular Bleeding Hemorrhage from a colonic diverticulum is the most common cause of hematochezia in patients >60 years, yet only 20% of patients with diverticulosis will have gastrointestinal bleeding. Patients at increased risk for bleeding tend to be hypertensive, have atherosclerosis, and regularly use aspirin and nonsteroidal anti-inflammatory agents. Most bleeds are self-limited and stop spontaneously with bowel rest. The lifetime risk of rebleeding is 25%.

Initial localization of diverticular bleeding may include colonoscopy, multiplanar computed tomography (CT) angiogram, or nuclear medicine tagged red cell scan. If the patient is stable, ongoing bleeding is best managed by angiography. If mesenteric angiography can localize the bleeding site, the vessel can be occluded successfully with a coil in

80% of cases. The patient can then be followed closely with repetitive colonoscopy, if necessary, looking for evidence of colonic ischemia. Alternatively, a segmental resection of the colon can be undertaken to eliminate the risk of further bleeding. This may be advantageous in patients on chronic anticoagulation. However, with highly selective coil embolization, the rate of colonic ischemia is <10% and the risk of acute rebleeding is <25%. Long-term results (40 months) indicate that >50% of patients with acute diverticular bleeds treated with highly selective angiography have had definitive treatment. As another alternative, a selective infusion of vasopressin can be given to stop the hemorrhage, although this has been associated with significant complications, including myocardial infarction and intestinal ischemia. Furthermore, bleeding recurs in 50% of patients once the infusion is stopped.

If the patient is unstable or has had a 6-unit bleed within 24 h, current recommendations are that surgery should be performed. If the bleeding has been localized, a segmental resection can be performed. If the site of bleeding has not been definitively identified, a subtotal colectomy may be required. In patients without severe comorbidities, surgical resection can be performed with a primary anastomosis. A higher anastomotic leak rate has been reported in patients who received >10 units of blood.

Presentation, Evaluation, and Staging of Diverticulitis

Acute uncomplicated diverticulitis (also known as Symptomatic Uncomplicated Diverticular Disease, SUDD) characteristically presents with fever, anorexia, left lower quadrant abdominal pain, and obstipation **(Table 321-1)**. In <25% of cases, patients may present with generalized peritonitis indicating the presence of a diverticular perforation. If a pericolonic abscess has formed, the patient may have abdominal distention and signs of localized peritonitis. Laboratory investigations will demonstrate a leukocytosis. Rarely, a patient may present with an air-fluid level in the left lower quadrant on plain abdominal film. This is a giant diverticulum of the sigmoid colon and is managed with resection to avoid impending perforation.

The diagnosis of diverticulitis is best made on CT with the following findings: sigmoid diverticula, thickened colonic wall >4 mm, and inflammation within the periodic fat ± the collection of contrast material or fluid. In up to 20% of patients, an abdominal abscess may be present. Symptoms of irritable bowel syndrome **(Chap. 320)** may mimic those of diverticulitis. Therefore, suspected diverticulitis that does not meet CT criteria or is not associated with a leukocytosis or fever is not diverticular disease. Other conditions that can mimic diverticular disease include an ovarian cyst, endometriosis, acute appendicitis, and pelvic inflammatory disease.

Although the benefit of colonoscopy in the evaluation of patients with diverticular disease has been called into question, its use is still considered important in the exclusion of colorectal cancer. The parallel epidemiology of colorectal cancer and diverticular disease provides enough concern for an endoscopic evaluation before operative management. Therefore, a colonoscopy should be performed ~6 weeks after an attack of diverticular disease.

Complicated diverticular disease is defined as diverticular disease associated with an abscess or perforation and less commonly with a fistula (Table 321-1). Perforated diverticular disease is staged using the Hinchey classification system **(Fig. 321-2)**. This staging system was developed to predict outcomes following the surgical management of complicated diverticular disease. In recent years, the Hinchey staging system has been modified to include the development of a phlegmon or early abscess (Hinchey stage Ia). A pericolic abscess is then considered Hinchey stage Ib. In complicated diverticular disease with fistula formation, common locations include cutaneous, vaginal, or vesicle fistulas. These conditions present with either passage of stool through the skin or vagina or the presence of air in the urinary stream (pneumaturia). Colovaginal fistulas are more common in women who have undergone a hysterectomy.

TABLE 321-1 Presentation of Diverticular Disease

Uncomplicated Diverticular Disease—75%
Abdominal pain
Fever
Leukocytosis
Anorexia/obstipation
Complicated Diverticular Disease—25%
Abscess 16%
Perforation 10%
Stricture 5%
Fistula 2%

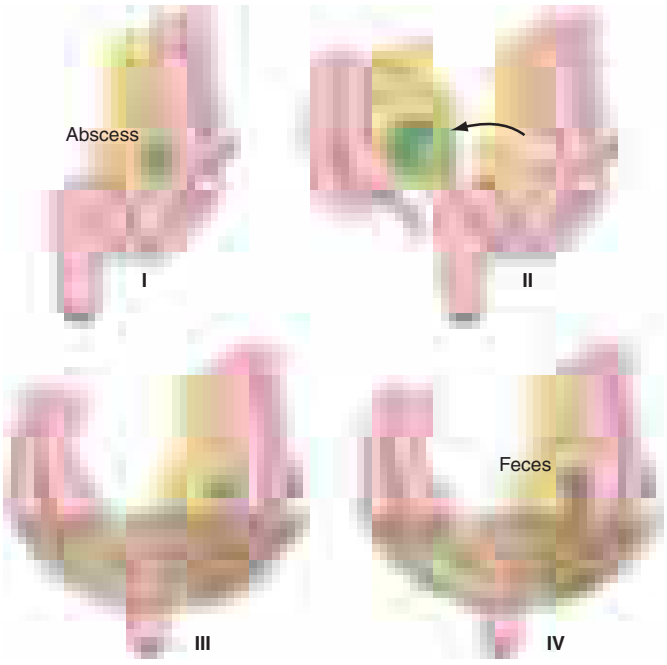


FIGURE 321-2 Hinchey classification of diverticulitis. Stage I: Perforated diverticulitis with a confined paracolic abscess. Stage II: Perforated diverticulitis that has closed spontaneously with distant abscess formation. Stage III: Noncommunicating perforated diverticulitis with fecal peritonitis (the diverticular neck is closed off, and, therefore, contrast will not freely expel on radiographic images). Stage IV: Perforation and free communication with the peritoneum, resulting in fecal peritonitis.

TREATMENT

Diverticular Disease

MEDICAL MANAGEMENT

Asymptomatic diverticular disease discovered on imaging studies or at the time of colonoscopy is best managed by lifestyle changes. Although the data regarding dietary risks and symptomatic diverticular disease are limited **(see Table 321-2)**, patients may benefit from a fiber-enriched diet that includes 30 g of fiber each day. Supplementary fiber products such as Metamucil, Fibercon, or Citrucel are useful. The use of fiber increases colonic transit time, and, therefore, preventing increased intraluminal pressure leading to the development of diverticulosis. The incidence of complicated diverticular disease appears to also be increased in patients who smoke. Therefore, patients should be encouraged to refrain from smoking. The historical recommendation to avoid eating nuts is based on no more than anecdotal data.

SUDD with confirmation of inflammation and infection within the colon should be treated initially with bowel rest. The routine use of antibiotics in uncomplicated diverticular disease did not demonstrate any benefit in time to symptom resolution, complications, or risk of recurrence. However, the data are limited and antibiotics remain in the treatment paradigm. Hospitalization is recommended if the patient is unable to take oral therapy, affected by several comorbidities, fails to improve with outpatient therapy,

TABLE 321-2 The Use of Fiber in the Management of Diverticular Disease

JOURNAL, STUDY YEAR	PATIENTS (n)	INTERVENTION	STUDY LENGTH	FINDINGS
Lancet, 1977	18	Wheat or bran crisp bread	3 months	Significant reduction of symptoms score
BMJ, 1981	58	Bran, ispaghula, placebo	16 weeks	No difference
J Gastroenterol, 1977	30	Methylcellulose	3 months	Significant reduction in symptoms
BMJ, 2011	47,033	Vegetarian vs nonvegetarian	11.6 years	Vegetarians had a 31% lower risk of DD
Gastroenterology, 2012	2104	Fiber consumption	12 years	Fiber associated with great risk of DD
Jama, 2008	47,288	Nut, corn, popcorn consumption	18 years	Higher nut, corn, and popcorn had lower risk of recurrence
Ann R Coll Surg Engl, 1985	56	Fiber consumption	66 months	Higher fiber associated with 19% reduction in symptom recurrence

Modified from A Turis, A Papa, S Danese: Review Article: The pathophysiology and medical management of diverticulosis and diverticular disease of the colon. Aliment Parmacol Ther 42:664, 2015.

and if the patient is affected by complicated diverticulitis. Nearly 75% of patients hospitalized for acute diverticulitis will respond to nonoperative treatment with a suitable antimicrobial regimen. The current recommended antimicrobial coverage is a third-generation cephalosporin or ciprofloxacin and metronidazole targeting aerobic gram-negative rods and anaerobic bacteria. Unfortunately, these agents do not cover enterococci, and the addition of ampicillin to this regimen for nonresponders is recommended. Alternatively, single-agent therapy with a third-generation penicillin such as IV piperacillin or oral penicillin/clavulanic acid may be effective. The usual course of antibiotics is 7–10 days, although this length of time is being investigated. Patients should remain on a limited diet until their pain resolves.

Once the acute attack has resolved, the mainstay medical management of diverticular disease to prevent symptoms has evolved. Established risk factors for symptomatic recurrence include younger age, the formation of a diverticular abscess, and more frequent attacks (>2 per year). Newer directions are targeted at colonic inflammation and dysbiosis. Diverticular disease is now considered a functional bowel disorder associated with low-grade inflammation. However, the use of anti-inflammatory medications (mesalazine) in randomized clinical trials has not demonstrated any effect on recurrence rates over placebo alone. Some authors have suggested that the use of anti-inflammatory medications is most helpful in patients with diverticular disease who also have segmental colitis (Segmental Colitis-Associated Diverticular Disease [SCADD]). Treatment strategies targeting dysbiosis in diverticular disease have also been evaluated using polymerase chain reaction (PCR) on stool specimens. Stool samples from consumers of a high-fiber diet have different bacterial content than stool samples from consumers of a low-fiber, high-fat diet. Probiotics are increasingly used by gastroenterologists for multiple bowel disorders and may prevent recurrence of diverticulitis. Specifically, probiotics containing *Lactobacillus acidophilus* and *Bifidobacterium* strains may be beneficial, however, a recent systematic review was unable to show any benefit to the use of probiotics. Rifaximin (a poorly absorbed broad-spectrum antibiotic), when compared to fiber alone for the treatment of SUDD, is associated with 30% less frequent recurrent symptoms from uncomplicated diverticular disease.

SURGICAL MANAGEMENT

Preoperative risk factors influencing postoperative mortality rates include higher American Society of Anesthesiologists (ASA) physical status class **(Table 321-3)** and preexisting organ failure. In patients who are low risk (ASA P1 and P2), surgical therapy can be offered to those who do not rapidly improve on medical therapy. For uncomplicated diverticular disease, medical therapy can be continued beyond two attacks without an increased risk of perforation requiring a colostomy. However, patients on immunosuppressive therapy, in chronic renal failure, or with a collagen-vascular disease have a fivefold greater risk of perforation during recurrent attacks. Surgical therapy is indicated in all low-surgical-risk patients with complicated diverticular disease.

The goals of surgical management of diverticular disease include controlling sepsis, eliminating complications such as fistula or obstruction, removing the diseased colonic segment, and restoring intestinal continuity. These goals must be obtained while minimizing morbidity rate, length of hospitalization, and cost in addition to maximizing survival and quality of life. **Table 321-4** lists the operations most commonly indicated based on the Hinchey classification and the predicted postoperative outcomes. The current options for uncomplicated diverticular disease include an open or a laparoscopic resection of the diseased area with reanastomosis to the rectosigmoid. Preservation of portions of the sigmoid colon may lead to early recurrence of the disease. The benefits of laparoscopic resection over open surgical techniques include early discharge (by at least 1 day), less narcotic use, less postoperative complications, and an earlier return to work.

The options for the surgical management of complicated diverticular disease **(Fig. 321-3)** include the following open or laparoscopic procedures: (1) proximal diversion of the fecal stream with an ileostomy or colostomy and sutured omental patch with drainage, (2) resection with colostomy and mucous fistula or closure of distal bowel with formation of a Hartmann's pouch (Hartmann's procedure), (3) resection with anastomosis (coloproctostomy), or (4) resection with anastomosis and diversion (coloproctostomy with loop ileostomy or colostomy). (5) Laparoscopic technique of washout and drainage without diversion has been described for Hinchey III patients; however, a threefold increased risk of recurrent peritonitis requiring reoperation with washout alone has been reported.

Patients with Hinchey stage Ia are managed with antibiotic therapy only followed by resection with anastomosis at 6 weeks. Patients with Hinchey stages Ib and II disease are managed with percutaneous drainage followed by resection with anastomosis about 6 weeks later. Current guidelines put forth by the American Society of Colon and Rectal Surgeons suggest, in addition to antibiotic therapy, CT-guided percutaneous drainage of diverticular abscesses that are >3 cm and have a well-defined wall. Abscesses that are <5 cm may resolve with antibiotic therapy alone. Contraindications to percutaneous drainage are no percutaneous access route, pneumoperitoneum, and fecal peritonitis. Drainage of a

TABLE 321-3 American Society of Anesthesiologists Physical Status Classification System

P1	A normal healthy patient
P2	A patient with mild systemic disease
P3	A patient with severe systemic disease
P4	A patient with severe systemic disease that is a constant threat to life
P5	A moribund patient who is not expected to survive without the operation
P6	A declared brain-dead patient whose organs are being removed for donor purposes

TABLE 321-4 Outcome Following Surgical Therapy for Complicated Diverticular Disease Based Upon Modified Hinchey Staging

HINCHEY STAGE	OPERATIVE PROCEDURE	ANASTOMOTIC LEAK RATE, %	OVERALL MORBIDITY RATE, %
Ia (pericolic phlegmon)	Laparoscopic or open colon resection	43	15
Ib (pericolic abscess)	Percutaneous drainage followed by laparoscopic or open colon resection	3	15
II	Percutaneous drainage followed by laparoscopic or open colon resection +/− proximal diversion with an ostomy	3	15
III	Laparoscopic washout and drainage *or* Laparoscopic or open resection with proximal diversion (ostomy) *or* Hartmann's procedure	3	30% risk of peritonitis requiring reoperation if no resection is performed. Overall morbidity 50% Overall mortality 15%
IV	Hartmann's procedure *or* Washout with proximal diversion	—	Overall morbidity 50% Overall mortality 15%

diverticular abscess is associated with a 20–25% failure rate. Urgent operative intervention is undertaken if percutaneous drainage fails and patients develop generalized peritonitis, and most will need to be managed with a Hartmann's procedure (resection of the sigmoid colon with end colostomy and rectal stump). In selected cases, nonoperative therapy may be considered. In one nonrandomized study, nonoperative management of isolated paracolic abscesses (Hinchey stage I) was associated with only a 20% recurrence rate at 2 years. More than 80% of patients with distant abscesses (Hinchey stage II) required surgical resection for recurrent symptoms.

The management of Hinchey stage III disease is under debate. In this population of patients, no fecal peritonitis is present and it is presumed that the perforation has sealed. Historically, Hinchey stage III has been managed with a Hartmann's procedure or with primary anastomosis and proximal diversion. Several studies have examined short- and long-term outcomes for laparoscopic peritoneal lavage to remove the peritoneal contamination and place drainage catheters should a communication to the bowel still exist. However, this procedure has been associated with an increased risk of requiring reoperation for ongoing peritonitis. Overall, ostomy rates are lower with the use of laparoscopic peritoneal lavage. No anastomosis of any type should be attempted in Hinchey stage IV disease, or the presence of fecal peritonitis. A limited approach to these patients is associated with a decreased mortality rate.

Recurrent Symptoms Recurrent abdominal symptoms following surgical resection for diverticular disease occur in 10% of patients. Recurrent diverticular disease develops in patients following inadequate surgical resection. A retained segment of diseased rectosigmoid colon is associated with twice the incidence of recurrence. The presence of irritable bowel syndrome may also cause recurrence of initial symptoms. Patients undergoing surgical resection for presumed diverticulitis and symptoms of chronic abdominal cramping and irregular loose bowel movements consistent with irritable bowel syndrome have poorer functional outcomes.

COMMON DISEASES OF THE ANORECTUM

RECTAL PROLAPSE (PROCIDENTIA)

Incidence and Epidemiology Rectal prolapse is six times more common in women than in men. The incidence of rectal prolapse peaks in women >60 years. Women with rectal prolapse have a higher incidence of associated pelvic floor disorders including urinary incontinence, rectocele, cystocele, and enterocele. About 20% of children with rectal prolapse will have cystic fibrosis. All children presenting with prolapse should undergo a sweat chloride test. Less common associations include Ehlers-Danlos syndrome, solitary rectal ulcer syndrome, congenital hypothyroidism, Hirschsprung's disease, dementia, mental retardation, and schizophrenia.

Anatomy and Pathophysiology Rectal prolapse (procidentia) is a circumferential, full-thickness protrusion of the rectal wall through the anal orifice. It is often associated with a redundant sigmoid colon, pelvic laxity, and a deep rectovaginal septum (pouch of Douglas). Initially, rectal prolapse was felt to be the result of early internal rectal intussusception, which occurs in the upper to mid rectum. This was considered to be the first step in an inevitable progression to full-thickness external prolapse. However, only 1 of 38 patients with internal prolapse followed for >5 years developed full-thickness prolapse. Others have suggested that full-thickness prolapse is the result of damage to the nerve supply to the pelvic floor muscles or pudendal nerves from repeated stretching with straining to defecate. Damage to the pudendal nerves would weaken the pelvic floor muscles, including

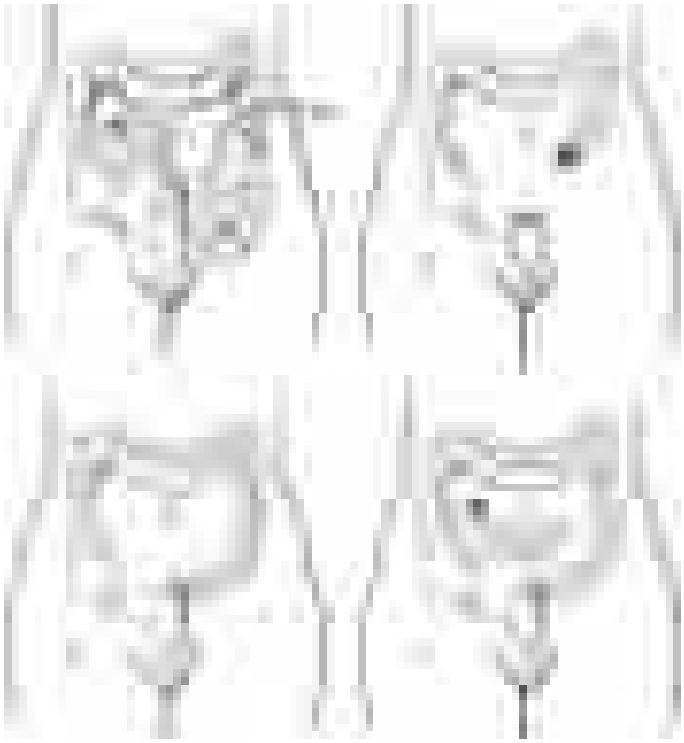


FIGURE 321-3 Methods of surgical management of complicated diverticular disease. *1.* Drainage, omental pedicle graft, and proximal diversion. ***2.*** Hartmann's procedure. ***3.*** Sigmoid resection with coloproctostomy. ***4.*** Sigmoid resection with coloproctostomy and proximal diversion.

the external anal sphincter muscles. Bilateral pudendal nerve injury is more significantly associated with prolapse and incontinence than unilateral injury.

Presentation and Evaluation In external prolapse, the majority of patient complaints include anal mass, bleeding per rectum, and poor perianal hygiene. Prolapse of the rectum usually occurs following defecation and will spontaneously reduce or require the patient to manually reduce the prolapse. Constipation occurs in ~30–67% of patients with rectal prolapse. Differing degrees of fecal incontinence occur in 50–70% of patients. Patients with internal rectal prolapse will present with symptoms of both constipation and incontinence. Other associated findings include outlet obstruction (anismus) in 30%, colonic inertia in 10%, and solitary rectal ulcer syndrome in 12%.

Office evaluation is best performed after the patient has been given an enema, which enables the prolapse to protrude. An important distinction should be made between full-thickness rectal prolapse and isolated mucosal prolapse associated with hemorrhoidal disease (Fig. 321-4). Mucosal prolapse is known for radial grooves rather than circumferential folds around the anus and is due to increased laxity of the connective tissue between the submucosa and underlying muscle of the anal canal. The evaluation of prolapse should also include cystoproctography and colonoscopy. These examinations evaluate for associated pelvic floor disorders and rule out a malignancy or a polyp as the lead point for prolapse. If rectal prolapse is associated with chronic constipation, the patient should undergo a defecating proctogram and a sitzmark study. This will evaluate for the presence of anismus or colonic inertia. Anismus is the result of attempting to defecate against a closed pelvic floor and is also known as *nonrelaxing puborectalis*. This can be seen when straightening of the rectum fails to occur on fluoroscopy while the patient is attempting to defecate. In colonic inertia, a sitzmark study will demonstrate retention of >20% of markers on abdominal x-ray 5 days after swallowing. For patients with fecal incontinence, endoanal ultrasound and manometric evaluation, including pudendal nerve testing of their anal sphincter muscles, may be performed before surgery for prolapse (see "Fecal Incontinence," below).

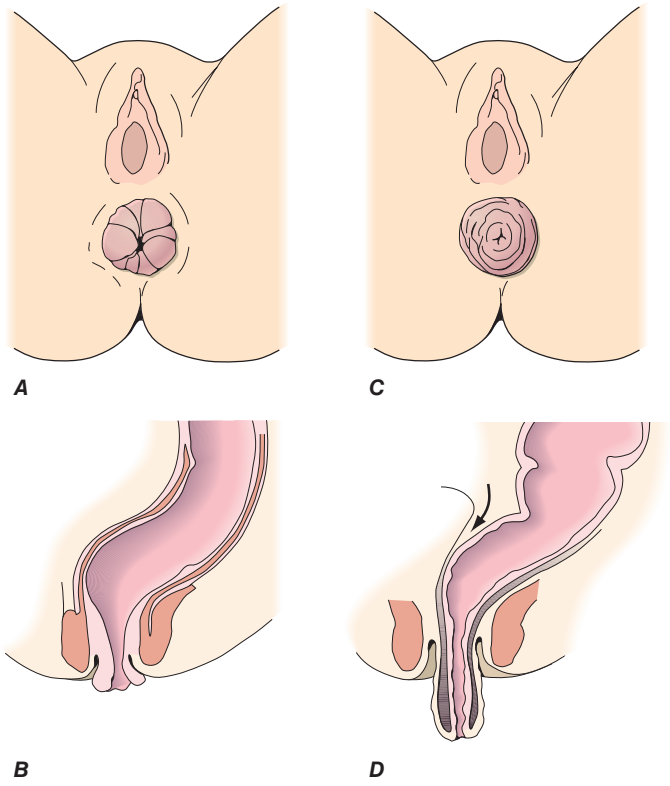


FIGURE 321-4 Degrees of rectal prolapse. Mucosal prolapse only (***A, B,*** sagittal view). Full-thickness prolapse associated with redundant rectosigmoid and deep pouch of Douglas (***C, D,*** sagittal view).

TREATMENT

Rectal Prolapse

The medical approach to the management of rectal prolapse is limited and includes stool-bulking agents or fiber supplementation to ease the process of evacuation. Surgical correction of rectal prolapse is the mainstay of therapy. Two approaches are commonly considered, transabdominal and transperineal. Transabdominal approaches have been associated with lower recurrence rates, but some patients with significant comorbidities are better served by a transperineal approach.

Common transperineal approaches include a transanal proctectomy (Altmeier procedure), mucosal proctectomy (Delorme procedure), or placement of a Tirsch wire encircling the anus. The goal of the transperineal approach is to remove the redundant rectosigmoid colon. Common transabdominal approaches include presacral suture or mesh rectopexy (Ripstein) with (Frykman-Goldberg) or without resection of the redundant sigmoid. Colon resection, in general, is reserved for patients with constipation and outlet obstruction. Ventral rectopexy is an effective method of abdominal repair of full-thickness prolapse that does not require sigmoid resection (see description below). This repair may have improved functional results over other abdominal repairs. Transabdominal procedures can be performed effectively with laparoscopic and, more recently, robotic techniques without increased incidence of recurrence. The goal of the transabdominal approach is to restore normal anatomy by removing redundant bowel and reattaching the supportive tissue of the rectum to the presacral fascia. The final alternative is abdominal proctectomy with end-sigmoid colostomy. If total colonic inertia is present, as defined by a history of constipation and a positive sitzmark study, a subtotal colectomy with an ileosigmoid or rectal anastomosis may be required at the time of rectopexy.

Previously, the presence of internal rectal prolapse identified on imaging studies has been considered a nonsurgical disorder, and biofeedback was recommended. However, only one-third of patients will have successful resolution of symptoms from biofeedback. Two surgical procedures more effective than biofeedback are the Stapled Transanal Rectal Resection (STARR) and the Laparoscopic Ventral Rectopexy (LVR). The STARR procedure (Fig. 321-5) is performed through the anus in patients with internal prolapse. A circular stapling device is inserted through the anus; the internal prolapse is identified and ligated with the stapling device. LVR (Fig. 321-6) is performed through an abdominal approach. An opening in the

FIGURE 321-5 Stapled transanal rectal resection. Schematic of placement of the circular stapling device.

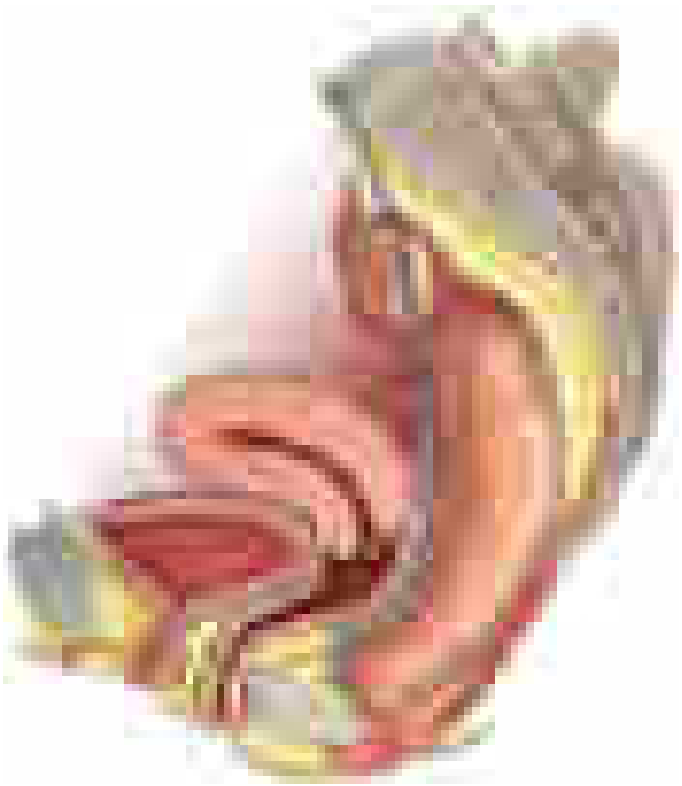


FIGURE 321-6 Laparoscopic ventral rectopexy (LVR). To reduce the internal prolapse and close any rectovaginal septal defect, the pouch of Douglas is opened and mesh is secured to the anterolateral rectum, vaginal fornix, and sacrum. *(From A D'Hoore et al: Br J Surg 91:1500, 2004.)*

peritoneum is created on the left side of the rectosigmoid junction, and this opening continues down anterior on the rectum into the pouch of Douglas. No rectal mobilization is performed, thus avoiding any autonomic nerve injury. Mesh is secured to the anterior and lateral portion of the rectum, the vaginal fornix, and the sacral promontory, allowing for closure of the rectovaginal septum and correction of the internal prolapse. In both procedures, recurrence at 1 year was low (<10%) and symptoms improved in more than three-fourths of patients.

FECAL INCONTINENCE

Incidence and Epidemiology Fecal incontinence is the involuntary passage of fecal material for at least 1 month in an individual with a developmental age of at least 4 years. The prevalence of fecal incontinence in the United States is 0.5–11%. The majority of patients are women and aged >65. A higher incidence of incontinence is seen among parous women. One-half of patients with fecal incontinence also suffer from urinary incontinence. The majority of incontinence is a result of obstetric injury to the pelvic floor, either while carrying a fetus or during the delivery. An anatomic sphincter defect may occur in up to 32% of women following childbirth regardless of visible damage to the perineum. Risk factors at the time of delivery include prolonged labor, the use of forceps, and the need for an episiotomy. Symptoms of incontinence can present after two or more decades following obstetric injury. Medical conditions known to contribute to the development of fecal incontinence are listed in Table 321-5.

Anatomy and Pathophysiology The anal sphincter complex is made up of the internal and external anal sphincter. The internal sphincter is smooth muscle and a continuation of the circular fibers of the rectal wall. It is innervated by the intestinal myenteric plexus and is therefore not under voluntary control. The external anal sphincter is formed in continuation with the levator ani muscles and is under voluntary control. The pudendal nerve supplies motor innervation to the external anal sphincter. Obstetric injury may result in tearing of the muscle fibers anteriorly at the time of the delivery. This results in an obvious anterior defect on endoanal ultrasound. Injury may also be the result of stretching of the pudendal nerves during pregnancy or delivery of the fetus through the birth canal.

TABLE 321-5 Medical Conditions That Contribute to Symptoms of Fecal Incontinence

Neurologic Disorders
• Dementia
• Brain tumor
• Stroke
• Multiple sclerosis
• Tabes dorsalis
• Cauda equina lesions
Skeletal Muscle Disorders
• Myasthenia gravis
• Myopathies, muscular dystrophy
Miscellaneous
• Hypothyroidism
• Irritable bowel syndrome
• Diabetes
• Severe diarrhea
• Scleroderma

Presentation and Evaluation Patients may suffer with varying degrees of fecal incontinence. Minor incontinence includes incontinence to flatus and occasional seepage of liquid stool. Major incontinence is frequent inability to control solid waste. As a result of fecal incontinence, patients suffer from poor perianal hygiene. Beyond the immediate problems associated with fecal incontinence, these patients are often withdrawn and suffer from depression. For this reason, quality-of-life measures are an important component in the evaluation of patients with fecal incontinence.

The evaluation of fecal incontinence should include a thorough history and physical examination including digital rectal examination (DRE). Weak sphincter tone on DRE and loss of the "anal wink" reflex (S1-level control) may indicate a neurogenic dysfunction. Perianal scars may represent surgical injury. Other studies helpful in the diagnosis of fecal incontinence include anal manometry, pudendal nerve terminal motor latency (PNTML), and endoanal ultrasound. Centers that care for patients with fecal incontinence will have an anorectal physiology laboratory that uses standardized methods of evaluating anorectal physiology. Anorectal manometry (ARM) measures resting and squeeze pressures within the anal canal using an intraluminal water-perfused catheter. Current methods of ARM include use of a three-dimensional, high-resolution system with a 12-catheter perfusion system, which allows physiologic delineation of anatomic abnormalities. Pudendal nerve studies evaluate the function of the nerves innervating the anal canal using a finger electrode placed in the anal canal. Stretch injuries to these nerves will result in a delayed response of the sphincter muscle to a stimulus, indicating a prolonged latency. Finally, endoanal ultrasound will evaluate the extent of the injury to the sphincter muscles before surgical repair. Unfortunately, all of these investigations are user-dependent, and very few studies demonstrate that these studies predict outcome following an intervention. Magnetic resonance imaging (MRI) has been used, but its routine use for imaging in fecal incontinence is not well established.

Rarely does a pelvic floor disorder exist alone. The majority of patients with fecal incontinence will have some degree of urinary incontinence. Similarly, fecal incontinence is a part of the spectrum of pelvic organ prolapse. For this reason, patients may present with symptoms of obstructed defecation as well as fecal incontinence. Careful evaluation including dynamic MRI or cinedefecography should be performed to search for other associated defects. Surgical repair of incontinence without attention to other associated defects may decrease the success of the repair.

TREATMENT

Fecal Incontinence

Medical management of fecal incontinence includes strategies to bulk up the stool, which help in increasing fecal sensation. These include fiber supplementation, loperamide, diphenoxylate, and bile acid binders. These agents harden the stool and delay frequency of bowel movements and are helpful in patients with minimal to mild symptoms. Furthermore, patients can be offered a form of physical therapy called biofeedback. This therapy helps strengthen the external sphincter muscle while training the patient to relax with defecation to avoid unnecessary straining and further injury to the sphincter muscles. Biofeedback has had variable success and is dependent on the motivation of the patient. At a minimum, biofeedback is risk-free. Most patients will have some improvement. For this reason, it should be incorporated into the initial recommendation to all patients with fecal incontinence.

Historically, the "gold standard" for the treatment of fecal incontinence with an isolated sphincter defect has been the overlapping sphincteroplasty. The external anal sphincter muscle and scar tissue as well as any identifiable internal sphincter muscle are dissected free from the surrounding adipose and connective tissue and then an overlapping repair is performed in an attempt to rebuild the muscular ring and restore its function. However, long-term results following overlapping sphincteroplasty have been poor with a 50% failure rate over 5 years.

Alternative therapies such as Sacral Nerve Stimulation (SNS), collagen-enhancing injectables, and magnetic "Fenix" ring are other options. SNS is an adaptation of a procedure developed for the management of urinary incontinence. SNS is ideally suited for patients with intact but weak anal sphincters. A temporary nerve stimulator is placed on the third sacral nerve. If there is at least a 50% improvement in symptoms, a permanent nerve stimulator is placed under the skin. Long-term results for sacral stimulation have been promising, with nearly 80% of patients having a reduction in incontinence episodes by at least 50%. This reduction has been sustainable in studies out to 5 years. Collagen-enhancing injectables have been around for several years. More than 50% of incontinent patients treated with nonanimal stabilized hyaluronic acid (NASHA/DX) achieved a 50% reduction in incontinence episodes, and these results were sustainable up to 2 years. Currently, this injectable is not universally available. The Fenix is a magnetic ring that is implanted around the anal sphincter muscles. Its long-term outcomes are still being studied and it is currently only available for compassionate use.

Finally, the use of stem cells to increase the bulk of the sphincter muscles is currently being tested. Stem cells can be harvested from the patient's own muscle, grown, and then implanted into their sphincter complex. Concern for cost and the need for an additional procedure dampen enthusiasm. Trial results are awaited.

HEMORRHOIDAL DISEASE

Incidence and Epidemiology Symptomatic hemorrhoids affect >1 million individuals in the Western world per year. The prevalence of hemorrhoidal disease is not selective for age or sex. However, age is known to be a risk factor. The prevalence of hemorrhoidal disease is less in underdeveloped countries. The typical low-fiber, high-fat Western diet is associated with constipation and straining and the development of symptomatic hemorrhoids.

Anatomy and Pathophysiology Hemorrhoidal cushions are a normal part of the anal canal. The vascular structures contained within this tissue aid in continence by preventing damage to the sphincter muscle. Three main hemorrhoidal complexes traverse the anal canal—the left lateral, the right anterior, and the right posterior. Engorgement and straining lead to prolapse of this tissue into the anal canal. Over time, the anatomic support system of the hemorrhoidal complex weakens, exposing this tissue to the outside of the anal canal where it is susceptible to injury. Hemorrhoids are commonly classified as external or internal. External hemorrhoids originate below the dentate line and are covered with squamous epithelium and are associated with an internal component. External hemorrhoids are painful when thrombosed. Internal hemorrhoids originate above the dentate line and are covered with mucosa and transitional zone epithelium and represent the majority of hemorrhoids. The standard classification of hemorrhoidal disease is based on the progression of the disease from their normal internal location to the prolapsing external position **(Table 321-6)**.

TABLE 321-6 The Staging and Treatment of Hemorrhoids

STAGE	DESCRIPTION OF CLASSIFICATION	TREATMENT
I	Enlargement with bleeding	Fiber supplementation Short course of cortisone suppository Sclerotherapy Infared coagulation
II	Protrusion with spontaneous reduction	Fiber supplementation Short course of cortisone suppository Sclerotherapy Infared coagulation
III	Protrusion requiring manual reduction	Fiber supplementation Short course of cortisone suppository Rubber band ligation Operative hemorrhoidectomy
IV	Irreducible protrusion	Fiber supplementation Cortisone suppository Operative hemorrhoidectomy

Presentation and Evaluation Patients commonly present to a physician for two reasons: bleeding and protrusion. Pain is less common than with fissures and, if present, is described as a dull ache from engorgement of the hemorrhoidal tissue. Severe pain may indicate a thrombosed hemorrhoid. Hemorrhoidal bleeding is described as painless bright red blood seen either in the toilet or upon wiping. Occasional patients can present with significant bleeding, which may be a cause of anemia; however, the presence of a colonic neoplasm must be ruled out in anemic patients. Patients who present with a protruding mass complain about inability to maintain perianal hygiene and are often concerned about the presence of a malignancy.

The diagnosis of hemorrhoidal disease is made on physical examination. Inspection of the perianal region for evidence of thrombosis or excoriation is performed, followed by a careful digital examination. Anoscopy is performed paying particular attention to the known position of hemorrhoidal disease. The patient is asked to strain. If this is difficult for the patient, the maneuver can be performed while sitting on a toilet. The physician is notified when the tissue prolapses. It is important to differentiate the circumferential appearance of a full-thickness rectal prolapse from the radial nature of prolapsing hemorrhoids (see "Rectal Prolapse," above). The stage and location of the hemorrhoidal complexes are defined.

TREATMENT

Hemorrhoidal Disease

The treatment for bleeding hemorrhoids is based on the stage of the disease (Table 321-6). In all patients with bleeding, the possibility of other causes must be considered. In young patients without a family history of colorectal cancer, the hemorrhoidal disease may be treated first and a colonoscopic examination performed if the bleeding continues. Older patients who have not had colorectal cancer screening should undergo colonoscopy or flexible sigmoidoscopy.

With rare exceptions, the acutely thrombosed hemorrhoid can be excised within the first 72 h by performing an elliptical excision. Sitz baths, fiber, and stool softeners are prescribed. Additional therapy for bleeding hemorrhoids includes the office procedures of rubber band ligation, infrared coagulation, and sclerotherapy. Sensation begins at the dentate line; therefore, all procedures can be performed without discomfort either endoscopically or in the office. Bands are placed around the engorged tissue, causing ischemia and fibrosis. This aids in fixing the tissue proximally in the anal canal. Patients may complain of a dull ache for 24 h following band application. During sclerotherapy, 1–2 mL of a sclerosant (usually sodium tetradecyl sulfate) is injected using a 25-gauge needle into the submucosa of the hemorrhoidal complex. Care must be taken not to inject the anal canal circumferentially, or stenosis may occur.

For surgical management of hemorrhoidal disease, excisional hemorrhoidectomy, transhemorrhoidal dearterialization (THD), or stapled hemorrhoidectomy ("the procedure for prolapse or hemorrhoids" [PPH]) is the procedure of choice. All surgical methods of management are equally effective in the treatment of symptomatic third- and fourth-degree hemorrhoids. However, because the sutured hemorrhoidectomy involves the removal of redundant tissue down to the anal verge, unpleasant anal skin tags are removed as well. The stapled hemorrhoidectomy is associated with less discomfort; however, this procedure does not remove anal skin tags and an increased number of complications are associated with use of the stapling device. THD uses ultrasound guidance to ligate the blood supply to the anal tissue, hence reducing hemorrhoidal engorgement. No procedures on hemorrhoids should be done in patients who are immunocompromised or who have active proctitis. Furthermore, emergent hemorrhoidectomy for bleeding hemorrhoids is associated with a higher complication rate.

Acute complications associated with the treatment of hemorrhoids include pain, infection, recurrent bleeding, and urinary retention. Care should be taken to place bands properly and to avoid overhydration in patients undergoing operative hemorrhoidectomy. Late complications include fecal incontinence as a result of injury to the sphincter during the dissection. Anal stenosis may develop from overzealous excision, with loss of mucosal skin bridges for reepithelialization. Finally, an *ectropion* (prolapse of rectal mucosa from the anal canal) may develop. Patients with an ectropion complain of a "wet" anus as a result of inability to prevent soiling once the rectal mucosa is exposed below the dentate line.

ANORECTAL ABSCESS

Incidence and Epidemiology The development of a perianal abscess is more common in men than women by a ratio of 3:1. The peak incidence is in the third to fifth decade of life. Perianal pain associated with the presence of an abscess accounts for 15% of office visits to a colorectal surgeon. The disease is more prevalent in immunocompromised patients such as those with diabetes, hematologic disorders, or inflammatory bowel disease (IBD) and persons who are HIV positive. These disorders should be considered in patients with recurrent perianal infections.

Anatomy and Pathophysiology An anorectal abscess is an abnormal fluid-containing cavity in the anorectal region. Anorectal abscess results from an infection involving the glands surrounding the anal canal. Normally, these glands release mucus into the anal canal, which aids in defecation. When stool accidentally enters the anal glands, the glands become infected and an abscess develops. Anorectal abscesses are perianal in 40–50% of patients, ischiorectal in 20–25%, intersphincteric in 2–5%, and supralevator in 2.5% **(Fig. 321-7)**.

Presentation and Evaluation Perianal pain and fever are the hallmarks of an abscess. Patients may have difficulty voiding and have blood in the stool. A prostatic abscess may present with similar complaints, including dysuria. Patients with a prostatic abscess will often have a history of recurrent sexually transmitted diseases. On physical examination, a large fluctuant area is usually readily visible. Routine laboratory evaluation shows an elevated white blood cell count. Diagnostic procedures are rarely necessary unless evaluating a recurrent abscess. A CT scan or MRI has an accuracy of 80% in determining incomplete drainage. If there is a concern about the presence of IBD, a rigid or flexible sigmoidoscopic examination may be done at the time of drainage to evaluate for inflammation within the rectosigmoid region. A more complete evaluation for Crohn's disease would include a full colonoscopy and small-bowel series.

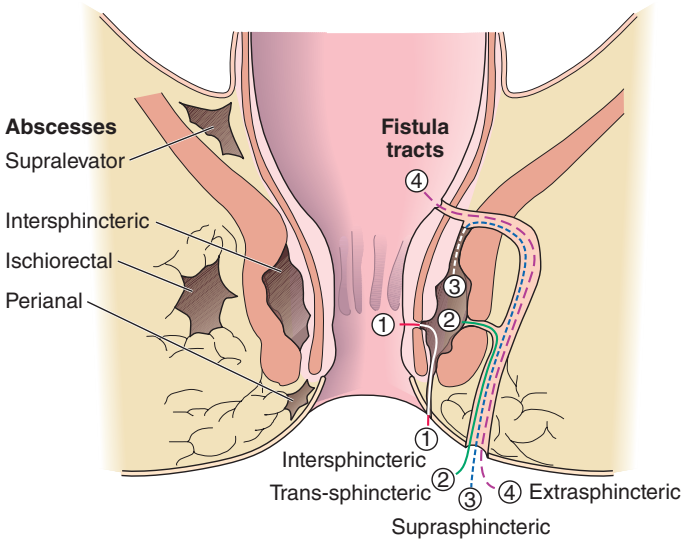


FIGURE 321-7 Common locations of anorectal abscess (*left*) and fistula in ano (*right*).

TREATMENT

Anorectal Abscess

As with all abscesses, the "gold standard" is drainage. Office drainage of an uncomplicated anorectal abscess may suffice. A small incision close to the anal verge is made, and a Mallenkot drain is advanced into the abscess cavity. For patients who have a complicated abscess or who are diabetic or immunocompromised, drainage should be performed in an operating room under anesthesia. These patients are at greater risk for developing necrotizing fasciitis. The role of antibiotics in the management of anorectal abscesses is limited. Antibiotics are only warranted in patients who are immunocompromised or have prosthetic heart valves, artificial joints, diabetes, or IBD.

FISTULA IN ANO

Incidence and Epidemiology The incidence and prevalence of fistulating perianal disease parallel the incidence of anorectal abscess, estimating to be 1 in 10,000 individuals. Some 30–40% of abscesses will give rise to fistula in ano. Although the majority of the fistulas are cryptoglandular in origin, 10% are associated with IBD, tuberculosis, malignancy, and radiation.

Anatomy and Pathophysiology A fistula in ano is defined as a communication of an abscess cavity with an identifiable internal opening within the anal canal. This identifiable opening is most commonly located at the dentate line where the anal glands enter the anal canal. Patients experiencing continuous drainage following the treatment of a perianal abscess likely have a fistula in ano. These fistulas are classified by their relationship to the anal sphincter muscles, with 70% being intersphincteric, 23% transsphincteric, 5% suprasphincteric, and 2% extrasphincteric (Fig. 321-7).

Presentation and Evaluation A patient with a fistula in ano will complain of constant drainage from the perianal region associated with a firm mass. The drainage may increase with defecation. Perianal hygiene is difficult to maintain. Examination under anesthesia is the

best way to evaluate a fistula. At the time of the examination, anoscopy is performed to look for an internal opening. Diluted hydrogen peroxide will aid in identifying such an opening. In lieu of anesthesia, MRI with an endoanal coil will also identify tracts in 80% of the cases. After drainage of an abscess with insertion of a Mallenkot catheter, a fistulagram through the catheter can be obtained in search of an occult fistula tract. Goodsall's rule states that a posterior external fistula will enter the anal canal in the posterior midline, whereas an anterior fistula will enter at the nearest crypt. A fistula exiting >3 cm from the anal verge may have a complicated upward extension and may not obey Goodsall's rule.

TREATMENT

Fistula in Ano

A newly diagnosed draining fistula is best managed with placement of a seton, a vessel loop or silk tie placed through the fistula tract, which maintains the tract open and quiets down the surrounding inflammation that occurs from repeated blockage of the tract. Once the inflammation is less, the exact relationship of the fistula tract to the anal sphincters can be ascertained. A simple fistulotomy can be performed for intersphincteric and low (less than one-third of the muscle) transsphincteric fistulas without compromising continence. For a higher transsphincteric fistula, an anorectal advancement flap in combination with a drainage catheter or fibrin glue may be used. Very long (>2 cm) and narrow tracts respond better to fibrin glue than shorter tracts. Simple ligation of the internal fistula tract (LIFT procedure) has also been used in the management of simple fistula with good success.

Patients should be maintained on stool-bulking agents, nonnarcotic pain medication, and sitz baths following surgery for a fistula. Early complications from these procedures include urinary retention and bleeding. Later complications are rare (<10%) and include temporary and permanent incontinence. Recurrence is 0–18% following fistulotomy and 20–30% following anorectal advancement flap and the LIFT procedure. The use of stem cell implants at the time of repair for recalcitrant fistulizing disease of the anus is being studied.

ANAL FISSURE

Incidence and Epidemiology Anal fissures occur at all ages but are more common in the third through the fifth decades. A fissure is the most common cause of rectal bleeding in infancy. The prevalence is equal in males and females. It is associated with constipation, diarrhea, infectious etiologies, perianal trauma, and Crohn's disease.

Anatomy and Pathophysiology Trauma to the anal canal occurs following defecation. This injury occurs in the anterior or, more commonly, the posterior anal canal. Irritation caused by the trauma to the anal canal results in an increased resting pressure of the internal sphincter. The blood supply to the sphincter and anal mucosa enters laterally. Therefore, increased anal sphincter tone results in a relative ischemia in the region of the fissure and leads to poor healing of the anal injury. A fissure that is not in the posterior or anterior position should raise suspicion for other causes, including tuberculosis, syphilis, Crohn's disease, and malignancy.

Presentation and Evaluation A fissure can be easily diagnosed on history alone. The classic complaint is pain, which is strongly associated with defecation and is relentless. The bright red bleeding that can be associated with a fissure is less extensive than that associated with hemorrhoids. On examination, most fissures are located in either the posterior or anterior position. A lateral fissure is worrisome because it may have a less benign nature, and systemic disorders should be ruled out. A chronic fissure is indicated by the presence of a hypertrophied anal papilla at the proximal end of the fissure and a sentinel pile or skin tag at the distal end. Often the circular fibers of the hypertrophied internal sphincter are visible within the base of the fissure. If anal manometry is performed, elevation in anal resting pressure and a sawtooth deformity with paradoxical contractions of the sphincter muscles are pathognomonic.

TREATMENT

Anal Fissure

The management of the acute fissure is conservative. Stool softeners for those with constipation, increased dietary fiber, topical anesthetics, glucocorticoids, and sitz baths are prescribed and will heal 60–90% of fissures. Chronic fissures are those present for >6 weeks. These can be treated with modalities aimed at decreasing the anal canal resting pressure including nifedipine ointment applied three times a day and botulinum toxin type A, up to 20 units, injected into the internal sphincter on each side of the fissure. Surgical management includes anal dilatation and lateral internal sphincterotomy. Usually, one-third of the internal sphincter muscle is divided; it is easily identified because it is hypertrophied. Recurrence rates from medical therapy are higher, but this is offset by a risk of incontinence following sphincterotomy. Lateral internal sphincterotomy may lead to incontinence more commonly in women.

Acknowledgment

We would like to thank Cory Sandore for providing some illustrations for this chapter. Gregory Bulkley, MD, contributed to this chapter in an earlier edition and some of that material has been retained here.

FURTHER READING

Bharucha AE , Rao SSC, Shin AS: Surgical interventions and the use of device-aided therapy for the treatment of fecal incontinence and defecatory disorders. Clin Gastroenterol Hepatol 15:1844, 2017.

Cotter TG et al: Approach to the patient with hematochezia. Mayo Clin Proc 92:797, 2017.

Guttenplan M: The evaluation and office management of hemorrhoids for the gastroenterologist. Curr Gastroenterol Rep 19:30, 2017.

Prichard D, Bharucha AE: Management of pelvic floor disorders: Biofeedback and more. Curr Treat Options Gastroenterol 12:456, 2014.

Salfity HV et al: Minimally invasive incision and drainage technique in the treatment of simple subcutaneous abscess in adults. Am Surg 83:699, 2017.

Sugrue J et al: Sphincter-sparing anal fistula repair: Are we getting better? Dis Colon Rectum 60:1071, 2017.

Tursi A: Dietary pattern and colonic diverticulosis. Curr Opin Clin Nutr Metab Care 20:409, 2017.

322 Mesenteric Vascular Insufficiency

Satinderjit Locham, Mahmoud Malas

INTESTINAL ISCHEMIA

INCIDENCE AND EPIDEMIOLOGY

Intestinal ischemia occurs when splanchnic perfusion fails to meet the metabolic demands of the intestines, resulting in ischemic tissue injury. Mesenteric ischemia affects 2–3 people per 100,000, with an increasing incidence in the aging population. Delay in diagnosis and management results in a high mortality, and prompt interventions may be lifesaving. Intestinal ischemia is further classified based on etiology, which dictates management: (1) arterioocclusive mesenteric ischemia, (2) nonocclusive mesenteric ischemia, and (3) mesenteric venous thrombosis.

Risk factors for arterioocclusive mesenteric ischemia are generally acute in onset that include atrial fibrillation, recent myocardial infarction, valvular heart disease, and recent cardiac or vascular catheterization, all of which result in embolic clots reaching the mesenteric circulation. Nonocclusive mesenteric ischemia, also known as "intestinal angina," is generally more insidious and often seen in the aging population affected by atherosclerotic disease. Patients with chronic atherosclerotic disease could also suffer an acute insult from emboli leading to complete occlusion. Nonocclusive mesenteric ischemia is also seen in patients receiving high-dose vasopressor infusions, patients presenting with cardiogenic or septic shock, and cocaine overdose. It is the most prevalent gastrointestinal disease complicating cardiovascular surgery. The incidence of ischemic colitis following elective aortic repair is 5–9%, and the incidence triples in patients following emergent repair. Mesenteric venous thrombosis is less common and is associated with the presence of a hypercoagulable state including protein C or S deficiency, antithrombin III deficiency, polycythemia vera, and carcinoma.

ANATOMY AND PATHOPHYSIOLOGY

The blood supply to the intestines is depicted in Fig. 322-1. To prevent ischemic injury, extensive collateralization occurs between major mesenteric trunks and branches of the mesenteric arcades. Collateral vessels within the small bowel are numerous and meet within the duodenum and the bed of the pancreas. Collateral vessels within the colon meet at the splenic flexure and descending/sigmoid colon. These areas, which are inherently at risk for decreased blood flow, are known as *Griffiths' point* and *Sudeck's point*, respectively, and are the most common locations for colonic ischemia (Fig. 322-1, shaded areas). The splanchnic circulation can receive up to 30% of the cardiac output. Protective responses to prevent intestinal ischemia include abundant collateralization, autoregulation of blood flow, and the ability to increase oxygen extraction from the blood.

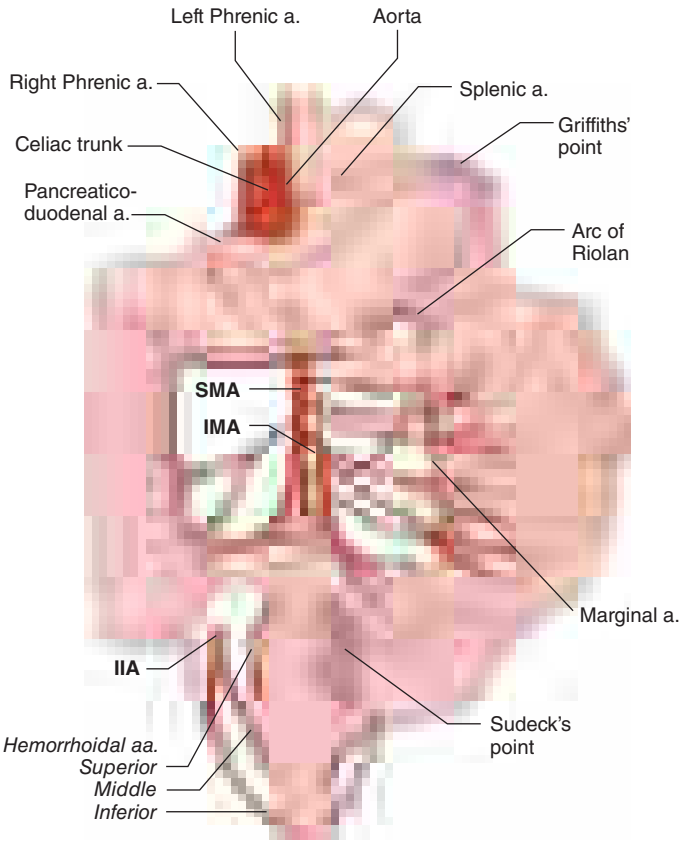


FIGURE 322-1 Blood supply to the intestines includes the celiac artery, superior mesenteric artery (SMA), inferior mesenteric artery (IMA), and branches of the internal iliac artery (IIA). Griffiths' and Sudeck's points, indicated by shaded areas, are watershed areas within the colonic blood supply and common locations for ischemia.

Occlusive ischemia is a result of disruption of blood flow by an embolus or progressive thrombosis in a major artery supplying the intestine. In >75% of cases, emboli originate from the heart and preferentially lodge in the superior mesenteric artery (SMA) just distal to the origin of the middle colic artery. Progressive thrombosis of at least two of the major vessels supplying the intestine is required for the development of chronic intestinal angina. Nonocclusive ischemia is disproportionate mesenteric vasoconstriction (arteriolar vasospasm) in response to a severe physiologic stress such as shock. If left untreated, early mucosal stress ulceration will progress to full-thickness injury. Even in the early stages of ischemia, there is translocation of bacteria across the intestinal mucosa, resulting in bacteremia that can lead to sepsis.

PRESENTATION, EVALUATION, AND MANAGEMENT

Intestinal ischemia remains one of the most challenging diagnoses. The mortality rate is >50%. The most significant indicator of survival is the timeliness of diagnosis and treatment. An overview of diagnosis and management of each form of intestinal ischemia is given in Table 322-1.

Acute mesenteric ischemia resulting from arterial embolus or thrombosis presents with severe acute, nonremitting abdominal pain strikingly out of proportion to the physical findings. Associated symptoms may include nausea and vomiting, transient diarrhea, anorexia, and bloody stools. With the exception of minimal abdominal distention and hypoactive bowel sounds, early abdominal examination is unimpressive. Later findings will demonstrate peritonitis and cardiovascular collapse. In the evaluation of acute intestinal ischemia, routine laboratory tests should be obtained, including complete blood count, serum chemistry, coagulation profile, arterial blood gas, amylase, lipase, lactic acid, blood type and cross match, and cardiac enzymes. Regardless of the need for urgent surgery, emergent admission to a monitored bed or intensive care unit is recommended for resuscitation and further evaluation. If the diagnosis of intestinal ischemia is being considered, consultation with a surgical service is necessary. Often the decision to operate is made on a high index of suspicion from the history and physical examination despite normal laboratory findings.

Other diagnostic modalities that may be useful in diagnosis but should not delay surgical therapy include electrocardiogram (ECG), echocardiogram, abdominal radiographs, computed tomography (CT), and mesenteric angiography. More recently, mesentery duplex scanning and visible light spectroscopy during colonoscopy have been demonstrated to be beneficial. The ECG may demonstrate an arrhythmia, indicating the possible source of the emboli. A plain abdominal film may show evidence of free intraperitoneal air, indicating a perforated viscus and the need for emergent exploration. Earlier features of intestinal ischemia seen on abdominal radiographs include bowel-wall edema, known as "thumbprinting." If the ischemia progresses, air can be seen within the bowel wall (pneumatosis intestinalis) and within the portal venous system. Other features include calcifications of the aorta and its tributaries, indicating atherosclerotic disease. With the administration of oral and IV contrast, dynamic CT angiography with three-dimensional reconstruction is a highly sensitive test for intestinal ischemia. In acute embolic disease, mesenteric angiography is best performed intraoperatively. A mesenteric duplex scan demonstrating a high peak velocity of flow in the SMA is associated with an ~80% positive predictive value of mesenteric ischemia. More significantly, a negative duplex scan virtually precludes the diagnosis of mesenteric ischemia. Duplex imaging serves as a screening test; however, further investigations with angiography are usually needed. One of the biggest limitations of duplex scanning is patients' body habitus. The duplex imaging yields poor results in obese patients. Nevertheless, "food fear" in patients with chronic disease often leads to a decreased appetite and lower abdominal fat, thus, yielding high duplex imaging results. The endoscopic techniques such as visible light spectroscopy can also be used in the diagnosis of chronic ischemia. When suspecting mesenteric ischemia involving the colon, performing an endoscopy to evaluate up to the splenic flexure is high yield. This is often an excellent diagnostic tool in patients with chronic renal insufficiency who cannot tolerate IV contrast.

TABLE 322-1 Overview of the Management of Acute Intestinal Ischemia

CONDITION	KEY TO EARLY DIAGNOSIS	TREATMENT OF UNDERLYING CAUSE	TREATMENT OF SPECIFIC LESION	TREATMENT OF SYSTEMIC CONSEQUENCE
Arterioocclusive mesenteric ischemia 1. Arterial embolus	Computed tomography (CT) angiography Early laparotomy	Anticoagulation Cardioversion Proximal thrombectomy	Laparotomy Embolectomy Vascular bypass Assess viability and resect dead bowel	Ensure hydration Give antibiotics Reverse acidosis Optimize oxygen delivery Avoid vasoconstrictors
2. Arterial thrombosis	Duplex ultrasound Angiography	Anticoagulation Hydration	Endovascular approach: thrombolysis, angioplasty and stenting Endarterectomy/thrombectomy or vascular bypass Assess viability and resect dead bowel	Give antibiotics Reverse acidosis Optimize oxygen delivery Support cardiac output Avoid vasoconstrictors
Mesenteric venous thrombosis Venous thrombosis	Spiral CT Angiography with venous phase	Anticoagulation Massive hydration	Anticoagulation ± laparotomy/ thrombectomy/catheter-directed thrombolysis Assess viability and resect dead bowel	Give antibiotics Reverse acidosis Optimize oxygen delivery Support cardiac output Avoid vasoconstrictors
Nonocclusive mesenteric ischemia	Vasospasm: Angiography Hypoperfusion: Spiral CT or colonoscopy	Ensure hydration Support cardiac output Avoid vasoconstrictors	Vasospasm Intraarterial vasodilators Hypoperfusion Delayed laparotomy Assess viability and resect dead bowel	Ensure hydration Give antibiotics Reverse acidosis Optimize oxygen delivery Support cardiac output Avoid vasoconstrictors

Source: Modified from GB Bulkley, in JL Cameron (ed): *Current Surgical Therapy*, 2nd ed. Toronto, BC Decker, 1986.

The "gold standard" for the diagnosis of acute arterial occlusive disease is angiography, and management is laparotomy. Surgical exploration should not be delayed if suspicion of acute occlusive mesenteric ischemia is high or evidence of clinical deterioration or frank peritonitis is present. The goal of operative exploration is to resect compromised bowel and restore blood supply. The entire length of the small and large bowel beginning at the ligament of Treitz should be evaluated. The pattern of intestinal ischemia may indicate the level of arterial occlusion. In the case of SMA occlusion where the embolus usually lies just proximal to the origin of the middle colic artery, the proximal jejunum is often spared while the remainder of the small bowel up to the transverse colon will be ischemic. The surgical management of acute mesenteric ischemia of the small bowel is embolectomy via arteriotomy; a small incision is made in the artery through which the clot is retrieved. Another way to manage acute thrombosis is thrombolysis therapy and angioplasty with stent placement. However, this approach is more commonly applied to treat chronic mesenteric ischemia. If this is unsuccessful, a bypass from the aorta or iliac artery to the SMA is performed.

Nonocclusive or vasospastic mesenteric ischemia presents with generalized abdominal pain, anorexia, bloody stools, and abdominal distention. Often these patients are obtunded, and physical findings may not assist in the diagnosis. The presence of a leukocytosis, metabolic acidosis, elevated amylase or creatinine phosphokinase levels, and/or lactic acidosis is useful in support of the diagnosis of advanced intestinal ischemia; however, these markers may not be indicative of either reversible ischemia or frank necrosis. Investigational markers for intestinal ischemia include D-dimer, glutathione *S*-transferase, platelet-activating factor (PAF), and mucosal pH monitoring. Regardless of the need for urgent surgery, emergent admission to a monitored bed or intensive care unit is recommended for resuscitation and further evaluation. Early manifestations of intestinal ischemia include fluid sequestration within the bowel wall leading to a loss of interstitial volume. Aggressive fluid resuscitation may be necessary. To optimize oxygen delivery, nasal O_2 and blood transfusions may be given. Broad-spectrum antibiotics should be given to provide sufficient coverage for enteric pathogens, including gram-negative and anaerobic organisms. Frequent monitoring of the patient's vital signs, urine output, blood gases, and lactate levels is paramount, as is frequent abdominal examination. All vasoconstricting agents should be avoided; fluid resuscitation is the intervention of choice to maintain hemodynamics.

If ischemic colitis is a concern, colonoscopy should be performed to assess the integrity of the colon mucosa. Visualization of the rectosigmoid region may demonstrate decreased mucosal integrity, associated more commonly with nonocclusive mesenteric ischemia, or, on occasion, occlusive disease as a result of acute loss of inferior mesenteric arterial flow following aortic surgery. Ischemia of the colonic mucosa is graded as *mild* with minimal mucosal erythema or as *moderate* with pale mucosal ulcerations and evidence of extension to the muscular layer of the bowel wall. *Severe* ischemic colitis presents with severe ulcerations resulting in black or green discoloration of the mucosa, consistent with full-thickness bowel-wall necrosis. The degree of reversibility can be predicted from the mucosal findings: mild erythema is nearly 100% reversible, moderate is ~50% reversible, and frank necrosis is simply dead bowel. Follow-up colonoscopy can be performed to rule out progression of ischemic colitis.

Laparotomy for nonocclusive mesenteric ischemia is warranted in patients with signs of peritonitis or worsening endoscopic findings and if the patient's condition does not improve with aggressive resuscitation. Ischemic colitis is optimally treated with resection of the ischemic bowel and formation of a proximal stoma. Primary anastomosis should not be performed in patients with acute intestinal ischemia.

Patients with mesenteric venous thrombosis may present with a gradual or sudden onset of pain. Symptoms include vague abdominal pain, nausea, and vomiting. Physical examination findings include abdominal distention with mild to moderate tenderness and signs of dehydration. The diagnosis of mesenteric thrombosis is frequently made on abdominal spiral CT with oral and IV contrast. Findings on CT angiography with venous phase include bowel-wall thickening and ascites. IV contrast will demonstrate a delayed arterial phase and clot within the superior mesenteric vein. The goal of management is to optimize hemodynamics and correct electrolyte abnormalities

with massive fluid resuscitation. Intravenous antibiotics as well as anticoagulation should be initiated. If laparotomy is performed and mesenteric venous thrombosis is suspected, heparin anticoagulation is immediately initiated, and compromised bowel is resected. Of all acute intestinal disorders, mesenteric venous insufficiency is associated with the best prognosis.

Chronic intestinal ischemia presents with intestinal angina or postprandial abdominal pain associated with increased need of blood flow to the intestine following meals. Patients report abdominal cramping and pain following ingestion of a meal. Weight loss and chronic diarrhea may also be noted. Abdominal pain without weight loss is not chronic mesenteric angina. Physical examination will often reveal a malnourished patient with an abdominal bruit as well as other manifestations of atherosclerosis. Duplex ultrasound evaluation of the mesenteric vessels has gained in popularity. It is important to perform the test while the patient is fasting because the presence of increased bowel gas prevents adequate visualization of flow disturbances within the vessels or the lack of a vasodilation response to feeding during the test. This tool is frequently used as a screening test for patients with symptoms suggestive of chronic mesenteric ischemia. The gold standard for confirmation of mesenteric arterial occlusion is mesenteric angiography. Evaluation with mesenteric angiography allows for identification and possible intervention for the treatment of atherosclerosis within the vessel lumen and will also evaluate the patency of remaining mesenteric vessels. The use of mesenteric angiography may be limited in the presence of renal failure or contrast allergy. Magnetic resonance angiography is an alternative if the administration of contrast dye is contraindicated.

The management of chronic intestinal ischemia includes medical management of atherosclerotic disease by exercise, cessation of smoking, and antiplatelet and lipid-lowering medications. A full cardiac evaluation should be performed before intervention on chronic mesenteric ischemia. Newer endovascular procedures may avoid an operative intervention in selected patient populations. Angioplasty with endovascular stenting in the treatment of chronic mesenteric ischemia is associated with an 80% long-term success rate. In patients requiring surgical exploration, the approach used is determined by findings of the mesenteric angiogram. The entire length of the small and large bowel should be evaluated, beginning at the ligament of Treitz. Restoration of blood flow at the time of laparotomy is accomplished with mesenteric vessel endarterectomy or bypass.

Determination of intestinal viability intraoperatively in patients with suspected intestinal ischemia can be challenging. After revascularization, peristalsis and return of a pink color of the bowel wall should be observed. Palpation of major arterial mesenteric vessels can be performed, as well as applying a Doppler flowmeter to the antimesenteric border of the bowel wall, but neither is a definitive indicator of viability. In equivocal cases, 1 g of IV sodium fluorescein is administered, and the pattern of bowel reperfusion is observed under ultraviolet illumination with a standard (3600 A) Wood's lamp. An area of nonfluorescence >5 mm in diameter suggests nonviability. If doubt persists, reexploration performed 24–48 h following surgery will allow demarcation of nonviable bowel. Primary intestinal anastomosis in patients with ischemic bowel is always worrisome; thus, delayed bowel reconstruction and reanastomosis should be deferred to the time of second-look laparotomy.

ACKNOWLEDGMENTS

We thank Cory Sandore for providing the illustration for this chapter. Susan Gearhart contributed to this chapter in the 18th edition and Rizwan Ahmed contributed to the 19th edition.

FURTHER READING

Deng QW et al: Risk factors for postoperative acute mesenteric ischemia among adult patients undergoing cardiac surgery: A systematic review and meta-analysis. J Crit Care 42:294, 2017.

Salsano G et al: What is the best revascularization strategy for acute occlusive arterial mesenteric ischemia: Systematic review and meta-analysis. Cardiovasc Intervent Radiol 41:27, 2018.

Sise MJ: Acute mesenteric ischemia. Surg Clin North Am 94:165–81, 2014.

323 Acute Intestinal Obstruction

Danny O. Jacobs

EPIDEMIOLOGY

Morbidity and mortality from acute intestinal obstruction have been decreasing over the past several decades. Nevertheless, the diagnosis can still be challenging, and the type of complications that patients suffer has not changed significantly. The extent of mechanical obstruction is typically described as partial, high-grade, or complete—generally correlating with the risk of complications and the urgency with which the underlying disease process must be addressed. Obstruction is also commonly described as being either "simple" or, alternatively, "strangulated" if vascular insufficiency and intestinal ischemia are evident.

Acute intestinal obstruction occurs either *mechanically* from blockage or from intestinal dysmotility when there is no blockage. In the latter instance, the abnormality is described as being *functional*. Mechanical bowel obstruction may be caused by extrinsic processes, intrinsic abnormalities of the bowel wall, or intraluminal abnormalities **(Table 323-1)**. Within each of these broad categories are many diseases that can impede intestinal propulsion. Intrinsic diseases that can cause intestinal obstruction are usually congenital, inflammatory, neoplastic, or traumatic in origin, although intussusception and radiation injury can also be etiologic.

Acute intestinal obstruction accounts for ~1–3% of all hospitalizations and a quarter of all urgent or emergent general surgery admissions. Approximately 80% of cases involve the small bowel, and about one-third of these patients show evidence of significant ischemia. The mortality rate for patients with strangulation who are operated on within 24–30 h of the onset of symptoms is ~8% but triples shortly thereafter.

Extrinsic diseases most commonly cause mechanical obstruction of the small intestine. In the United States and Europe, almost all cases are caused by postoperative adhesions, carcinomatosis, or herniation of the anterior abdominal wall. Carcinomatosis most often originates from the ovary, pancreas, stomach, or colon, although rarely, metastasis from distant organs like the breast and skin can occur. Adhesions are responsible for the majority of cases of early postoperative obstruction that require intervention. It is important to note many patients who are successfully treated for adhesive small-bowel instruction will recur. Approximately 20% of patients who were treated conservatively and

TABLE 323-1 Most Common Causes of Acute Intestinal Obstruction

Extrinsic Disease
Adhesions (especially due to previous abdominal surgery), internal or external hernias, neoplasms (including carcinomatosis and extraintestinal malignancies, mostly commonly ovarian), endometriosis or intraperitoneal abscesses, and idiopathic sclerosis
Intrinsic Disease
Congenital (e.g., malrotation, atresia, stenosis, intestinal duplication, cyst formation, and congenital bands—the latter rarely in adults)
Inflammation (e.g., inflammatory bowel disease, especially Crohn's disease, but also diverticulitis, radiation, tuberculosis, lymphogranuloma venereum, and schistosomiasis)
Neoplasia (note: primary small-bowel cancer is rare; obstructive colon cancer may mimic small-bowel obstruction if the ileocecal valve is incompetent)
Traumatic (e.g., hematoma formation, anastomotic strictures)
Other, including intussusception (where the lead point is typically a polyp or tumor in adults), volvulus, obstruction of duodenum by superior mesenteric artery, radiation or ischemic injury, and aganglionosis, which is Hirschsprung's disease
Intraluminal Abnormalities
Bezoars, feces, foreign bodies including inspissated barium, gallstones (entering the lumen via a cholecystoenteric fistula), enteroliths

TABLE 323-2 Acute Small-Intestinal and Colonic Obstruction Incidences

CAUSE	INCIDENCE
Postoperative adhesions	>50%
Neoplasms	~20%
Hernias (especially ventral or internal types, where the risk of strangulation is increased)	~10%
Inflammatory bowel disease, other inflammation (obstruction may resolve if acute inflammation and edema subside)	~5%
Intussusception, volvulus, other miscellaneous diseases	<15%

between 5 and 30% of patients who were managed operatively will require readmission within 10 years.

Open operations of the lower abdomen, including appendectomy and colorectal and gynecologic procedures, are especially likely to create adhesions that can cause bowel obstruction **(Table 323-2)**. The risk of internal herniation is increased by abdominal procedures such as laparoscopic or open Roux-en-Y gastric bypass. Although laparoscopic procedures may generate fewer postoperative adhesions compared with open surgery, the risk of obstructive adhesion formation is not eliminated.

Volvulus, which occurs when bowel twists on its mesenteric axis, can cause partial or complete obstruction and vascular insufficiency. The sigmoid colon is most commonly affected, accounting for approximately two-thirds of all cases of volvulus and 4% of all cases of large-bowel obstruction. The cecum and terminal ileum can also volvulize, or the cecum alone may be involved as a cecal bascule. Risk factors include institutionalization, the presence of neuropsychiatric conditions requiring psychotropic medication, chronic constipation, and aging; patients typically present in their seventies or eighties.

Colonic volvulus is more common in Eastern Europe, Russia, and Africa than it is in the United States. It is rare for adhesions or hernias to obstruct the colon. Cancer of the descending colon and rectum is responsible for approximately two-thirds of all cases, followed by diverticulitis and volvulus.

Functional obstruction, also known as *ileus* and *pseudo-obstruction*, is present when dysmotility prevents intestinal contents from being propelled distally and no mechanical blockage exists. Ileus that occurs after intraabdominal surgery is the most commonly identified form of functional bowel obstruction, but there are many other causes **(Table 323-3)**. Although postoperative ileus is most often transient, it is often the most common reason why hospital discharge is delayed. Pseudo-obstruction of the colon, also known as Ogilvie's syndrome, is a relatively rare disease. Some patients with Ogilvie's syndrome have colonic dysmotility due to abnormalities of their autonomic nervous system that may be inherited.

TABLE 323-3 Most Common Causes of Ileus (Functional or Pseudo-Obstruction of the Intestine)

Intraabdominal procedures, lumbar spinal injuries, or surgical procedures on the lumbar spine and pelvis
Metabolic or electrolyte abnormalities, especially hypokalemia and hypomagnesemia, but also hyponatremia, uremia, and severe hyperglycemia
Drugs such as opiates, antihistamines, and some psychotropic (e.g., haloperidol, tricyclic antidepressants) and anticholinergic agents
Intestinal ischemia
Intraabdominal or retroperitoneal inflammation or hemorrhage
Lower lobe pneumonias
Intraoperative radiation (likely due to muscle damage)
Systemic sepsis
Hyperparathyroidism
Pseudo-obstruction (Ogilvie's syndrome)
Ileus secondary to hereditary or acquired visceral myopathies and neuropathies that disrupt myocellular neural coordination
Some collagen vascular diseases such as lupus erythematosus or scleroderma

PATHOPHYSIOLOGY

The manifestations of acute intestinal obstruction depend on the nature of the underlying disease process, its location, and changes in blood flow **(Fig. 323-1)**. Increased intestinal contractility, which occurs proximally *and* distal to the obstruction, is a characteristic response. Subsequently, intestinal peristalsis slows as the intestine or stomach proximal to the point of obstruction dilates and fills with gastrointestinal secretions and swallowed air. Although swallowed air is the primary contributor to intestinal distension, intraluminal air may also accumulate from fermentation, local carbon dioxide production, and altered gaseous diffusion.

Intraluminal dilation also increases intraluminal pressure. When luminal pressure exceeds venous pressure, venous and lymphatic drainage is impeded. Edema ensues, and the bowel wall proximal to the site of blockage may become hypoxemic. Epithelial necrosis can be identified within 12 h of obstruction. Ultimately, arterial blood supply may become so compromised that full-thickness ischemia, necrosis, and perforation result. Stasis increases the bacteria counts within the jejunum and ileum. Bacteria like *Escherichia coli*, *Streptococcus faecalis*, and *Klebsiella*, and other pathogens may be recovered from intestinal cultures, mesenteric lymph nodes, the blood stream, and other sites.

Other manifestations depend on the degree of hypovolemia, the patient's metabolic response, and the presence or absence of associated intestinal ischemia. Inflammatory edema eventually increases the production of reactive oxygen species and activates neutrophils and macrophages, which accumulate within the bowel wall. Their accumulation, along with changes in innate immunity, disrupts secretory and neuromotor processes. Dehydration is caused by loss of the normal intestinal absorptive capacity as well as fluid accumulation in the gastric or intestinal wall and intraperitoneally.

Anorexia and emesis tend to exacerbate intravascular volume depletion. In the worst case scenario that is most commonly identified after high-grade distal obstruction, emesis leads to losses of gastric potassium, hydrogen, and chloride, while dehydration stimulates proximal renal tubule bicarbonate reabsorption. Intraperitoneal fluid accumulation, especially in patients with severe distal bowel obstruction, may increase intraabdominal pressure enough to elevate the diaphragm, inhibit respiration and to impede systemic venous return and promote vascular instability. Severe hemodynamic compromise may elicit a systemic inflammatory response and generalized microvascular leakage.

Closed-loop obstruction results when the proximal and distal openings of a given bowel segment are both occluded, for example, due to volvulus or a hernia. It is the most common precursor for strangulation, but not every closed loop strangulates. The risk of vascular insufficiency, systemic inflammation, hemodynamic compromise, and irreversible intestinal ischemia is much greater in patients with closed-loop obstruction. Pathologic changes may occur more rapidly, and emergency intervention is indicated. Irreversible bowel ischemia may progress to transmural necrosis even if obstruction is relieved. It is also important to remember that patients with high-grade distal colonic obstruction who have competent ileocecal valves may present with closed-loop obstruction. In the latter instance, the cecum may progressively dilate such that ischemic necrosis results in perforation especially when the cecal diameter exceeds 10–12 cm, as informed by Laplace's law. Patients with distal colonic obstruction whose ileocecal valves are incompetent tend to present later in the course of disease and mimic patients with distal small-bowel obstruction.

HISTORY AND PHYSICAL FINDINGS

Even though the presenting signs and symptoms can be misleading, many patients with acute obstruction can be accurately diagnosed after a thorough history and physical examination is performed. However, small-bowel obstruction with strangulation can be especially difficult to diagnosis promptly. Early recognition allows earlier treatment that decreases the risk of progression or other excess morbidity.

The cardinal signs are colicky abdominal pain, abdominal distention, emesis, and obstipation. More intraluminal fluid accumulates

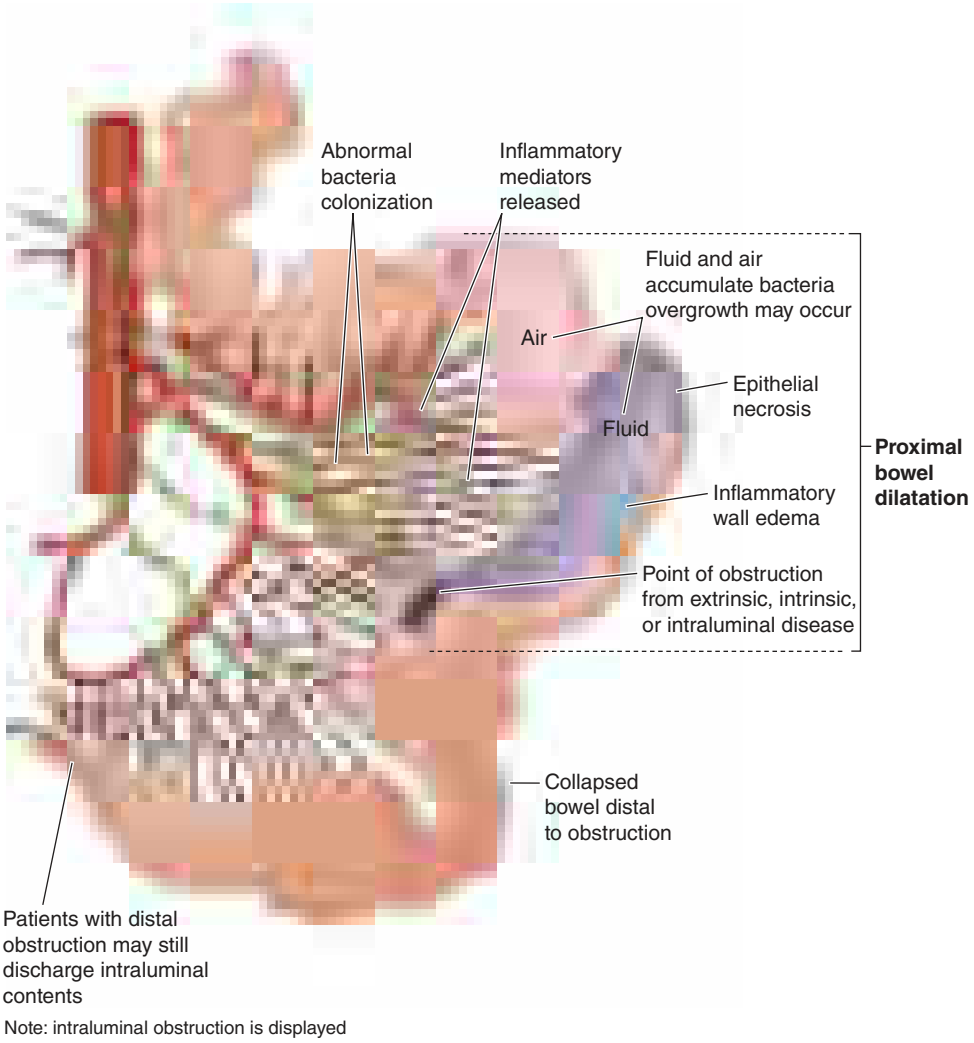


FIGURE 323-1 Pathophysiologic changes of small-bowel obstruction.

in patients with distal obstruction, which typically leads to greater distention, more discomfort, and delayed emesis. This emesis is feculent when there is bacterial overgrowth. Patients with more proximal obstruction commonly present with less abdominal distention but more pronounced vomiting. Elements of the history that might be helpful include any prior history of surgery, including herniorrhaphy, as well as any history of cancer or inflammatory bowel disease.

Most patients, even those with simple obstruction, appear to be critically ill. Many may be oliguric, hypotensive, and tachycardic because of severe intravascular volume depletion. Fever is worrisome for strangulation or systemic inflammation. Bowel sounds and bowel functional activity are notoriously difficult to interpret. Classically, many patients with early small-bowel obstruction will have high-pitched, "musical" tinkling bowel sounds and peristaltic "rushes" known as borborygmi. Later in the course of disease, the bowel sounds may be absent or hypoactive as peristaltic activity decreases. This is in contrast to the common findings in patients with ileus or pseudo-obstruction where bowel sounds are typically absent or hypoactive from the beginning. Lastly, patients with partial blockage may continue to pass flatus and stool, and those with complete blockage may evacuate bowel contents present downstream beyond their obstruction.

All surgical incisions should be examined and the presence of a tender abdominal or groin mass strongly suggests that an incarcerated hernia may be the cause of obstruction. The presence of tenderness should increase the concern about the presence of complications such as ischemia, necrosis, or peritonitis. Severe pain with localization or signs of peritoneal irritation is suspicious for strangulated or closed-loop obstruction. It is important to remember that the discomfort may be out of proportion to physical findings mimicking the complaints of patients with acute mesenteric ischemia. Patients with colonic volvulus present with the classic manifestations of closed-loop obstruction: severe abdominal pain, vomiting, and obstipation. Asymmetrical abdominal distension and a tympanic mass may be evident.

Patients with ileus or pseudo-obstruction may have signs and symptoms similar to those of bowel obstruction. Although abdominal distention is present, colicky abdominal pain is typically absent, and patients may not have nausea or emesis. Ongoing, regular discharge of stool or flatus can sometimes help distinguish patients with ileus from those with complete mechanical bowel obstruction.

LABORATORY AND IMAGING STUDIES

Laboratory testing should include a complete blood count and serum electrolyte and creatinine measurements. Serial assessments are often useful. Mild hemoconcentration and slight elevation of the white blood cell count commonly occur after simple bowel obstruction. Emesis and dehydration may cause hypokalemia, hypochloremia, elevated blood urea nitrogen–to–creatinine ratios, and metabolic alkalosis. Patients may be hyponatremic on admission because many have attempted to rehydrate themselves with hypotonic fluids. The presence of guaiac-positive stools and iron-deficiency anemia are strongly suggestive of malignancy.

Higher white blood cell counts with the presence of immature forms or the presence of metabolic acidosis are worrisome for severe volume depletion or ischemic necrosis and sepsis. Presently, there are no laboratory tests that are especially useful for identifying the presence

of simple or strangulated obstruction, although increases in serum D-lactate, creatine kinase BB isoenzymes, or intestinal fatty acid binding protein levels may be suggestive of the latter.

Recommendations for diagnostic imaging continue to evolve. In all cases, the key is not to operative intervention unnecessarily when the patient's signs or symptoms strongly suggest that high-grade or complete obstruction or bowel compromise is present. Abdominal radiography, which must include upright or cross-table lateral views, can be completed quickly and may indicate the need for emergency surgical intervention in patients who are not in the immediate postoperative period. A "staircasing" pattern of dilated air and fluid-filled small-bowel loops >2.5 cm in diameter with little or no air seen in the colon are classical findings in patients with small-bowel obstruction, although findings may be equivocal in some patients with documented disease. Little bowel gas appears in patients with proximal bowel obstruction or in patients whose intestinal lumens are filled with fluid. Upright plain films of the abdomen of patients with large-bowel obstruction typically show colon dilatation. Small-bowel air-fluid levels will not be obvious if the ileocecal valve is competent. Although it can be difficult to distinguish from ileus, small-bowel obstruction is more likely when air-fluid levels are seen without significant colonic distension. Free air suggests that perforation has occurred in patients who have not recently undergone surgical procedures. A gas-filled, "coffee bean"–shaped dilated shadow may be seen in patients with volvulus.

More sophisticated imaging, which may be unnecessarily time consuming and expensive, can nevertheless be beneficial when the diagnosis is unclear. Computed tomography (CT) is the most commonly used imaging modality. Its sensitivity for detecting bowel obstruction is ~95% (78–100%) in patients with high-grade obstruction, with a specificity of 96% and an accuracy of ≥95%. Its accuracy in diagnosing closed-loop obstruction is much lower (60%). CT may also provide useful information regarding location or to identify particular circumstances where surgical intervention is needed urgently. Patients who have evidence of contrast appearing within the cecum within 4–24 h of oral administration of water-soluble contrast can be expected to improve with high sensitivity and specificity (~95% each). For example, contrast studies may demonstrate a "bird's beak," a "c-loop," or "whorl" deformity on CT imaging at the site where twisting obstructs the lumen when a colonic volvulus is present. Although abdominal radiography is usually the initial examination, unlike CT imaging, it may not accurately distinguish obstruction from other causes of colonic dysmotility. Examples of some CT images are reproduced in **Fig. 323-2**.

Ultrasonographic evaluations are especially difficult to interpret but may be sensitive and appropriate studies to evaluate patients who are pregnant or for whom x-ray exposure is otherwise contraindicated or inappropriate.

CT imaging with enteral and IV contrast can also identify ischemia. Altered bowel wall enhancement is the most specific early finding, but its sensitivity is low. Mesenteric venous gas, pneumoperitoneum, and pneumatosis intestinalis are late findings indicating the presence of bowel necrosis. CT scanning after a water-soluble contrast enema may help distinguish ileus or pseudo-obstruction from distal large-bowel obstruction in patients who present with evidence of small-bowel and colonic distention. CT enteroclysis, though rarely performed, can accurately identify neoplasia as a cause of bowel obstruction. Contrast enemas or colonoscopies are almost always needed to identify causes of acute colonic obstruction.

Barium studies are generally contraindicated in patients with firm evidence of complete or high-grade bowel obstruction, especially when they present acutely. Barium should never be given orally to a patient with possible obstruction until that diagnosis has been excluded. In every other case, such investigations should only be performed in exceptional circumstances and with great caution because patients with significant obstruction may develop barium concretions as an additional source of blockage and some who would have otherwise recovered will require operative intervention. Barium opacification also renders cross-sectional imaging studies or angiography uninterpretable.

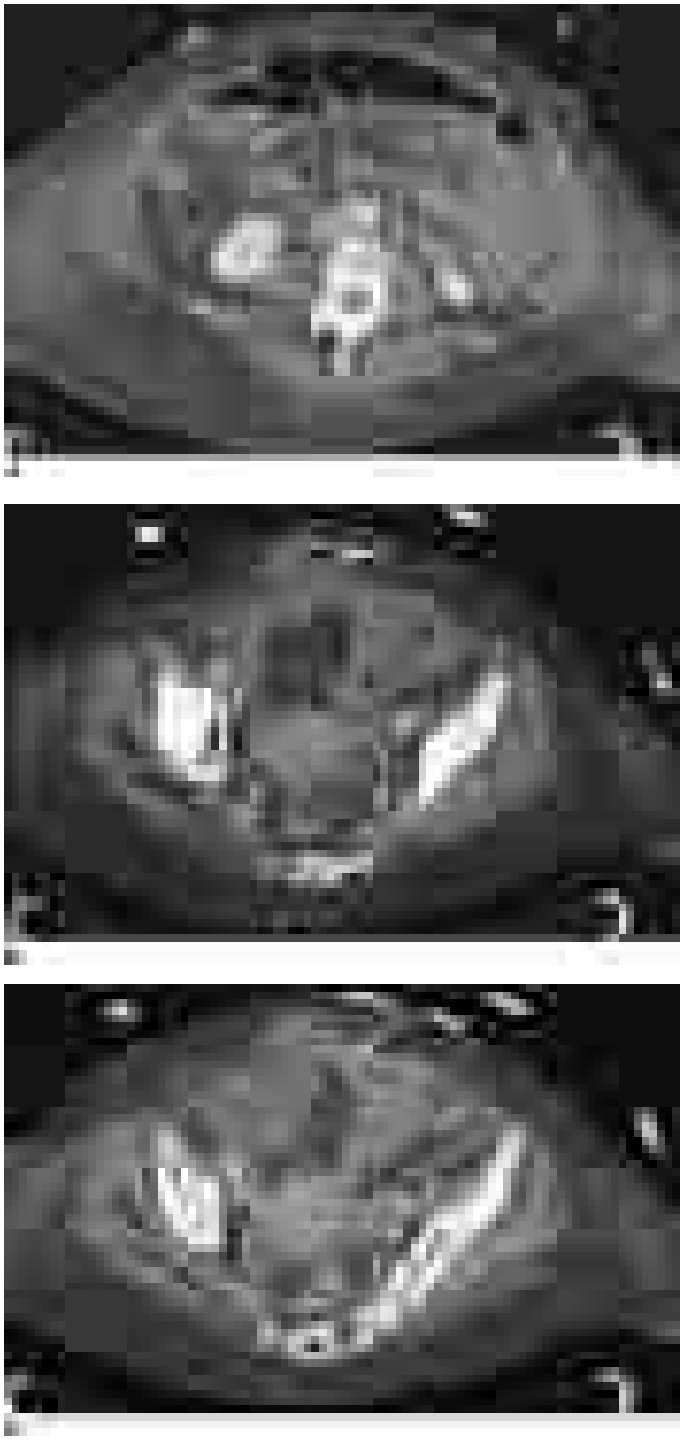


FIGURE 323-2 Computed tomography with oral and intravenous contrast demonstrating (***A***) evidence of small-bowel dilatation with air-fluid levels consistent with a small-bowel obstruction; (***B***) a partial small-bowel obstruction from an incarcerated ventral hernia (*arrow*); and (***C***) decompressed bowel seen distal to the hernia (*arrow*). *(From W Silen: Acute intestinal obstruction, in DL Longo et al [eds]: Harrison's Principles of Internal Medicine, 18th ed. New York, McGraw-Hill, 2012.)*

TREATMENT

Acute Intestinal Obstruction

An improved understanding of the pathophysiology of bowel obstruction and the importance of fluid resuscitation, electrolyte repletion, intestinal decompression, and the selected use of antibiotics have likely contributed to a reduction in mortality from acute bowel obstruction. Every patient should be stabilized as quickly

as possible. Nasogastric tube suction decompresses the stomach, minimizes further distention from swallowed air, improves patient comfort, and reduces the risk of aspiration. Urine output should be assessed using a Foley catheter. In some cases, for example, in patients with cardiac disease, central venous pressures should be monitored. The use of antibiotics is controversial, although prophylactic administration may be warranted if operation is anticipated. Complete bowel obstruction is an indication for intervention. Stenting may be possible and warranted for some patients with high-grade obstruction due to unresectable stage IV malignancy. Stenting may also allow elective mechanical bowel preparation before surgery is undertaken. Because treatment options are so variable, it is helpful to make as precise a diagnosis as possible preoperatively.

ILEUS

Patients with ileus are treated supportively with intravenous fluids and nasogastric decompression while any underlying pathology is treated. Pharmacologic therapy is not yet proven to be efficacious or cost-effective. However, peripherally active μ-opioid receptor antagonists (e.g., alvimopan and methylnaltrexone) may accelerate gastrointestinal recovery in some patients who have undergone abdominal surgery.

COLONIC PSEUDO-OBSTRUCTION (OGILVIE'S DISEASE)

Neostigmine is an acetylcholinesterase inhibitor that increases cholinergic (parasympathetic) activity, which can stimulate colonic motility. Some studies have shown it to be moderately effective in alleviating acute colonic pseudo-obstruction. It is the most common therapeutic approach and can be used once it is certain that there is no mechanical obstruction. Cardiac monitoring is required, and atropine should be immediately available. Intravenous administration induces defecation and flatus within 10 min in the majority of patients who will respond. Sympathetic blockade by epidural anesthesia can successfully ameliorate pseudo-obstruction in some patients.

VOLVULUS

Patients with sigmoid volvulus can often be decompressed using a flexible tube inserted through a rigid proctoscope or using a flexible sigmoidoscope. Successful decompression results in sudden release of gas and fluid with evidence of decreased abdominal distension and allows definitive correction to be scheduled electively. Cecal volvulus most often requires laparotomy or laparoscopic correction.

INTRAOPERATIVE STRATEGIES

Approximately 60–80% of selected patients with mechanical bowel obstruction can be successfully treated conservatively. Indeed, most cases of radiation-induced obstruction should also be managed nonoperatively if possible. In most circumstances, early consultation with a surgeon is prudent when there is concern about strangulation obstruction or other abnormality that needs to be addressed urgently. Deterioration signifies a need for intervention. At this time, the decision as to whether the patient can continue to be treated nonoperatively can only be based on clinical judgment, although, as described earlier, imaging studies can sometimes be helpful. The frequency of major complications after operation ranges from 12 to 47%, with greater risk being attributed to resection therapies and the patient's overall health. Risk is increased for patients with American Society of Anesthesiologists (ASA) class III or higher.

At operation, dilation proximal to the site of blockage with distal collapse is a defining feature of bowel obstruction. Intraoperative strategies depend on the underlying problem and range from lysis of adhesions to resection with or without diverting ostomy to primary resection with anastomosis. Resection is warranted when there is concern about the bowel's viability after the obstructive process is relieved. Laparoscopic approaches can be useful for patients with early obstruction when extensive adhesions are not expected to be present. Some patients with high-grade obstruction secondary to malignant disease that is not amendable to resection will benefit from bypass procedures.

ADULT INTUSSUSCEPTION AND GALLSTONE ILEUS

Primary resection is prudent. Careful manual reduction of any involved bowel may limit the amount of intestine that needs to be removed. A proximal ostomy may be required if unprepped colon is involved. The most common site of intestinal obstruction in patients with gallstone "ileus" is the ileum (60% of patients). The gallstone enters the intestinal tract most often via a cholecystoduodenal fistula. It can usually be removed by operative enterolithotomy. Addressing the gallbladder disease during urgent or emergent surgery is not recommended.

POSTOPERATIVE BOWEL OBSTRUCTION

Early postoperative mechanical bowel obstruction is that which occurs within the first 6 weeks of operation. Most are partial and can be expected to resolve spontaneously. It tends to respond and behave differently from classic mechanical bowel obstruction and may be very difficult to distinguish from postoperative ileus. A higher index of suspicion for a definitive site of obstruction is warranted for patients who undergo laparoscopic surgical procedures. Patients who first had ileus and then subsequently develop obstructive symptoms after an initial return of normal bowel function are more likely to have true postoperative small-bowel obstruction. The longer it takes for a patient's obstructive symptoms to resolve after hospitalization, the more likely the patient is to require surgical intervention.

Acknowledgment

The wisdom and expertise of Dr. William Silen are gratefully acknowledged.

FURTHER READING

Catena F et al: Adhesive small bowel adhesions obstruction: Evolutions in diagnosis, management and prevention. World J Gastrointest Surg 27:222, 2016.

Ferrada P et al: Surgery or stenting for colonic obstruction: A practice management guideline from the Eastern Association for the Surgery of Trauma. J Trauma Acute Care Surg 80:659, 2016.

Jaffe T, Thompson WM: Large-bowel obstruction in the adults: Classic radiographic and CT findings, etiology and mimics. Radiology 275:651, 2015.

Paulson EK, Thompson WM: Review of small-bowel obstruction: The diagnosis and when to worry. Radiology 275:332, 2015.

Perry H et al: Relative accuracy of emergency CT in adults with non-traumatic abdominal pain. Brit Inst Rad 89:20150416, 2016.

Taylor MR, Lalani N: Adult small bowel obstruction. Acad Emerg Med 20:528, 2013.

324 Acute Appendicitis and Peritonitis

Danny O. Jacobs

ACUTE APPENDICITIS

INCIDENCE AND EPIDEMIOLOGY

Appendicitis occurs more frequently in Westernized societies but its incidence is decreasing for uncertain reasons. Nevertheless, acute appendicitis remains the most common emergency general surgical disease affecting the abdomen, with a rate of ~100 per 100,000 person-years in Europe and the Americas or about 11 cases per 10,000 people annually. Approximately 9% of men and 7% of women will experience an episode during their lifetime. Appendicitis occurs most commonly in 10- to 19-year-olds; however, the average age at diagnosis appears to be gradually increasing, as is the frequency of the

disease in African Americans, Asians, and Native Americans. Overall, 70% of patients are <30 years old and most are men.

One of the more common complications and most important causes of excess morbidity and mortality is perforation, whether it is contained and localized or unconstrained within the peritoneal cavity. In contrast to the trend observed for appendicitis and appendectomy, the incidence of perforated appendicitis (~20 cases per 100,000 person-years) is increasing. The explanation for this trend is unknown. Approximately 20% of all patients will present with evidence of perforation, but the percentage risk is much higher in patients under 5 or over 65 years of age.

PATHOGENESIS OF APPENDICITIS AND APPENDICEAL PERFORATION

Appendicitis was first described in 1886 by Reginald Fitz. Its etiology is still not completely understood. Fecaliths, incompletely digested food residue, lymphoid hyperplasia, intraluminal scarring, tumors, bacteria, viruses, and inflammatory bowel disease have all been associated with inflammation of the appendix and appendicitis.

Although not proven, obstruction of the appendiceal lumen is believed to be an important step in the development of appendicitis—at least in some cases. Here, obstruction leads to bacterial overgrowth and luminal distension, with an increase in intraluminal pressure that can inhibit the flow of lymph and blood. Then, vascular thrombosis and ischemic necrosis with perforation of the distal appendix may occur. Therefore, perforation that occurs near the base of the appendix should raise concerns about another disease process. Most patients who will perforate do so before they are evaluated by surgeons.

Appendiceal fecaliths (or appendicoliths) are found in ~50% of patients with gangrenous appendicitis who perforate but are rarely identified in those who have simple disease. As mentioned earlier, the incidence of perforated, but not simple, appendicitis is increasing. The rate of perforated and nonperforated appendicitis is correlated in men but not in women. Together these observations suggest that the underlying pathophysiologic processes are different and that simple appendicitis does not always progress to perforation. It appears that some cases of simple acute appendicitis may resolve spontaneously or with antibiotic therapy with limited risk of recurrent disease. The use of antibiotics to treat uncomplicated appendicitis is currently being studied intensively. Preliminary data indicate that as many as 70% of patients who present with uncomplicated appendicitis based on computed tomography (CT) and who are treated with antibiotics alone will be free of recurrent disease for at least a year. These findings highlight the importance of clinical decision-making and risk assessment when deciding and discussing treatment options with patients who presumably have simple disease, for example, deciding who is an appropriate candidate for non-operative management and who is not. The latter is especially pertinent given the difficulty in assessing which patients might progress to perforation and which will not.

Increasingly it appears that there are two broad categories of patients with appendicitis—those with complicated disease like gangrene or perforation and those without. When perforation occurs, the resultant leak may be contained by the omentum or other surrounding tissues to form an abscess. Free perforation normally causes severe peritonitis. These patients may also develop infective suppurative thrombosis of the portal vein and its tributaries along with intrahepatic abscesses. The prognosis of the very unfortunate patients who develop this rare but dreaded complication is very poor.

CLINICAL MANIFESTATIONS

Improved diagnosis, supportive care, and surgical interventions are likely responsible for the remarkable decrease in the risk of mortality from simple appendicitis to currently <1%. Nevertheless, it is still important to identify patients who might have appendicitis as early as possible. Patients who have persistent symptoms that haven't improved over 48 h may be more likely to perforate or develop other complications.

Appendicitis should be included in the differential diagnosis of abdominal pain for every patient in any age group unless it is certain that the organ has been previously removed **(Table 324-1)**.

TABLE 324-1 Some Conditions That Mimic Appendicitis

Crohn's disease	Meckel's diverticulitis
Cholecystitis or other gallbladder disease	Mittelschmerz
Diverticulitis	Mesenteric adenitis
Ectopic pregnancy	Omental torsion
Endometriosis	Pancreatitis
Gastroenteritis or colitis	Lower lobe pneumonia
Gastric or duodenal ulceration	Pelvic inflammatory disease
Hepatitis	Ruptured ovarian cyst or other cystic disease of the ovaries
Kidney disease, including nephrolithiasis	Small-bowel obstruction
Liver abscess	Urinary tract infection

The appendix's anatomical location, which varies, may directly influence how the patient presents. Where the appendix can be "found" ranges from local differences in how the appendiceal body and tip lie relative to its attachment to the cecum **(Figs. 324-1 and 324-2)**, to where the appendix is actually situated in the peritoneal cavity—for example, from its typical location in the right lower quadrant, to the pelvis, right flank, right upper quadrant (as may be observed during pregnancy), or even the left side of the abdomen for patients with malrotation or who have severely redundant colons.

Because the differential diagnosis of appendicitis is so extensive, deciding if a patient has appendicitis can be difficult **(Table 324-2)**. Many patients may not present with the classically described history or physical findings and some may not have any abdominal discomfort early in the disease process. Soliciting an appropriate history requires detecting and evaluating symptoms that might suggest alternative diagnoses.

What is the classic history? Nonspecific complaints occur first. Patients may notice changes in bowel habits or malaise and vague, perhaps intermittent, crampy, abdominal pain in the epigastric or periumbilical region. The pain subsequently migrates to the right lower quadrant over 12–24 h, where it is sharper and can be definitively localized as transmural inflammation when the appendix irritates the parietal peritoneum. Parietal peritoneal irritation may be associated with local muscle rigidity and stiffness. Patients with appendicitis will most often observe that their nausea, if present, followed the development of abdominal pain, which can help distinguish them from patients with gastroenteritis, for example, where nausea occurs first. Emesis, if present, also occurs after the onset of pain and is typically mild and scant. Thus, timing of the onset of symptoms and the characteristics of the patient's pain and any associated findings must be rigorously assessed. Anorexia is so common that the diagnosis of appendicitis should be questioned in its absence.

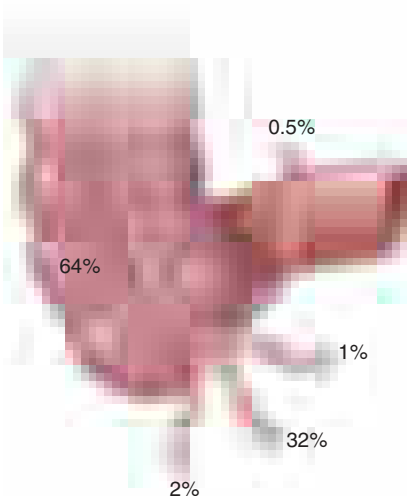


FIGURE 324-1 Regional anatomical variations of the appendix.

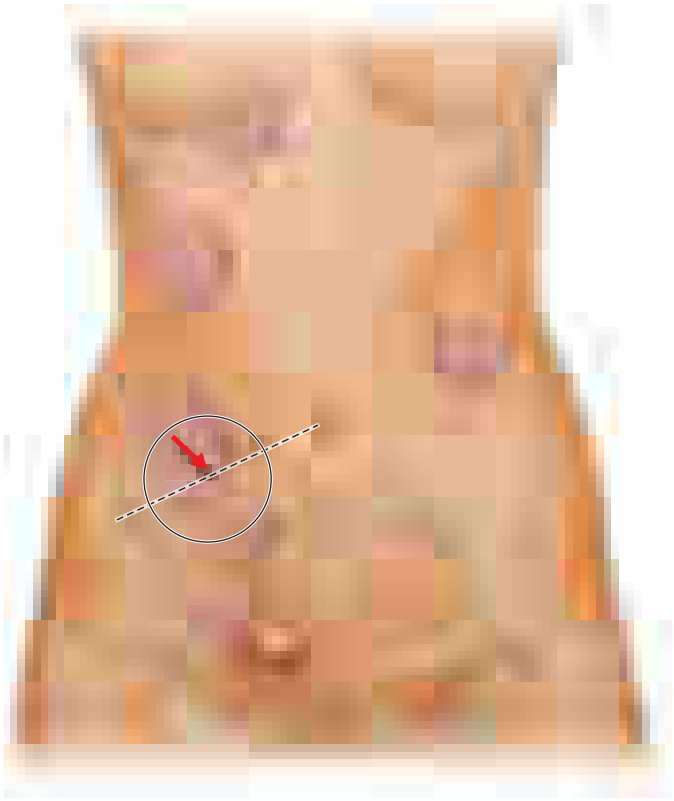


FIGURE 324-2 Locations of the appendix and cecum.

Arriving at the correct diagnosis is even more challenging when the appendix is not located in the right lower quadrant, in women of childbearing age, and in the very young or elderly. Because the differential diagnosis of appendicitis is so broad, often the key question to answer expeditiously is whether the patient has appendicitis or some other condition that requires immediate operative intervention. A major concern is that the likelihood of a delay in diagnosis is greater if the appendix is unusually positioned. All patients should undergo a rectal examination. An inflamed appendix located behind the cecum or below the pelvic brim may prompt very little tenderness of the anterior abdominal wall.

Patients with pelvic appendicitis are more likely to present with dysuria, urinary frequency, diarrhea, or tenesmus. They may only experience pain in the suprapubic region on palpation or on rectal or pelvic examination. A pelvic examination in women is mandatory to rule out conditions affecting urogynecologic organs that can cause abdominal pain and mimic appendicitis such as pelvic inflammatory disease, ectopic pregnancy, and ovarian torsion. Interest in the ability of various clinical scoring systems to predict appendicitis or the need for imaging studies continues. However, none of the currently available decision tools yet appear to be able to circumvene or obviate the need for expert clinical opinion. The relative frequencies of some presenting signs are displayed in Table 324-3.

TABLE 324-2 Relative Frequency of Common Presenting Symptoms

SYMPTOMS	FREQUENCY
Abdominal pain	>95%
Anorexia	>70%
Constipation	4–16%
Diarrhea	4–16%
Fever	10–20%
Migration of pain to right lower quadrant	50–60%
Nausea	>65%
Vomiting	50–75%

TABLE 324-3 Relative Frequency of Some Presenting Signs

SIGNS	FREQUENCY (%)
Abdominal tenderness	>95%
Right lower quadrant tenderness	>90%
Rebound tenderness	30–70%
Rectal tenderness	30–40%
Cervical motion tenderness	30%
Rigidity	~10%
Psoas sign	3–5%
Obturator sign	5–10%
Rovsing's sign	5%
Palpable mass	<5%

Patients with simple appendicitis normally only appear mildly ill with a pulse and temperature that are usually only slightly above normal. The provider should be concerned about other disease processes beside appendicitis or the presence of complications such as perforation, phlegmon, or abscess formation if the temperature is >38.3°C (~101°F) and if there are rigors.

Patients with appendicitis will be found to lie quite still to avoid peritoneal irritation caused by movement, and some will report discomfort caused by a bumpy car ride on the way to the hospital or clinic, coughing, sneezing, or other actions that replicate a Valsalva maneuver. The entire abdomen should be examined systematically starting in an area where the patient does not report discomfort if possible. Classically, maximal tenderness is identified in the right lower quadrant at or near McBurney's point, which is located approximately one-third of the way along a line originating at the anterior iliac spine and running to the umbilicus. Gentle pressure in the left lower quadrant may elicit pain in the right lower quadrant if the appendix is located there. This is Rovsing's sign (Table 324-4). Evidence of parietal peritoneal irritation is often best elicited by gentle abdominal percussion, jiggling the patient's gurney or bed, or mildly bumping the feet.

Atypical presentation and pain patterns are common, especially in the very old or the very young. Diagnosing appendicitis in children can be especially challenging because they tend to respond so dramatically to stimulation and obtaining an accurate history may be difficult. In addition, it is important to remember that the smaller omentum found in children may be less likely to wall off an appendiceal perforation. Observing the child in a quiet surrounding may be helpful.

Signs and symptoms of appendicitis can be subtle in the elderly who may not react as vigorously to appendicitis as younger people. Pain, if noticed, may be minimal and have originated in the right lower quadrant or, otherwise, where the appendix is located. It may never have been noticed to be intermittent, or there may only be significant discomfort with deep palpation. Nausea, anorexia, and emesis may be the predominant complaints. The rare patient may even present with signs and symptoms of distal bowel obstruction secondary to appendiceal inflammation and phlegmon or abscess formation.

LABORATORY TESTING

Laboratory testing does not identify patients with appendicitis. The white blood cell count is only mildly to moderately elevated in ~70% of patients

TABLE 324-4 Classic Signs of Appendicitis in Patients with Abdominal Pain

MANEUVER	FINDINGS
Rovsing's sign	Palpating in the left lower quadrant causes pain in the right lower quadrant
Obturator sign	Internal rotation of the hip causes pain, suggesting the possibility of an inflamed appendix located in the pelvis
Iliopsoas sign	Extending the right hip causes pain along posterolateral back and hip, suggesting retrocecal appendicitis

with simple appendicitis (with a leukocytosis of 10,000–18,000 cells/μL). A "left shift" toward immature polymorphonuclear leukocytes is present in >95% of cases. A sickle cell preparation may be prudent to obtain in those of African, Spanish, Mediterranean, or Indian ancestry. Serum amylase and lipase levels should be measured.

Urinalysis is indicated to help exclude genitourinary conditions that may mimic acute appendicitis, but a few red or white blood cells may be present as a nonspecific finding. However, an inflamed appendix that abuts the ureter or bladder may cause sterile pyuria or hematuria. Every woman of childbearing age should have a pregnancy test. Cervical cultures are indicated if pelvic inflammatory disease is suspected. Anemia and guaiac-positive stools should raise concern about the presence of other diseases or complications such as cancer.

IMAGING

Plain films of the abdomen are rarely helpful and so are not routinely obtained unless the clinician is worried about other conditions such as intestinal obstruction, perforated viscus, or ureterolithiasis. Less than 5% of patients will present with an opaque fecalith in the right lower quadrant. The presence of a fecalith is not diagnostic of appendicitis, although its presence in an appropriate location where the patient complains of pain is suggestive and is associated with a greater likelihood of complications.

The effectiveness of ultrasonography as a tool to diagnosis appendicitis is highly operator dependent. Even in very skilled hands, the appendix may not be visualized. Its overall sensitivity is 0.86, with a specificity of 0.81. Ultrasonography, especially intravaginal techniques, appears to be most useful for identifying pelvic pathology in women. Ultrasonographic findings suggesting the presence of appendicitis include wall thickening, an increased appendiceal diameter, and the presence of free fluid. Current practice in many institutions is to first perform ultrasonography and progress to other imaging studies only if the findings are equivocal.

The sensitivity and specificity of CT are at least 0.94 and 0.95, respectively. Thus, CT imaging, given its high negative predictive value, may be helpful if the diagnosis is in doubt, although studies performed early in the course of disease may not have any typical radiographic findings. In patients where the diagnosis is uncertain, delaying operation at the time of presentation to obtain CT does not appear to increase the risk of perforation. CT scanning is a superior method for assessing the severity of acute appendicitis in the absence of peritoneal findings indicative of perforation, abscess, or suspicion of an associated malignancy.

Suggestive findings on CT examination include dilatation >6 mm with wall thickening, a lumen that does not fill with enteric contrast, and fatty tissue stranding or air surrounding the appendix, which suggests inflammation **(Figs. 324-3 and 324-4)**. The presence of luminal air or contrast is not consistent with a diagnosis of appendicitis. Furthermore, nonvisualization of the appendix is a nonspecific finding that should not be used to rule out the presence of appendiceal or periappendiceal inflammation.

SPECIAL PATIENT POPULATIONS

Appendicitis in the most common extrauterine general surgical emergency observed during pregnancy. Early symptoms of appendicitis such as nausea and anorexia may be overlooked. Diagnosing appendicitis in pregnant patients may be especially difficult because as the uterus enlarges the appendix may be pushed higher along the right flank even to the right upper quadrant or because the gravid uterus may obscure typical physical findings. Ultrasonography may facilitate early diagnosis. A high index of suspicion is required because of the effects of unrecognized and untreated appendicitis on the fetus. For example, the fetal mortality rate is four times greater (from 5 to 20%) in patients with perforation.

Immunocompromised patients may present with only mild tenderness and may have many other disease processes in their differential diagnosis, including atypical infections from mycobacteria, *Cytomegalovirus*, or other fungi. Enterocolitis is a concern and may be present in patients who present with abdominal pain, fever, and neutropenia due to chemotherapy. CT imaging may be very helpful, although it is important not to be overly cautious and delay operative intervention for those patients who are believed to have appendicitis.

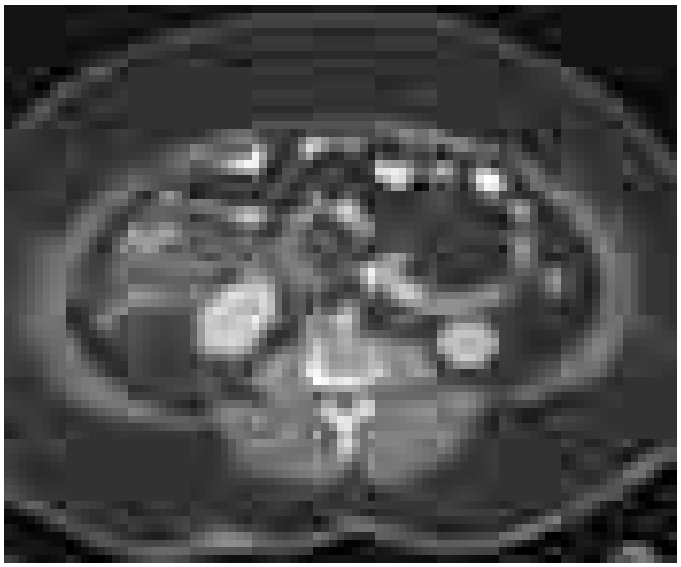


FIGURE 324-3 Computed tomography with oral and intravenous contrast of acute appendicitis. There is thickening of the wall of the appendix and periappendiceal stranding (*arrow*).

TREATMENT

Acute Appendicitis

In the absence of contraindications, most patients who have strongly suggestive medical histories and physical examinations with supportive laboratory findings are candidates for appendectomy. In many instances, imaging studies are not required but are often obtained before surgical consultation is requested. Certainly, imaging and further study is appropriate in patients whose evaluations are suggestive but not convincing.

CT may accurately indicate the presence of appendicitis or other intraabdominal processes that warrant intervention. Whenever the diagnosis is uncertain, it is prudent to observe the patient and repeat the abdominal examination over 6–8 h. Any evidence of progression

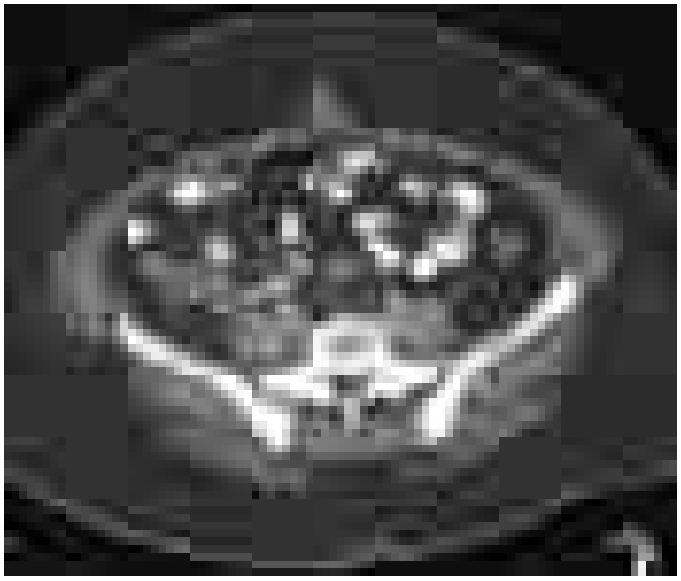


FIGURE 324-4 Appendiceal fecalith (*arrow*).

is an indication for operation. Narcotics can be given to patients with severe discomfort.

All patients should be fully prepared for surgery and have any fluid and electrolyte abnormalities corrected. Either laparoscopic or open appendectomy is a satisfactory choice for patients with uncomplicated appendicitis though most procedures are performed in a minimally invasive fashion. Management of those who present with a mass representing a phlegmon or abscess can be more difficult. Such patients are best served by treatment with broad-spectrum antibiotics, drainage if there is an abscess >3 cm in diameter, and parenteral fluids and bowel rest if they appear to respond to conservative management. The appendix can then be more safely removed 6–12 weeks later when inflammation has diminished.

Laparoscopic appendectomy now accounts for the majority of all appendectomies performed in Western cultures and is associated with less postoperative pain, shorter lengths of stay, faster return to normal activity and likely fewer superficial wound complications—although the risk of intraabdominal abscess formation may be higher.

A laparoscopic approach may also be useful when the exact diagnosis is uncertain. A laparoscopic approach may also facilitate exposure in those who are very obese. Absent complications, most patients can be discharged within 24–40 h of operation. The most common postoperative complications are fever and leukocytosis. Continuation of these findings beyond 5 days should raise concern for the presence of an intraabdominal abscess. The mortality rate for uncomplicated, nonperforated appendicitis is 0.1–0.5%, which approximates the overall risk of general anesthesia. The mortality rate for perforated appendicitis or other complicated disease is much higher, ranging from 3% overall to a high as 15% in the elderly.

ACUTE PERITONITIS

Acute peritonitis, or inflammation of the visceral and parietal peritoneum, is most often but not always infectious in origin, resulting from perforation of a hollow viscus. This is called *secondary peritonitis*, as opposed to *primary* or *spontaneous peritonitis*, when a specific intraabdominal source cannot be identified. In either instance, the inflammation can be localized or diffuse.

ETIOLOGY

Infective organisms may contaminate the peritoneal cavity after spillage from a hollow viscus, because of a penetrating wound of the abdominal wall, or because of the introduction of a foreign object like a peritoneal dialysis catheter or port that becomes infected. Secondary peritonitis most commonly results from perforation of the appendix, colonic diverticuli, or the stomach and duodenum. It may also occur as a complication of bowel infarction or incarceration, cancer, inflammatory bowel disease, and intestinal obstruction or volvulus. Conditions that may cause secondary bacterial peritonitis and their mechanisms are listed in Table 324-5. Over 90% of the cases of primary or spontaneous bacterial peritonitis occur in patients with ascites or hypoproteinemia (<1 g/L).

Aseptic peritonitis is most commonly caused by the abnormal presence of physiologic fluids like gastric juice, bile, pancreatic enzymes, blood, or urine. It can also be caused by the effects of normally sterile foreign bodies like surgical sponges or instruments. More rarely, it occurs as a complication of systemic diseases like lupus erythematosus, porphyria, and familial Mediterranean fever. The chemical irritation caused by stomach acid and activated pancreatic enzymes is extreme and secondary bacterial infection may occur.

CLINICAL FEATURES

The cardinal signs and symptoms of peritonitis are acute, typically severe, abdominal pain with tenderness and fever. How the patient's complaints of pain are manifested depends on their overall physical health and whether the inflammation is diffuse or localized. Elderly and immunosuppressed patients may not respond as aggressively to the irritation. Diffuse, generalized peritonitis is most often recognized as diffuse abdominal tenderness with local guarding, rigidity, and other evidence of parietal peritoneal irritation. Physical findings may only be identified in a specific region of the abdomen if the intraperitoneal inflammatory process is limited or otherwise contained as may occur in patients with uncomplicated appendicitis or diverticulitis. Bowel sounds are usually absent to hypoactive.

Most patients present with tachycardia and signs of volume depletion with hypotension. Laboratory testing typically reveals a significant leukocytosis, and patients may be severely acidotic. Radiographic studies may show dilatation of the bowel and associated bowel wall edema. Free air, or other evidence of leakage, requires attention and could represent a surgical emergency. In stable patients in whom ascites is present, diagnostic paracentesis is indicated, where the fluid is tested for protein and lactate dehydrogenase and the cell count is measured.

TABLE 324-5 Conditions Leading to Secondary Bacterial Peritonitis

Bowel perforation	Perforation or leakage of other organs
Appendicitis trauma (blunt or penetrating)	Biliary leakage (e.g., after liver biopsy)
Anastomotic leakage	Cholecystitis
Adhesion	Intraperitoneal bleeding
Diverticulitis	Pancreatitis
Iatrogenic (including endoscopic perforation)	Salpingitis
Ingested foreign body	Traumatic or other rupture of urinary bladder
Inflammation	**Loss of peritoneal integrity**
Intussusception	Intraperitoneal chemotherapy
Neoplasms	Iatrogenic (e.g., postoperative foreign body)
Obstruction	Perinephric abscess
Peptic ulcer disease	Peritoneal dialysis or other indwelling devices
Strangulated hernia	Trauma
Vascular (including ischemia or embolus)	

THERAPY AND PROGNOSIS

Whereas mortality rates can be <10% for reasonably healthy patients with relatively uncomplicated, localized peritonitis, mortality rates >40% have been reported for the elderly or immunocompromised. Successful treatment depends on correcting any electrolyte abnormalities, restoration of fluid volume and stabilization of the cardiovascular system, appropriate antibiotic therapy, and surgical correction of any underlying abnormalities.

Acknowledgment

The wisdom and expertise of Dr. William Silen is gratefully acknowledged in this updated chapter on acute appendicitis and peritonitis.

FURTHER READING

Andersson RE: Short-term complications and long-term morbidity of laparoscopic and open appendicectomy in a national cohort. Br J Surg 101:1135, 2014.

Buckius MT et al: Changing epidemiology of acute appendicitis in the United States: Study period 1993–2008. J Surg Res 175:185, 2012.

Drake FT et al: Time to appendectomy and risk of perforation in acute appendicitis. JAMA Surg 149:837, 2014.

Flum DR: Acute appendicitis—appendectomy of the "antibiotics first" strategy. N Engl J Med 372:1937, 2015.

Ohle R et al: The Alvarado score for predicting acute appendicitis: A systematic review. BMC Med 9:139, 2011.

Salminen P et al: Antibiotic therapy versus appendectomy for treatment of uncomplicated acute appendicitis: The APPAC Randomized Clinical Trial. JAMA 313:2340, 2015.

Vons C et al: Amoxicillin plus clavulanic acid versus appendicectomy for treatment of acute uncomplicated appendicitis: An open-label, non-inferiority, randomised controlled trial. Lancet 377:1573, 2011.

325 Nutrient Requirements and Dietary Assessment

Johanna Dwyer

Nutrients are substances that are not synthesized in sufficient amounts in the body and therefore must be supplied by the diet. Nutrient requirements for groups of healthy persons have been determined experimentally. The absence of essential nutrients leads to growth impairment, organ dysfunction, and failure to maintain nitrogen balance or adequate status of protein and other nutrients. For good health, we require energy-providing nutrients (protein, fat, and carbohydrate), vitamins, minerals, and water. Requirements for organic nutrients include 9 essential amino acids, several fatty acids, glucose, 4 fat-soluble vitamins, 10 water-soluble vitamins, dietary fiber, and choline. Several inorganic substances, including 4 minerals, 7 trace minerals, 3 electrolytes, and the ultratrace elements, must also be supplied by diet.

The amounts of essential nutrients required by individuals differ by their age and physiologic state. Conditionally essential nutrients are not required in the diet but must be supplied to certain individuals who do not synthesize them in adequate amounts, such as those with genetic defects, those with pathologies such as infection, disease or trauma with nutritional implications, and developmentally immature infants. For example, inositol, taurine, arginine, and glutamine may be needed by premature infants. Many other organic and inorganic compounds that are present in foods, such as pesticides, lead, phytochemicals, zoochemicals, and microbial products may also have health effects.

ESSENTIAL NUTRIENT REQUIREMENTS

Energy For weight to remain stable, energy intake must match energy output. The major components of energy output are resting energy expenditure (REE) and physical activity; minor components include the energy cost of metabolizing food (thermic effect of food, or specific dynamic action) and shivering thermogenesis (e.g., cold-induced thermogenesis). The average energy intake is ~2600 kcal/d for American men and ~1800 kcal/d for American women, although these estimates vary with body size and activity level. Formulas for roughly estimating REE are useful in assessing the energy needs of an individual whose weight is stable. Thus, for males, REE = 900 + 10*m*, and for females, REE = 700 + 7*m*, where *m* is mass in kilograms. The calculated REE is then adjusted for physical activity level by multiplying by 1.2 for sedentary, 1.4 for moderately active, or 1.8 for very active individuals. The final figure, the estimated energy requirement (EER), provides an approximation of total caloric needs in a state of energy balance for a person of a certain age, sex, weight, height, and physical activity level. **For further discussion of energy balance in health and disease, see Chap. 327.**

Protein Dietary protein consists of both essential and nonessential amino acids that are required for protein synthesis. The nine essential amino acids are histidine, isoleucine, leucine, lysine, methionine/cystine, phenylalanine/tyrosine, threonine, tryptophan, and valine. Certain amino acids, such as alanine, can also be used for energy and gluconeogenesis. When energy intake is inadequate, protein intake must be increased, because ingested amino acids are diverted into pathways of glucose synthesis and oxidation. In extreme energy deprivation, protein-calorie malnutrition may ensue **(Chap. 327)**.

For adults, the recommended dietary allowance (RDA) for protein is ~0.8 g/kg desirable body mass per day, assuming that energy needs are met and that the protein is of relatively high biologic value. Current recommendations for a healthy diet call for at least 10–14% of calories from protein. Most American diets provide at least those amounts. Biological value tends to be highest for animal proteins, followed by proteins from legumes (beans), cereals (rice, wheat, corn), and roots. Combinations of plant proteins that complement one another in their essential amino acid profiles or combinations of animal and plant proteins can increase biological value and lower total protein intakes necessary to meet requirements. In healthy people with adequate diets, the timing of protein intake over the course of the day has little effect.

Protein needs increase during growth, pregnancy, lactation, and rehabilitation after injury or malnutrition. Tolerance to dietary protein is decreased in renal insufficiency (with consequent uremia) and in liver failure. Usual protein intakes can precipitate encephalopathy in patients with cirrhosis of the liver.

Fat and Carbohydrate Fats are a concentrated source of energy and constitute, on average, 34% of calories in U.S. diets. However, for optimal health, fat intake should total no more than 30% of calories. Saturated fat and trans fat should be limited to <10% of calories and polyunsaturated fats to <10% of calories, with monounsaturated fats accounting for the remainder of fat intake. At least 45–55% of total calories should be derived from carbohydrates. The brain requires ~100 g of glucose per day for fuel; other tissues use about 50 g/d. Some tissues (e.g., brain and red blood cells) rely on glucose supplied either exogenously or from muscle proteolysis. Over time, during hypocaloric states, adaptations in carbohydrate needs are possible. Like fat (9 kcal/g), carbohydrate (4 kcal/g), and protein (4 kcal/g), alcohol (ethanol) provides energy (7 kcal/g). However, it is not a nutrient.

Water For adults, 1–1.5 mL of water per kilocalorie of energy expenditure is sufficient under usual conditions to allow for normal variations in physical activity, sweating, and solute load of the diet. Water losses include 50–100 mL/d in the feces; 500–1000 mL/d by evaporation or exhalation; and, depending on the renal solute load, ≥1000 mL/d in the urine. If external losses increase, intakes must increase accordingly to avoid underhydration. Fever increases water losses by ~200 mL/d per °C; diarrheal losses vary but may be as great as 5 L/d in severe diarrhea. Heavy sweating, vigorous exercise, and vomiting also increase water losses. When renal function is normal and solute intakes are adequate, the kidneys can adjust to increased water intake by excreting up to 18 L of excess water per day **(Chap. 374)**. However, obligatory urine outputs can compromise hydration status when there is inadequate water intake or when losses increase in disease or kidney damage.

Infants have high requirements for water because of their large surface area to volume ratios, their inability to communicate their thirst, and the limited capacity of the immature kidney to handle high renal solute loads. Increased water needs during pregnancy are ~30 mL/d. During lactation, milk production increases daily water requirements so that ~1000 mL of additional water is needed, or 1 mL for each milliliter of milk produced. Special attention must also be paid to the water needs of the elderly, who have reduced total body water and blunted thirst sensation and are more likely to be taking medications such as diuretics.

Other Nutrients **See Chap. 326 for detailed descriptions of vitamins and minerals.**

DIETARY REFERENCE INTAKES AND RDAS

Fortunately, human life and well-being can be maintained within a fairly wide range with most nutrient intakes. However, the capacity for adaptation is not infinite—too much, as well as too little, intake of a nutrient can have adverse effects or alter the health benefits conferred by another nutrient. Therefore, benchmark recommendations regarding nutrient intakes have been developed to guide clinical practice. These quantitative estimates of nutrient intakes are collectively referred to as the *dietary reference intakes* (DRIs). The DRIs have supplanted the RDAs—the single reference values used in the United States until the early 1990s. DRIs include an *estimated average requirement* (EAR) for nutrients as well as other reference values used for dietary planning:

the RDA, the *adequate intake* (AI), and the tolerable *upper level* (UL). The DRIs also include acceptable macronutrient distribution ranges (AMDRs) for protein, fat, and carbohydrate. The current DRIs for vitamins and elements are provided in **Tables 325-1 and 325-2**, respectively. **Table 325-3** provides DRIs for water and macronutrients. **EERs are discussed in Chap. 327 on energy balance in health and disease.**

Estimated Average Requirement When florid manifestations of the classic dietary-deficiency diseases such as rickets (deficiency of vitamin D and calcium), scurvy (deficiency of vitamin C), xerophthalmia (deficiency of vitamin A), and protein-calorie malnutrition were common, nutrient adequacy was inferred from the absence of their clinical deficiency signs. Later, biochemical and other changes were used that became evident long before the deficiency was clinically apparent. Consequently, criteria of adequacy are now based on biological markers when they are available. Priority is given to sensitive biochemical, physiologic, or behavioral tests that reflect early changes in regulatory processes; maintenance of body stores of nutrients; or, if available, the amount of a nutrient that minimizes the risk of chronic degenerative disease. Current efforts focus on this last variable, but relevant markers often are not available, and long time lags between intake and disease outcomes further complicate the picture.

The types of evidence and criteria used to establish nutrient requirements vary by nutrient, age, and physiologic group. The EAR is the amount of a nutrient estimated to be adequate for half of the healthy individuals of a specific age and sex. The EAR is not an effective estimate of nutrient adequacy in individuals because it is a median requirement for a group; 50% of individuals in a group fall below the requirement and 50% fall above it. Thus, a person with a usual intake at the EAR has a 50% risk of inadequate intake. For these reasons the other standards described below are more useful for clinical purposes.

Recommended Dietary Allowances The RDA, the nutrient intake goal for planning diets of individuals, is the average daily dietary intake level that meets the nutrient requirements of nearly all healthy persons of a specific sex, age, life stage, or physiologic condition (e.g., pregnancy or lactation). It is defined statistically as two standard deviations above the EAR to ensure that the needs of any given individual are met. The online tool at *http://fnic.nal.usda.gov/interactiveDRI/* allows health professionals to calculate individualized daily nutrient recommendations for dietary planning based on the DRIs. The RDAs are used to formulate food guides such as the U.S. Department of Agriculture (USDA) MyPlate Food Guide for individuals (*www.supertracker.usda.gov/default.aspx*), to create food-exchange lists for therapeutic diet planning, and as a standard for describing the nutritional content of foods and nutrient-containing dietary supplements on labels.

The risk of dietary inadequacy increases as one's intake falls below the RDA. However, the RDA is an overly generous criterion for evaluating nutrient adequacy. For example, by definition, the RDA exceeds the actual requirements of all but ~2–3% of the population. Therefore, many people whose intake fall below the RDA are still getting enough of the nutrient. On food labels, the nutrient content in a food is stated by weight or as a percent of the daily value (DV), a variant of the RDA used on the nutrition facts panel that, for an adult, represents the highest RDA for an adult consuming 2000 kcal.

Adequate Intake It is not possible to set an RDA for some nutrients that lack an established EAR. In this circumstance, the AI is based on observed or experimentally determined approximations of nutrient intakes in healthy people. In the DRIs, AIs rather than RDAs are proposed for nutrients consumed by infants (up to age 1 year) as well as for chromium, fluoride, manganese, sodium, potassium, pantothenic acid, biotin, choline, and water consumed by persons of all ages. Vitamin D and calcium recommendations were recently revised, and more precise estimates are now available.

Tolerable Upper Levels of Nutrient Intake Healthy individuals gain no established benefit from consuming nutrient levels above the RDA or AI. In fact, excessive nutrient intake can disturb body functions and cause acute, progressive, or permanent disabilities. The tolerable UL is the highest level of chronic nutrient intake (usually daily) that is unlikely to pose a risk of adverse health effects for most of the population. Data on the adverse effects of large amounts of many nutrients are unavailable or too limited to establish a UL. Therefore, the lack of a UL does *not* mean that the risk of adverse effects from high intake is nonexistent. Nutrient levels in commonly eaten foods rarely exceed the UL. However, very highly fortified foods and dietary supplements provide more concentrated amounts of nutrients per serving and thus pose a potential risk of toxicity. Nutrient dietary supplements are labeled with Supplement Facts that express the amount of nutrient in absolute units or as the percentage of the DV provided per recommended serving size. Total nutrient consumption, including that in foods, supplements, and over-the-counter medications (e.g., antacids), should not exceed RDA levels.

Acceptable Macronutrient Distribution Ranges The AMDRs are not experimentally determined; rather they are rough ranges for energy-providing macronutrient intakes (protein, carbohydrate, and fat) that the National Academy of Medicine's (formerly Institute of Medicine, IOM) Food and Nutrition Board considers to be healthful. These ranges are 10–35% of calories for protein, 20–35% of calories for fat, and 45–65% of calories for carbohydrate. Alcohol, which also provides energy, is not a nutrient; therefore, no recommendations are not provided.

FACTORS ALTERING NUTRIENT NEEDS

The DRIs are affected by age, sex, growth rate, pregnancy, lactation, physical activity level, concomitant diseases, drugs, and dietary composition. If requirements for nutrient sufficiency are close to intake levels indicating excess of a nutrient, dietary planning is difficult.

Physiologic Factors Growth, strenuous physical activity, pregnancy, and lactation all increase needs for energy and several essential nutrients. Energy needs rise during pregnancy due to fetal growth demands and increased energy required for milk production during lactation. Energy needs decrease with loss of lean body mass, the major determinant of REE. The energy needs of older persons, especially those aged >70 years, tend to be lower than those of younger persons because lean tissue, physical activity, and health often decline with age.

Dietary Composition Dietary composition affects the biological availability and use of nutrients. For example, iron absorption may be impaired by large amounts of calcium or lead; likewise, non-heme iron uptake may be impaired by a lack of ascorbic acid and amino acids in the meal. Bodily protein may be decreased when essential amino acids are not present in sufficient amounts—a rare scenario in U.S. diets. Animal foods, such as milk, eggs, and meat, have high biologic values, with most of the needed amino acids present in adequate amounts. Plant proteins in corn (maize), soy, rice, and wheat have lower biological values and must be combined with other plant or animal proteins or fortified with the amino acids that are deficient to achieve optimal use by the body.

Route of Intake The RDAs apply only to oral intakes. When nutrients are administered parenterally, similar values can sometimes be used for amino acids, glucose (carbohydrate), fats, sodium, chloride, potassium, and most vitamins because their intestinal absorption rate is nearly 100%. However, the oral bioavailability of most mineral elements may be only half that obtained by parenteral administration. For some nutrients that are not readily stored in the body or that cannot be stored in large amounts, timing of administration may also be important. For example, amino acids cannot be used for protein synthesis if they are not supplied together; instead, they will be used for energy production, although in healthy individuals eating adequate diets, the distribution of protein intake over the course of the day has little effect on health.

Disease Dietary deficiency diseases include protein-calorie malnutrition, iron-deficiency anemia, goiter (due to iodine deficiency), rickets and osteomalacia (vitamin D deficiency), and xeropthalmia (vitamin A deficiency), megaloblastic anemia (vitamin B_{12} or folic acid deficiency), scurvy (vitamin C/ascorbic acid deficiency), beriberi (thiamin deficiency),

TABLE 325-1 Dietary Reference Intakes (DRIs): Recommended Dietary Allowances and Adequate Intakes for Vitamins

LIFE-STAGE GROUP	VITAMIN A (μg/d)[a]	VITAMIN C (mg/d)	VITAMIN D (μg/d)[b,c]	VITAMIN E (mg/d)[d]	VITAMIN K (μg/d)	THIAMIN (mg/d)	RIBOFLAVIN (mg/d)	NIACIN (mg/d)[e]	VITAMIN B_6 (mg/d)	FOLATE (μg/d)[f]	VITAMIN B_{12} (μg/d)	PANTOTHENIC ACID (mg/d)	BIOTIN (μg/d)	CHOLINE (mg/d)[g]
Infants														
Birth to 6 mo	400*	40*	10	4*	2.0*	0.2*	0.3*	2*	0.1*	65*	0.4*	1.7*	5*	125*
6–12 mo	500*	50*	10	5*	2.5*	0.3*	0.4*	4*	0.3*	80*	0.5*	1.8*	6*	150*
Children														
1–3 y	**300**	**15**	**15**	**6**	30*	**0.5**	**0.5**	**6**	**0.5**	**150**	**0.9**	2*	8*	200*
4–8 y	**400**	**25**	**15**	**7**	55*	**0.6**	**0.6**	**8**	**0.6**	**200**	**1.2**	3*	12*	250*
Males														
9–13 y	**600**	**45**	**15**	**11**	60*	**0.9**	**0.9**	**12**	**1.0**	**300**	**1.8**	4*	20*	375*
14–18 y	**900**	**75**	**15**	**15**	75*	**1.2**	**1.3**	**16**	**1.3**	**400**	**2.4**	5*	25*	550*
19–30 y	**900**	**90**	**15**	**15**	120*	**1.2**	**1.3**	**16**	**1.3**	**400**	**2.4**	5*	30*	550*
31–50 y	**900**	**90**	**15**	**15**	120*	**1.2**	**1.3**	**16**	**1.3**	**400**	**2.4**	5*	30*	550*
51–70 y	**900**	**90**	**15**	**15**	120*	**1.2**	**1.3**	**16**	**1.7**	**400**	**2.4**[h]	5*	30*	550*
>70 y	**900**	**90**	**20**	**15**	120*	**1.2**	**1.3**	**16**	**1.7**	**400**	**2.4**[h]	5*	30*	550*
Females														
9–13 y	**600**	**45**	**15**	**11**	60*	**0.9**	**0.9**	**12**	**1.0**	**300**	**1.8**	4*	20*	375*
14–18 y	**700**	**65**	**15**	**15**	75*	**1.0**	**1.0**	**14**	**1.2**	**400**[i]	**2.4**	5*	25*	400*
19–30 y	**700**	**75**	**15**	**15**	90*	**1.1**	**1.1**	**14**	**1.3**	**400**[i]	**2.4**	5*	30*	425*
31–50 y	**700**	**75**	**15**	**15**	90*	**1.1**	**1.1**	**14**	**1.3**	**400**[i]	**2.4**	5*	30*	425*
51–70 y	**700**	**75**	**15**	**15**	90*	**1.1**	**1.1**	**14**	**1.5**	**400**	**2.4**[h]	5*	30*	425*
>70 y	**700**	**75**	**20**	**15**	90*	**1.1**	**1.1**	**14**	**1.5**	**400**	**2.4**[h]	5*	30*	425*
Pregnant Women														
14–18 y	**750**	**80**	**15**	**15**	75*	**1.4**	**1.4**	**18**	**1.9**	**600**[j]	**2.6**	6*	30*	450*
19–30 y	**770**	**85**	**15**	**15**	90*	**1.4**	**1.4**	**18**	**1.9**	**600**[j]	**2.6**	6*	30*	450*
31–50 y	**770**	**85**	**15**	**15**	90*	**1.4**	**1.4**	**18**	**1.9**	**600**[j]	**2.6**	6*	30*	450*
Lactating Women														
14–18 y	**1200**	**115**	**15**	**19**	75*	**1.4**	**1.6**	**17**	**2.0**	**500**	**2.8**	7*	35*	550*
19–30 y	**1300**	**120**	**15**	**19**	90*	**1.4**	**1.6**	**17**	**2.0**	**500**	**2.8**	7*	35*	550*
31–50 y	**1300**	**120**	**15**	**19**	90*	**1.4**	**1.6**	**17**	**2.0**	**500**	**2.8**	7*	35*	550*

Note: This table (taken from the DRI reports; see *www.nap.edu*) presents recommended dietary allowances (RDAs) in **bold type** and adequate intakes (AIs) in ordinary type followed by an asterisk (*). An RDA is the average daily dietary intake level sufficient to meet the nutrient requirements of nearly all healthy individuals (97–98%) in a group. The RDA is calculated from an estimated average requirement (EAR). If sufficient scientific evidence is not available to establish an EAR and thus to calculate an RDA, an AI is usually developed. For healthy breast-fed infants, an AI is the mean intake. The AI for other life-stage and sex-specific groups is believed to cover the needs of all healthy individuals in those groups, but lack of data or uncertainty in the data makes it impossible to specify with confidence the percentage of individuals covered by this intake.

[a]As retinol activity equivalents (RAEs). 1 RAE = 1 μg retinol, 12 μg β-carotene, 24 μg α-carotene, or 24 μg β-cryptoxanthin. The RAE for dietary provitamin A carotenoids is twofold greater than the retinol equivalent (RE), whereas the RAE for preformed vitamin A is the same as the RE. [b]As cholecalciferol. 1 μg cholecalciferol = 40 IU vitamin D. [c]Under the assumption of minimal sunlight. [d]As α-tocopherol. α-Tocopherol includes *RRR*-α-tocopherol, the only form of α-tocopherol that occurs naturally in foods, and the *2R*-stereoisomeric forms of α-tocopherol (*RRR*-, *RSR*-, *RRS*-, and *RSS*-α-tocopherol) that occur in fortified foods and supplements. It does not include the *2S*-stereoisomeric forms of α-tocopherol (*SRR*-, *SSR*-, *SRS*-, and *SSS*-α-tocopherol) also found in fortified foods and supplements. [e]As niacin equivalents (NEs). 1 mg of niacin = 60 mg of tryptophan; 0–6 months = preformed niacin (not NE). [f]As dietary folate equivalents (DFEs). 1 DFE = 1 μg food folate = 0.6 μg of folic acid from fortified food or as a supplement consumed with food = 0.5 μg of a supplement taken on an empty stomach. [g]Although AIs have been set for choline, there are few data to assess whether a dietary supply of choline is needed at all stages of the life cycle, and it may be that the choline requirement can be met by endogenous synthesis at some of these stages. [h]Because 10–30% of older people may malabsorb food-bound B12, it is advisable for those >50 years of age to meet their RDA mainly by consuming foods fortified with B_{12} or a supplement containing B_{12}. [i]In view of evidence linking inadequate folate intake with neural tube defects in the fetus, it is recommended that all women capable of becoming pregnant consume 400 μg of folate from supplements or fortified foods in addition to intake of food folate from a varied diet. [j]It is assumed that women will continue consuming 400 μg from supplements or fortified food until their pregnancy is confirmed and they enter prenatal care, which ordinarily occurs after the end of the periconceptional period—the critical time for formation of the neural tube.

Source: Food and Nutrition Board, Institute of Medicine, National Academies (*http://www.iom.edu/Activities/Nutrition/SummaryDRIs/DRI-Tables.aspx*).

TABLE 325-2 Dietary Reference Intakes (DRIs): Recommended Dietary Allowances and Adequate Intakes for Elements

LIFE-STAGE GROUP	CALCIUM (mg/d)	CHROMIUM (μg/d)	COPPER (μg/d)	FLUORIDE (mg/d)	IODINE (μg/d)	IRON (mg/d)	MAGNESIUM (mg/d)	MANGANESE (mg/d)	MOLYBDENUM (μg/d)	PHOSPHORUS (mg/d)	SELENIUM (μg/d)	ZINC (mg/d)	POTASSIUM (g/d)	SODIUM (g/d)	CHLORIDE (g/d)
Infants															
Birth to 6 mo	200*	0.2*	200*	0.01*	110*	0.27*	30*	0.003*	2*	100*	15*	2*	0.4*	0.12*	0.18*
6–12 mo	260*	5.5*	220*	0.5*	130*	**11**	75*	0.6*	3*	275*	20*	**3**	0.7*	0.37*	0.57*
Children															
1–3 y	**700**	11*	**340**	0.7*	**90**	**7**	**80**	1.2*	**17**	**460**	**20**	**3**	3.0*	1.0*	1.5*
4–8 y	**1000**	15*	**440**	1*	**90**	**10**	**130**	1.5*	**22**	**500**	**30**	**5**	3.8*	1.2*	1.9*
Males															
9–13 y	**1300**	25*	**700**	2*	**120**	**8**	**240**	1.9*	**34**	**1250**	**40**	**8**	4.5*	1.5*	2.3*
14–18 y	**1300**	35*	**890**	3*	**150**	**11**	**410**	2.2*	**43**	**1250**	**55**	**11**	4.7*	1.5*	2.3*
19–30 y	**1000**	35*	**900**	4*	**150**	**8**	**400**	2.3*	**45**	**700**	**55**	**11**	4.7*	1.5*	2.3*
31–50 y	**1000**	35*	**900**	4*	**150**	**8**	**420**	2.3*	**45**	**700**	**55**	**11**	4.7*	1.5*	2.3*
51–70 y	**1000**	30*	**900**	4*	**150**	**8**	**420**	2.3*	**45**	**700**	**55**	**11**	4.7*	1.3*	2.0*
>70 y	**1200**	30*	**900**	4*	**150**	**8**	**420**	2.3*	**45**	**700**	**55**	**11**	4.7*	1.2*	1.8*
Females															
9–13 y	**1300**	21*	**700**	2*	**120**	**8**	**240**	1.6*	**34**	**1250**	**40**	**8**	4.5*	1.5*	2.3*
14–18 y	**1300**	24*	**890**	3*	**150**	**15**	**360**	1.6*	**43**	**1250**	**55**	**9**	4.7*	1.5*	2.3*
19–30 y	**1000**	25*	**900**	3*	**150**	**18**	**310**	1.8*	**45**	**700**	**55**	**8**	4.7*	1.5*	2.3*
31–50 y	**1000**	25*	**900**	3*	**150**	**18**	320	1.8*	**45**	**700**	**55**	**8**	4.7*	1.5*	2.3*
51–70 y	**1200**	20*	**900**	3*	**150**	**8**	**320**	1.8*	**45**	**700**	**55**	**8**	4.7*	1.3*	2.0*
>70 y	**1200**	20*	**900**	3*	**150**	**8**	**320**	1.8*	**45**	**700**	**55**	**8**	4.7*	1.2*	1.8*
Pregnant Women															
14–18 y	**1300**	29*	**1000**	3*	**220**	**27**	**400**	2.0*	**50**	**1250**	**60**	**12**	4.7*	1.5*	2.3*
19–30 y	**1000**	30*	**1000**	3*	**220**	**27**	**350**	2.0*	**50**	**700**	**60**	**11**	4.7*	1.5*	2.3*
31–50 y	**1000**	30*	**1000**	3*	**220**	**27**	**360**	2.0*	**50**	**700**	**60**	**11**	4.7*	1.5*	2.3*
Lactating Women															
14–18 y	**1300**	44*	**1300**	3*	**290**	**10**	**360**	2.6*	**50**	**1250**	**70**	**13**	5.1*	1.5*	2.3*
19–30 y	**1000**	45*	**1300**	3*	**290**	**9**	**310**	2.6*	**50**	**700**	**70**	**12**	5.1*	1.5*	2.3*
31–50 y	**1000**	45*	**1300**	3*	**290**	**9**	**320**	2.6*	**50**	**700**	**70**	**12**	5.1*	1.5*	2.3*

Note: This table (taken from the DRI reports; see *www.nap.edu*) presents recommended dietary allowances (RDAs) in **bold type** and adequate intakes (AIs) in ordinary type followed by an asterisk (*). An RDA is the average daily dietary intake level sufficient to meet the nutrient requirements of nearly all healthy individuals (97–98%) in a group. The RDA is calculated from an estimated average requirement (EAR). If sufficient scientific evidence is not available to establish an EAR and thus to calculate an RDA, an AI is usually developed. For healthy breast-fed infants, an AI is the mean intake. The AI for other life-stage and sex-specific groups is believed to cover the needs of all healthy individuals in those groups, but lack of data or uncertainty in the data makes it impossible to specify with confidence the percentage of individuals covered by this intake.

Sources: Food and Nutrition Board, Institute of Medicine, National Academies (*http://www.iom.edu/Activities/Nutrition/SummaryDRIs/DRI-Tables.aspx*), based on: *Dietary Reference Intakes for Calcium, Phosphorus, Magnesium, Vitamin D, and Fluoride* (1997); *Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B_6, Folate, Vitamin B_{12}, Pantothenic Acid, Biotin, and Choline* (1998); *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids* (2000); and *Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc* (2001); *Dietary Reference Intakes for Water, Potassium, Sodium, Chloride, and Sulfate* (2005); and *Dietary Reference Intakes for Calcium and Vitamin D* (2011). These reports can be accessed via *www.nap.edu*.

TABLE 325-3 Dietary Reference Intakes (DRIs): Recommended Dietary Allowances and Adequate Intakes for Total Water and Macronutrients

LIFE-STAGE GROUP	TOTAL WATER[a] (L/d)	CARBOHYDRATE (g/d)	TOTAL FIBER (g/d)	FAT (g/d)	LINOLEIC ACID (g/d)	α-LINOLENIC ACID (g/d)	PROTEIN[b] (g/d)
Infants							
Birth to 6 mo	0.7*	60*	ND[c]	31*	4.4*	0.5*	9.1*
6–12 mo	0.8*	95*	ND	30*	4.6*	0.5*	**11.0**
Children							
1–3 y	1.3*	**130**	19*	ND	7*	0.7*	**13**
4–8 y	1.7*	**130**	25*	ND	10*	0.9*	**19**
Males							
9–13 y	2.4*	**130**	31*	ND	12*	1.2*	**34**
14–18 y	3.3*	**130**	38*	ND	16*	1.6*	**52**
19–30 y	3.7*	**130**	38*	ND	17*	1.6*	**56**
31–50 y	3.7*	**130**	38*	ND	17*	1.6*	**56**
51–70 y	3.7*	**130**	30*	ND	14*	1.6*	**56**
>70 y	3.7*	**130**	30*	ND	14*	1.6*	**56**
Females							
9–13 y	2.1*	**130**	26*	ND	10*	1.0*	**34**
14–18 y	2.3*	**130**	26*	ND	11*	1.1*	**46**
19–30 y	2.7*	**130**	25*	ND	12*	1.1*	**46**
31–50 y	2.7*	**130**	25*	ND	12*	1.1*	**46**
51–70 y	2.7*	**130**	21*	ND	11*	1.1*	**46**
>70 y	2.7*	**130**	21*	ND	11*	1.1*	**46**
Pregnant Women							
14–18 y	3.0*	**175**	28*	ND	13*	1.4*	**71**
19–30 y	3.0*	**175**	28*	ND	13*	1.4*	**71**
31–50 y	3.0*	**175**	28*	ND	13*	1.4*	**71**
Lactating Women							
14–18	3.8*	**210**	29*	ND	13*	1.3*	**71**
19–30 y	3.8*	**210**	29*	ND	13*	1.3*	**71**
31–50 y	3.8*	**210**	29*	ND	13*	1.3*	**71**

Note: This table (taken from the DRI reports; see *www.nap.edu*) presents recommended dietary allowances (RDAs) in **bold type** and adequate intakes (AIs) in ordinary type followed by an asterisk (*). An RDA is the average daily dietary intake level sufficient to meet the nutrient requirements of nearly all healthy individuals (97–98%) in a group. The RDA is calculated from an estimated average requirement (EAR). If sufficient scientific evidence is not available to establish an EAR and thus to calculate an RDA, an AI is usually developed. For healthy breast-fed infants, an AI is the mean intake. The AI for other life-stage and sex-specific groups is believed to cover the needs of all healthy individuals in those groups, but lack of data or uncertainty in the data make it impossible to specify with confidence the percentage of individuals covered by this intake.

[a]Total water includes all water contained in food, beverages, and drinking water. [b]Based on grams of protein per kilogram of body weight for the reference body weight (e.g., for adults: 0.8 g/kg body weight for the reference body weight). [c]Not determined.

Source: Food and Nutrition Board, Institute of Medicine, National Academies (*http://www.iom.edu/Activities/Nutrition/SummaryDRIs/DRI-Tables.aspx*), based on: *Dietary Reference Intakes for Energy, Carbohydrate, Fiber, Fat, Fatty Acids, Cholesterol, Protein, and Amino Acids* (2002/2005) and *Dietary Reference Intakes for Water, Potassium, Sodium, Chloride, and Sulfate* (2005). These reports can be accessed via *www.nap.edu*.

and pellagra (niacin and tryptophan deficiency) **(Chaps. 326 and 327)**. Each deficiency disease is characterized by imbalances at the cellular level between the supply of nutrients or energy and the body's nutritional needs for growth, maintenance, and other functions. Imbalances and excesses in nutrient intakes are recognized as risk factors for certain chronic degenerative diseases, such as saturated fat and cholesterol in coronary artery disease; sodium in hypertension; obesity in hormone-dependent cancers (endometrial and breast); and ethanol in alcoholism. Diet is only one of many risk factors because the etiology and pathogenesis of these disorders are multifactorial. Osteoporosis, for example, is associated with calcium deficiency, sometimes secondary to vitamin D deficiency, as well as with environment related risk factors (e.g., smoking, sedentary lifestyle), physiology (e.g., estrogen deficiency), genetic determinants (e.g., defects in collagen metabolism), and drug use (chronic steroid and aromatase inhibitors) **(Chap. 404)**.

DIETARY ASSESSMENT

Nutrition assessment in clinical situations is an iterative process that involves: (1) screening for malnutrition, (2) assessing the diet and other data to establish either the absence or the presence of malnutrition and its possible causes, (3) planning and implementing the most appropriate nutritional therapy, and (4) reassessing intakes to make sure that they have been consumed. Some disease states affect the bioavailability, requirements, use, or excretion of specific nutrients. In these circumstances, specific measurements of various nutrients or their biomarkers may be required to ensure adequate replacement **(Chap. 326)**.

Most health care facilities have nutrition-screening processes in place for identifying possible malnutrition after hospital admission. Nutritional screening is required by the Joint Commission, which accredits and certifies health care organizations in the United States. However, no universally recognized or validated standards exist. The factors that are usually assessed include abnormal weight for height or body mass index (e.g., BMI <19 or >25); reported weight change (involuntary loss or gain of >5 kg in the past 6 months) **(Chap. 43)**; diagnoses with known nutritional implications (e.g., metabolic disease, any disease affecting the gastrointestinal tract, alcoholism); present therapeutic dietary prescription; chronic poor appetite; presence of chewing and swallowing problems or major food intolerances; need for assistance with preparing or shopping for food, eating, or other aspects of self-care; and social isolation. The nutritional status of hospitalized patients should be reassessed periodically—at least once every week.

A more complete dietary assessment is indicated for patients who exhibit a high risk of or frank malnutrition on nutritional screening. The type of assessment varies with the clinical setting, the severity of the patient's illness, and the stability of the patient's condition.

Acute-Care Settings In acute-care settings, anorexia, various other diseases, test procedures, and medications can compromise dietary intake. Under such circumstances, the goal is to identify and avoid inadequate intake and to assure appropriate alimentation. Dietary assessment focuses on what patients are currently eating, whether or not they are able and willing to eat, and whether or not they experience any problems with eating. Dietary intake assessment is based on information from observed intakes; medical records; history; clinical examination; and anthropometric, biochemical, and functional status evaluations. The objective is to gather enough information to establish the likelihood of malnutrition due to poor dietary intake or other causes in order to assess whether nutritional therapy is indicated **(Chap. 328)**.

Simple observations may suffice to suggest inadequate oral intake. These include dietitians' and nurses' notes; observation of a patient's frequent refusal to eat or the amount of food eaten on trays; the frequent performance of tests and procedures that are likely to cause meals to be skipped; adherence to nutritionally inadequate diet orders (e.g., clear liquids or full liquids) for more than a few days; the occurrence of fever, gastrointestinal distress, vomiting, diarrhea, or a comatose state; and the presence of diseases or use of treatments that involve any part of the alimentary tract. Acutely ill patients with diet-related diseases such as diabetes need assessment because an inappropriate diet may exacerbate these conditions and adversely affect other therapies. Abnormal biochemical values (serum albumin levels <35 g/L [<3.5 mg/dL]; serum cholesterol levels <3.9 mmol/L [<150 mg/dL]) are nonspecific but may indicate a need for further nutritional assessment.

Most therapeutic diets offered in hospitals are calculated to meet individual nutrient requirements and the RDA *if they are eaten*. Exceptions include clear liquids, some full-liquid diets, and test diets (such as those adhered to in preparation for gastrointestinal procedures), which are inadequate for several nutrients and should not be used, if possible, for more than 24 h. However, because as much as half of the food served to hospitalized patients is not eaten, it cannot be assumed that the intakes of hospitalized patients are adequate. Dietary assessment should compare how much and what kinds of food the patient has consumed with the diet that has been provided. Major deviations in intakes of energy, protein, fluids, or other nutrients of special concern for the patient's illness should be noted and corrected, especially for long-staying patients.

Nutritional monitoring is especially important for patients who are very ill and who have extended lengths of hospital stay. Patients who are fed by enteral and parenteral routes also require special nutritional assessment and monitoring by physicians and/or dietitians with certification in nutritional support **(Chap. 328)**.

Ambulatory Settings The aim of dietary assessment in the outpatient setting is to determine whether or not the patient's usual diet is a health risk in itself or if it contributes to existing chronic disease-related problems. Dietary assessment also provides the basis for planning a diet that fulfills therapeutic goals while ensuring patient adherence. The outpatient's dietary assessment should review the adequacy of present and usual food intakes, including vitamin and mineral supplements, oral nutritional supplements, medical foods, other dietary supplements, medications, and alcohol, because all of these may affect the patient's nutritional status. The assessment should focus on the dietary constituents that are most likely to be involved or compromised by a specific diagnosis as well as on any comorbidities that are present. More than one day's intake should be reviewed to provide a better representation of the usual diet, upon which personalized dietary recommendations can be based.

There are many ways to assess the adequacy of a patient's habitual diet. These include use of a food guide, a food-exchange list, a diet history, or a food-frequency questionnaire. A commonly used food guide for healthy persons is the USDA's Choose My Plate, which is useful as a rough guide for avoiding inadequate intakes of essential nutrients as well as likely excesses in the amounts of fat (especially saturated and trans fats), sodium, sugar, and alcohol consumed **(Table 325-4)**. The Choose My Plate graphic emphasizes a balance between calories and nutritional needs, encouraging increased intake of fruits and vegetables, whole grains, and low-fat milk in conjunction with reduced intake of sodium and high-calorie sugary drinks. The Web version of the guide provides a calculator that tailors the number of servings suggested for healthy patients of different weights, sexes, ages, and life-cycle stages to help them to meet their needs while avoiding excess (*http://www.supertracker.usda.gov/default.aspx* and *www.ChooseMyPlate.gov*). Patients who follow ethnic or unusual dietary patterns may need extra instruction on how foods should be categorized and on the appropriate portion sizes that constitute a serving. The process of reviewing the guide with patients helps them transition to healthier dietary patterns and identifies food groups eaten in excess of recommendations or in insufficient quantities. For persons on therapeutic diets, assessment against food-exchange lists may be useful. These include, for example, American Diabetes Association food-exchange lists for diabetes and the Academy of Nutrition and Dietetics food-exchange lists for renal disease.

TABLE 325-4 Choose My Plate: A Guide to Individualized Dietary Planning

	EXAMPLES OF STANDARD PORTION SIZES AT INDICATED ENERGY LEVEL		
DIETARY FACTOR, UNIT OF MEASURE (ADVICE)	LOWER: 1600 kcal	MODERATE: 2200 kcal	HIGHER: 2800 kcal
Fruits, cups (Focus on fruits.)	1.5	2	2.5
Vegetables, cups (Vary vegetables.)	2	3	3.5
Grains, oz eq (Make at least half of grains whole.)[a]	5	7	10
Protein foods, oz eq (Go lean with protein.)[b]	5	6	7
Dairy, cups or oz[c] (Choose calcium-rich foods.)	3	3	3
"Empty" calories, kcal[d]	120	260	400
Sodium, mg	<2300 at all energy levels		
Physical activity, min	At least 150 min vigorous physical activity per week at all energy levels		

Note: Oils (formerly listed with portions of 5, 6, and 8 teaspoons for the lower, moderate, and higher energy levels, respectively) are no longer singled out in Choose My Plate, but rather are included in the empty calories/added sugar category with SOFAS (calories from solid fats and added sugars). The limit is the remaining number of calories in each food pattern above after intake of the recommended amounts of the nutrient-dense foods.

[a]For example, 1 serving equals 1 slice bread, 1 cup ready-to-eat cereal, or 0.5 cup cooked rice, pasta, or cooked cereal. [b]For example, 1 serving equals 1 oz lean meat, poultry, or fish; 1 egg; 1 tablespoon peanut butter; 0.25 cup cooked dry beans; or 0.5 oz nuts or seeds. [c]For example, 1 serving equals 1 cup milk or yogurt, 1.5 oz natural cheese, or 2 oz processed cheese. [d]Formerly called "discretionary calorie allowance." Portions are calculated as the number of calories remaining after all of the above allotments are accounted for.

Abbreviation: oz eq, ounce equivalent.

Source: Data from U.S. Department of Agriculture (*http://www.Choosemyplate.gov*).

NUTRITIONAL STATUS ASSESSMENT

Full nutritional status assessment is reserved for seriously ill patients and those at very high nutritional risk when the cause of malnutrition is still uncertain after the initial clinical evaluation and dietary assessment. It involves multiple dimensions, including documentation of dietary intake, anthropometric measurements, biochemical measurements of blood and urine, clinical examination, health history elicitation, and functional status evaluation. Therapeutic dietary prescriptions and menu plans for most diseases are available from most hospitals and from the Academy of Nutrition and Dietetics. **For further discussion of nutritional assessment, see Chap. 327.**

GLOBAL CONSIDERATIONS

The DRIs (e.g., the EAR, the UL, and energy needs) are estimates of physiologic requirements based on experimental evidence. Assuming that appropriate adjustments are made for age, sex, body size, and physical activity level, these estimates

should be applicable to individuals in most parts of the world. However, other values are not transportable. The AIs are based on customary and adequate intakes in U.S. and Canadian populations, which appear to be compatible with good health, rather than on a large body of direct experimental evidence. Similarly, the AMDRs represent expert opinion regarding the approximate intakes of energy-providing nutrients that are healthful in these North American populations. Thus these measures should be used with caution in other settings. Nutrient-based standards like the DRIs have also been developed by the World Health Organization/Food and Agricultural Organization of the United Nations and are available on the Web (*http://www.who.int/nutrition/topics/nutrecomm/en/index.html*). The European Food Safety Authority (EFSA) Panel on Dietetic Products, Nutrition and Allergies periodically publishes its recommendations in the EFSA's on-line *journal*. Other countries have promulgated similar recommendations. The different standards have many similarities in their basic concepts, definitions, and nutrient recommendation levels, but there are some differences from the DRIs as a result of the functional criteria chosen, environmental differences, the timeliness of the evidence reviewed, and expert judgment.

FURTHER READING

Brannon PM et al: Scanning for new evidene to prioritize updates to the Dietary Reference Intakes: Case studies for thiamin and phosphorus. Am J Clin Nur 104:1366, 2016.

Forster H et al: Personalised nutrition: The role of new dietary assessment methods. Proc Nutr Soc 75:96, 2016.

Gibson RS: *Principles of Nutritional Assessment*, 2nd ed. Oxford, Oxford University Press, 2005.

Lupton JR et al: Nutrient reference value: Non-communicable disease endpoints—A conference report. Eur J Nutr 55 SUPPL 1:S1–10, 2016.

Ocke MC: Evaluation of methodologies for assessing the overall diet: Dietary quality scales and dietary pattern analysis. Proc Nutr Soc 72:191, 2013.

Otten JJ et al: *Dietary Reference Intakes: The Essential Guide to Nutrient Requirements*. Washington, DC, National Academies Press, 2006.

Potischman N, Freudenheim JL: Biomarkers of nutritional exposure and nutritional status: An overview. J Nutr 133:873S, 2003.

Report of the Subcommittee on Interpretation and Uses of Dietary Reference Intakes and Upper Reference Levels of Nutrients, and the Steering Committtee on the Scientific Evaluation of Dietary Reference Intakes, Food and Nutrition Board: Dietary Reference Intakes: Applications in Dietary Assessment. Washington, DC, National Academies Press, 2008.

Willett WC: *Nutritonal Epidemiology*, 3rd ed. Oxford, Oxford University Press, 2012.

Yetley EA et al: Options for basing Dietary Reference Intakes (DRIs) on chronic disease endpoints report from a joint US-/Canadian-sponsored working group. Am J Clin Nutr 105:249S, 2017.

326 Vitamin and Trace Mineral Deficiency and Excess

Paolo M. Suter, Robert M. Russell

Vitamins are required constituents of the human diet since they are synthesized inadequately or not at all in the human body. Only small amounts of these substances are needed to carry out essential biochemical reactions (e.g., by acting as coenzymes or prosthetic groups). Overt vitamin or trace mineral deficiencies are rare in Western countries because of a plentiful, varied, and inexpensive food supply; food fortification; and use of supplements. However, multiple nutrient deficiencies may appear together in persons who are chronically ill or alcoholic. After gastric bypass surgery, patients are at high risk for multiple nutrient deficiencies. Moreover, subclinical vitamin and trace mineral deficiencies, as diagnosed by laboratory testing, are quite common in the normal population, especially in the geriatric age group. Conversely, because of the widespread use of nutrient supplements, nutrient toxicities are gaining pathophysiologic and clinical importance.

Victims of famine, emergency-affected and displaced populations, and refugees are at increased risk for protein-energy malnutrition and classic micronutrient deficiencies (vitamin A, iron, iodine) as well as for overt deficiencies in thiamine (beriberi), riboflavin, vitamin C (scurvy), and niacin (pellagra).

Body stores of vitamins and minerals vary tremendously. For example, stores of vitamins B_{12} and A are large, and an adult may not become deficient until ≥1 year after beginning to eat a deficient diet. However, folate and thiamine may become depleted within weeks among those eating a deficient diet. Therapeutic modalities can deplete essential nutrients from the body; for example, hemodialysis or diuretics remove water-soluble vitamins, which must be replaced by supplementation.

Vitamins and trace minerals play several roles in diseases: (1) Deficiencies of vitamins and minerals may be caused by disease states such as malabsorption; (2) either deficiency or excess of vitamins and minerals can cause disease in and of itself (e.g., vitamin A intoxication and liver disease); and (3) vitamins and minerals in high doses may be used as drugs (e.g., niacin for hypercholesterolemia). Since they are covered elsewhere, the hematologic-related vitamins and minerals **(Chaps. 93 and 95)** either are not considered or are considered only briefly in this chapter, as are the bone-related vitamins and minerals (vitamin D, calcium, phosphorus, magnesium; **Chap. 402**).

VITAMINS

See also Table 326-1 and Fig. 326-1.

THIAMINE (VITAMIN B_1)

Thiamine was the first B vitamin to be identified and therefore is referred to as vitamin B_1. Thiamine functions in the decarboxylation of α-ketoacids (e.g., pyruvate α-ketoglutarate) and branched-chain amino acids and thus is essential for energy generation. In addition, thiamine pyrophosphate acts as a coenzyme for a transketolase reaction that mediates the conversion of hexose and pentose phosphates. It has been postulated that thiamine plays a role in peripheral nerve conduction, although the exact chemical reactions underlying this function are not known.

Food Sources The median intake of thiamine in the United States from food alone is 2 mg/d. Primary food sources for thiamine include yeast, organ meat, pork, legumes, beef, whole grains, and nuts. Milled rice and grains contain little thiamine. Thiamine deficiency is therefore more common in cultures that rely heavily on a rice-based diet. Tea, coffee (regular and decaffeinated), raw fish, and shellfish contain thiaminases, which can destroy the vitamin. Thus, drinking large amounts of tea or coffee could theoretically lower thiamine body stores.

Deficiency Most dietary deficiency of thiamine worldwide is the result of poor dietary intake. In Western countries, the primary causes of thiamine deficiency are alcoholism and chronic illnesses such as cancer. Alcohol interferes directly with the absorption of thiamine and with the synthesis of thiamine pyrophosphate, and it increases urinary excretion. Thiamine should always be replenished when a patient with alcoholism is being refed, as carbohydrate repletion without adequate thiamine can precipitate acute thiamine deficiency with lactic acidosis. Other at-risk populations are women with prolonged hyperemesis gravidarum and anorexia, patients with overall poor nutritional status who are receiving parenteral glucose, patients who have had bariatric/metabolic surgery (*bariatric Wernicke*), and patients receiving chronic diuretic therapy (e.g., in hypertension or heart failure) due to increased urinary thiamine losses. Maternal thiamine deficiency can lead to infantile beriberi in breast-fed children. Thiamine deficiency could be an underlying factor in motor vehicle accidents and could be overlooked in the setting of head injury.

Thiamine deficiency in its early stage induces anorexia and nonspecific symptoms (e.g., irritability, decrease in short-term memory). Prolonged thiamine deficiency causes *beriberi*, which is classically

TABLE 326-1 Principal Clinical Findings of Vitamin Malnutrition

NUTRIENT	CLINICAL FINDING	DIETARY LEVEL PER DAY ASSOCIATED WITH OVERT DEFICIENCY IN ADULTS	CONTRIBUTING FACTORS TO DEFICIENCY
Thiamine	Beriberi: neuropathy, muscle weakness and wasting, cardiomegaly, edema, ophthalmoplegia, confabulation	<0.3 mg/1000 kcal	Alcoholism, chronic diuretic use, hyperemesis, thiaminases in food
Riboflavin	Magenta tongue, angular stomatitis, seborrhea, cheilosis, ocular symptoms, corneal vascularization	<0.4 mg	Alcoholism, individuals with poor diets and low intake of milk products
Niacin	Pellagra: pigmented rash of sun-exposed areas, bright red tongue, diarrhea, apathy, memory loss, disorientation	<9.0 niacin equivalents	Alcoholism, vitamin B_6 deficiency, riboflavin deficiency, tryptophan deficiency
Vitamin B_6	Seborrhea, glossitis, convulsions, neuropathy, depression, confusion, microcytic anemia	<0.2 mg	Alcoholism, isoniazid
Folate	Megaloblastic anemia, atrophic glossitis, depression, ↑ homocysteine	<100 μg/d	Alcoholism, sulfasalazine, pyrimethamine, triamterene
Vitamin B_{12}	Megaloblastic anemia, loss of vibratory and position sense, abnormal gait, dementia, impotence, loss of bladder and bowel control, ↑ homocysteine, ↑ methylmalonic acid	<1.0 μg/d	Gastric atrophy (pernicious anemia), terminal ileal disease, strict vegetarianism, acid-reducing drugs (e.g., H_2 blockers), metformin
Vitamin C	Scurvy: petechiae, ecchymosis, coiled hairs, inflamed and bleeding gums, joint effusion, poor wound healing, fatigue	<10 mg/d	Smoking, alcoholism
Vitamin A	Xerophthalmia, night blindness, Bitot's spots, follicular hyperkeratosis, impaired embryonic development, immune dysfunction	<300 μg/d	Fat malabsorption, infection, measles, alcoholism, protein-energy malnutrition
Vitamin D	Rickets: skeletal deformation, rachitic rosary, bowed legs; osteomalacia	<2.0 μg/d	Aging, lack of sunlight exposure, fat malabsorption, deeply pigmented skin
Vitamin E	Peripheral neuropathy, spinocerebellar ataxia, skeletal muscle atrophy, retinopathy	Not described unless underlying contributing factor is present	Occurs only with fat malabsorption or genetic abnormalities of vitamin E metabolism/transport
Vitamin K	Elevated prothrombin time, bleeding	<10 μg/d	Fat malabsorption, liver disease, antibiotic use

categorized as wet or dry although there is considerable overlap between the two categories. In either form of *beriberi*, patients may complain of pain and paresthesia. *Wet beriberi* presents primarily with cardiovascular symptoms that are due to impaired myocardial energy metabolism and dysautonomia; it can occur after 3 months of a thiamine-deficient diet. Patients present with an enlarged heart, tachycardia, high-output congestive heart failure, peripheral edema, and peripheral neuritis. Patients with *dry beriberi* present with a symmetric peripheral neuropathy of the motor and sensory systems, with diminished reflexes. The neuropathy affects the legs most markedly, and patients have difficulty rising from a squatting position.

Alcoholic patients with chronic thiamine deficiency also may have central nervous system (CNS) manifestations known as *Wernicke's encephalopathy*, which consists of horizontal nystagmus, ophthalmoplegia (due to weakness of one or more extraocular muscles), cerebellar ataxia, and mental impairment (Chap. 445). When there is an additional loss of memory and a confabulatory psychosis, the syndrome is known as *Wernicke-Korsakoff syndrome*. Despite the typical clinical picture and history, Wernicke-Korsakoff syndrome is underdiagnosed.

The laboratory diagnosis of thiamine deficiency usually is made by a functional enzymatic assay of transketolase activity measured before and after the addition of thiamine pyrophosphate. A >25% stimulation in response to the addition of thiamine pyrophosphate (i.e., an activity coefficient of 1.25) is interpreted as abnormal. Thiamine or the phosphorylated esters of thiamine in serum or blood also can be measured by high-performance liquid chromatography to detect deficiency.

TREATMENT

Thiamine Deficiency

In acute thiamine deficiency with either cardiovascular or neurologic signs, 200 mg of thiamine three times daily should be given intravenously until there is no further improvement in acute symptoms; oral thiamine (10 mg/d) should subsequently be given until recovery is complete. Cardiovascular and ophthalmoplegic improvement occurs within 24 h. Other manifestations gradually clear, although psychosis in Wernicke-Korsakoff syndrome may be permanent or may persist for several months. Other nutrient deficiencies should be corrected concomitantly.

Toxicity Although anaphylaxis has been reported after high intravenous doses of thiamine, no adverse effects have been recorded from either food or supplements at high doses. Thiamine supplements may be bought over the counter in doses of up to 50 mg/d.

RIBOFLAVIN (VITAMIN B_2)

Riboflavin is important for the metabolism of fat, carbohydrate, and protein, acting as a respiratory coenzyme and an electron donor. Enzymes that contain flavin adenine dinucleotide (FAD) or flavin mononucleotide (FMN) as prosthetic groups are known as *flavoenzymes* (e.g., succinic acid dehydrogenase, monoamine oxidase, glutathione reductase). FAD is a cofactor for methyltetrahydrofolate reductase and therefore modulates homocysteine metabolism. The vitamin also plays a role in drug and steroid metabolism, including detoxification reactions.

Although much is known about the chemical and enzymatic reactions of riboflavin, the clinical manifestations of riboflavin deficiency are nonspecific and are similar to those of other deficiencies of B vitamins. Riboflavin deficiency is manifested principally by lesions of the mucocutaneous surfaces of the mouth and skin. In addition, corneal vascularization, anemia, and personality changes have been described with riboflavin deficiency.

Deficiency and Excess Riboflavin deficiency almost always is due to dietary deficiency. Milk, other dairy products, and enriched breads and cereals are the most important dietary sources of riboflavin in the United States, although lean meat, fish, eggs, broccoli, and legumes are also good sources. Riboflavin is extremely sensitive to light, and milk should be stored in containers that protect against photodegradation. Laboratory diagnosis of riboflavin deficiency can be made by determination of red blood cell or urinary riboflavin concentrations or by measurement of erythrocyte glutathione reductase activity, with and without added FAD. Because the capacity of the gastrointestinal tract to absorb riboflavin is limited (~20 mg after one oral dose), riboflavin toxicity has not been described.

NIACIN (VITAMIN B_3)

The term *niacin* refers to nicotinic acid and nicotinamide and their biologically active derivatives. Nicotinic acid and nicotinamide serve as precursors of two coenzymes, nicotinamide adenine dinucleotide (NAD) and NAD phosphate (NADP), which are important in

numerous oxidation and reduction reactions in the body. In addition, NAD and NADP are active in adenine diphosphate–ribose transfer reactions involved in DNA repair and calcium mobilization.

Metabolism and Requirements Nicotinic acid and nicotinamide are absorbed well from the stomach and small intestine. The bioavailability of niacin from beans, milk, meat, and eggs is high; bioavailability from cereal grains is lower. Since flour is enriched with "free" niacin (i.e., the non-coenzyme form), bioavailability is excellent. Median intakes of niacin in the United States considerably exceed the recommended dietary allowance (RDA).

The amino acid tryptophan can be converted to niacin with an efficiency of 60:1 by weight. Thus, the RDA for niacin is expressed in niacin equivalents. A lower-level conversion of tryptophan to niacin occurs in vitamin B_6 and/or riboflavin deficiencies and in the presence of

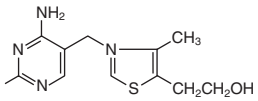
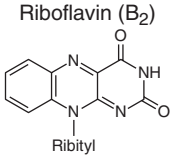
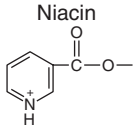
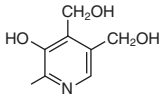
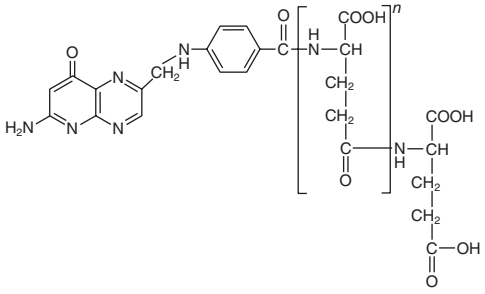
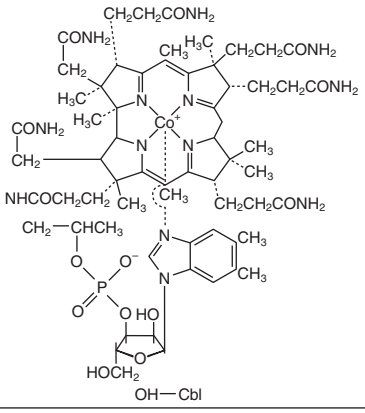
Vitamin	Active derivative or cofactor form	Principal function
Thiamine (B_1) 	Thiamine pyrophosphate	Coenzyme for cleavage of carbon-carbon bonds; amino acid and carbohydrate metabolism
Riboflavin (B_2) 	Flavin mononucleotide (FMN) and flavin adenine dinucleotide (FAD)	Cofactor for oxidation, reduction reactions, and covalently attached prosthetic groups for some enzymes
Niacin 	Nicotinamide adenine dinucleotide phosphate (NADP) and nicotinamide adenine dinucleotide (NAD)	Coenzymes for oxidation and reduction reactions
Vitamin B_6 	Pyridoxal phosphate	Cofactor for enzymes of amino acid metabolism
Folate 	Polyglutamate forms of (5, 6, 7, 8) tetrahydrofolate with carbon unit attachments	Coenzyme for one carbon transfer in nucleic acid and amino acid metabolism
Vitamin B_{12} 	Methylcobalamine Adenosylcobalamin	Coenzyme for methionine synthase and L-methylmalonyl-CoA mutase

FIGURE 326-1 Structures and principal functions of vitamins associated with human disorders.

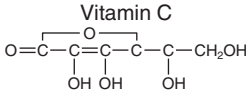
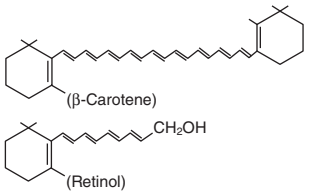
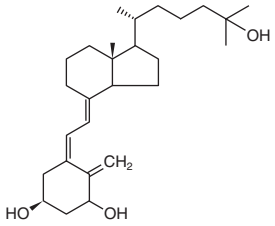
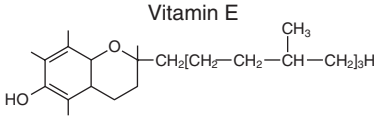
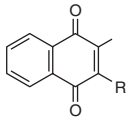
Vitamin	Active derivative or cofactor form	Principal function
Vitamin C 	Ascorbic acid and dehydroascorbic acid	Participation as a redox ion in many biologic oxidation and hydrogen transfer reactions
Vitamin A 	Retinol, retinaldehyde, and retinoic acid	Formation of rhodopsin (vision) and glycoproteins (epithelial cell function); also regulates gene transcription
Vitamin D 	1,25-Dihydroxyvitamin D	Maintenance of blood calcium and phosphorus levels; antiproliferative hormone
Vitamin E 	Tocopherols and tocotrienols	Antioxidants
Vitamin K 	Vitamin K hydroquinone	Cofactor for posttranslation carboxylation of many proteins including essential clotting factors

FIGURE 326-1 (*Continued*)

isoniazid. The urinary excretion products of niacin include 2-pyridone and 2-methyl nicotinamide, measurements of which are used in the diagnosis of niacin deficiency.

Deficiency Niacin deficiency causes *pellagra*, which is found mostly among people eating corn-based diets in parts of China, Africa, and India. Pellagra in North America is found mainly among alcoholics; among patients with congenital defects of intestinal and kidney absorption of tryptophan (Hartnup disease; **Chap. 413**); and among patients with carcinoid syndrome **(Chap. 80)**, in which there is increased conversion of tryptophan to serotonin. The antituberculosis drug isoniazid is a structural analog of niacin and can precipitate pellagra. In the setting of famine or population displacement, pellagra results from the absolute lack of niacin but also from the deficiency of micronutrients required for the conversion of tryptophan to niacin (e.g., iron, riboflavin, and pyridoxine). The early symptoms of pellagra include loss of appetite, generalized weakness and irritability, abdominal pain, and vomiting. Bright red glossitis then ensues and is followed by a characteristic skin rash that is pigmented and scaling, particularly in skin areas exposed to sunlight. This rash is known as *Casal's necklace* because it forms a ring around the neck; it is seen in advanced cases. Vaginitis and esophagitis also may occur. Diarrhea (due in part to proctitis and in part to malabsorption), depression, seizures, and dementia are also part of the pellagra syndrome. The primary manifestations of this syndrome are sometimes referred to as "the four *D*s": *d*ermatitis, *d*iarrhea, and *d*ementia leading to *d*eath.

TREATMENT

Pellagra

Treatment of pellagra consists of oral supplementation with 100–200 mg of nicotinamide or nicotinic acid three times daily for 5 days. High doses of nicotinic acid (2 g/d in a time-release form) are used for the treatment of elevated cholesterol and triglyceride levels and/or low high-density lipoprotein cholesterol levels without, however, a proven benefit on cardiovascular endpoints **(Chap. 400)**.

Toxicity Prostaglandin-mediated flushing due to binding of the vitamin to a G protein–coupled receptor has been observed at daily nicotinic acid doses as low as 30 mg taken as a supplement or as therapy for dyslipidemia. There is no evidence of toxicity from niacin that is derived from food sources. Flushing always starts in the face and may be accompanied by skin dryness, itching, paresthesia, and headache. Flushing is subject to tachyphylaxis and often improves with time; premedication with aspirin may alleviate these symptoms. Nausea, vomiting, and abdominal pain also occur at similar doses of niacin. Hepatic toxicity is the most serious toxic reaction caused by sustained-release niacin and may present as jaundice with elevated aspartate aminotransferase (AST) and alanine aminotransferase (ALT) levels. A few cases of fulminant hepatitis requiring liver transplantation have been reported at doses of 3–9 g/d. Other toxic reactions include glucose intolerance, hyperuricemia, macular edema, and macular cysts.

The combination of nicotinic acid preparations for dyslipidemia with 3-hydroxy-3-methylglutaryl coenzyme A (HMG-CoA) reductase inhibitors may increase the risk of rhabdomyolysis. The upper limit for daily niacin intake has been set at 35 mg. However, this upper limit does not pertain to the therapeutic use of niacin.

PYRIDOXINE (VITAMIN B_6)

Vitamin B_6 refers to a family of compounds that includes pyridoxine, pyridoxal, pyridoxamine, and their 5′-phosphate derivatives. 5′-Pyridoxal phosphate (PLP) is a cofactor for >100 enzymes involved in amino acid metabolism. Vitamin B_6 also is involved in heme and neurotransmitter synthesis and in the metabolism of glycogen, lipids, steroids, sphingoid bases, and several vitamins, including the conversion of tryptophan to niacin.

Dietary Sources Plants contain vitamin B_6 in the form of pyridoxine, whereas animal tissues contain PLP and pyridoxamine phosphate. The vitamin B_6 contained in plants is less bioavailable than that in animal tissues. Rich food sources of vitamin B_6 include legumes, nuts, wheat bran, and meat, although it is present in all food groups.

Deficiency Symptoms of vitamin B_6 deficiency include epithelial changes, as seen frequently with other B vitamin deficiencies. In addition, severe vitamin B_6 deficiency can lead to peripheral neuropathy, abnormal electroencephalograms, and personality changes that include depression and confusion. In infants, diarrhea, seizures, and anemia have been reported. Microcytic hypochromic anemia is due to diminished hemoglobin synthesis, since the first enzyme involved in heme biosynthesis (aminolevulinate synthase) requires PLP as a cofactor (Chap. 93). In some case reports, platelet dysfunction has been reported. Since vitamin B_6 is necessary for the conversion of homocysteine to cystathionine, it is possible that chronic low-grade vitamin B_6 deficiency may result in hyperhomocysteinemia and increased risk of cardiovascular disease (Chap. 413). Independent of homocysteine, low levels of circulating vitamin B_6 have been associated with inflammation and elevated levels of C-reactive protein.

Certain medications, such as isoniazid, L-dopa, penicillamine, and cycloserine, interact with PLP due to a reaction with carbonyl groups. Pyridoxine should be given concurrently with isoniazid to avoid neuropathy. The increased ratio of AST to ALT seen in alcoholic liver disease reflects the relative vitamin B_6 dependence of ALT. Vitamin B_6 dependency syndromes that require pharmacologic doses of vitamin B_6 are rare; they include cystathionine β-synthase deficiency, pyridoxine-responsive (primarily sideroblastic) anemias, and gyrate atrophy with chorioretinal degeneration due to decreased activity of the mitochondrial enzyme ornithine aminotransferase. In these situations, 100–200 mg/d of oral vitamin B_6 is required for treatment.

Severe nausea and vomiting in pregnancy might respond to pyridoxine combined with doxylamine. High doses of vitamin B_6 have been used to treat carpal tunnel syndrome, premenstrual syndrome, schizophrenia, autism, and diabetic neuropathy but have not been found to be effective.

The laboratory diagnosis of vitamin B_6 deficiency is generally based on low plasma PLP values (<20 nmol/L). Vitamin B_6 deficiency is treated with 50 mg/d; higher doses of 100–200 mg/d are given if the deficiency is related to medication use. Vitamin B_6 should not be given with L-dopa, since the vitamin interferes with the action of this drug.

Toxicity The safe upper limit for vitamin B_6 has been set at 100 mg/d, although no adverse effects have been associated with high intakes of vitamin B_6 from food sources only. When toxicity occurs, it causes severe sensory neuropathy, leaving patients unable to walk. Some cases of photosensitivity and dermatitis have been reported.

FOLATE (VITAMIN B_{12})

See Chap. 95.

VITAMIN C

Both ascorbic acid (only the L-isomer) and its oxidized product dehydroascorbic acid are biologically active. Actions of vitamin C include antioxidant activity, promotion of nonheme iron absorption, carnitine biosynthesis, conversion of dopamine to norepinephrine, and synthesis of many peptide hormones. Vitamin C is also important for connective tissue metabolism and cross-linking (proline hydroxylation), and it is a component of many drug-metabolizing enzyme systems, particularly the mixed-function oxidase systems.

Absorption and Dietary Sources Vitamin C is almost completely absorbed if <100 mg is administered in a single dose; however, only ≤50% is absorbed at doses >1 g. Enhanced degradation and fecal and urinary excretion of vitamin C occur at higher intake levels.

Good dietary sources of vitamin C include citrus fruits, green vegetables (especially broccoli), tomatoes, and potatoes. Consumption of five servings of fruits and vegetables a day provides vitamin C in excess of the RDA of 90 mg/d for men and 75 mg/d for women. In addition, ~40% of the U.S. population consumes vitamin C as a dietary supplement in which "natural forms" of the vitamin are no more bioavailable than synthetic forms. Smoking, hemodialysis, pregnancy, lactation, and stress (e.g., infection, trauma) appear to increase vitamin C requirements.

Deficiency Vitamin C deficiency causes scurvy. In the United States, this condition is seen primarily among the poor and the elderly, in alcoholics who consume <10 mg/d of vitamin C, and in individuals consuming macrobiotic diets. Vitamin C deficiency also can occur in young adults who eat severely unbalanced diets. In addition to generalized fatigue, symptoms of scurvy primarily reflect impaired formation of mature connective tissue and include bleeding into the skin (petechiae, ecchymoses, perifollicular hemorrhages); inflamed and bleeding gums; and manifestations of bleeding into joints, the peritoneal cavity, the pericardium, and the adrenal glands. In children, vitamin C deficiency may cause impaired bone growth. Laboratory diagnosis of vitamin C deficiency is based on low plasma or leukocyte levels.

Administration of vitamin C (200 mg/d) improves the symptoms of scurvy within several days. High-dose vitamin C supplementation (e.g., 0.2 up to several grams per day) may slightly decrease the symptoms and duration of upper respiratory tract infections. Vitamin C supplementation has also been reported to be useful in Chédiak-Higashi syndrome (Chap. 60) and osteogenesis imperfecta (Chap. 406). Diets high in vitamin C have been claimed to lower the incidence of certain cancers, particularly esophageal and gastric cancers. If proved, this effect may be because vitamin C can prevent the conversion of nitrites and secondary amines to carcinogenic nitrosamines. Evidence for a potential role of pro-oxidative effects of parenteral ascorbic acid in the treatment of advanced cancers is emerging in laboratory studies.

Toxicity Taking >2 g of vitamin C in a single dose may result in abdominal pain, diarrhea, and nausea. Since vitamin C may be metabolized to oxalate, it is feared that chronic high-dose vitamin C supplementation could result in an increased prevalence of kidney stones. However, except in patients with preexisting renal disease, this association has not been borne out in several trials. Nevertheless, it is reasonable to advise patients with a history of kidney stones and renal insufficiency not to take large doses of vitamin C. There is also an unproven but possible risk that chronic high doses of vitamin C could promote iron overload and iron toxicity. High doses of vitamin C can induce hemolysis in patients with glucose-6-phosphate dehydrogenase deficiency, and doses >1 g/d can cause false-negative guaiac reactions and interfere with tests for urinary glucose. High doses may interfere with the activity of certain drugs (e.g., bortezomib in myeloma patients).

BIOTIN

Biotin is a water-soluble vitamin that plays a role in gene expression, gluconeogenesis, and fatty acid synthesis and serves as a CO_2 carrier on the surface of both cytosolic and mitochondrial carboxylase enzymes. The vitamin also functions in the catabolism of specific amino acids (e.g., leucine) and in gene regulation by histone biotinylation. Excellent food sources of biotin include organ meat such as liver or kidney, soy and other beans, yeast, and egg yolks; however, egg white contains

the protein avidin, which strongly binds the vitamin and reduces its bioavailability.

Biotin deficiency due to low dietary intake is rare; rather, deficiency is due to inborn errors of metabolism. Biotin deficiency has been induced by experimental feeding of egg white diets and by biotin-free parenteral nutrition in patients with short bowels. In adults, biotin deficiency results in mental changes (depression, hallucinations), paresthesia, anorexia, and nausea. A scaling, seborrheic, and erythematous rash may occur around the eyes, nose, and mouth as well as on the extremities. In infants, biotin deficiency presents as hypotonia, lethargy, and apathy. In addition, infants may develop alopecia and a characteristic rash that includes the ears. The laboratory diagnosis of biotin deficiency can be established on the basis of a decreased concentration of urinary biotin (or its major metabolites), increased urinary excretion of 3-hydroxyisovaleric acid after a leucine challenge, or decreased activity of biotin-dependent enzymes in lymphocytes (e.g., propionyl-CoA carboxylase). Treatment requires pharmacologic doses of biotin, that is, up to 10 mg/d. No toxicity is known.

PANTOTHENIC ACID (VITAMIN B_5)

Pantothenic acid is a component of coenzyme A and phosphopantetheine, which are involved in fatty acid metabolism and the synthesis of cholesterol, steroid hormones, and all compounds formed from isoprenoid units. In addition, pantothenic acid is involved in the acetylation of proteins. The vitamin is excreted in the urine, and the laboratory diagnosis of deficiency is based on low urinary vitamin levels.

The vitamin is ubiquitous in the food supply. Liver, yeast, egg yolks, whole grains, and vegetables are particularly good sources. Human pantothenic acid deficiency has been demonstrated only by experimental feeding of diets low in pantothenic acid or by administration of a specific pantothenic acid antagonist. The symptoms of pantothenic acid deficiency are nonspecific and include gastrointestinal disturbance, depression, muscle cramps, paresthesia, ataxia, and hypoglycemia. Pantothenic acid deficiency is believed to have caused the "burning feet syndrome" seen in prisoners of war during World War II. No toxicity of this vitamin has been reported.

CHOLINE

Choline is a precursor for acetylcholine, phospholipids, and betaine. Choline is necessary for the structural integrity of cell membranes, cholinergic neurotransmission, lipid and cholesterol metabolism, methyl-group metabolism, and transmembrane signaling. Recently, a recommended adequate intake was set at 550 mg/d for men and 425 mg/d for women, although certain genetic polymorphisms can increase an individual's requirement. Choline is thought to be a "conditionally essential" nutrient in that its de novo synthesis occurs in the liver and results in lesser-than-used amounts only under certain stress conditions (e.g., alcoholic liver disease). The dietary requirement for choline depends on the status of other nutrients involved in methyl-group metabolism (folate, vitamin B_{12}, vitamin B_6, and methionine) and thus varies widely. Choline is widely distributed in food (e.g., egg yolks, wheat germ, organ meat, milk) in the form of lecithin (phosphatidylcholine). Choline deficiency has occurred in patients receiving parenteral nutrition devoid of choline. Deficiency results in fatty liver, elevated aminotransferase levels, and skeletal muscle damage with high creatine phosphokinase values. The diagnosis of choline deficiency is currently based on low plasma levels, although nonspecific conditions (e.g., heavy exercise) may also suppress plasma levels.

Toxicity from choline results in hypotension, cholinergic sweating, diarrhea, salivation, and a fishy body odor. The upper limit for choline intake has been set at 3.5 g/d. Because of its ability to lower cholesterol and homocysteine levels, choline treatment has been suggested for patients with dementia and patients at high risk of cardiovascular disease. However, the benefits of such treatment have not been firmly documented. Choline- and betaine-restricted diets are of therapeutic value in trimethylaminuria ("fish odor syndrome").

FLAVONOIDS

Flavonoids constitute a large family of polyphenols that contribute to the aroma, taste, and color of fruits and vegetables. Major groups of dietary flavonoids include anthocyanidins in berries; catechins in green tea and chocolate; flavonols (e.g., quercetin) in broccoli, kale, leeks, onions, and the skins of grapes and apples; and isoflavones (e.g., genistein) in legumes. Isoflavones have a low bioavailability and are partially metabolized by the intestinal flora. The dietary intake of flavonoids is estimated at 10–100 mg/d; this figure is almost certainly an underestimate attributable to a lack of information on their concentrations in many foods. Several flavonoids have antioxidant activity and affect cell signaling. From observational epidemiologic studies and limited clinical (human and animal) studies, flavonoids have been postulated to play a role in the prevention of several chronic diseases, including neurodegenerative disease, diabetes, and osteoporosis. The ultimate importance and usefulness of these compounds against human disease have not been consistently demonstrated.

VITAMIN A

Vitamin A, in the strictest sense, refers to retinol. However, the oxidized metabolites retinaldehyde and retinoic acid are also biologically active compounds. The term *retinoids* includes all molecules (including synthetic molecules) that are chemically related to retinol. Retinaldehyde (11-*cis*) is the form of vitamin A that is required for normal vision, whereas retinoic acid is necessary for normal morphogenesis, growth, and cell differentiation. Retinoic acid does not function directly in vision and, in contrast to retinol, is not involved in reproduction. Vitamin A also plays a role in iron utilization, humoral immunity, T cell–mediated immunity, natural killer cell activity, and phagocytosis.

Vitamin A is found in the human food supply in two forms: preformed as esters and provitamin A in carotenoids. There are >600 carotenoids in nature, ~50 of which can be metabolized to vitamin A. β-Carotene is the most prevalent carotenoid with provitamin A activity in the food supply. In humans, significant fractions of carotenoids are absorbed intact and are stored in liver and fat. It is estimated that ≥12 μg (range, 4–27 μg) of dietary all-*trans* β-carotene is equivalent to 1 μg of retinol activity, whereas the figure is ≥24 μg for other dietary provitamin A carotenoids (e.g., cryptoxanthin, α-carotene). The vitamin A equivalency for a β-carotene supplement in an oily solution is 2:1.

Metabolism The liver contains ~90% of the vitamin A reserves and secretes vitamin A in the form of retinol, which is bound in the circulation to retinol-binding protein. Once binding has occurred, the retinol-binding protein complex interacts with a second protein, transthyretin. This trimolecular complex functions to prevent vitamin A from being filtered by the kidney glomerulus, thus protecting the body against the toxicity of retinol and allowing retinol to be taken up by specific cell-surface receptors that recognize retinol-binding protein. A certain amount of vitamin A enters peripheral cells even if it is not bound to retinol-binding protein. After retinol is internalized by the cell, it becomes bound to a series of cellular retinol-binding proteins, which function as sequestering and transporting agents as well as co-ligands for enzymatic reactions. Certain cells also contain retinoic acid–binding proteins, which have sequestering functions but also shuttle retinoic acid to the nucleus and enable its metabolism.

Retinoic acid is a ligand for certain nuclear receptors that act as transcription factors. Two families of receptors (retinoic acid receptors [RARs] and retinoid X receptors [RXRs]) are active in retinoid-mediated gene transcription. Retinoid receptors regulate transcription by binding as dimeric complexes to specific DNA sites—the retinoic acid response elements—in target genes (Chap. 370). The receptors can either stimulate or repress gene expression in response to their ligands. RARs bind all-*trans* retinoic acid and 9-*cis*-retinoic acid, whereas RXRs bind only 9-*cis*-retinoic acid.

The retinoid receptors play an important role in controlling cell proliferation and differentiation. RXRs dimerize with other nuclear receptors to function as coregulators of genes responsive to retinoids, but also to thyroid hormone and calcitriol. RXR agonists induce insulin sensitivity experimentally, perhaps because RXRs are cofactors for the peroxisome proliferator-activated receptors, which mediate also fatty

acid and carbohydrate metabolism and are targets for different drugs including thiazolidinedione drugs (e.g., rosiglitazone and pioglitazone) (Chap. 397).

Dietary Sources The retinol activity equivalent (RAE) is used to express the vitamin A value of food: 1 RAE is defined as 1 µg of retinol (0.003491 mmol), 12 µg of β-carotene, and 24 µg of other provitamin A carotenoids. In older literature, vitamin A often was expressed in international units (IUs), with 1 µg of retinol equal to 3.33 IU of retinol and 20 IU of β-carotene. Although these IUs are no longer in scientific use, they can still be found in reports of the food industry and in public health interventions in low-income countries.

Liver, fish, and eggs are excellent food sources for preformed vitamin A; vegetable sources of provitamin A carotenoids include dark green and deeply colored fruits and vegetables. Moderate cooking of vegetables enhances carotenoid release for uptake in the gut. Carotenoid absorption is also aided by some fat in a meal. Exclusive breastfeeding can cover the vitamin A needs of infants if the mother has an adequate vitamin A status and a large enough volume of milk. If the nursing mother has inadequate vitamin A intake, concomitant diseases, or her infant was a preterm delivery, breast milk probably will not supply enough vitamin A to prevent deficiency. In developing countries, chronic dietary deficiency is the main cause of vitamin A deficiency and is exacerbated by infection. In early childhood, low vitamin A status results from inadequate intakes of animal food sources and edible oils, both of which are expensive, coupled with seasonal unavailability of vegetables and fruits and lack of marketed fortified food products. Factors that interfere with vitamin A metabolism may also affect status or function. For example, concurrent zinc deficiency can interfere with the mobilization of vitamin A from liver stores. Alcohol interferes with the conversion of retinol to retinaldehyde in the eye by competing for alcohol (retinol) dehydrogenase. Drugs that interfere with the absorption of vitamin A include mineral oil, neomycin, and cholestyramine.

Deficiency Vitamin A deficiency is endemic in areas where diets are chronically poor, especially in southern Asia, sub-Saharan Africa, some parts of Latin America, and the western Pacific, including parts of China. Vitamin A status is usually assessed by measuring serum retinol (normal range, 1.05–3.50 µmol/L [30–100 µg/dL]) or blood-spot retinol or by tests of dark adaptation. Stable isotopic or invasive liver biopsy methods are available to estimate total body stores of vitamin A. As judged by deficient serum retinol (<0.70 µmol/L [20 µg/dL]), vitamin A deficiency worldwide is present in 190 million preschool-age children, among whom >5 million have an ocular manifestation of deficiency termed *xerophthalmia*. This condition includes milder stages of night blindness and conjunctival *xerosis* (dryness) with *Bitot's spots* (white patches of keratinized epithelium appearing on the sclera) that may affect 1–5% of children in deficient populations as well as rare, potentially blinding corneal ulceration and necrosis. *Keratomalacia* (softening of the cornea) leads to corneal scarring that blinds an estimated quarter of a million children each year and is associated with fatality rates of 4–25%. However, vitamin A deficiency severe enough to cause any clinical stage poses an increased risk of death from diarrhea, dysentery, measles, malaria, or respiratory disease. This is because vitamin A deficiency can compromise barrier, innate, and acquired immune defenses to infection. In areas where deficiency is widely prevalent, vitamin A supplementation can markedly reduce the risk of childhood mortality (by 23–34%, on average). About 10% of pregnant women in undernourished settings also develop night blindness (assessed by history) during the latter half of pregnancy; this level of moderate to severe vitamin A deficiency is associated with an increased risk of maternal infection and death. Maternal vitamin A deficiency may also exacerbate already low vitamin A nutrition and associated risks for the newborn. In South Asia, where maternal deficiency is prominent, giving infants a single oral dose (50,000 IU) of vitamin A shortly after birth has reduced infant mortality by ≥10%, whereas in African settings less affected by maternal vitamin A deficiency, no effect has been noted, revealing differences in risk of deficiency and benefit of supplementation across regions.

TREATMENT

Vitamin A Deficiency

Vitamin A is commercially available for treatment and prevention in esterified forms (e.g., acetate, palmitate), which are more stable than other forms. Any stage of xerophthalmia should be treated with 60 mg (or RAE) or 200,000 IU of vitamin A in oily solution, usually contained in a soft-gel capsule. The same dose is repeated 1 and 14 days later. Doses should be reduced by half for patients 6–11 months of age. Mothers with night blindness or Bitot's spots should be given vitamin A orally 3 mg daily for at least 3 months. These regimens are efficacious, and they are far less expensive and more widely available than injectable water-miscible vitamin A. A common approach to prevention is to provide vitamin A supplementation every 4–6 months to young children 6 months to 5 years of age (both HIV-positive and HIV-negative) in high-risk areas. For prevention, infants 6–11 months of age should receive 30 mg vitamin A; children 12–59 months of age, 60 mg. For reasons that are not clear, while early neonatal vitamin A may reduce infant mortality, vitamin A given between 1 and 5 months of age has not proven effective in improving survival in high-risk settings.

Uncomplicated vitamin A deficiency is rare in industrialized countries. One high-risk group—extremely low-birth-weight (<1000-g) infants—is likely to be vitamin A–deficient and should receive a supplement of 1500 µg (or RAE) three times a week for 4 weeks. Severe measles in any society can lead to secondary vitamin A deficiency. Children hospitalized with measles should receive two 60-mg doses of vitamin A on two consecutive days. Vitamin A deficiency most often occurs in patients with malabsorptive diseases (e.g., celiac sprue, short-bowel syndrome) who have abnormal dark adaptation or symptoms of night blindness without other ocular changes. Typically, such patients are diagnosed in advanced care settings where they are treated for 1 month with 15 mg/d of a water-miscible preparation of vitamin A. This treatment is followed by a lower maintenance dose, with the exact amount determined by monitoring serum retinol. Finding application elsewhere in medicine, retinoic acid is useful in the treatment of promyelocytic leukemia (Chap. 100) and also is used in the treatment of cystic acne because it inhibits keratinization, decreases sebum secretion, and possibly alters the inflammatory reaction (Chap. 53).

No specific signs or symptoms result from carotenoid deficiency. It was postulated that β-carotene would be an effective chemopreventive agent for cancer because numerous epidemiologic studies had shown that diets high in β-carotene were associated with lower incidences of cancers of the respiratory and digestive systems. However, intervention studies in smokers found that treatment with high doses of β-carotene actually resulted in more lung cancers than did treatment with placebo. Non–provitamin A carotenoids such as lutein and zeaxanthin have been suggested to confer protection against macular degeneration, and one large-scale intervention study did not show a beneficial effect except in those with a low lutein status. The use of the non–provitamin A carotenoid lycopene to protect against prostate cancer has been proposed. Again, however, the effectiveness of these agents has not been proved by intervention studies, and the mechanisms underlying these purported biologic actions are unknown.

Selective plant-breeding techniques that lead to a higher provitamin A carotenoid content in staple foods may decrease vitamin A malnutrition in low-income countries. Moreover, a recently developed genetically modified food (Golden Rice) has an improved β-carotene–to–vitamin A conversion ratio of ~3:1.

Toxicity The acute toxicity of vitamin A was first noted in Arctic explorers who ate polar bear liver and has also been seen after administration of 150 mg to adults or 100 mg to children. Acute toxicity is manifested by increased intracranial pressure, vertigo, diplopia, bulging fontanels (in children), seizures, and exfoliative dermatitis; it

may result in death. Among children being treated for vitamin A deficiency according to the protocols outlined above, transient bulging of fontanels occurs in 2% of infants, and transient nausea, vomiting, and headache occur in 5% of preschoolers. Chronic vitamin A intoxication is largely a concern in industrialized countries and has been seen in otherwise healthy adults who ingest 15 mg/d and children who ingest 6 mg/d over a period of several months. Manifestations include dry skin, cheilosis, glossitis, vomiting, alopecia, bone demineralization and pain, hypercalcemia, lymph node enlargement, hyperlipidemia, amenorrhea, and features of pseudotumor cerebri with increased intracranial pressure and papilledema. Liver fibrosis with portal hypertension may also result from chronic vitamin A intoxication. Provision of vitamin A in excess to pregnant women has resulted in spontaneous abortion and in congenital malformations, including craniofacial abnormalities and valvular heart disease. In pregnancy, the daily dose of vitamin A should not exceed 3 mg. Commercially available retinoid derivatives are also toxic, including 13-*cis*-retinoic acid, which has been associated with birth defects. Thus contraception should be continued for at least 1 year and possibly longer in women who have taken 13-*cis*-retinoic acid.

In malnourished children, vitamin A supplements (30–60 mg), in amounts calculated as a function of age and given in several rounds over 2 years, are considered to amplify nonspecific effects of vaccines. However, for unclear reasons, in one African setting there has been a negative effect on mortality rates in incompletely vaccinated girls.

High doses of carotenoids do not result in toxic symptoms but should be avoided in smokers due to an increased risk of lung cancer. Very high doses of β-carotene (~200 mg/d) have been used to treat or prevent the skin rashes of erythropoietic protoporphyria. Carotenemia, which is characterized by a yellowing of the skin (increases of the palms and soles) but not the sclerae, may follow ingestion of >30 mg of β-carotene daily. Hypothyroid patients are particularly susceptible to the development of carotenemia due to impaired breakdown of carotene to vitamin A. Reduction of carotenes in the diet results in the disappearance of skin yellowing and carotenemia over a period of 30–60 days.

VITAMIN D

The metabolism of the fat-soluble vitamin D is described in detail in Chap. 402. The biologic effects of this vitamin are mediated by vitamin D receptors, which are found in most tissues; binding with these receptors potentially expands vitamin D actions to many different cell systems and organs (e.g., immune cells, brain, breast, colon, and prostate) in addition to the classic endocrine effects on calcium and phosphate metabolism and bone health. Vitamin D is thought to be important for maintaining normal function of many nonskeletal tissues such as muscle (including heart muscle), for immune function, and for inflammation as well as for cell proliferation and differentiation. Studies have shown that vitamin D may be useful as adjunctive treatment for tuberculosis, psoriasis, and multiple sclerosis or for the prevention of certain cancers. Vitamin D insufficiency may increase the risk of type 1 diabetes mellitus, cardiovascular disease (insulin resistance, hypertension, or low-grade inflammation), or brain dysfunction (e.g., depression). However, the exact physiologic roles of vitamin D in these nonskeletal diseases and the importance of these roles have not been clarified.

The skin is a major source of vitamin D, which is synthesized upon skin exposure to ultraviolet B radiation (UV-B; wavelength, 290–320 nm). Except for fish, food (unless fortified) contains only limited amounts of vitamin D. Vitamin D_2 (ergocalciferol) is obtained from plant sources and is the chemical form found in some supplements.

Deficiency Vitamin D status has been assessed by measuring serum levels of 25-dihydroxyvitamin D (25[OH] vitamin D); however, there is no consensus on a uniform assay or on optimal serum levels. The optimal level might, in fact, differ according to the targeted disease entity. Epidemiologic and experimental data indicate that a 25(OH) vitamin D level of >20 ng/mL (≥50 nmol/L; to convert ng/mL to nmol/L, multiply by 2.496) is sufficient for good bone health. Some experts, however, advocate higher serum levels (e.g., >30 ng/mL) for other desirable endpoints of vitamin D action. There is insufficient evidence to recommend combined vitamin D and calcium supplementation as a primary preventive strategy (as opposed to secondary prevention) for reduction of the incidence of fractures in healthy men and premenopausal women.

Risk factors for vitamin D deficiency are old age, lack of sun exposure, dark skin (especially among residents of northern latitudes), fat malabsorption, and obesity. *Rickets* represents the classic disease of vitamin D deficiency. Signs of deficiency are muscle soreness, weakness, and bone pain. Some of these effects are independent of calcium intake.

The U.S. National Academy of Sciences recently advised that the majority of adult North Americans should receive 600 IU/d of vitamin D (RDA = 15 μg/d or 600 IU/d; Chap. 325). However, for people aged >70 years, the RDA is set at 20 μg/d (800 IU/d). The consumption of fortified or enriched foods as well as suberythemal sun exposure should be encouraged for people at risk for vitamin D deficiency. If adequate intake is impossible, vitamin D supplements should be taken, especially during the winter months. Vitamin D deficiency can be treated by the oral administration of 50,000 IU/week for 6–8 weeks followed by a maintenance dose of 800 IU/d (20 μg/d) from food and supplements once normal plasma levels have been attained. There is uncertainty regarding the optimal therapeutic dosage (high vs low) for elderly at risk of falls. The physiologic effects of vitamin D_2 and vitamin D_3 are similar when these vitamins are ingested over long periods.

Toxicity The upper limit of intake has been set at 4000 IU/d. Contrary to earlier beliefs, acute vitamin D intoxication is rare and usually is caused by the uncontrolled and excessive ingestion of supplements or by faulty food fortification practices. High plasma levels of $1,25(OH)_2$ vitamin D and calcium are central features of toxicity and mandate discontinuation of vitamin D and calcium supplements; in addition, treatment of hypercalcemia may be required.

VITAMIN E

Vitamin E is the collective designation for all stereoisomers of tocopherols and tocotrienols, although only the RR tocopherols meet human requirements. Vitamin E acts as a chain-breaking antioxidant and is an efficient peroxyl radical scavenger that protects low-density lipoproteins and polyunsaturated fats in membranes from oxidation. A network of other antioxidants (e.g., vitamin C, glutathione) and enzymes maintains vitamin E in a reduced state. Vitamin E also inhibits prostaglandin synthesis and the activities of protein kinase C and phospholipase A_2.

Absorption and Metabolism After absorption, vitamin E is taken up from chylomicrons by the liver, and a hepatic α-tocopherol transport protein mediates intracellular vitamin E transport and incorporation into very low density lipoprotein. The transport protein has a particular affinity for the RRR isomeric form of α-tocopherol; thus, this natural isomer has the most biologic activity.

Requirement Vitamin E is widely distributed in the food supply, with particularly high levels in sunflower oil, safflower oil, and wheat germ oil; γ-tocotrienols are notably present in soybean and corn oils. Vitamin E is also found in meats, nuts, and cereal grains, and small amounts are present in fruits and vegetables. Vitamin E pills containing doses of 50–1000 mg are ingested by ~10% of the U.S. population. The RDA for vitamin E is 15 mg/d (34.9 μmol or 22.5 IU) for all adults. Diets high in polyunsaturated fats may necessitate a slightly higher intake of vitamin E.

Dietary deficiency of vitamin E does not exist. Vitamin E deficiency is seen only in severe and prolonged malabsorptive diseases, such as celiac disease, or after small-intestinal resection or bariatric surgery. Children with cystic fibrosis or prolonged cholestasis may develop vitamin E deficiency characterized by areflexia and hemolytic anemia. Children with abetalipoproteinemia cannot absorb or transport vitamin E and become deficient quite rapidly. A familial form of isolated vitamin E deficiency also exists; it is due to a defect in the α-tocopherol transport protein. Vitamin E deficiency causes axonal degeneration of the large myelinated axons and results in posterior column and spinocerebellar symptoms. Peripheral neuropathy is initially characterized by areflexia, with progression to an ataxic gait, and by decreased vibration and position sensations. Ophthalmoplegia, skeletal myopathy, and

pigmented retinopathy may also be features of vitamin E deficiency. A deficiency of either vitamin E or selenium in the host has been shown to increase certain viral mutations and, therefore, virulence. The laboratory diagnosis of vitamin E deficiency is based on low blood levels of α-tocopherol (<5 μg/mL, or <0.8 mg of α-tocopherol per gram of total lipids).

TREATMENT

Vitamin E Deficiency

Symptomatic vitamin E deficiency should be treated with 800–1200 mg of α-tocopherol per day. Patients with abetalipoproteinemia may need as much as 5000–7000 mg/d. Children with symptomatic vitamin E deficiency should be treated orally with water-miscible esters (400 mg/d); alternatively, 2 mg/kg/d may be administered intramuscularly. Vitamin E in high doses may protect against oxygen-induced retrolental fibroplasia and bronchopulmonary dysplasia as well as intraventricular hemorrhage of prematurity. Vitamin E has been suggested to increase sexual performance, treat intermittent claudication, and slow the aging process, but convincing evidence for these properties is lacking. When given in combination with other antioxidants, vitamin E may help prevent macular degeneration. Vitamin E may have favorable therapeutic effects in noncirrhotic nondiabetic patients with NASH (nonalcoholic steatohepatitis). High doses (60–800 mg/d) of vitamin E have been shown in controlled trials to improve parameters of immune function and reduce colds in nursing home residents, but intervention studies using vitamin E to prevent cardiovascular disease or cancer have not shown efficacy, and, at doses >400 mg/d, vitamin E may even increase all-cause mortality rates and prostate cancer risk.

Toxicity All forms of vitamin E are absorbed and could contribute to toxicity; however, the toxicity risk seems to be rather low as long as liver function is normal. High doses of vitamin E (>800 mg/d) may reduce platelet aggregation and interfere with vitamin K metabolism and are therefore contraindicated in patients taking warfarin and antiplatelet agents (such as aspirin or clopidogrel). Nausea, flatulence, and diarrhea have been reported at doses >1 g/d.

VITAMIN K

There are two natural forms of vitamin K: vitamin K_1, also known as *phylloquinone*, from vegetable sources, and vitamin K_2, or *menaquinones*, which are synthesized by bacterial flora and found in hepatic tissue. Phylloquinone can be converted to menaquinone in some organs.

Vitamin K is required for the posttranslational carboxylation of glutamic acid, which is necessary for calcium binding to γ-carboxylated proteins such as prothrombin (factor II); factors VII, IX, and X; protein C; protein S; and proteins found in bone (osteocalcin) and vascular smooth muscle (e.g., matrix Gla protein). However, the importance of vitamin K for bone mineralization and prevention of vascular calcification is not known. Warfarin-type drugs inhibit γ-carboxylation by preventing the conversion of vitamin K to its active hydroquinone form.

Dietary Sources Vitamin K is found in green leafy vegetables such as kale and spinach, and appreciable amounts are also present in margarine and liver. Vitamin K is present in vegetable oils; olive, canola, and soybean oils are particularly rich sources. The average daily intake by Americans is estimated to be ~100 μg/d.

Deficiency The symptoms of vitamin K deficiency are due to hemorrhage; newborns are particularly susceptible because of low fat stores, low breast milk levels of vitamin K, relative sterility of the infantile intestinal tract, liver immaturity, and poor placental transport. Intracranial bleeding as well as gastrointestinal and skin bleeding can occur in vitamin K–deficient infants 1–7 days after birth. Thus, vitamin K (0.5–1 mg IM) is given prophylactically at delivery.

Vitamin K deficiency in adults may be seen in patients with chronic small-intestinal disease (e.g., celiac disease, Crohn's disease), in those with obstructed biliary tracts, or after small-bowel resection. Broad-spectrum antibiotic treatment can precipitate vitamin K deficiency by reducing numbers of gut bacteria, which synthesize menaquinones, and by inhibiting the metabolism of vitamin K. In patients with warfarin therapy, the antiobesity drug orlistat can lead to changes in international normalized ratio due to vitamin K malabsorption. Vitamin K deficiency usually is diagnosed on the basis of an elevated prothrombin time or reduced clotting factors, although vitamin K may also be measured directly by high-pressure liquid chromatography. Vitamin K deficiency is treated with a parenteral dose of 10 mg. For patients with chronic malabsorption, 1–2 mg/d should be given orally or 1–2 mg per week can be taken parenterally. Patients with liver disease may have an elevated prothrombin time because of liver cell destruction as well as vitamin K deficiency. If an elevated prothrombin time does not improve during vitamin K therapy, it can be deduced that this abnormality is not the result of vitamin K deficiency.

Toxicity Toxicity from dietary phylloquinones and menaquinones has not been described. High doses of vitamin K can impair the actions of oral vitamin K antagonist anticoagulants.

MINERALS

See also Table 326-2.

CALCIUM

See Chap. 402.

ZINC

Zinc is an integral component of many metalloenzymes in the body; it is involved in the synthesis and stabilization of proteins, DNA, and RNA, and plays a structural role in ribosomes and membranes. Zinc is necessary for the binding of steroid hormone receptors and several other transcription factors to DNA. Zinc is absolutely required for normal spermatogenesis, fetal growth, and embryonic development.

Absorption The absorption of zinc from the diet is inhibited by dietary phytate, fiber, oxalate, iron, and copper as well as by certain drugs, including penicillamine, sodium valproate, and ethambutol. Meat, shellfish, nuts, and legumes are good sources of bioavailable zinc, whereas zinc in grains and legumes is less available for absorption.

Deficiency Mild zinc deficiency has been described in many diseases, including diabetes mellitus, HIV/AIDS, cirrhosis, alcoholism, inflammatory bowel disease, malabsorption syndromes, and sickle cell disease. In these diseases, mild chronic zinc deficiency can cause stunted growth in children, decreased taste sensation (*hypogeusia*), and impaired immune function. Severe chronic zinc deficiency has been described as a cause of hypogonadism and dwarfism in several Middle Eastern countries. In these children, hypopigmented hair is also part of the syndrome. Acrodermatitis enteropathica is a rare autosomal recessive disorder characterized by abnormalities in zinc absorption. Clinical manifestations include diarrhea, alopecia, muscle wasting, depression, irritability, and a rash involving the extremities, face, and perineum. The rash is characterized by vesicular and pustular crusting with scaling and erythema. Occasional patients with Wilson's disease have developed zinc deficiency as a consequence of penicillamine therapy (Chap. 408).

Zinc deficiency is prevalent in many developing countries and usually coexists with other micronutrient deficiencies (especially iron deficiency). Zinc (20 mg/d until recovery) may be an effective adjunctive therapeutic strategy for diarrheal disease and pneumonia in children ≥6 months of age.

The diagnosis of zinc deficiency is usually based on a serum zinc level <12 μmol/L (<70 μg/dL). Pregnancy and birth control pills may cause a slight depression in serum zinc levels, and hypoalbuminemia from any cause can result in hypozincemia. In acute stress situations (illness but also post-exercise recovery), zinc may be redistributed from serum into tissues. Zinc deficiency may be treated with 60 mg of elemental zinc taken by mouth twice a day. Zinc gluconate lozenges (13 mg of elemental zinc every 2 h while awake) have been reported to reduce the duration and symptoms of the common cold in adults, but study results are conflicting.

Toxicity Acute zinc toxicity after oral ingestion causes nausea, vomiting, and fever. Zinc fumes from welding may also be toxic and cause fever, respiratory distress, excessive salivation, sweating, and headache. Chronic large doses of zinc may depress immune function and cause hypochromic anemia as a result of copper deficiency. Intranasal zinc preparations should be avoided because they may lead to irreversible damage of the nasal mucosa and anosmia.

COPPER

Copper is an integral part of numerous enzyme systems, including amine oxidases, ferroxidase (ceruloplasmin), cytochrome c oxidase, superoxide dismutase, and dopamine hydroxylase. Copper is also a component of ferroprotein, a transport protein involved in the basolateral transfer of iron during absorption from the enterocyte. As such, copper plays a role in iron metabolism, melanin synthesis, energy production, neurotransmitter synthesis, and CNS function; the synthesis and cross-linking of elastin and collagen; and the scavenging of superoxide radicals. Dietary sources of copper include shellfish, liver, nuts, legumes, bran, and organ meats.

Deficiency Dietary copper deficiency is relatively rare, although it has been described in premature infants who are fed milk diets and in infants with malabsorption (Table 326-2). Copper-deficiency anemia (refractory to therapeutic iron) has been reported in patients with malabsorptive diseases and nephrotic syndrome and in patients treated for Wilson's disease with chronic high doses of oral zinc, which can interfere with copper absorption. *Menkes kinky hair syndrome* is an X-linked metabolic disturbance of copper metabolism characterized by mental retardation, hypocupremia, and decreased circulating ceruloplasmin (Chap. 406). This syndrome is caused by mutations in the copper-transporting *ATP7A* gene. Children with this disease often die within 5 years because of dissecting aneurysms or cardiac rupture. Aceruloplasminemia is a rare autosomal recessive disease characterized by tissue iron overload, mental deterioration, microcytic anemia, and low serum iron and copper concentrations.

The diagnosis of copper deficiency is usually based on low serum levels of copper (<65 μg/dL) and low ceruloplasmin levels (<20 mg/dL). Serum levels of copper may be elevated in pregnancy or stress conditions since ceruloplasmin is an acute-phase reactant and 90% of circulating copper is bound to ceruloplasmin.

Toxicity Copper toxicity is usually accidental (Table 326-2). In severe cases, kidney failure, liver failure, and coma may ensue. In Wilson's disease, mutations in the copper-transporting *ATP7B* gene lead to accumulation of copper in the liver and brain, with low blood levels due to decreased ceruloplasmin (Chap. 408).

SELENIUM

Selenium, in the form of selenocysteine, is a component of the enzyme glutathione peroxidase, which serves to protect proteins, cell membranes, lipids, and nucleic acids from oxidant molecules. As such, selenium is being actively studied as a chemopreventive agent against certain cancers, such as prostate cancer. However, it remains unclear whether selenium is effective as a chemopreventive

TABLE 326-2 Deficiencies and Toxicities of Metals

ELEMENT	DEFICIENCY	TOXICITY	TOLERABLE UPPER (DIETARY) INTAKE LEVEL
Boron	No biologic function determined	Developmental defects, male sterility, testicular atrophy	20 mg/d (extrapolated from animal data)
Calcium	Reduced bone mass, osteoporosis	Renal insufficiency (milk-alkali syndrome), nephrolithiasis, impaired iron absorption, thiazide diuretics	2500 mg/d (milk-alkali)
Copper	Anemia, growth retardation, defective keratinization and pigmentation of hair, hypothermia, degenerative changes in aortic elastin, osteopenia, mental deterioration	Nausea, vomiting, diarrhea, hepatic failure, tremor, mental deterioration, hemolytic anemia, renal dysfunction	10 mg/d (liver toxicity)
Chromium	Impaired glucose tolerance	*Occupational:* Renal failure, dermatitis, pulmonary cancer	Not determined
Fluoride	↑ Dental caries	Dental and skeletal fluorosis, osteosclerosis	10 mg/d (fluorosis)
Iodine	Thyroid enlargement, ↓ T_4, cretinism	Thyroid dysfunction, acne-like eruptions	1100 μg/d (thyroid dysfunction)
Iron	Muscle abnormalities, koilonychia, pica, anemia, ↓ work performance, impaired cognitive development, premature labor, ↑ perinatal maternal death	Gastrointestinal effects (nausea, vomiting, diarrhea, constipation), iron overload with organ damage, acute and chronic systemic toxicity, increased susceptibility to malaria, increased risk association with certain chronic diseases (e.g., diabetes)	45 mg/d of elemental iron (gastrointestinal side effects)
Manganese	Impaired growth and skeletal development, reproduction, lipid and carbohydrate metabolism; upper body rash	*General:* Neurotoxicity, Parkinson-like symptoms *Occupational:* Encephalitis-like syndrome, Parkinson-like syndrome, psychosis, pneumoconiosis	11 mg/d (neurotoxicity)
Molybdenum	Severe neurologic abnormalities	Reproductive and fetal abnormalities	2 mg/d (extrapolated from animal data)
Selenium	Cardiomyopathy, heart failure, striated muscle degeneration	*General:* Alopecia, nausea, vomiting, abnormal nails, emotional lability, peripheral neuropathy, lassitude, garlic odor to breath, dermatitis *Occupational:* Lung and nasal carcinomas, liver necrosis, pulmonary inflammation	400 μg/d (hair, nail changes)
Phosphorus	Rickets (osteomalacia), proximal muscle weakness, rhabdomyolysis, paresthesia, ataxia, seizure, confusion, heart failure, hemolysis, acidosis	Hyperphosphatemia	4000 mg/d
Zinc	Growth retardation, ↓ taste and smell, alopecia, dermatitis, diarrhea, immune dysfunction, failure to thrive, gonadal atrophy, congenital malformations	*General:* Reduced copper absorption, gastritis, sweating, fever, nausea, vomiting *Occupational:* Respiratory distress, pulmonary fibrosis	40 mg/d (impaired copper metabolism)

agent or whether it increases cancer risk (e.g., prostate cancer). Selenocysteine is also found in the deiodinase enzymes, which mediate the deiodination of thyroxine to triiodothyronine (Chap. 375). Rich dietary sources of selenium include seafood, muscle meat, and cereals, although the selenium content of cereal is determined by the soil concentration. Countries with low soil concentrations include parts of Scandinavia, China, and New Zealand. *Keshan disease* is an endemic cardiomyopathy found in children and young women residing in regions of China where dietary intake of selenium is low (<20 μg/d). Concomitant deficiencies of iodine and selenium may worsen the clinical manifestations of cretinism. Chronic ingestion of large amounts of selenium leads to selenosis, characterized by hair and nail brittleness and loss, garlic breath odor, skin rash, myopathy, irritability, and other abnormalities of the nervous system.

CHROMIUM

Chromium potentiates the action of insulin in patients with impaired glucose tolerance, presumably by increasing insulin receptor–mediated signaling, although its usefulness in treating type 2 diabetes is uncertain. In addition, improvement in blood lipid profiles has been reported in some patients. The usefulness of chromium supplements in muscle building has not been substantiated. Rich food sources of chromium include yeast, meat, and grain products. Chromium in the trivalent state is found in supplements and is largely nontoxic; however, chromium-6 is a product of stainless steel welding and is a known pulmonary carcinogen as well as a cause of liver, kidney, and CNS damage.

MAGNESIUM

See Chap. 402.

FLUORIDE, MANGANESE, AND ULTRATRACE ELEMENTS

An essential function for fluoride in humans has not been described, although it is useful for the maintenance of structure in teeth and bones. Adult fluorosis results in mottled and pitted defects in tooth enamel as well as brittle bone (skeletal fluorosis).

Manganese and molybdenum deficiencies have been reported in patients with rare genetic abnormalities and in a few patients receiving prolonged total parenteral nutrition. Several manganese-specific enzymes have been identified (e.g., manganese superoxide dismutase). Deficiencies of manganese have been reported to result in bone demineralization, poor growth, ataxia, disturbances in carbohydrate and lipid metabolism, and convulsions.

Ultratrace elements are defined as those needed in amounts <1 mg/d. Essentiality has not been established for most ultratrace elements, although selenium, chromium, and iodine are clearly essential (Chap. 375). Molybdenum is necessary for the activity of sulfite and xanthine oxidase, and molybdenum deficiency may result in skeletal and brain lesions.

FURTHER READING

Boyce SG et al: Can composite nutritional supplement based on the current guidelines prevent vitamin and mineral deficiency after weight loss surgery? Obes Surg 26:966, 2016.

Colleen R et al: Revisiting vitamin C and cancer. Science 350:1317, 2015.

Day E et al: Thiamine for prevention and treatment of Wernicke-Korsakoff syndrome in people who abuse alcohol. Cochrane Database Syst Rev 7:CD004033, 2013.

Imdad A et al: Vitamin A supplementation for preventing morbidity and mortality in children from six months to five years of age. Cochrane Database Syst Rev 3:CD008524, 2017.

Stevens GA et al: Trends and mortality effects of vitamin A deficiency in children in 138 low-income and middle-income countries between 1991 and 2013: A pooled analysis of population-based surveys. Lancet Glob Health 3:e528, 2015.

Vinceti M, Rothman KJ: More results but no clear conclusion on selenium and cancer. Am J Clin Nutr 104:245, 2016.

WHO: Guideline: Vitamin A supplementation in pregnant women. Geneva, World Health Organization, 2011.

327 Malnutrition and Nutritional Assessment

Gordon L. Jensen

Malnutrition occurs in 30–50% of hospitalized patients depending upon the setting and criteria that are used. Poor wound healing, compromised immune status, impaired organ function, increased length of hospital stay, and increased mortality are among the notable adverse outcomes associated with malnutrition. It is now widely appreciated that acute or chronic inflammation contribute to the pathophysiology of disease-related or injury-related malnutrition. The presence of inflammation can also render historic nutrition assessment indicators, like albumin and prealbumin, unreliable and inflammation diminishes favorable responses to nutrition therapies. In order to guide appropriate care, it is necessary to properly assess and diagnose malnutrition. Nutrition assessment is a comprehensive evaluation to diagnose a malnutrition syndrome and to guide intervention and expected outcomes. Patients are often targeted for assessment after being identified at nutritional risk based upon screening procedures conducted by nursing or nutrition personnel within 24 h of hospital admission. Screening tends to focus explicitly on a few risk variables like weight loss, compromised dietary intake, and high risk medical/surgical diagnoses. Preferably, health professionals complement this screening with a systematic approach to comprehensive nutrition assessment that incorporates an appreciation for the contributions of inflammation that serve as the basis for new approaches to the diagnosis and management of malnutrition syndromes.

MALNUTRITION SYNDROMES

Famine and starvation have long been leading causes of malnutrition and remain so in developing countries. However, with improvements in agriculture, education, public health, healthcare, and living standards, malnutrition in the settings of disease, surgery, and injury has become a prevalent concern throughout the world. Malnutrition now encompasses the full continuum of undernutrition and over-nutrition (obesity). For the objectives of this chapter, we will focus upon the former. Historic definitions for malnutrition syndromes are problematic in their use of diagnostic criteria that suffer poor sensitivity, sensitivity, and inter-observer reliability. Definitions overlap and confusion and misdiagnosis are frequent. In addition, some approaches do not recognize undernutrition in obese persons. While the historic syndromes of marasmus, kwashiorkor, and protein-calorie malnutrition remain in use, this chapter will instead highlight new insights to the diagnosis of malnutrition syndromes.

The Subjective Global Assessment, a comprehensive nutrition assessment that included a metabolic stress of disease component, was described and validated in the 1980s. In 2010, an International Consensus Guideline Committee incorporated a new appreciation for the role of inflammatory response into their proposed nomenclature for nutrition diagnosis in adults in the clinical practice setting. *Starvation-associated malnutrition*, when there is chronic starvation without inflammation (anorexia nervosa or major depression with lack of interest in eating), *chronic disease-associated malnutrition*, when inflammation is chronic and of mild to moderate degree (e.g., organ failure, pancreatic cancer, or sarcopenic obesity), and *acute disease or injury-associated malnutrition*, when inflammation is acute and of severe degree (e.g., major infection, burns, trauma, or closed head injury). In 2012, the Academy of Nutrition and Dietetics and the American Society for Parenteral and Enteral Nutrition (ASPEN) extended this approach with corresponding nomenclature that included malnutrition in the setting of social and environmental circumstances, malnutrition in the setting of chronic illness, and malnutrition in the setting of acute illness or injury. Clinical characteristics were proposed to support a diagnosis that encompasses the presence of illness or injury, poor food intake, weight loss, and physical findings of fat loss, muscle loss, edema, or reduced grip strength. In 2016, the European Society for Parenteral and Enteral Nutrition (ESPEN) formally

adopted an inflammation-based construct similar to these earlier approaches. Also in 2016, the Global Leadership Initiative on Malnutrition, a collaborative effort of ASPEN, ESPEN, the Latin American Federation of Parenteral and Enteral Nutrition, the Parenteral and Enteral Society of Asia, and other nutrition societies embarked on an effort to identify evidence-based criteria that will be disseminated throughout the world for use as dictated by regional preference.

Recent studies suggest that these newer approaches to diagnosis of malnutrition have similar utility in predicting adverse outcomes. This is not surprising since they share a number of common criteria including a metabolic stress of disease component that is a proxy indicator of inflammation. Irrespective of the approach that is selected, assessment of patients can be facilitated using the indicators of malnutrition and inflammation described below.

NUTRITION ASSESSMENT

There is unfortunately no single clinical or laboratory indicator of comprehensive nutritional status. Assessment therefore requires systematic integration of data from a variety of sources. Micronutrient deficiencies of clinical relevance may be detected in association with any of the malnutrition syndromes, but a detailed discussion of their assessment is beyond the scope of this chapter **(see Chap. 326)**. Physical findings characteristic of micronutrient deficiencies are however summarized in **Table 327-1**.

TABLE 327-1 History and Physical Examination Elements

ELEMENT	NOTES
Historical Data	
Body weight	Ask about usual weight, peak weight, and deliberate weight loss. A 4.5 kg (10-lb) weight loss over 6 months is noteworthy and a weight loss of >10% of usual body weight is prognostic of clinical outcomes. Use medical records, family, and caregivers as information resources.
Medical and surgical conditions; chronic disease	Look for medical or surgical conditions or chronic disease that can place one at nutritional risk secondary to increased requirements, or compromised intake or assimilation like: critical illness, severe burns, major abdominal surgery, multi-trauma, closed head injury, previous gastrointestinal surgery, severe gastrointestinal hemorrhage, enterocutaneous fistula, gastrointestinal obstruction, mesenteric ischemia, severe acute pancreatitis, chronic pancreatitis, inflammatory bowel disease, celiac disease, bacterial overgrowth, solid or hematologic malignancy, bone marrow transplant, acquired immune deficiency syndrome, and organ failure/transplant—kidney, liver, heart, lung, or gut. A number of conditions or diseases are characterized by severe acute inflammatory response including critical illness, major infection/sepsis, adult respiratory distress syndrome, systemic inflammatory response syndrome, severe burns, major abdominal surgery, multi-trauma, and closed head injury. Many conditions or diseases are more typically associated with mild to moderate chronic inflammatory response. Examples include cardiovascular disease, congestive heart failure, cystic fibrosis, inflammatory bowel disease, celiac disease, chronic pancreatitis, rheumatoid arthritis, solid tumors, hematologic malignancies, sarcopenic obesity, diabetes mellitus, metabolic syndrome, cerebrovascular accident, neuromuscular disease, dementia, organ failure/transplant—kidney, liver, heart, lung, or gut, periodontal disease, pressure wounds, and chronic obstructive pulmonary disease. Note that acute exacerbations, infections, or other complications may superimpose acute inflammatory response on such conditions or diseases. Examples of starvation-associated conditions that generally have little or no inflammatory component include anorexia nervosa or compromised intake in the setting of major depression.
Constitutional signs/symptoms	Fever or hypothermia can indicate active inflammatory response. Tachycardia is also common. Anorexia is another manifestation of inflammatory response and is also often a side effect of treatments and medications.
Eating difficulties/gastrointestinal complaints	Poor dentition or problems swallowing can compromise oral intake. Vomiting, nausea, abdominal pain, abdominal distension, diarrhea, constipation, and gastrointestinal bleeding can be signs of gastrointestinal pathology that may place one at nutritional risk.
Eating disorders	Look for distorted body image, compulsive exercise, amenorrhea, vomiting, tooth loss, dental caries, and use of laxatives, diuretics or Ipecac.
Medication use	Many medications can adversely affect nutrient intake or assimilation. Review potential drug–drug and drug–nutrient interactions. A pharmacist consultant can be helpful.
Dietary practices and supplement use	Look for dietary practices including therapeutic, weight reduction, vegetarian, macrobiotic, and fad diets. Also record use of dietary supplements, including vitamins, minerals, and herbals. Ask about dietary intake. Recall, record, and food frequency tools are available. It is estimated that 50% or more of adults take dietary supplements.
Influences on nutritional status	Ask about factors such as living environment, functional status (activities of daily living and instrumental activities of daily living), dependency, caregiver status, resources, dentition, alcohol or substance abuse, mental health (depression or dementia), and lifestyle.
Physical Examination Data	
Body mass index (BMI)	BMI = weight in kg/(height in meters)^2 BMI <18.5 kg/m^2 proposed screen for malnutrition per National Institutes of Health guidelines. BMI ≤15 kg/m^2 or less is associated with increased mortality. Comparison with ideal body weight for stature can also be determined from reference tables. Note hydration status and edema at the time body weight is determined.
Weight loss	Look for loss of muscle mass and subcutaneous fat. Temporal and neck muscle wasting may be readily observed. Anthropometrics including skin-folds and circumferences can be useful but require training to achieve reliability.
Weakness/loss of strength	Decreased hand-grip and leg extensor strength have been related to loss of muscle mass in malnourished states. Lower extremity weakness may be observed in thiamine deficiency.
Peripheral edema	Peripheral edema may confound weight measurements and is often observed with reduced visceral proteins as well as inflammatory states. Edema may also be observed with thiamine deficiency.
Hair examination	Hair findings are indicative of certain nutrient deficiencies. Loss: protein, B12, folate Brittle: biotin Color change: zinc Dry: vitamins A and E Easy pluckability: protein, biotin, zinc Coiled, corkscrew: vitamins A and C Alopecia is common in severely malnourished persons. Ask about excessive hair loss on pillow or when combing hair.

(*Continued*)

TABLE 327-1 History and Physical Examination Elements (Continued)

ELEMENT	NOTES
Skin examination	Skin findings are indicative of certain nutrient deficiencies. Desquammation: riboflavin Petechiae: vitamins A and C Perifollicular hemorrhage: vitamin C Ecchymosis: vitamins C and K Xerosis, bran-like desquamation: essential fatty acid Pigmentation, cracking, crusting: niacin Acneiform lesions, follicular keratosis, xerosis: vitamin A Acro-orificial dermatitis, erythematous, vesiculbullous, and pustular: zinc Characteristic nutritional dermatitis and skin findings may be observed with a number of nutrient deficiencies. Wounds and pressure sores should also be noted as indicators of compromised nutritional status.
Eye examination	Ocular findings are indicative of certain nutrient deficiencies. Bitot's spots: vitamin A Xerosis: vitamin A Angular palpebritis: riboflavin Also ask about difficulties with night vision/night blindness; indicates vitamin A deficiency.
Perioral examination	Perioral findings are indicative of certain nutrient deficiencies. Angular stomatitis and cheilosis: B complex, iron, and protein Glossitis: niacin, folate, and vitamin B12 Magenta tongue: riboflavin Bleeding gums, gingivitis, tooth loss: vitamin C Angular stomatitis, cheilosis, and glossitis are associated with vitamin and mineral deficiencies. Note poor dentition, caries, and tooth loss. Difficulty swallowing and impairment of gag should also be recognized.
Extremity examination	Extremity findings indicate certain nutrient deficiencies Arthralgia: vitamin C Calf pain: thiamine Extremities may also exhibit loss of muscle mass and/or peripheral edema. Neurological findings in the extremities may also result from deficiencies described below.
Mental status/ nervous system examination	Mental and nervous system findings indicate certain nutrient deficiencies. Ophthalmoplegia and foot drop: thiamine Paresthesia: thiamine, vitamin B12, and biotin Depressed vibratory and position senses: vitamin B12 Anxiety, depression, and hallucinations: niacin Memory disturbance: vitamin B12 Hyporeflexia, loss of lower extremity deep tendon reflexes: thiamine and vitamin B12 Conduct formal cognitive and depression assessments as appropriate. Dementia and depression are common causes of malnutrition among older persons. Wernicke-Korsakoff syndrome may be observed with severe thiamine deficiency.
Functional assessment	Observe and test physical performance as indicated: gait, chair stands, stair steps, and balance. These provide complex measures of integrated neurological status, coordination, and strength.

Source: Adapted with permission from G Jensen: *Nutritional Syndromes. Smart Medicine/PIER.* Philadelphia, American College of Physicians, 2013.

Medical/Surgical History and Clinical Diagnosis Knowledge of a patient's medical/surgical history and associated clinical diagnoses is especially helpful in discerning the likelihood of malnutrition and inflammation. Non-volitional weight loss is a well validated nutrition assessment indicator and is often also associated with underlying disease or inflammatory condition. The degree and duration of weight loss determine its clinical significance. A 10% loss of body weight over 6-months is of clinical relevance, while a 30% loss of body weight over the same duration is severe and life-threatening. Since weight loss history is often unavailable or unreliable, one should query the patient as well as the medical records, family, and caregivers as appropriate to secure a valid weight trajectory.

A number of conditions or diseases are characterized by severe acute inflammatory response whereas others are more typically associated with a chronic inflammatory response that is mild to moderate in severity and may be relapsing and remitting (Table 327-1). It is also common for acute inflammatory events to be superimposed on those with chronic conditions; for example, a patient with chronic renal disease is admitted to the hospital with sepsis. The inflammatory milieu, especially when severe, may modify nutrient requirements by elevating resting energy expenditure and promoting muscle catabolism and nitrogen losses. Inflammation also promotes anorexia, decreasing food intake and further compromising nutritional status. A deteriorating course may result because the presence of inflammation may reduce the benefit of nutritional interventions and the associated malnutrition may in turn diminish the effectiveness of medical therapies. It is also imperative to recognize medical/surgical conditions or diseases that place the patient at increased risk to become malnourished because they have increased nutritional requirements, or compromised intake or assimilation (Table 327-1).

Nutrition assessment should also include a review of medications with attention to undesirable side effects including anorexia, xerostomia, nausea, diarrhea, and constipation. Potential drug/nutrient interactions should also be identified.

Clinical Signs and Physical Examination Nonspecific clinical indicators of inflammation include fever, hypothermia, and tachycardia. The nutrition-focused physical examination should identify edema as well as signs of weight gain/loss and specific nutrient deficiencies. Thorough examination should be particularly directed to those parts of the body where high cell turnover occurs (e.g., hair, skin, mouth, tongue) as they are most likely to exhibit observable signs of

nutritional deficiencies (Table 327-1). Physical findings of weight loss associated with decreased muscle and subcutaneous fat mass should not be overlooked, but when appreciable edema is present, these changes may not be readily appreciated.

Anthropometric Data Body weight measurements are recommended with each clinic visit or hospitalization so that a reliable weight change trajectory may be monitored. Patients should be weighed in a consistent manner without over-garments or shoes. In order to secure valid measurements, calibration of scales and appropriate staff training are essential. Chair or bed scales may be used for those who cannot stand. For those who are able, height should be measured in a standing position without shoes using a stadiometer. If an adult cannot safely stand, height can be estimated by doubling the arm span measurement (from the patient's sternal notch to the end of the longest finger). Stature of frail older persons can also be estimated from measurement of knee height using a caliper device.

Body weight is often standardized for height to obtain an ideal weight for comparison, but available reference tables require subjective assessment of frame size and offer limited reference data for many relevant population groups, including older persons. A simple measure of body size and an indirect measure of body fatness is provided by body mass index (BMI), defined as weight (kg) / height (m^2). The National Institutes of Health BMI categories for adults are: BMI <18.5 = underweight, BMI 18.5–24.9 = desirable, BMI 25.0–29.9 = overweight, and BMI ≥30 = obese. A higher desirable BMI range for persons 65 years of age and older has been proposed by the Centers for Medicare and Medicaid Services in its quality indicators system: BMI ≥23 and <30.

While classical anthropometric measurements including skin-folds and circumferences can be helpful, their utility in routine patient care has been limited because practitioner training is required to achieve suitable reliability. Body composition assessment methodologies include bioelectrical impedance analysis (BIA), dual-energy x-ray absorptiometry (DEXA), computed tomography (CT), and magnetic resonance imaging (MRI). The imaging modalities have become the state of the art for precise measurements of muscle mass. It is possible to take advantage of CT or MRI studies that are being done for other clinical purposes to evaluate musculature.

Laboratory Indicators Laboratory findings **(Table 327-2)** are but one part of the comprehensive nutrition assessment and must be used in combination with other domains of assessment to appropriately diagnose a malnutrition syndrome. Although serum albumin or prealbumin are often measured in patients with suspected malnutrition, their utility is limited due to their poor sensitivity and specificity as indicators of nutritional status. Patients with low albumin or prealbumin may or

TABLE 327-2 Body Composition, Laboratories, and Other Studies

TEST	NOTES
Body Composition Studies	
Anthropometrics	Skin folds and circumferences require training for reliability. Typical coefficient of variation is ≥10%.
Bioelectrical impedance	Based upon differential resistance of body tissues. Equipment easily portable. Good measure of body water. Requires population specific validation of regression equations.
Water displacement	Impractical for most clinical settings. Weighed in water tank. Historic reference measure.
Whole body counting and isotope dilution techniques	Research methodologies. Naturally occurring ^{40}K isotope to measure body cell mass by whole body counting. Total body water measurement by dilution volume of tritium, deuterium, or ^{18}O-labeled water.
Air plethysmography	Research methodology. Subject sits inside moderately sized BodPod chamber. Validated against water displacement and impedance.
Dual energy x-ray absorptiometry (DEXA)	Often used for bone density but can be used for soft tissue measurements with appropriate software. Can compare truncal and appendicular components. Modest x-ray exposure.
Imaging with computed tomography (CT) or magnetic resonance imaging (MRI)	State of the art research methods for visualizing body tissue compartments. Can quantify visceral fat. Costly and CT entails X-ray exposure.
Laboratories and Other Studies	
Albumin	Lacks sensitivity and specificity for malnutrition. Potent risk indicator for morbidity and mortality. Proxy measure for underlying injury, disease or inflammation. Half-life is 14–20 days. Also consider liver disease, nephrotic syndrome, and protein-wasting enteropathy.
Prealbumin	Sensitive to short-term changes in inflammation and protein nutrition with half-life of 2–3 days. Otherwise suffers the same limitations of albumin with limited sensitivity and specificity for malnutrition. Levels may be decreased in liver failure and increased in renal failure.
Transferrin	Acute phase reactant also altered by perturbation in iron status. Half-life is 8–10 days. Lacks sensitivity and specificity for malnutrition.
Retinol-binding protein	Responds to very short-term changes in nutritional status but utility is also limited by response to stress and inflammation. Half-life is 12 h. Also affected by vitamin A deficiency and renal disease.
C-reactive protein	C-reactive protein is a positive acute phase reactant. It is generally elevated if an active inflammatory process is manifest.
Cholesterol	Low cholesterol (<160 mg/dL) is often observed in malnourished persons with serious underlying disease. It is unrelated to dietary intake in many clinical settings. Increased complications and mortality are observed. It appears that low cholesterol is again a nonspecific feature of poor health status that reflects cytokine-mediated inflammatory condition. Vegans and patients with hyperthyroidism may also exhibit low cholesterol.
Carotene	Nonspecific indicator of malabsorption and poor nutritional intake.
Cytokines	Research is exploring prognostic use of cytokine measurements as indicators of inflammatory status.
Electrolytes, blood urea nitrogen, creatinine, and glucose	Monitor for abnormalities consistent with under- or over-hydration status and purging (contraction alkalosis). BUN may also be low in the setting of markedly reduced body cell mass. Blood urea nitrogen and creatinine are elevated in renal failure. Hyperglycemia may be nonspecific indicator of inflammatory response.
Complete blood count with differential	Screen for nutritional anemias (iron, B12, and folate), lymphopenia (malnutrition) and thrombocytopenia (vitamin C and folate). Leukocytosis may be observed with inflammatory response.
Total lymphocyte count	Relative lymphopenia (total lymphocyte count <1200/mm^3) is a nonspecific marker for malnutrition.
Helper/suppressor T cell ratio	Ratio may be reduced in severely undernourished patients. Not specific for nutritional status.

(*Continued*)

TABLE 327-2 Body Composition, Laboratories, and Other Studies (Continued)

TEST	NOTES
Nitrogen balance	24-h urine can be analyzed for urine urea nitrogen (UUN) to determine nitrogen balance and give indication of degree of catabolism and adequacy of protein replacement. Requires accurate urine collection and normal renal function. Nitrogen balance = (protein/6.25) – (UUN+4). Generally negative in the setting of acute severe inflammatory response.
Urine 3-methylhistidine	Indicator of muscle catabolism and protein sufficiency. Released upon breakdown of myofibrillar protein and excreted without reutilization. Urine measurement requires a meat-free diet for 3 days prior to collection.
Creatinine height index (CHI)	CHI = (24-h urinary creatinine excretion/ideal urinary creatinine for gender and height) × 100. Indicator of muscle depletion. Requires accurate urine collection and normal renal function.
Prothrombin time/international normalized ratio (INR)	Nonspecific indicator of vitamin K status. Prolonged in liver failure.
Specific micronutrients	When suspected a variety of specific micronutrient levels may be measured: thiamine, riboflavin, niacin, folate, pyridoxine, vitamins A, C, D, E, B12, zinc, iron, selenium, carnitine, and homocysteine—indicator of B12, folate, and pyridoxine status.
Skin testing—recall antigens	Delayed hypersensitivity testing. While malnourished patients are often anergic, this is not specific for nutritional status.
Electrocardiogram	Severely malnourished patients with reduced body cell mass may exhibit low voltage and prolonged QT interval. These findings are not specific for malnutrition.
Video fluoroscopy	Helpful to evaluate suspected swallowing disorders.
Endoscopic and x-ray studies of gastrointestinal tract	Useful to evaluate impaired function, motility, and obstruction.
Fat absorption	72-h fecal fat can be used to quantitate degree of malabsorption.
Schilling test	Identify the cause for impaired vitamin B12 absorption.
Indirect calorimetry	Metabolic cart can be used to determine resting energy expenditure (REE) for accurate estimation of energy needs. Elevated REE is a sign of systemic inflammatory response.

Source: Adapted with permission from G Jensen: *Nutritional Syndromes. Smart Medicine/PIER*. Philadelphia, American College of Physicians, 2013.

may not prove to be malnourished when evaluated by comprehensive nutrition assessment because these proteins are readily reduced by the systemic response to injury, disease or inflammation. C-reactive protein is a positive acute phase reactant that may be measured to help discern whether active inflammation is manifest. If C-reactive protein is increased, and albumin or prealbumin decreased, then inflammation is likely to be a contributing factor. Since it is recognized that C-reactive protein suffers limitations as a point in time measure, trends in levels over the clinical course may be helpful. Research suggests that interleukin 6, and perhaps other cytokines, may also offer promise as indicators of inflammatory status. Nonspecific laboratory indicators that are often associated with inflammatory response include leukocytosis and hyperglycemia. Additional tests that may be obtained to help confirm the presence of inflammatory response include 24-h urine urea nitrogen and indirect calorimetry. In the setting of severe acute systemic inflammatory response, negative nitrogen balance and elevated resting energy expenditure are anticipated.

Dietary Assessment Dietary assessment can be used to detect inadequate or imbalanced food or nutrient intakes. While dietary assessment in patient care settings can be quite challenging, 24-h recall and modified diet history approaches are sometimes used. A modified diet history is targeted to query types and frequencies of intake of specific foods of interest. It is often necessary to access diverse resources for diet history information including the patient, medical records, family, and caregivers. Consultation of a registered dietitian nutritionist is highly recommended. Dietary practices and supplements should be carefully reviewed for potential inadequacies and toxicities. Since patients will often present to healthcare practitioners with acute medical events superimposed upon chronic health conditions, it is common for patients to have had decreased food intakes and malnutrition for extended periods prior to assessment. It is therefore imperative that compromised dietary intake should not be overlooked so that appropriate intervention may be undertaken.

Ongoing assessment is indicated when parenteral or enteral feedings are initiated, because it is necessary to discern what amount of formula is actually being administered to and received by the patient. Enteral feedings, in particular, often interrupted or held for procedures, tolerance issues, and feeding tube displacements. It is therefore not unusual for such patients to be appreciably underfed for extended periods. When a patient is beginning to transition to oral feedings, it is imperative to monitor quantities of food and/or supplements that are actually consumed as well as patient tolerance to feeding. Meals are often delayed or missed for tests or procedures. If possible, the patient should be queried about intake since tray inspection is notoriously unreliable as an indicator of consumption.

Functional Outcomes Advanced malnutrition is accompanied by declines in muscle mass and function that can be detected by strength and physical performance measures. Hand-grip strength measured with a simple handgrip dynamometer is the most practical routine clinical assessment. Physical performance tests like timed gait, chair stands, and stair steps are used in the comprehensive assessment of integrated functions in frail older persons.

The decline in overall functional status observed in advanced malnutrition is associated with nutrient deficiencies and impairment of organ system functions. Poor wound healing and immune compromise are examples of such impairments. Improved wound healing parameters and restored responsiveness to recall antigens by delayed hypersensitivity testing may be measured to demonstrate improvements with nutritional repletion, though it must be appreciated that these are multivariable outcomes for which improved nutritional status is but one variable.

FURTHER READING

Cederholm T et al: ESPEN guidelines on definitions and terminology of clinical nutrition. Clin Nutr 36:49, 2017.

Cederholm T, Jensen GL: To create a consensus on malnutrition diagnostic criteria. JPEN J Parenter Enteral Nutr 41:311, 2017.

Detsky AS et al: What is Subjective Global Assessment of nutritional status? J Parenter Enteral Nutr 11:8, 1987.

Guerra RS et al: Usefulness of six diagnostic and screening measures for undernutrition in predicting length of hospital stay: A comparative analysis. J Acad Nutr Diet 115:927, 2015.

Jensen GL: Inflammation as the key interface of the medical and nutrition universes: A provocative examination of the future of clinical nutrition and medicine. J Parenter Enteral Nutr 30:453, 2006.

Jensen GL: Malnutrition and inflammation—"Burning down the house": Inflammation as an adaptive physiologic response versus self-destruction? J Parenter Enteral Nutr 39:56, 2015.

Jensen GL: Global leadership conversation: Addressing malnutrition. J Parenter Enteral Nutr 40:455, 2016.

Jensen GL, Hsiao PY, Wheeler D: Adult nutrition assessment tutorial. J Parenter Enteral Nutr 36:267, 2012.
Jensen GL et al: Adult starvation and disease-related malnutrition: A proposal for etiology-based diagnosis in the clinical practice setting from the International Consensus Guideline Committee. J Parenter Enteral Nutr 34:156, 2010.
White JV et al: Consensus statement: Academy of Nutrition and Dietetics and American Society for Parenteral and Enteral Nutrition. Characteristics recommended for the identification and documentation of adult malnutrition (under-nutrition). J Parenter Enteral Nutr 36:275, 2012.

328 Enteral and Parenteral Nutrition

L. John Hoffer, Bruce R. Bistrian, David F. Driscoll

There are three kinds of specialized nutritional support (SNS): (1) optimized voluntary nutritional support, which is used when a patient's barriers to adequate nutrition can be overcome by special attention to the details of how their food is constituted, prepared, served, and its consumption monitored, (2) forced enteral nutrition (EN), in which a liquid nutrient formula is delivered through a tube placed in the stomach or small intestine, and (3) parenteral nutrition (PN), in which a nutritionally complete mixture of crystalline amino acids, dextrose, triglyceride emulsions, minerals, electrolytes, and micronutrients is infused directly into the bloodstream.

When does a hospitalized patient need SNS? When SNS is indicated, how should it be provided? This chapter reviews the physiological principles that underlie the correct use of SNS, and provides practical information about the diagnosis and management of nutritional disorders in adult hospitalized patients.

The management of in-hospital nutritional disorders follows 3 steps: (1) screening and diagnosis; (2) determination of the severity and urgency of treating a diagnosed nutritional disorder in its overall clinical context; and (3) selection of the modality of SNS, its composition, and the details of providing it. To follow these steps effectively, physicians require a general understanding of nutritional physiology, nutrient requirements, the pathophysiology and diagnosis of the nutritional disorders, and familiarity with the indications, advantages, risks, and administration of the different kinds of SNS. Because most physicians are incompletely trained in clinical nutrition, they must collaborate with clinical dietitians and specialized pharmacists when ordering EN and PN.

NUTRITIONAL PHYSIOLOGY

(See Chaps. 325–327)

Energy Total energy expenditure (TEE) is comprised of resting energy expenditure (REE, ~24 kcal/kg normal adult body weight/day), activity energy expenditure (~12 kcal/kg in healthy sedentary individuals) and the thermic effect of food (10% of TEE). The TEE of a healthy adult is ~36 kcal/kg. REE can be measured by indirect calorimetry or estimated using a variety of predictive equations that input weight, height, age, sex, and sometimes disease-related factors. Fever and some forms of critical illness increase REE. Prolonged semi-starvation normally triggers an adaptive reduction in REE, voluntary physical activity, and the thermic effect of food. Broadly speaking, a patient's TEE identifies the amount of dietary energy they have to consume and metabolize to maintain their existing store of body fat (and protein). The amount of energy they actually *require* may be less than TEE (as in obesity therapy) or greater than TEE (when rehabilitating nutritionally depleted patients).

Protein and Amino Acids Protein is an essential nutrient because whole body protein turnover—a continuous process of protein synthesis and breakdown to its constituent amino acids—is associated with obligatory amino acid catabolism to carbon dioxide, water, carbohydrates, ammonium, urea, and sulfuric acid. Amino acid catabolism can be adaptively reduced when protein intake decreases, but not below a certain lowest rate known as the protein minimum. After adjustment for the inefficiency of exogenous protein retention, this rate defines the minimum dietary protein intake necessary to maintain whole body protein homeostasis and zero nitrogen (N) balance. Proteins of nutritional interest, including the ones that form the metabolically active tissues of the body known as the body cell mass (BCM), are 16% nitrogen (N) by weight. Skeletal muscle makes up ~80% of the BCM and 40–50% of the body weight of normal-weight young adults. The excretion of 1 g N from the body—mostly as urinary urea N—implies the loss of 6.25 g formed protein and (since the BCM is ~80% tissue water) ~31 g of BCM, almost of all it as muscle tissue.

The lowest daily protein intake compatible with zero N balance—approximately 0.65 g/kg body weight—is the average minimum protein requirement of a healthy adult. To account for inter-individual variability, two standard deviations are added to calculate the "safe" or "recommended" daily minimum protein intake of 0.80 g/kg normal body weight. Protein intakes greater than the minimum requirement are usually inconsequential, as surfeit amino acids are normally easily catabolized. The average protein consumption in wealthy societies is approximately twice the average minimum requirement.

Many diseases (or their treatment) increase the dietary protein requirement, by (1) increasing amino acid loss from the body (as in malabsorption and protein loss via wound exudates, fistulas or inflammatory diarrhea), removing amino acids from the circulation (renal replacement therapy), or (2) increasing muscle protein catabolism (as with high-dose glucocorticoid therapy). Protein-catabolic critically ill patients often excrete 15 g N/day or more in their urine in the absence of dietary protein provision. This rate of N loss implies a loss of 15 × 6.25 = 94 g muscle protein—equivalent to one pound of muscle lost from the body every day. Sufficiently generous protein provision can minimize this kind of muscle atrophy, and there is widespread agreement that protein-catabolic patients require much more dietary protein requirement than healthy people. The exact magnitude of the increase has not yet been determined, but the most frequent recommendation for patients with protein-catabolic diseases is 1.5 g protein/kg normal body weight/day, close to the habitual protein intake of healthy people in wealthy societies.

Protein-Energy Interaction Energy deficiency—whether deliberate, as in weight reduction therapy, or inadvertent, as frequently occurs in hospitalized patients—increases amino acid loss from muscle and increases the dietary protein requirement. The mechanism responsible for energy deficiency's protein-wasting effect differs from the one that mediates inflammation-induced muscle atrophy, and the interaction between these different processes is imperfectly understood. It does appear, however, that systemic inflammation diminishes, but doesn't prevent the protein-sparing effect of generous protein provision as long as the protein is combined with some exogenous energy (e.g., 50% of TEE). Energy provision more generous than ~50–70% of TEE exerts little further protein-sparing effect in this situation, and the additional amounts of glucose and fluid volume required to provide it often have adverse effects.

Micronutrients Minimum amounts of nine water-soluble vitamins (the B vitamins and vitamin C) and four fat-soluble vitamins (A, D, E, and K), eight minerals (calcium, phosphate, potassium, sodium, chloride, magnesium, zinc, and iron), essential fatty acids, and several essential trace elements (notably including selenium, copper, and iodine) are required throughout life to avoid deficiency diseases and death. Patients who have been hospitalized for more than a few days commonly consume inadequate amounts of food and the micronutrients it provides.

Overt deficiencies of potassium, sodium, magnesium, phosphate, and iron occur so often in hospitalized patients that it is standard

practice to monitor for and correct them. Many drugs used in acute-care medicine induce renal potassium, magnesium, or zinc wasting that necessitate appropriate increases in their provision. Gastrointestinal losses from nasogastric drainage tubes or intestinal losses from fistulas or diarrhea incur losses of potassium, sodium, calcium, magnesium, and zinc which add to their normal daily requirement.

Less studied, but nonetheless common, are subclinical or unrecognized deficiencies of calcium, zinc, vitamin D, vitamin C, and possibly other micronutrients. Physicians often assume that consumption of the regular hospital diet will protect patients from these deficiencies. This assumption is not warranted when the patient's nutritional status was borderline or deficient when they were admitted to hospital and remains inadequate throughout their hospital stay. These patients are at risk of a variety of micronutrient deficiency diseases in addition to the symptoms and disability created by continuing in-hospital starvation.

PROTEIN-ENERGY MALNUTRITION AND ITS VARIANTS

The decision to embark on SNS should be justified by a well-formulated nutritional diagnosis and explicitly defined therapeutic goal. This chapter focuses on the diagnosis, treatment, and prevention of in-hospital *protein-energy malnutrition* (PEM). PEM is the disease caused by prolonged inadequate energy and protein consumption—starvation—with consequent depletion of the BCM and body fat. The pathologic features that differentiate PEM—which is a disease—from the physiological process of starvation that leads to it, emerge when the BCM has become depleted seriously enough to impair specific physiological functions. There are many synonyms for simple, starvation-induced PEM: starvation disease, hunger disease, inanition and, most recently, starvation-related malnutrition (SM).

The body normally adapts to starvation by reducing energy expenditure and curtailing protein catabolism, partly by hormone- and nervous system-regulated alterations in cellular metabolism, and partly by reducing its muscle mass. These adaptations enable prolonged survival during sub-lethal starvation, but survival comes at a cost that includes lethargy, a tendency to hypothermia, muscle atrophy (including of the cardiac and respiratory muscles), skin thinning, and functional disability. The cardinal diagnostic features of PEM—generalized muscle atrophy and subcutaneous adipose tissue depletion—are easy to detect by simple physical examination. Too often, however, in-hospital PEM remains undiagnosed, partly because of health care worker unawareness and inattention, and partly because PEM overlaps and is often confused with other common conditions that also cause muscle atrophy.

Terminology The terms used to describe nutritional disorders are often ambiguous and confusing. "Malnutrition" is a blanket term that indiscriminately refers to the sum total of the nutritional environments that give rise to starvation, the physiological process of starvation, and the PEM that may result from it. Until this terminology is better standardized, we suggest that health care workers explicitly distinguish among (1) situations that create a risk of inadequate nutrient intake, (2) situations in which inadequate intake of specific nutrients actually occurs and creates a discernable risk of developing a specific nutritional disease, and (3) the specific diseases themselves, as enumerated below.

Starvation-Related Malnutrition (Uncomplicated Protein-Energy Malnutrition) PEM is diagnosed when the physical examination reveals generalized muscle atrophy and diminished (but not necessarily depleted) subcutaneous adipose tissue due to prolonged inadequate food consumption or malabsorption. PEM is always associated with weight loss, but weight loss alone may not reveal its full severity, because extracellular fluid (ECF) volume and hence body weight increase, sometimes seriously enough to cause edema. SM is "uncomplicated" PEM due solely to prolonged starvation. Adaptation is an important feature of SM that increases the likelihood of survival by reducing energy expenditure and slowing body protein turnover, thus reducing or halting the loss of body protein and fat. The risk of complications, including death, increases as the depletion of BCM worsens. A 50% depletion of BCM puts otherwise uncompromised young adults at the cusp of survival; older patients with co-morbidities are at even greater risk. People with SM feel unwell, lack strength, are frail, and are at risk of hypothermia. The severity of SM is revealed by physical examination and an evaluation of the patient's strength and physical function.

The main cause of SM worldwide is involuntary food deprivation. The causes of SM in hospitalized patients are many. They include inadvertent or physician-ordered food deprivation, psychologic depression or distress, poorly controlled pain or nausea, badly presented unappealing food, communication barriers, anorexia nervosa, physical or sensory disability (including dysgeusia), thrush, dysphagia and other mechanical difficulties ingesting food, partial obstruction of the esophagus, stomach or intestinal tract, intestinal angina, and very commonly, combinations of these causes.

Chronic Disease-Related Malnutrition and Cachexia These two terms refer to starvation-induced PEM that is complicated by chronic systemic inflammation. Chronic disease-related malnutrition (CDM) is highly prevalent among patients with chronic infections, inflammatory autoimmune diseases, chronic severe hepatic, renal, cardiac and pulmonary disease, and certain inflammatory cancers. CDM both causes and is worsened by anorexia: a strong disinclination to eat much food even when there is no physical barrier to eating. CDM is characterized by a moderately increased rate of muscle protein catabolism, muscle atrophy and weakness, fatigue and reduced voluntary activity, and a subverted adaption to starvation, all of which contribute to a vicious cycle of worsening CDM. Fortunately, the nutrient deficit on the input side (anorexia-driven inadequate food consumption) is often a stronger driver of the patient's CDM than increased nutrient loss on the output side (increased amino acid catabolism and REE), thus making CDM amenable to nutritional intervention while effective treatment of the primary disease is underway. The anorexia of CDM is less inhibiting to people who have a hearty pre-morbid appetite and obesity-prone phenotype than it is to previously thin, habitual under-eaters.

Acute Disease-Related Malnutrition The term acute disease-related malnutrition (ADM) refers to a specific metabolic-nutritional environment that creates a very high risk of severe PEM, rather than an existing disease entity. A synonym for ADM is "catabolic critical illness." The intense systemic inflammation that accompanies severe tissue injury and sepsis causes extremely rapid and severe muscle atrophy (and increases REE to a variable extent) in a setting in which voluntary food intake is impossible. People with ADM are usually treated in intensive care units. They may or may not have PEM at the onset of ADM, but it will surely develop within days to a few weeks unless their medical or surgical disease is rapidly and effectively treated and SNS is appropriately provided.

NUTRITIONAL DIAGNOSIS

The cardinal anatomic features of PEM—generalized muscle atrophy and diminished body fat—are revealed by a discerning physical examination, but what ought to be an easy diagnosis is commonly missed. This section explains the details and pitfalls of diagnosing PEM and judging its severity.

Muscle Mass Once the examiner's attention has been drawn to it, generalized muscle atrophy is easy to identify and its severity determinable almost at a glance. The problem with diagnosing PEM in the hospital setting, apart from simple inattention, is that generalized muscle atrophy has many causes: old age-related muscle atrophy (sarcopenia), disuse muscle atrophy due to reduced mobility and bed rest, high-dose glucocorticoid therapy, certain endocrine diseases, and primary muscle or neuromuscular diseases. Indeed, muscle atrophy is so common among hospitalized patients that health care workers often mistakenly regard it is a usual, even defining, characteristic of the patient's primary disease. In reality, muscle atrophy is very commonly at least partly due to SM or CDM. As such, it represents a potentially remediable *complication* of the patient's primary disease rather than a necessary *feature* of it. Whenever a health care worker detects

generalized muscle atrophy, they should review the patient's overall situation and identify which of its potential causes are pertinent, and in particular, which of them are treatable and reversible. The combination of old age, disuse muscle atrophy, and starvation is very common. Old age is irreversible, but adequate protein and energy provision combined with physical rehabilitation can be lifesaving.

Muscle atrophy, no matter what its cause, is especially dangerous in ADM, because patients in this situation are closer to the cliff-edge of lethal BCM depletion. As well, their reduced muscle protein mass is unable to release amino acids into the circulation at a rate sufficient to meet the need for protein synthesis at sites of injury and healing, and within the central protein pool to regulate the immuno-inflammatory process.

Subcutaneous Adipose Tissue Severe adipose tissue depletion is sufficient to diagnose PEM, but it is not a *necessary* criterion. The modern obesity epidemic has created a population of obese patients with chronic inflammation and starvation whose muscle atrophy outpaces their fat loss. A targeted physical examination easily differentiates these patients' muscle groups from their subcutaneous fat, and reveals their moderate and sometimes severe PEM.

ECF Volume The ECF volume normally represents ~20% of body weight. Chronic starvation increases the ECF volume, occasionally enough to cause dependent edema ("starvation edema"). Hospitalized patients with CDM commonly have other edema-causing conditions, including hypoalbuminemia with its associated reduction in plasma oncotic pressure. Unless recognized and accounted for, increased ECF volume can mask the true extent of muscle and adipose tissue depletion in patients with PEM.

Body Mass Index Body mass index (BMI) is defined as body weight (kg) divided by the square of height (m^2). Normal BMI ranges from 20 to 25 kg/m^2. BMI >25 usually indicates increased body fat; BMI <20 indicates decreased muscle mass and body fat. Survival during prolonged, severe starvation depends both on fat and protein stores. A BMI of 11–13 is usually incompatible with life. Some guidelines and clinical trial enrollment criteria define "malnutrition"—in this context a synonym for PEM—as a BMI <16 or 17. This oversimplification can lead to serious error. A BMI of 17 certainly is consistent with PEM, because the body architecture of such a BMI can only be created by jettisoning a large amount of BCM and adipose tissue. But a BMI >17 does not rule out PEM. Many patients with PEM have normal or above-normal BMIs due to residual obesity or an expanded ECF volume.

Visual BMI After some practice, health care workers can accurately predict the BMI of non-obese, non-edematous patients by examining their muscular architecture. Visual BMI is useful for quantifying and communicating the severity of a patient's PEM. Once acquired, this skill can be used to estimate the severity of PEM in obese or edematous patients—in whom measured BMI is unreliable—by focusing attention on their muscular architecture while simultaneously discounting their subcutaneous fat and edema. Visual BMI may also be used to estimate a patient's "normalized dry body weight." The normalized dry body weight of a 1.75 m tall patient with visual BMI 17 = $1.75^2 \times 17 = 52$ kg. Since protein and energy targets are based on normal body weight, this calculation is useful in situations in which actual body weight is unreliable or difficult to measure.

Laboratory and Technical Assessment Laboratory measurements have three main purposes in the evaluation and management of PEM.

MUSCLE MASS Bedside ultrasound is a potentially valuable technique for quantifying muscle mass at specific body sites, but it need not, nor should it, replace the immediate and comprehensive evaluation provided by the eyes, hands and mind of the discerning bedside examiner.

SYSTEMIC INFLAMMATION The presence or absence of systemic inflammation distinguishes SM from CDM/cachexia. The most useful laboratory indicators of systemic inflammation are a reduced serum albumin concentration and an increased serum C-reactive protein concentration. Systemic inflammation increases the permeability of capillary walls to large molecules; the resulting osmotic shift increases the ECF volume. Intravascular albumin pool redistributes into this large volume, decreasing the serum albumin concentration. Increased albumin catabolism likely also contributes. Muscle atrophy and dietary protein deficiency perpetuate inflammation-induced hypoalbuminemia, because muscle protein and the diet provide the amino acids required for hepatic albumin synthesis.

Hypoalbuminemia is often claimed to indicate or diagnose "malnutrition." This is incorrect. Hypoalbuminemia indicates the presence of systemic inflammation, which, by inducing anorexia and increasing muscle catabolism, creates a high risk that PEM could develop; but PEM may or may not exist at the present moment. Hypoalbuminemia will not improve as long as systemic inflammation persists, even with prolonged optimal nutritional therapy. After systemic inflammation has subsided, several weeks of optimal nutrition may be required for serum albumin concentrations to renormalize.

PROTEIN-CATABOLIC INTENSITY The defining feature of protein-catabolic disease (as occurs most intensely in ADM, but also in CDM) is increased muscle amino acid catabolism. Conditions that increase body protein loss can be identified by measuring the rate of body N loss. Most N leaves the body in the urine (almost all of it in urea, ammonium, and creatinine), the feces, skin, and by other minor routes. Total N is not usually measured in hospital laboratories, but urinary urea concentrations are routinely available. Urea normally accounts for ~85% of urinary N. Formulas are available that estimate that total N loss solely from 24-h urinary urea excretion. A recent, validated formula estimates daily total N loss (g) = g N in urinary urea/0.85 + 2.

Net muscle protein catabolism follows approximately first-order ("decay") kinetics, such that the rate of N loss from muscle is proportional to the existing total amount of N available to be lost. Muscle atrophic, protein-catabolic patients lose less body N/day in absolute terms than an equivalently catabolic patients with normal muscle mass, but they are at nevertheless at greater risk of succumbing to their critical illness. The interpretation of a patient's rate of N loss should be tempered by a consideration of their existing muscle mass.

Instrumental Nutritional Assessment Many nutritional assessment instruments claim to identify "malnutrition" by enumerating and summing a list of risk factors, laboratory results, and physical findings. These tools are often hindered by ambiguity about the intended meaning of "malnutrition" and failure to distinguish between screening and diagnosis. Diagnosis is the process of identifying a known pathological entity in a particular patient—SM or CDM, for example—by considering the patient's medical history, pertinent findings on physical examination, and laboratory or imaging reports. Diagnosis also involves an estimation of the probability that the diagnosis is correct and a judgment about its severity. By contrast, screening is the application of a test that identifies people at sufficiently high risk of a certain disease to warrant carrying out definitive procedures to establish the diagnosis or rule it out, or which identifies people at sufficient risk of developing the disease to warrant specific preventive interventions. Screening tools and risk predictors are useful, but it's a mistake to confuse them with clinical diagnosis.

Subjective Global Assessment The best-validated and most useful formal bedside instrument for diagnosing SM and CDM (and judging their severity) is subjective global assessment. With this method the examiner reflects on the totality of (1) the patient's history (for evidence of inadequate food intake, weight loss, and the presence of factors, such as gastrointestinal disease, and systemic inflammation, that strongly predict diminished ability consume enough food), (2) the patient's current body composition (muscle mass, subcutaneous fat, and ECF volume), and (3) their functional status (strength and mobility), then takes a moment to form an intuitive judgment as to whether the patient has (A) no SM or CDM, (B) is in the gray zone of possible or mild SM or CDM, or (C) definitely has SM or CDM.

A judgment is also reached as to how urgently nutritional intervention is required.

SPECIALIZED NUTRITION SUPPORT

Optimized Voluntary Nutritional Support When feasible, this is the approach of choice because it engages and empowers the patient, encourages mobilization and reconditioning, is consistent with the objectives of patient-centered medicine, and is risk free. Its disadvantage is that it is time-consuming and labor-intensive, and demands interest in and attention to the specific needs of individual patients.

Enteral Nutrition This is nutrition provided through a feeding tube placed through the nose into the stomach or beyond it into the duodenum, via a mini-surgical procedure in which a feeding tube is inserted through the abdominal wall into the stomach or beyond it into the jejunum using an endoscope, or by an open surgical approach to access the stomach or small intestine. EN is the treatment of choice when optimized voluntary nutritional support is impossible or has failed. It is relatively simple, safe, inexpensive, and maintains the digestive, absorptive, and immunologic barrier functions of the gastrointestinal tract. EN allows the delivery of accurately known amounts of nutrients. Pliable, small-bore feeding tubes make placement relatively easy and acceptable to most patients. Constant-rate infusion pumps increase the reliability of nutrient delivery. Patients are candidates for EN when optimized voluntary nutrition is not feasible or has failed, and their GI tract is adequately functional and can be accessed.

EN Products The commonest forms of EN used are commercially manufactured formulas with defined compositions.

STANDARD POLYMERIC FORMULAS These are the most widely used sources of EN. They are available in a wide variety of formats that generally meet the nutritional requirements of a normal, healthy person. Carbohydrates provide most of the energy. The proteins (from casein, whey, or soy) are intact and require normal pancreatic enzyme function for digestion and absorption. These products are isotonic or nearly so, and provide from 1000 to 2000 kcal and 50–70 g protein/L.

POLYMERIC FORMULAS WITH FIBER The addition of dietary fiber to formulas sometimes improves bowel function and feeding tolerance. Fermentable (soluble) fibers such as pectin and guar are metabolized by colonic bacteria, yielding short-chain fatty acids that fuel colonocytes. Nonfermentable (insoluble) fibers increase fecal bulk, improve peristalsis, and may improve diarrhea.

ELEMENTAL AND SEMI-ELEMENTAL FORMULAS The macronutrients in these formulas are partially or completely hydrolyzed. They are primarily designed for patients with known maldigestion and malabsorption, but they are sometimes used empirically for patients who have had prolonged bowel rest or are critically ill without strong evidence of their superiority, or when a patient is intolerant of a standard polymeric formula.

IMMUNE-ENHANCING FORMULAS In addition to providing macronutrients and conventional amounts of micronutrients, these products (IEFs) contain large amounts of certain nutrients designed to favorably modulate the immune response: arginine and n-3 fatty acids especially, but also various combinations of glutamine, nucleotides, and antioxidants. It is difficult to evaluate the IEFs because there are many different formulations on the market, and considerable heterogeneity in the patient populations studied. Good evidence of benefit has emerged from clinical trials of perioperative IEFs in patients undergoing elective gastrointestinal surgery and patients with traumatic brain injury. EIFs have not yet demonstrated benefit in other kinds of critical illness.

PROTEIN-ENRICHED FORMULAS Most EN formulas provide calories and protein in a ratio appropriate for a healthy person, whereas protein-enriched formulas provide ~90 g protein and 1000 kcal/L. Originally marketed to meet increased protein requirement associated with hypocaloric nutrition in obese patients, these products are increasingly used to provide protein-catabolic patients an appropriately generous amount of protein without calorie-overfeeding.

OTHER FORMULAS A wide variety of disease-specific EN products are available for patients with diabetes, hepatic, and renal or pulmonary disease. The details and evidence related to these products go beyond the scope of this chapter.

Parenteral Nutrition PN delivers a complete nutritional regimen directly into the bloodstream in the form of crystalline amino acids, dextrose, triglyceride emulsions, minerals (calcium, phosphate, magnesium, and zinc), electrolytes, and micronutrients. Because of its high osmolarity (>1200 mOsm/L) and often large volume, PN is infused into a central vein in adults. Ready-to-use PN admixtures typically containing 4–7% hydrous amino acids and 20–25% dextrose (with or without electrolytes) are available in 2-chamber (amino acids and dextrose) or 3-chamber (amino acids, dextrose, and lipid) bags that are intermixed and vitamins, trace minerals and additional electrolytes added just prior to infusion. Although convenient and cost-effective, these products have fixed nutrient composition and thus are dosed according to the volume required to meet calorie requirements. In some situations—especially ADM—a costlier approach is justified that uses a computer-controlled sterile compounder to generate combinations of amino acid and dextrose that meet the precise protein and energy requirements of individual patients.

For example, 1 L of a standard ready-to-use admixture of 5% amino acids and 25% dextrose provides 50 g of amino acids (equivalent to 41.5 g protein substrate) and 1000 kcal; the use of this solution to meet the 1.5–2.0-g/kg protein requirement of an acutely ill 70-kg patient requires the infusion of 2.5–3.4 L of fluid and a potentially excessive amount of energy (2500–3300 kcal). When the patient has an adequate fat store, clinical evidence increasingly supports the safety and efficacy of high-protein, moderately hypocaloric nutrition in ADM. A sterile compounder can accurately generate an appropriate recipe for such a patient. For example, 1 L of an admixture of 600 mL of 15% amino acids, 300 mL of 70% dextrose, and 100 mL of electrolyte/micronutrient mix contains 90 g amino acids (equivalent to 75 g of protein substrate) and 1020 kcal. (1 g mixed hydrated amino acids provides 3.3 kcal; 1 g dextrose provides 3.4 kcal; 1 g lipid emulsion provides ~10 kcal.) When a compounder is used, this patient's increased protein requirement can be met with less volume and avoid excessive calorie provision.

Amino Acids Protein synthesis requires 21 alpha-amino acids, 9 of which are essential and 11 are non-essential because they are readily synthesized from an essential amino acid (methionine or phenylalanine) or from widely available carbohydrate precursors and the amine groups provided by the body's large and rapidly interconverting pool of non-essential amino acids. One amino acid—arginine—is conditionally essential. PN amino acid admixtures vary somewhat but all of them provide appropriate amounts of the essential amino acids and arginine while compensating for their lack of glutamine (and sometimes glutamate or aspartate) by including large amounts of glycine or other non-essential amino acids. The specific contribution of each non-essential amino acid to a nutritional admixture is less important than the total amount of non-essential N it provides. The hydrated status of the free amino acids in PN reduces their calorie density from 4.0 to 3.3 kcal/g, and reduces the amount of protein substrate they provide by 17%. For example, 100 g of free mixed amino acids provide 83 g protein substrate and 340 kcal.

Carbohydrate and Lipids The carbohydrate in PN is dextrose monohydrate (3.4 kcal/g). Lipid emulsions provide calories and the essential n-6 and n-3 fatty acids. Traditional lipid emulsions are based solely on soy bean oil, but they are giving way to mixed emulsions that include medium-chain triglycerides, n-9 monounsaturated fatty acids, and n-3 fatty acids. Emulsions of pure soybean oil, a mixture of 80% olive oil and 20% soybean oil, and a mixture of 30% soybean oil, 30% medium chain triglycerides, 25% olive oil and 15% fish oil are available. The more complex lipid emulsions are more highly enriched in n-3 fatty acids and fewer n-6 polyunsaturated fatty acids than soybean

lipid, which is more prone to lipid peroxidation and could promote the formation of the pro-inflammatory n-6 derivatives. As a general rule, lipid infusion rates should not exceed 8 g/h (for a 70 kg patient) or 175 g (1925 kcal)/day.

Minerals, Micronutrients, and Trace Elements The default concentrations of electrolytes, minerals, and micronutrients in PN solutions are designed to meet normal requirements, and adjusted to meet the frequently abnormal and often-changing requirements of individual patients. Because they are unstable, multivitamin mixtures are injected into PN bags just prior to their delivery to the floor. Parenteral water-soluble vitamin requirements are greater than standard oral requirements because hospitalized patients often have vitamin deficiencies or increased requirements, and because intravenous administration of vitamins increases urinary losses. Vitamin C spontaneously degrades in PN solutions even when light-protected. The amount of vitamin D in currently used intravenous vitamin products is inadequate. Iron salts are incompatible with most PN solutions.

APPROACH TO THE PATIENT

Indications, Selection, and Provision of Specialized Nutritional Support

SNS is complicated, and the quantity and quality of the formal clinical evidence supporting its different uses is limited. Yet malnutrition-caused diseases are highly prevalent; they worsen clinical outcomes; and they are preventable and treatable. Physicians are charged with preventing, diagnosing, and treating these diseases, and they carry out their duty properly when guided by a sound understanding of nutritional principles, astute observation, rigorous clinical reasoning, and collaboration with dietitians and specialized clinical pharmacists.

Most hospitalized patients should not require SNS. Many of them will improve with appropriate management of their primary disease. Others have a terminal disease whose downward course will not be halted by SNS. Patients who can't eat enough hospital food and who have, or are at high risk of SM or CDM, are candidates for optimized voluntary nutrition support. When this approach is inappropriate or has failed, the pros and cons of invasive SNS are considered. The decision to provide or withhold EN or PN is based on a synthesis of four factors: (1) the determination that nutrient ingestion will likely continue to be inadequate for many days; (2) the patient has important muscle atrophy (of any cause) or fat depletion; (3) the patient's nutrient requirements are increased (as from inflammatory diarrhea, enterocutaneous fistulas or exudates, or a pronounced inflammatory protein-catabolic state); and (4) the reasoned judgment that SNS has a reasonable prospect of improving the patient's clinical outcome or quality of life. When the patient already has PEM (and treating it is likely to improve their well-being and clinical outcome), the decision tips more steeply in favor of intervention. It is important to formally diagnose and document PEM and its variants. Formal diagnosis focuses attention on the urgency of the situation and guides the selection, composition, and urgency of nutritional therapy.

EN THERAPY

EN is indicated when patients cannot eat enough food and unlikely to do so for a long time, their gastrointestinal tract is functional and accessible, and optimized voluntary nutrition is impossible or inadequate to meet the patient's nutritional needs. EN is most commonly used in settings of impaired consciousness, severe dysphagia, severe upper gastrointestinal tract dysfunction or obstruction, the requirement for mechanical ventilation, and critical illness in general. As well, situations frequently arise in which a patient's voluntary food intake is seriously curtailed by combinations of anorexia, nausea, vomiting, pain, delirium, depression, distress, chewing difficulties, undiagnosed thrush, unappealing food, and physical and sensory disability (including dysgeusia). In these complicated, difficult, and evolving situations, the diagnosis of SM or CDM should tip the decision towards EN.

Intestinal ischemia, mechanical obstruction, peritonitis, and gastrointestinal hemorrhage are contraindications to EN. Severe coagulopathy, esophageal varices, absent gag reflex, hypotension, adynamic ileus, pancreatitis, diarrhea, and nausea and vomiting are not absolute contraindications, but they increase the risk of complications and make it less likely that EN will succeed in achieving its nutritional goal.

Initiation, Progression, and Monitoring Feeding tube insertion and EN delivery are medical procedures that require voluntary informed consent. Nasogastric tube feeding may proceed when gastrointestinal function is adequate with regard to gastric contractility (e.g., nasogastric tube output <1200 mL/d), intestinal contractility (absence of a known or suspected intraabdominal pathologic process, non-distended abdomen and detectable bowel sounds, although the *absence* of bowel sounds is not, in itself, a contraindication), and adequate colonic function (passage of stools and flatus). After consent has been obtained and the appropriate feeding tube (usually a nasogastric tube for short-term feeding) has been placed and its position verified, the head of the patient's bed is raised to at least 30° and kept raised to reduce the risk of regurgitation. A standard polymeric formula is infused, usually at a starting rate of 50 mL/h and advanced by 25 mL/h every 4–8 h until the goal rate is met. Elemental formulas commence at a slower rate and progress more slowly. Intragastric feeding allows a higher formula osmolarity. Intragastric bolus feeding is an option (200–400 mL feeding solution infused over 15–60 min at regular intervals with verification of residual gastric contents every 4 h). Bolus feeding is not possible with jejunal feeding, which requires the patient to be monitored for abdominal pain and abdominal distention and bowel sounds every 4 h.

Complications and Their Management The most common complications of EN are aspiration of regurgitated or vomited formula, diarrhea, fluid volume and electrolyte derangements, hyperglycemia, nausea, abdominal pain, constipation, and failure to achieve the nutritional goal.

Aspiration Debilitated patients with delayed gastric emptying, impaired gag reflex, and ineffective cough are at high risk of aspiration pneumonia. Aspiration is particularly common in mechanically ventilated patients. Ventilator-associated pneumonia is mostly caused by aspiration of microbial pathogens in the mouth and throat past the cuffs of endotracheal or tracheostomy tubes, but tracheal suctioning induces coughing and gastric regurgitation. Measures to prevent ventilator-associated pneumonia include elevation of the head of the bed, mouth hygiene and gastrointestinal decontamination, nurse-directed algorithms for formula advancement and, sometimes, post–pyloric feeding. EN does not have to be held for gastric residual volumes less than 300 to 400 mL in the absence of other signs of gastrointestinal intolerance (nausea, vomiting, severe abdominal pain, abdominal distention). Continuous EN is often tolerated better than bolus feeding, and it is the only option with jejunal feeding.

Diarrhea Diarrhea is common when bowel function is compromised by disease or drugs (most often, broad-spectrum antibiotics). Once infectious and inflammatory causes have been ruled out, EN-associated diarrhea may be controlled by the use of a fiber-containing formula or the addition of an antidiarrheal agent to it. H2 blockers or proton pump inhibitors may help reduce the net volume of fluid presented to the colon. Diarrhea does not usually impair macronutrient absorption, since amino acids, lipids, and glucose are mostly absorbed in the proximal-to-middle small intestine. Since luminal nutrients have trophic effects on the intestinal mucosa, it is often appropriate to persist with tube feeding despite moderate, tolerable diarrhea, even it necessitates supplemental parenteral fluid support. Except for patients with markedly impaired small-intestinal absorptive function, there are no well-established indications for elemental formulas, but they may be used empirically

when diarrhea persists despite the use of fiber-enriched formulas and antidiarrheal agents.

Gastrointestinal Intolerance Abnormally high gastric residual volumes, abdominal distention, pain, and nausea are distressing for patients, increase the nursing workload, and delay the progression of EN. These problems can be avoided or minimized by ensuring normal fluid and electrolyte balance, preventing severe hyperglycemia, and, when a patient experiences nausea, vomiting or abdominal distention, by the judicious use of antiemetic and prokinetic drugs (and sometimes proton pump inhibitors) on a regular—rather than as needed—basis. Patients with gastroparesis require post-pyloric feeding.

Fluid Volume, Electrolyte, and Blood Glucose Abnormalities The essential purpose of EN is to provide macronutrients at an appropriate rate. EN formulas provide standard amounts of fluid, electrolytes, minerals, and micronutrients. They are not designed to manage abnormal fluid volume, electrolyte, and mineral requirements, which vary considerably among different patients and can change rapidly. Altered fluid volume requirements can, to a certain extent, be accommodated by selecting EN formulas with appropriate osmolarities, but medically active patients have widely varying and rapidly changing requirements for fluid, electrolyte, and glucose control. Blood glucose status must be monitored regularly and measures—including intravenous fluid, electrolyte, and insulin therapy—taken to maintain homeostasis.

Failure to Reach the Nutritional Goal EN is frequently delayed or interrupted in medically active patients. The reasons are many: diagnostic tests and procedures (including dialysis), physical or occupational therapy, a clogged or pulled out tube, and intolerance to EN. When the flow rate is low, formula at the tip of the feeding tube may be precipitated by gastric acid and cause an occlusion. Inadequate tube flushing, dense formulas, and the introduction of inadequately homogenized solid medications also cause tube clogging. The end result is prolonged delay in the progression of EN and failure to meet the patient's nutrient requirements. Brief periods of inadequate calorie provision are usually inconsequential in patients who have adequate fat reserves, but deficient provision of protein is a serious but, unfortunately, under-investigated problem. The protein-to-calorie ratio of most EN formulas is appropriate for healthy people and hence too low for patients whose protein requirement is increased. High-protein formulas are appropriate when EN is progressing too slowly or the patient has an increased protein requirement, existing PEM, or is obese and will directly benefit from hypocaloric nutrition.

EN in the Intensive Care Unit Most critically ill patients cannot eat anything—they depend entirely on SNS. EN has two purposes in the intensive care unit. The first purpose is to meet the patient's macronutrient requirements—especially their often dramatically increased protein requirement. The second purpose is to infuse nutrients into the intestines at a rate that sustains normal intestinal barrier and immunological functions in the face of a systemic inflammatory response that threatens intestinal integrity and immune function. Current guidelines recommend that EN commence as soon as possible after a critically ill patient has been fluid resuscitated and stabilized. The initial rate is 10–20 kcal/h. This rate of EN delivery provides ~25 g protein and ~500 kcal/d to patients who may require >100 g protein and >1800 kcal/d. Once EN is underway, the rate of delivery is increased as tolerated toward the patient's nutritional goal. Too often, however, EN falls far short of the protein provision target, even after a week or longer in the intensive care unit. Newer, high-protein EN products may reduce the severity of this protein shortfall.

PN THERAPY

PN is costlier, more resource-intensive, potentially riskier, and requires more expertise than EN. It is used when invasive SNS is indicated and EN is impossible, inappropriate, or unable to meet the patient's nutritional needs. The risks of PN are those of inserting and maintaining a central venous catheter (traumatic injury from the insertion, serious infection, and venous thrombosis), allergy to some of its components, glucose, electrolyte, magnesium, phosphate, and acid-base balance abnormalities, the adverse effects of the large intravenous fluid volumes. Prolonged PN—especially when it delivers excess calories—can lead to hepatic dysfunction.

INITIATION, PROGRESSION, MONITORING, AND DISCONTINUATION

PN may commence after the patient has been hemodynamically resuscitated, glucose, electrolyte, and acid-base homeostasis are established, and the patient is able to tolerate the fluid volumes involved. In adults, the high osmolarity of PN solutions, fluid demands, and need for strict sterility require their infusion through a dedicated port in a central venous catheter. Peripherally inserted central catheters (PICCs) are increasingly favored, although they limit monitoring for catheter infection by wire exchange. Jugular or femoral vein catheters are discouraged because it is very difficult to maintain a dry, sterile dressing over the insertion site. The initial dose of dextrose should not exceed 200 g/day to avoid hyperglycemia (or, in susceptible patients, the refeeding syndrome). On the other hand, the full requirement dose of amino acids can be administered from the first day onwards (this option is unavailable when premixed PN solutions are used). The glucose dose increases on a daily basis until it approaches the patient's energy requirement. Lipid emulsions are added after the first week.

Capillary blood glucose is monitored several times daily and subcutaneous regular insulin is added to the PN admixture as required to maintain average serum glucose concentrations <140 mg/dL and >80 mg/dL. The dose of regular insulin required to do this on a given day can be added to the following day's PN solution. The insulin dose increases roughly proportionately to the increasing glucose dose. Certain benchmarks are useful. Basal endogenous insulin secretion is ~30 units/day in normal people. When insulin is required for non-diabetic, non-catabolic patients, 10 units of regular insulin will roughly cover 100 g infused dextrose. Patients with non-insulin dependent diabetes require ~20 units/100 g dextrose. Non-catabolic patients with insulin-dependent diabetes usually require approximately twice the at-home insulin dose, because parenteral glucose is a more potent insulin secretagogue than oral carbohydrate and because some insulin adheres to the infusion bag. As a general rule (for a patient with dry body weight 70 kg) the glucose infusion rate should not exceed ~500 g (1700 kcal)/day in non-critically ill patients, and should not exceed ~350 g (1200 kcal)/day in critically ill patients. Even lower glucose infusion rates (e.g., 200 g/day) are advisable if they prevent or minimize hyperglycemia in insulin-resistant patients.

In general, the benefits of constraining glucose, lipid, and fluid volume provision in ADM justify hypocaloric nutrition for the first 2 weeks of SNS as long as the calorie deficit is counterbalanced by generous amino acid provision. Lipids are commonly introduced after the first week of PN and can make up calorie shortfalls. Serum triglyceride concentrations are measured before commencing lipid infusions in order to detect preexisting hypertriglyceridemia (usually defined as >400 mg/dL), which is a relative contraindication. They may be infused daily or 2–3 times weekly. As appreciation or the adverse effects of excessive glucose administration has increased, interest has grown in the use of lipid emulsions as an energy source. Lipid infusion rates should not usually exceed ~50 g (500 kcal)/day in critical illness.

Biochemical Monitoring Serum biochemistry (urea, creatinine, electrolytes, glucose, magnesium, phosphate, calcium, and albumin) are measured prior to starting PN and followed daily for the first few days, then twice weekly or as required. Serum triglycerides and liver function tests (and often ferritin) are measured at baseline and again after PN is underway to confirm that the lipid infusions are well tolerated. N balance (calculated from 24-h urinary urea N excretion) is useful at the outset for evaluating the severity

of protein catabolism in patients with CDM or ADM to identify patients who require more generous amino acid provision, and during PN to determine whether the patient's N balance is improving with therapy. Serum ferritin should be measured every 2 months, although the duration of most in-hospital PN is shorter than this.

Discontinuation PN is tapered and discontinued as soon as the patient can be adequately nourished by the enteral route. The dose of PN is reduced as food intake increases. As a general rule, once a patient is tolerating one-half to two-thirds of their food requirement by the enteral route and there is no mechanical or other barrier to further improvements in intake, PN can be discontinued. The transition to oral nutrition is slow for many patients with CDM. Optimized voluntary nutrition is much preferred to invasive EN for these patients because it is safe, effective, fosters well-being, and prepares them for discharge. It is tempting to stop PN as a way to stimulate more food consumption; the temptation should be resisted. PN does not create anorexia, nor does discontinuing it stimulate appetite. Too-early discontinuation of PN may delay a patient's progression to full voluntary food consumption by inducing anxiety and recreating starvation conditions. PN is most successfully weaned by encouraging physical activity, optimizing voluntary nutrition (including food from home), emotional support, and having patience. On the other hand, patients who are at the cusp of adequate oral nutrition commonly benefit from hospital discharge, where the security and pleasure of being at home and eating home-made food are potent stimuli.

Drawbacks, Side Effects, and Complications • **Complications of Central Catheters** Patients receiving PN are at greater risk of bloodstream infections than other patients with central venous catheters. Proper aseptic insertion technique, meticulous dressing care, and one port dedicated solely to PN reduce this risk. Catheter-induced upper arm venous thrombosis is an uncommon but important complication.

Hyperglycemia The most frequent metabolic complication of PN is hyperglycemia in patients with insulin resistance due to non-insulin dependent diabetes mellitus, high-dose glucocorticoid therapy, or severe systemic inflammation; the problem is exacerbated by excessively high rates of glucose provision. Glucose concentrations are most easily kept at <140 mg/dL with the least risk of hypoglycemia by infusing hypocaloric amounts of glucose and, if necessary, meeting the patient's energy requirement with intravenous lipid. In ADM, the benefits of using the lowest insulin dose—minimal hyperinsulinemia and a reduced risk of hypoglycemia—almost always outweigh the doubtful goal of rapidly matching calorie provision to the patient's energy expenditure rate.

Hypoglycemia Reactive hypoglycemia is uncommon but may occur when high-dextrose, non-insulin containing PN is abruptly discontinued. It is prevented by slowing the PN infusion rate to 50 mL/h for 1 or 2 h prior to discontinuing it (or replacing it with 10% dextrose), or, when the oral route is available, providing a snack. More often, hypoglycemia occurs when the intensity of the patient's metabolic stress (or their glucocorticoid dose) decreases without an appropriate downward adjustment of the insulin dose. This problem is avoided by frequent capillary glucose determinations and careful attention to medication doses and the patient's general condition.

Artefactual Hyperglycemia and Hyperkalemia Blood samples must be carefully and appropriately drawn from dual-port PICC. Intermixing of the sample with even a tiny volume of PN solution will falsely indicate hyperglycemia and hyperkalemia, and may trigger a treatment error. The problem is identified when the patient's apparent serum glucose (and potassium) concentrations abruptly increase without reason and the apparently very high glucose concentration is out of keeping with concurrent capillary glucose readings.

Volume Overload Hypertonic intravenous dextrose stimulates a much more intense insulin response than oral glucose. A potent antinatriuretic and antidiuretic hormone, insulin potentiates sodium and water retention. In this setting net fluid retention is likely when total fluid provision exceeds 2 L/day in patients not experiencing large gastrointestinal losses. The problem of volume overload can be reduced by preparing PN solutions with a compounder, infusing glucose at a rate that minimizes the need for exogenous insulin therapy, and avoiding overfeeding in general. Sodium intake plays an important role in fluid retention. Net fluid retention can be minimized by limiting sodium delivery to 20–30 mmol/day (with an adjustment for gastric or intestinal sodium losses) since, when renal function is normal and antidiuretic hormone secretion is not increased, urinary sodium concentrations are usually <10 mmol/L.

Hypertriglyceridemia This complication occurs when the rate of lipid infusion exceeds plasma triglyceride clearance capacity. Renal failure, sepsis, excessive glucose (which stimulates lipogenesis), diabetes mellitus, high-dose glucocorticoid therapy, and multiple-organ failure reduce triglyceride clearance. An impaired immune response, increased risk of acute pancreatitis, and altered pulmonary hemodynamics are potential but not well documented complications of PN-induced severe hypertriglyceridemia.

Hepatic Dsyfunction Mild elevations of serum liver enzyme concentrations can occur within 2–4 weeks of initiating PN, but in most cases they return to normal even when PN is continued. Clinically important hepatic dysfunction, although common in children, is uncommon in adults as long as energy overfeeding and resultant fatty liver are avoided. Intrahepatic cholestasis occasionally occurs after many weeks of continuous PN and is most often multifactorial in origin. Cyclic PN—in which PN is infused for only 12 h/day—may prevent or reduce the severity of this complication.

PN IN THE INTENSIVE CARE UNIT

Current guidelines recommend starting EN as soon as a critically ill patient has been resuscitated, stabilized, and enteral access established. EN is then advanced over the following days to toward the patient's nutritional goal. If the goal has not been achieved after 7–10 days, amino acid-rich PN is recommended, especially when the patient remains protein catabolic. PN that is generous in amino acids (85–140 g protein substrate/day for a 70 kg patient) and hypocaloric (1200–1400 kcal/day) limits the risk and severity of hyperglycemia and volume overload and may improve clinical outcomes in this setting. Soy-based lipid emulsions should be avoided during the first week of PN in the intensive care unit; alternative lipid emulsions may prove to be safe and beneficial.

SPECIAL CLINICAL CONDITIONS

Old Age In addition to their other physiological frailties, elderly people commonly suffer from age-related muscle atrophy that is compounded by disuse muscle atrophy. These factors place them at high risk of PEM and make them plausible candidates for early SNS.

Inactivity Physical activity and adequate nutrition are closely interdependent. Reduced physical activity reduces appetite, and physical rehabilitation and its associated emotional benefits restore optimism and appetite. Full nutrient provision will maintain or normalize many physiological functions in bedridden patients, but they will not increase muscle mass.

Renal Failure Protein provision should not be reduced in the presence of renal injury unless renal replacement therapy is unavailable. Protein and vitamin C provision should increase when renal replacement therapy is used, for it removes large amounts of amino acids and vitamin C from the circulation.

Liver Failure Patients with severe hepatic disease are plausible candidates for SNS because they are relatively intolerant of starvation and commonly already have CDM when they are admitted to hospital. SNS should be generous in calories and especially protein, which should be provided despite an increased risk of hepatic encephalopathy. The risk of encephalopathy can be mitigated by meticulous attention to fluid balance and electrolyte status and by

spreading protein provision over the day to accommodate the liver's reduced capacity to clear amino acid-derived ammonia.

Perioperative SNS Patients awaiting elective major surgery benefit from 7 to 10 days of preoperative SNS if SM (or especially CDM) is present. When feasible and properly implemented, optimized voluntary nutrition is greatly preferred, but when a patient has been admitted to hospital their condition is by definition semi-urgent, and EN or PN will meet the patient's nutritional goal more quickly. Preoperative SNS improves immunity and reduces postoperative complications, but it will not increase serum albumin concentrations and it should not be provided for >7–10 days with that goal in mind. More prolonged preoperative EN or PN may confer slight additional nutritional benefits, but they are counterbalanced by its risks and the consequences of prolonged hospitalization and delayed surgery. Surgery should not be delayed for starving patients whose muscle mass is normal or only mildly depleted and who are not experiencing systemic inflammation, since they tolerate even major uncomplicated surgery well without preoperative SNS. The urgency of surgery often precludes otherwise indicated preoperative SNS. Early postoperative PN is usually indicated for these patients, for they are at increased risk of postoperative complications and are highly unlikely to consume an adequate amount of food voluntarily over the next many days. Patients with only mild muscle atrophy, no systemic inflammation, and no postoperative complications do not require postoperative PN unless (1) adequate feeding by mouth has not been achieved by day 5–7 after surgery or (2) there are indications that voluntary feeding will be further delayed. Warning indicators include inadequately controlled nausea or pain, impaired gastric, small intestinal or colonic function, serum electrolyte imbalance, altered mental status, inability to mobilize from the bed, or a suspected surgical-site infection or anastomotic leak. Perioperative immune-enhancing EN appears to reduce morbidity in patients undergoing major elective gastrointestinal surgery.

Iron and PN Iron deficiency is common in acutely ill hospitalized patients. Risk factors for in-hospital iron deficiency include inadequate nutritional intake, gastrointestinal disease, and frequent blood withdrawals. In-hospital iron deficiency is often missed because inflammation-associated anemia is much commoner and increases serum concentrations of ferritin, a positive acute-phase protein. The parenteral iron requirement is normally only ~1 mg/day, but since iron is a highly reactive catalyst of oxidative reactions it is not included in PN mixtures. Serum ferritin concentrations should usually be measured when PN commences and re-measured at ~8 week intervals. A falling mean red cell volume (even within the low-normal range) or an intermediate serum ferritin concentration in the presence of systemic inflammation strongly suggest iron deficiency. Intravenous iron should be ordered according to standard guidelines. A termination order should also be written to avoid the risk of inadvertent iron overdosing. Iron replacement should be avoided during the acute phase of critical illness because a substantial rise in the serum iron concentration could increase susceptibility to some bacterial infections, but it is indicated in SM and CDM both to improve iron-deficiency anemia and provide the non-hematologic benefits of adequate iron nutrition.

Zinc It is insufficiently appreciated that 1 L of secretory diarrhea contains ~12 mg of zinc. Patients with intestinal fistulas or high volume chronic diarrhea require this amount of zinc in addition to their daily requirement of 15 mg to avoid zinc deficiency. Zinc may be provided parenterally or enterally (because of its low bioavailability, 12 mg parenteral zinc is equivalent to 30 mg oral elemental zinc).

Cancer SNS plays a crucial role in cancer therapy. Many cancers (especially those that involve the gastrointestinal tract or induce systemic inflammation) and most cancer therapies create the conditions for starvation and commonly lead to SM or CDM. The prevention or treatment of these diseases may improve patients' quality of life and tolerance to anti-cancer therapy. As a general rule, EN and PN are not prescribed to patients who are not undergoing active anti-cancer therapy because the side effects and complications of invasive SNS are not counterbalanced by an improved disease trajectory. In some cases, a disease may be inexorably progressing but so slowly that the patient will die of starvation long before they would from the cancer. EN or PN is usually appropriate in these cases.

Advanced Dementia Optimized voluntary nutrition is the key approach in this situation, and it can be used to deal with problems such as disability and dysphagia in patients who show discernable pleasure from eating. There is no evidence that EN or PN improve quality or length of life in patients who have advanced dementia and show little or no interest in eating food, and their side effects and complications are unpleasant and sometimes dangerous.

REFEEDING SYNDROME

The refeeding syndrome can occur in starving patients during the first week of nutritional repletion if carbohydrates are introduced too rapidly. Carbohydrate provision stimulates insulin secretion, which, owing to its antinatriuretic effect, expands the ECF volume when excessive sodium is provided. Refeeding edema can be minimized by severely limiting sodium provision and increasing carbohydrate provision slowly. Carbohydrate refeeding may stimulate enough intracellular glucose-6-phosphate and glycogen synthesis to seriously lower serum phosphate concentrations. Refeeding increases the adaptively down-regulated REE of patients with SM and stimulates N retention, new cell synthesis, and cellular rehydration. Phosphate, potassium, magnesium, and zinc are common and dangerous during refeeding. Their status must be monitored and appropriate supplements provided. Acute thiamine deficiency is a devastating but preventable complication of refeeding, even with simple dextrose infusions. Left heart failure may occur in predisposed patients. The precipitants of left heart failure are an abrupt increase of intravascular volume due to the administration of fluids and of glucose, which stimulates insulin secretion with associated renal sodium retention; increased cardiac demand because of increased REE in a patient with an atrophic left ventricle and low stroke volume; and myocardial deficiencies of potassium, phosphorus, or magnesium. Cardiac arrhythmias may occur. Refeeding may precipitate acute wet beriberi in malnourished, poverty-stricken populations.

GLOBAL CONSIDERATIONS

The macro-nutritional diseases are the same worldwide, but their prevalence varies in accordance with regional variations in the prevalence of the primary medical and surgical diseases that induce them and the two most important causes of starvation: poverty and social iniquity. Longstanding borderline SM is so prevalent in some parts of the world that the World Health Organization has coined the term *chronic energy deficiency* (CED) to describe it. CED is classified into three grades of severity: Grade I: BMI 17.0–18.4, Grade II: BMI 16.0–16.9, and Grade III: BMI <16. In otherwise healthy adults, CED of Grades II and III are associated with an increasing probability of days of illness, reduced work capacity, poorer reproductive function, and poorer lactation performance. Voluntary physical activity is decreased in Grade III CED, which is equivalent to the diagnosis of SM.

FURTHER READING

Derenski K et al: Parenteral nutrition basics for the clinician caring for the Adult Patient. Nutr Clin Pract 31:578, 2016.

Esuro AA, Hummell AC: Enteral formulas in nutrition support practice: Is there a better choice for your patient? Nutr Clin Pract 31:709, 2016.

Lembeck ME et al: The role of intravenous fluids and enteral or parenteral nutrition in patients with life-limiting illness. Med Clin North Am 100:1131, 2016.

McClave SA et al: Guidelines for the provision and assessment of nutrition support therapy in the adult critically ill patient: Society of Critical Care Medicine (SCCM) and American Society for Parenteral and Enteral Nutrition (ASPEN). J Parenter Enteral Nutr 40: 159, 2016.

2332 White JV et al: Consensus statement: Academy of Nutrition and Dietetics and American Society for Parenteral and Enteral Nutrition: Characteristics recommended for the identification and documentation of adult malnutrition (undernutrition). J Parenter Enteral Nutr 36:275, 2012.

Section 3 Liver and Biliary Tract Disease

329 Approach to the Patient with Liver Disease

Marc G. Ghany, Jay H. Hoofnagle

A diagnosis of liver disease usually can be made accurately by careful elicitation of the patient's history, physical examination, and application of a few laboratory tests. In some circumstances, radiologic examinations are helpful or, indeed, diagnostic. Liver biopsy is considered the criterion standard in evaluation of liver disease, but is now needed less for diagnosis than for grading (activity) and staging (fibrosis) of disease. Non-invasive means of assessing fibrosis stage have become increasingly helpful and may allow for avoidance of biopsy in a proportion of patients. This chapter provides an introduction to diagnosis and management of liver disease, briefly reviewing the structure and function of the liver; the major clinical manifestations of liver disease; and the use of clinical history, physical examination, laboratory tests, imaging studies, and liver biopsy.

LIVER STRUCTURE AND FUNCTION

The liver is the largest organ of the body, weighing 1–1.5 kg and representing 1.5–2.5% of the lean body mass. The size and shape of the liver vary and generally match the general body shape—long and lean or squat and square. This organ is located in the right upper quadrant of the abdomen under the right lower rib cage against the diaphragm and projects for a variable extent into the left upper quadrant. It is held in place by ligamentous attachments to the diaphragm, peritoneum, great vessels, and upper gastrointestinal organs. The liver receives a dual blood supply; ~20% of the blood flow is oxygen-rich blood from the hepatic artery, and 80% is nutrient-rich blood from the portal vein arising from the stomach, intestines, pancreas, and spleen.

The majority of cells in the liver are hepatocytes, which constitute two-thirds of the organ's mass. The remaining cell types are Kupffer cells (members of the reticuloendothelial system), stellate (Ito or fat-storing) cells, endothelial and blood vessel cells, bile ductular cells, and cells of supporting structures. Viewed by light microscopy, the liver appears to be organized in lobules, with portal areas at the periphery and central veins in the center of each lobule. However, from a functional point of view, the liver is organized into acini, with both hepatic arterial and portal venous blood entering the acinus from the portal areas (zone 1) and then flowing through the sinusoids to the terminal hepatic veins (zone 3); the intervening hepatocytes constitute zone 2. The advantage of viewing the acinus as the physiologic unit of the liver is that this perspective helps to explain the morphologic patterns and zonality of many vascular and biliary diseases not explained by the lobular arrangement.

Portal areas of the liver consist of small veins, arteries, bile ducts, and lymphatics organized in a loose stroma of supporting matrix and small amounts of collagen. Blood flowing into the portal areas is distributed through the sinusoids, passing from zone 1 to zone 3 of the acinus and draining into the terminal hepatic veins ("central veins"). Secreted bile flows in the opposite direction—that is, in a counter-current pattern from zone 3 to zone 1. The sinusoids are lined by unique endothelial cells that have prominent fenestrae of variable sizes, allowing the free flow of plasma but not of cellular elements. The plasma is thus in direct contact with hepatocytes in the subendothelial space of Disse.

Hepatocytes have distinct polarity. The basolateral side of the hepatocyte lines the space of Disse and is richly lined with microvilli; it exhibits endocytotic and pinocytotic activity, with passive and active uptake of nutrients, proteins, and other molecules. The apical pole of the hepatocyte forms the canalicular membranes through which bile components are secreted. The canaliculi of hepatocytes form a fine network, which fuses into the bile ductular elements near the portal areas. Kupffer cells usually lie within the sinusoidal vascular space and represent the largest group of fixed macrophages in the body. The stellate cells are located in the space of Disse but are not usually prominent unless activated, when they produce collagen and matrix. Red blood cells stay in the sinusoidal space as blood flows through the lobules, but white blood cells can migrate through or around endothelial cells into the space of Disse and from there to portal areas, where they can return to the circulation through lymphatics.

Hepatocytes perform numerous and vital roles in maintaining homeostasis and health. These functions include the synthesis of most essential serum proteins (albumin, carrier proteins, coagulation factors, many hormonal and growth factors), the production of bile and its carriers (bile acids, cholesterol, lecithin, phospholipids), the regulation of nutrients (glucose, glycogen, lipids, cholesterol, amino acids), and the metabolism and conjugation of lipophilic compounds (bilirubin, anions, cations, drugs) for excretion in the bile or urine. Measurement of these activities to assess liver function is complicated by the multiplicity and variability of these functions. The most commonly used liver "function" tests are measurements of serum bilirubin, serum albumin, and prothrombin time. The serum bilirubin level is a measure of hepatic conjugation and excretion; the serum albumin level and prothrombin time are measures of protein synthesis. Abnormalities of bilirubin, albumin, and prothrombin time are typical of hepatic dysfunction. Frank liver failure is incompatible with life, and the functions of the liver are too complex and diverse to be subserved by a mechanical pump; a dialysis membrane; or a concoction of infused hormones, proteins, and growth factors.

LIVER DISEASES

While there are many causes of liver disease **(Table 329-1)**, these disorders generally present clinically in a few distinct patterns and are usually classified as hepatocellular, cholestatic (obstructive), or mixed. In *hepatocellular diseases* (such as viral hepatitis and alcoholic liver disease), features of liver injury, inflammation, and necrosis predominate. In *cholestatic diseases*, such as gallstone or malignant obstruction, primary biliary cholangitis (previously referred to as primary biliary cirrhosis), and some drug-induced liver diseases, features of inhibition of bile flow predominate. In a mixed pattern, features of both hepatocellular and cholestatic injury are present (such as in cholestatic forms of viral hepatitis and many drug-induced liver diseases). The pattern of onset and prominence of symptoms can rapidly suggest a diagnosis, particularly if major risk factors are considered, such as the age and sex of the patient and a history of exposure or risk behaviors.

Typical presenting symptoms of liver disease include jaundice, fatigue, itching, right-upper-quadrant pain, nausea, poor appetite, abdominal distention, and intestinal bleeding. At present, however, many patients are diagnosed with liver disease who have no symptoms and who have been found to have abnormalities in biochemical liver tests as a part of a routine physical examination or screening for blood donation or for insurance or employment. The wide availability of batteries of liver tests makes it relatively simple to demonstrate the presence of liver injury as well as to rule it out in someone in whom liver disease is suspected.

Evaluation of patients with liver disease should be directed at (1) establishing the etiologic diagnosis, (2) estimating disease severity (*grading*), and (3) establishing the disease stage (*staging*). *Diagnosis* should focus on the category of disease (hepatocellular, cholestatic, or mixed injury) as well as on the specific etiologic diagnosis. *Grading* refers

TABLE 329-1 Liver Diseases

Inherited hyperbilirubinemia
- Gilbert syndrome
- Crigler-Najjar syndrome, types I and II
- Dubin-Johnson syndrome
- Rotor syndrome

Viral hepatitis
- Hepatitis A
- Hepatitis B
- Hepatitis C
- Hepatitis D
- Hepatitis E
- Others (Epstein-Barr virus [mononucleosis] herpesvirus, cytomegalovirus, adenovirus hepatitis)
- Cryptogenic hepatitis

Immune and autoimmune liver diseases
- Primary biliary cholangitis
- Autoimmune hepatitis
- Sclerosing cholangitis
- Overlap syndromes
- Graft-versus-host disease
- Allograft rejection

Genetic liver diseases
- α_1 antitrypsin deficiency
- Hemochromatosis
- Wilson disease
- Benign recurrent intrahepatic cholestasis
- Progressive familial intrahepatic cholestasis, types I–III
- Others (galactosemia, tyrosinemia, cystic fibrosis, Niemann-Pick-disease, Gaucher's disease)

Alcoholic liver disease
- Acute fatty liver
- Acute alcoholic hepatitis
- Laënnec cirrhosis

Nonalcoholic fatty liver
- Steatosis
- Steatohepatitis

Acute fatty liver of pregnancy

Liver involvement in systemic diseases
- Sarcoidosis
- Amyloidosis
- Glycogen storage diseases
- Celiac disease
- Tuberculosis
- *Mycobacterium avium-intracellulare* infection

Cholestatic syndromes
- Benign postoperative cholestasis
- Jaundice of sepsis
- Total parenteral-nutrition–induced jaundice
- Cholestasis of pregnancy
- Cholangitis and cholecystitis
- Extrahepatic biliary obstruction (stone, stricture, cancer)
- Biliary atresia
- Caroli disease
- Cryptosporidiosis

Drug-induced liver disease
- Hepatocellular patterns (isoniazid, acetaminophen)
- Cholestatic patterns (methyltestosterone)
- Mixed patterns (sulfonamides, phenytoin)
- Micro- and macrovesicular steatosis (methotrexate, fialuridine)

Vascular injury
- Sinusoidal obstruction syndrome
- Budd-Chiari syndrome
- Ischemic hepatitis
- Passive congestion
- Portal vein thrombosis
- Nodular regenerative hyperplasia

Mass lesions
- Hepatocellular carcinoma
- Cholangiocarcinoma
- Adenoma
- Focal nodular hyperplasia
- Metastatic tumors
- Abscess
- Cysts
- Hemangioma

to assessment of the severity or activity of disease—active or inactive as well as mild, moderate, or severe. *Staging* refers to estimation of the point in the course of the natural history of the disease, whether early or late; or precirrhotic, cirrhotic, or end-stage. This chapter introduces general, salient concepts in the evaluation of patients with liver disease that help lead to the diagnoses discussed in subsequent chapters.

CLINICAL HISTORY

The clinical history should focus on the symptoms of liver disease—their nature, patterns of onset, and progression—and on potential risk factors for liver disease. The manifestations of liver disease include constitutional symptoms such as fatigue, weakness, nausea, poor appetite, and malaise and the more liver-specific symptoms of jaundice, dark urine, light stools, itching, abdominal pain, and bloating. Symptoms can also suggest the presence of cirrhosis, end-stage liver disease, or complications of cirrhosis such as portal hypertension. Generally, the constellation of symptoms and their patterns of onset rather than a specific symptom points to an etiology.

Fatigue is the most common and most characteristic symptom of liver disease. It is variously described as lethargy, weakness, listlessness, malaise, increased need for sleep, lack of stamina, and poor energy. The fatigue of liver disease typically arises after activity or exercise and is rarely present or severe after adequate rest; that is, it is "afternoon" rather than "morning" fatigue. Fatigue in liver disease is often intermittent and variable in severity from hour to hour and day to day. In some patients, it may not be clear whether fatigue is due to the liver disease or due to other problems such as stress, anxiety, sleep disturbance, or a concurrent illness.

Nausea occurs with more severe liver disease and may accompany fatigue or be provoked by smelling food odors or eating fatty foods. Vomiting can occur but is rarely persistent or prominent. Poor appetite with weight loss occurs frequently in acute liver disease, but is rare in chronic disease except when cirrhosis is present and advanced. Diarrhea is uncommon in liver disease except with severe jaundice, in which a lack of bile acids reaching the intestine can lead to steatorrhea.

Right-upper-quadrant discomfort or ache ("liver pain") occurs in many liver diseases and is usually marked by tenderness over the liver area. The pain arises from stretching or irritation of Glisson's capsule, which surrounds the liver and is rich in nerve endings. Severe pain is most typical of gallbladder disease, liver abscess, and severe sinusoidal obstruction syndrome (previously known as veno-occlusive disease) but is also an occasional accompaniment of acute hepatitis.

Itching occurs with acute liver disease, appearing early in obstructive jaundice (from biliary obstruction or drug-induced cholestasis) and somewhat later in hepatocellular disease (acute hepatitis). Itching also occurs in chronic liver diseases—typically the cholestatic forms such as primary biliary cholangitis and sclerosing cholangitis, in which it is often the presenting symptom, preceding the onset of jaundice. However, itching can occur in any liver disease, particularly once cirrhosis develops.

Jaundice is the hallmark symptom of liver disease and perhaps the most reliable marker of severity. Patients usually report darkening of the urine before they notice scleral icterus. Jaundice is rarely detectable with a bilirubin level <43 μmol/L (2.5 mg/dL). With severe cholestasis, there will also be lightening of the color of the stools and steatorrhea. Jaundice without dark urine usually indicates indirect (unconjugated) hyperbilirubinemia and is typical of hemolytic anemia and the genetic disorders of bilirubin conjugation, the common and benign form being Gilbert syndrome and the rare and severe form being Crigler-Najjar syndrome. Gilbert syndrome affects up to 5% of the general population; the jaundice in this condition is more noticeable after fasting and with stress.

Major risk factors for liver disease that should be sought in the clinical history include details of alcohol use, medication use (including herbal compounds, birth control pills, and over-the-counter medications), personal habits, sexual activity, travel, exposure to jaundiced or other high-risk persons, injection drug use, recent surgery, remote or recent transfusion of blood or blood products, occupation, accidental exposure to blood or needlestick, and familial history of liver disease.

For assessing the risk of viral hepatitis, a careful history of sexual activity is of particular importance and should include the number of lifetime sexual partners and, for men, a history of having sex with men. Sexual exposure is a common mode of spread of hepatitis B but is uncommon for hepatitis C. A family history of hepatitis, liver disease, and liver cancer is also important. Maternal-infant transmission occurs with both hepatitis B and C. Vertical spread of hepatitis B can now be prevented by passive and active immunization of the infant at birth, although addition of antiviral therapy during the third trimester of pregnancy is now recommended for mothers with levels of HBV DNA >200,000 IU/mL. Vertical spread of hepatitis C is uncommon, but there are no reliable means of prevention. Transmission is more common among HIV-co-infected mothers and is also linked to prolonged and difficult labor and delivery, early rupture of membranes, internal fetal monitoring, and a high viral load. A history of injection drug use, even in the remote past, is of great importance in assessing the risk for hepatitis B and C. Injection drug use is now the single most common risk factor for hepatitis C. Transfusion with blood or blood products

is no longer an important risk factor for acute viral hepatitis. However, blood transfusions received before the introduction of sensitive enzyme immunoassays for antibody to hepatitis C virus in 1992 is an important risk factor for chronic hepatitis C. Blood transfusion before 1986, when screening for antibody to hepatitis B core antigen was introduced, is also a risk factor for hepatitis B. Travel to a developing area of the world, exposure to persons with jaundice, and exposure to young children in day-care centers are risk factors for hepatitis A. Tattooing and body piercing (for hepatitis B and C) and eating shellfish (for hepatitis A) are frequently mentioned but are actually types of exposure that quite rarely lead to the acquisition of hepatitis.

Hepatitis E is one of the more common causes of jaundice in Asia and Africa but is uncommon in developed nations. In endemic areas transmission is usually through exposure to fecally contaminated water. Recently, non-travel-related (*autochthonous*) cases of hepatitis E have been described in developed countries, including the United States. These cases appear to be due to strains of hepatitis E virus that are endemic in swine and some wild animals (genotypes 3 and 4). While occasional cases are associated with eating raw or undercooked pork or game (deer and wild boars), most cases of hepatitis E occur without known exposure, predominantly in elderly men without typical risk factors for viral hepatitis. Hepatitis E infection can become chronic in immunosuppressed individuals (such as transplant recipients, patients receiving chemotherapy, or patients with HIV infection), in whom it presents with abnormal serum enzymes in the absence of markers of hepatitis B or C.

A history of alcohol intake is important in assessing the cause of liver disease and also in planning management and recommendations. In the United States, for example, at least 70% of adults drink alcohol to some degree, but significant alcohol intake is less common; in population-based surveys, only 5% of individuals have more than two drinks per day, the average drink representing 11–15 g of alcohol. Alcohol consumption associated with an increased rate of alcoholic liver disease is probably more than two drinks (22–30 g) per day in women and three drinks (33–45 g) in men. Most patients with alcoholic cirrhosis have a much higher daily intake and have drunk excessively for ≥10 years before onset of liver disease. In assessing alcohol intake, the history should also focus on whether alcohol abuse or dependence is present. Alcoholism is usually defined by the behavioral patterns and consequences of alcohol intake, not by the amount. *Abuse* is defined by a repetitive pattern of drinking alcohol that has adverse effects on social, family, occupational, or health status. *Dependence* is defined by alcohol-seeking behavior, despite its adverse effects. Many alcoholics demonstrate both dependence and abuse, and dependence is considered the more serious and advanced form of alcoholism. A clinically helpful approach to diagnosis of alcohol dependence and abuse is the use of the CAGE questionnaire **(Table 329-2)**, which is recommended for all medical history-taking.

Family history can be helpful in assessing liver disease. Familial causes of liver disease include Wilson disease; hemochromatosis and α_1 antitrypsin deficiency; and the more uncommon inherited pediatric liver diseases—that is, familial intrahepatic cholestasis, benign recurrent intrahepatic cholestasis, and Alagille syndrome. Onset of severe liver disease in childhood or adolescence in conjunction with a family history of liver disease or neuropsychiatric disturbance should lead to investigation for Wilson's disease. A family history of cirrhosis, diabetes, or endocrine failure and the appearance of liver disease in adulthood suggests hemochromatosis and should prompt investigation of iron status. Abnormal iron studies in adult patients warrant genotyping of the *HFE* gene for the C282Y and H63D mutations typical of genetic hemochromatosis. In children and adolescents with iron overload, other non-*HFE* causes of hemochromatosis should be sought. A family history of emphysema should lead to investigation of α_1 antitrypsin levels and, if levels are low, for protease inhibitor (Pi) genotype.

TABLE 329-2 CAGE Questions[a]

ACRONYM	QUESTION
C	Have you ever felt you ought to cut down on your drinking?
A	Have people annoyed you by criticizing your drinking?
G	Have you ever felt guilty or bad about your drinking?
E	Have you ever had a drink first thing in the morning to steady your nerves or get rid of a hangover (eye-opener)?

[a]One "yes" response should raise suspicion of an alcohol use problem, and more than one is a strong indication of abuse or dependence.

PHYSICAL EXAMINATION

The physical examination rarely uncovers evidence of liver dysfunction in a patient without symptoms or laboratory findings, nor are most signs of liver disease specific to one diagnosis. Thus, the physical examination complements rather than replaces the need for other diagnostic approaches. In many patients, the physical examination is normal unless the disease is acute or severe and advanced. Nevertheless, the physical examination is important in that it can yield the first evidence of hepatic failure, portal hypertension, and liver decompensation. In addition, the physical examination can reveal signs—related either to risk factors or to associated diseases or findings—that point to a specific diagnosis.

Typical physical findings in liver disease are icterus, hepatomegaly, hepatic tenderness, splenomegaly, spider angiomata, palmar erythema, and skin excoriations. Signs of advanced disease include muscle wasting, ascites, edema, dilated abdominal veins, hepatic fetor, asterixis, mental confusion, stupor, and coma. In male patients with cirrhosis, particularly that related to alcohol use, signs of hyperestrogenemia such as gynecomastia, testicular atrophy, and loss of male-pattern hair distribution may be found.

Icterus is best appreciated when the sclera is inspected under natural light. In fair-skinned individuals, a yellow tinge to the skin may be obvious. In dark-skinned individuals, examination of the mucous membranes below the tongue can demonstrate jaundice. Jaundice is rarely detectable if the serum bilirubin level is <43 μmol/L (2.5 mg/dL) but may remain detectable below this level during recovery from jaundice (because of protein and tissue binding of conjugated bilirubin).

Spider angiomata and palmar erythema occur in both acute and chronic liver disease; these manifestations may be especially prominent in persons with cirrhosis but can develop in normal individuals and are frequently found during pregnancy. Spider angiomata are superficial, tortuous arterioles, and—unlike simple telangiectases—typically fill from the center outward. Spider angiomata occur only on the arms, face, and upper torso; they can be pulsatile and may be difficult to detect in dark-skinned individuals.

Hepatomegaly is not a highly reliable sign of liver disease because of variability in the liver's size and shape and the physical impediments to assessment of liver size by percussion and palpation. Marked hepatomegaly is typical of cirrhosis, sinusoidal obstruction syndrome, infiltrative disorders such as amyloidosis, metastatic, or primary cancers of the liver, and alcoholic hepatitis. Careful assessment of the liver edge may also reveal unusual firmness, irregularity of the surface, or frank nodules. Perhaps the most reliable physical finding in the liver examination is hepatic tenderness. Discomfort when the liver is touched or pressed upon should be carefully sought with percussive comparison of the right and left upper quadrants.

Splenomegaly, which occurs in many medical conditions, can be a subtle but significant physical finding in liver disease. The availability of ultrasound (US) methods for assessment of the spleen allows confirmation of the physical finding.

Signs of advanced liver disease include muscle wasting and weight loss as well as hepatomegaly, bruising, ascites, and edema. Ascites is best appreciated by attempts to detect shifting dullness by careful percussion. US examination will confirm the finding of ascites in equivocal cases. Peripheral edema can occur with or without ascites. In patients with advanced liver disease, other factors frequently contribute to edema formation, including hypoalbuminemia, venous insufficiency, heart failure, and medications.

Hepatic failure is defined as the occurrence of signs or symptoms of hepatic encephalopathy in a person with severe acute or chronic liver

disease. The first signs of hepatic encephalopathy can be subtle and nonspecific—change in sleep patterns, change in personality, irritability, and mental dullness. Thereafter, confusion, disorientation, stupor, and eventually coma supervene. In acute liver failure, excitability and mania may be present. Physical findings include asterixis and flapping tremors of the body and tongue. *Fetor hepaticus* refers to the slightly sweet, ammoniacal odor that can develop in patients with liver failure, particularly if there is portal-venous shunting of blood around the liver. Other causes of coma and disorientation should be excluded, mainly electrolyte imbalances, sedative use, and renal or respiratory failure. The appearance of hepatic encephalopathy during acute hepatitis is the major criterion for diagnosis of fulminant hepatitis and indicates a poor prognosis. In chronic liver disease, encephalopathy is usually triggered by a medical complication such as gastrointestinal bleeding, over-diuresis, uremia, dehydration, electrolyte imbalance, infection, constipation, or use of narcotic analgesics.

A helpful measure of hepatic encephalopathy is a careful mental-status examination and use of the trail-making test, which consists of a series of 25 numbered circles that the patient is asked to connect as rapidly as possible using a pencil. The normal range for the connect-the-dot test is 15–30 sec; it is considerably longer in patients with early hepatic encephalopathy. Other tests include drawing of abstract objects or comparison of a signature to previous examples. More sophisticated testing—for example, with electroencephalography and visual evoked potentials—can detect mild forms of encephalopathy but are rarely clinically useful.

Other signs of advanced liver disease include umbilical hernia from ascites, hydrothorax, prominent veins over the abdomen, and *caput medusa*, a condition that consists of collateral veins radiating from the umbilicus and results from recanulation of the umbilical vein. Widened pulse pressure and signs of a hyperdynamic circulation can occur in patients with cirrhosis as a result of fluid and sodium retention, increased cardiac output, and reduced peripheral resistance. Patients with long-standing cirrhosis and portal hypertension are prone to develop the hepatopulmonary syndrome, which is defined by the triad of liver disease, hypoxemia, and pulmonary arteriovenous shunting. The hepatopulmonary syndrome is characterized by platypnea and orthodeoxia: shortness of breath and oxygen desaturation that occur paradoxically upon the assumption of an upright position. Measurement of oxygen saturation by pulse oximetry is a reliable screening test for hepatopulmonary syndrome.

Several skin disorders and changes are common in liver disease. Hyperpigmentation is typical of advanced chronic cholestatic diseases such as primary biliary cholangitis and sclerosing cholangitis. In these same conditions, xanthelasma and tendon xanthomata occur as a result of retention and high serum levels of lipids and cholesterol. Slate-gray pigmentation of the skin is also seen with hemochromatosis if iron levels are high for a prolonged period. Mucocutaneous vasculitis with palpable purpura, especially on the lower extremities, is typical of cryoglobulinemia of chronic hepatitis C but can also occur in chronic hepatitis B.

Some physical signs point to specific liver diseases. Kayser-Fleischer rings occur in Wilson disease and consist of a golden-brown copper pigment deposited in Descemet's membrane at the periphery of the cornea; they are best seen by slit-lamp examination. Dupuytren contracture and parotid enlargement are suggestive of chronic alcoholism and alcoholic liver disease. In metastatic liver disease or primary hepatocellular carcinoma, signs of cachexia and wasting as well as firm hepatomegaly and a hepatic bruit may be prominent.

DIAGNOSIS OF LIVER DISEASE

The major causes of liver disease and key diagnostic features are outlined in Table 329-3, and an algorithm for evaluation of the patient with suspected liver disease is shown in Fig. 329-1. Specifics of diagnosis are discussed in later chapters. The most common causes of acute liver disease are viral hepatitis (particularly hepatitis A, B, and C), drug-induced liver injury, cholangitis, and alcoholic liver disease. Liver biopsy usually is not needed in the diagnosis and management of acute liver disease, exceptions being situations where the diagnosis remains unclear despite thorough clinical and laboratory investigation. Liver biopsy can be helpful in diagnosing drug-induced liver disease and acute alcoholic hepatitis.

TABLE 329-3 Important Diagnostic Tests in Common Liver Diseases

DISEASE	DIAGNOSTIC TEST
Hepatitis A	Anti-HAV IgM
Hepatitis B	
Acute	HBsAg and anti-HBc IgM
Chronic	HBsAg and HBeAg and/or HBV DNA
Hepatitis C	Anti-HCV and HCV RNA
Hepatitis D (delta)	HBsAg and anti-HDV
Hepatitis E	Anti-HEV IgM and HEV RNA
Autoimmune hepatitis	ANA or SMA, elevated IgG levels, and compatible histology
Primary biliary cholangitis	Mitochondrial antibody, elevated IgM levels, and compatible histology
Primary sclerosing cholangitis	P-ANCA, cholangiography
Drug-induced liver disease	History of drug ingestion
Alcoholic liver disease	History of excessive alcohol intake and compatible histology
Nonalcoholic steatohepatitis	Ultrasound or CT evidence of fatty liver and compatible histology
α_1 antitrypsin disease	Reduced α_1 antitrypsin levels, phenotype PiZZ or PiSZ
Wilson disease	Decreased serum ceruloplasmin and increased urinary copper; increased hepatic copper level
Hemochromatosis	Elevated iron saturation and serum ferritin; genetic testing for *HFE* gene mutations
Hepatocellular cancer	Elevated α-fetoprotein level (to >500 ng/mL); ultrasound or CT image of mass

Abbreviations: ANA, antinuclear antibody; anti-HBc, antibody to hepatitis B core (antigen); HAV, HBV, HCV, HDV, HEV: hepatitis A, B, C, D, E virus; HBeAg, hepatitis B e antigen; HBsAg, hepatitis B surface antigen; P-ANCA, peripheral antineutrophil cytoplasmic antibody; SMA, smooth-muscle antibody.

The most common causes of chronic liver disease, in general order of frequency, are chronic hepatitis C, alcoholic liver disease, nonalcoholic steatohepatitis, chronic hepatitis B, autoimmune hepatitis, sclerosing cholangitis, primary biliary cholangitis, hemochromatosis, and Wilson disease. Hepatitis E virus is a rare cause of chronic hepatitis, with cases occurring mostly in persons who are immunosuppressed or immunodeficient. Strict diagnostic criteria have not been developed for most liver diseases, but liver biopsy plays an important role in the diagnosis of autoimmune hepatitis, primary biliary cholangitis, nonalcoholic and alcoholic steatohepatitis, and Wilson disease (with a quantitative hepatic copper level in the last instance).

Laboratory Testing Diagnosis of liver disease is greatly aided by the availability of reliable and sensitive tests of liver injury and function. A typical battery of blood tests used for initial assessment of liver disease includes measurement of levels of serum alanine (ALT) and aspartate aminotransferases (AST), alkaline phosphatase (AlkP), direct and total serum bilirubin and albumin, and prothrombin time. The pattern of abnormalities generally points to hepatocellular versus cholestatic liver disease and helps determine whether the disease is acute or chronic and whether cirrhosis and hepatic failure are present. On the basis of these results, further testing over time may be necessary. Other laboratory tests may be helpful, such as γ-glutamyl transpeptidase (γGT) to define whether AlkP elevations are due to liver disease; hepatitis serology to define the type of viral hepatitis; and autoimmune markers to diagnose primary biliary cholangitis (antimitochondrial antibody), sclerosing cholangitis (peripheral antineutrophil cytoplasmic antibody), and autoimmune hepatitis (antinuclear, smooth-muscle, and liver-kidney microsomal antibody). A simple delineation of laboratory abnormalities and common liver diseases is given in Table 329-3.

The use and interpretation of liver function tests are summarized in Chap. 330.

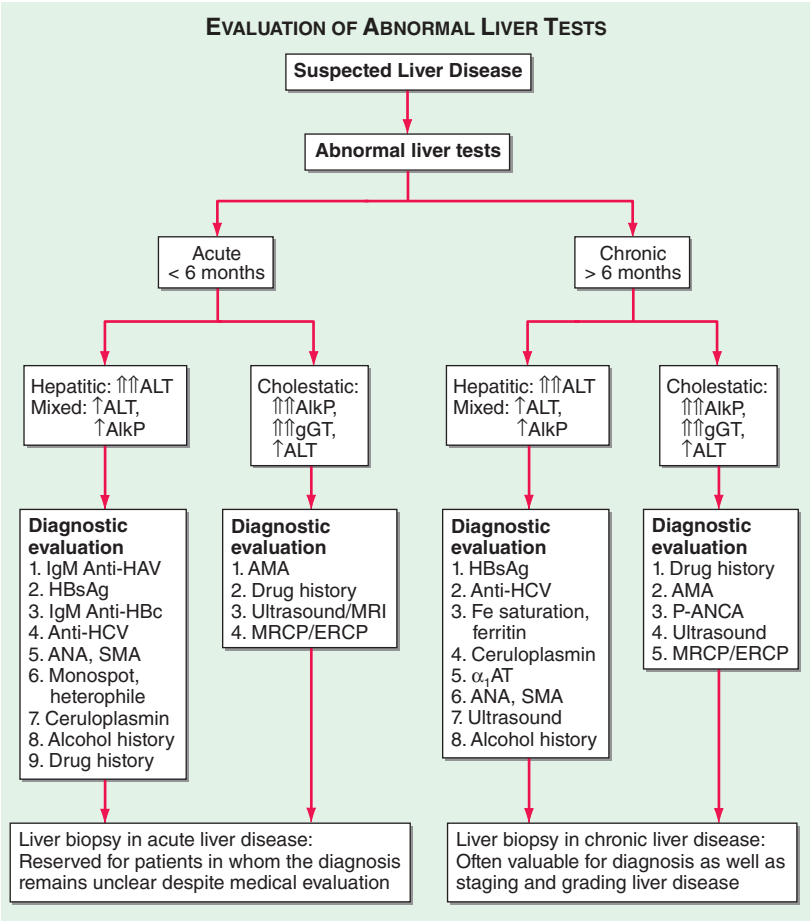


FIGURE 329-1 Algorithm for evaluation of abnormal liver tests. For patients with suspected liver disease, an appropriate approach to evaluation is initial routine liver testing—for example, measurement of serum bilirubin, albumin, alanine aminotransferase (ALT), AST, and AlkP. These results (sometimes complemented by testing of γ-glutamyl transpeptidase; gGT) will establish whether the pattern of abnormalities is hepatic, cholestatic, or mixed. In addition, the duration of symptoms or abnormalities will indicate whether the disease is acute or chronic. If the disease is acute and if history, laboratory tests, and imaging studies do not reveal a diagnosis, liver biopsy is appropriate to help establish the diagnosis. If the disease is chronic, liver biopsy can be helpful not only for diagnosis but also for grading of the activity and staging the progression of disease. This approach is generally applicable to patients without immune deficiency. In patients with HIV infection or recipients of bone marrow or solid organ transplants, the diagnostic evaluation should also include evaluation for opportunistic infections (e.g., with adenovirus, cytomegalovirus, *Coccidioides*, hepatitis E virus) as well as for vascular and immunologic conditions (venoocclusive disease, graft-versus-host disease). α_1 AT, α_1 antitrypsin; AMA; antimitochondrial antibody; ANA, antinuclear antibody; anti-HBc, antibody to hepatitis B core (antigen); ERCP, endoscopic retrograde cholangiopancreatography; HAV, hepatitis A virus; HBsAg, hepatitis B surface antigen; HCV, hepatitis C virus; MRCP, magnetic resonance cholangiopancreatography; P-ANCA, peripheral antineutrophil cytoplasmic antibody; SMA, smooth-muscle antibody.

Diagnostic Imaging Great advances have been made in hepatobiliary imaging, although no method is adequately accurate in demonstrating underlying cirrhosis in its early stages. Of the many modalities available for imaging the liver, US, computerized tomography (CT), and magnetic resonance imaging (MRI) are the most commonly employed and are complementary to one another. In general, US and CT are highly sensitive for detecting biliary duct dilation and are the first-line options for investigating cases of suspected obstructive jaundice. All three modalities can detect a fatty liver, which appears bright on imaging studies. Modifications of CT and MRI can be used to quantify liver fat, and this information may ultimately be valuable in monitoring therapy in patients with fatty liver disease. Magnetic resonance cholangiopancreatography (MRCP) and endoscopic retrograde cholangiopancreatography (ERCP) are the procedures of choice for visualization of the biliary tree. MRCP offers several advantages over ERCP: there is no need for contrast media or ionizing radiation, images can be acquired faster, the procedure is less operator dependent, and it carries no risk of pancreatitis. MRCP is superior to US and CT for detecting choledocholithiasis but is less specific. MRCP is useful in the diagnosis of bile duct obstruction and congenital biliary abnormalities, but ERCP is more valuable in evaluating ampullary lesions and primary sclerosing cholangitis. ERCP permits biopsy, direct visualization of the ampulla and common bile duct, and intraductal ultrasonography and brushings for cytological evaluation of malignancy. It also provides several therapeutic options in patients with obstructive jaundice, such as sphincterotomy, stone extraction, and placement of nasobiliary catheters and biliary stents. Doppler US and MRI are used to assess hepatic vasculature and hemodynamics and to monitor surgically or radiologically placed vascular shunts, including transjugular intrahepatic portosystemic shunts. Multidetector or spiral CT and MRI with contrast-enhancement are the procedures of choice for the identification and evaluation of hepatic masses, the staging of liver tumors, and preoperative assessment. With regard to mass lesions, the sensitivity of hepatic imaging continues to increase; unfortunately, specificity remains a problem, and often two and sometimes three studies are needed before a diagnosis can be reached. An emerging imaging modality for the investigation of hepatic lesions is contrast-enhanced US. This procedure permits enhancement of liver lesions in a similar fashion as contrast-enhanced, cross-sectional CT, or MR imaging. Major advantages are real-time assessment of liver perfusion throughout the vascular phases without risk of nephrotoxicity and radiation exposure. Other advantages are its widespread availability and lower cost. Limitations include body habitus of the patient and skill of the operator. US is the recommended modality for hepatocellular carcinoma screening. Contrast-enhanced US, CT, and MRI are appropriate for further investigation of lesions detected on screening US. The American College of Radiologists has developed a Liver Imaging Reporting and Data System (LI-RADS) to standardize the reporting and data collection of CT, MR, and contrast-enhanced US imaging for hepatocellular carcinoma (HCC). This system allows for more consistent reporting and reduces imaging interpretation variability and errors.

Recently, US transient elastography has been approved for the measurement of hepatic stiffness—providing an indirect assessment of fibrosis and cirrhosis; this technique can eliminate the need for liver biopsy if the only indication is the assessment of disease stage. MR elastography is more sensitive than US elastography, but is also more expensive and requires advanced scheduling and special equipment. Studies are ongoing to determine whether hepatic elastography is an appropriate means of monitoring fibrosis and disease progression. Finally, interventional radiologic techniques allow the biopsy of solitary lesions, the radiofrequency ablation and chemoembolization of cancerous lesions, the insertion of drains into hepatic abscesses, the measurement of portal pressure, and the creation of vascular shunts in patients with portal hypertension. Which modality to use depends on factors such as availability, cost, and experience of the radiologist with each technique.

Liver Biopsy Liver biopsy remains the gold standard in the evaluation of patients with liver disease, particularly chronic liver disease. Liver biopsy is necessary for diagnosis in selected instances but is more often useful for assessment of the severity (grade) and stage of liver damage, prediction of prognosis, and monitoring of the response to treatment. The size of the liver biopsy sample is an important determinant of reliability; a length of 1.5–2 cm is necessary for accurate assessment of fibrosis. Because liver biopsy is an invasive procedure and not without complications, it should be used only when it will contribute materially to decisions about management and therapy. In the future, noninvasive means of assessing disease activity (batteries of blood tests) and fibrosis

(elastography and fibrosis markers) may replace liver biopsy for the staging and grading of disease.

GRADING AND STAGING OF LIVER DISEASE

Grading refers to an assessment of the severity or activity of liver disease, whether acute or chronic; active or inactive; and mild, moderate, or severe. Liver biopsy is the most accurate means of assessing severity, particularly in chronic liver disease. Serum aminotransferase levels serve as convenient and noninvasive markers for disease activity but do not always reliably reflect disease severity. Thus, normal serum aminotransferase levels in patients with hepatitis B surface antigen in serum may indicate the inactive carrier state or may reflect mild chronic hepatitis B or hepatitis B with fluctuating disease activity. Serum testing for hepatitis B e antigen and hepatitis B virus DNA can help sort out these different patterns, but these markers can also fluctuate and change over time. Similarly, in chronic hepatitis C, serum aminotransferase levels can be normal despite moderate disease activity. Finally, in both alcoholic and nonalcoholic steatohepatitis, aminotransferase levels are quite unreliable in reflecting severity. In these conditions, liver biopsy is helpful in guiding management and identifying appropriate therapy, particularly if treatment is difficult, prolonged, and expensive, as is often the case in chronic viral hepatitis. Of the several well-verified numerical scales for grading activity in chronic liver disease, the most commonly used are the METAVIR, histology activity index and the Ishak fibrosis scale.

Liver biopsy is also the most accurate means of assessing stage of disease as early or advanced, precirrhotic, and cirrhotic. Staging of disease pertains largely to chronic liver diseases in which progression to cirrhosis and end-stage disease can occur but may require years or decades. Clinical features, biochemical tests, and hepatic imaging studies are helpful in assessing stage but generally become abnormal only in the middle to late stages of cirrhosis. Noninvasive tests that suggest advanced fibrosis include mild elevations of bilirubin, prolongation of prothrombin time, slight decreases in serum albumin, and mild thrombocytopenia (which is often the first indication of worsening fibrosis). Combinations of blood test results that include clinical features, routine laboratory tests, and special laboratory tests such as serum proteins or small molecules that are affected by or involved with fibrogenesis have been used to create models for predicting advanced liver disease, but these models are not reliable enough to use on a regular basis or for repeated measures and only separate advanced from early disease **(Table 329-4)**. Recently, elastography and noninvasive breath tests using ^{13}C-labeled compounds have been proposed as a means of detecting early stages of fibrosis and liver dysfunction, but their reliability and reproducibility remain to be proven. A major limitation of noninvasive markers is that they can be affected by disease activity. Even elastography is limited in this regard, in that it measures liver stiffness, not fibrosis per se, and can be affected by inflammation, edema, hepatocyte necrosis, and intrasinusoidal cellularity (inflammatory, malignant, or sickled cells). Thus, at present, mild to moderate stages of hepatic fibrosis are detectable only by liver biopsy. In the assessment of stage, the degree of fibrosis is usually used as the quantitative measure. The amount of fibrosis is generally staged on a scale of 0 to 4+ (METAVIR scale) or 0 to 6+ (Ishak scale). The importance of staging relates primarily to prognosis, recommendation of therapy and to optimal management of complications. Patients with cirrhosis are candidates for screening and surveillance for esophageal varices and hepatocellular carcinoma. Patients without advanced fibrosis need not undergo screening.

Once cirrhosis develops, other scoring systems are employed to assess compensated versus decompensated disease and prognosis. The initial staging system used for this purpose was the modified Child-Pugh classification, with a scoring system of 5–15: scores of 5 and 6 represent Child-Pugh class A (consistent with "compensated cirrhosis"), scores of 7–9 represent class B, and scores of 10–15 represent class C **(Table 329-5)**. This scoring system was initially devised to stratify patients with cirrhosis into risk groups before portal decompressive surgery. The Child-Pugh score is a reasonably reliable predictor of survival in many liver diseases and predicts the likelihood of major complications of cirrhosis, such as bleeding from varices and spontaneous bacterial peritonitis. This classification scheme was used to assess prognosis in cirrhosis and to provide standard criteria for listing a patient as a candidate for liver transplantation (Child-Pugh class B). More recently, the Child-Pugh system has been replaced by the Model for End-Stage Liver Disease (MELD) system for the latter purpose. The MELD score is a prospectively derived system designed to predict the prognosis of patients with liver disease and portal hypertension. Initially, this score was calculated from three noninvasive variables: the prothrombin time expressed as the international normalized ratio (INR), the serum bilirubin level, and the serum creatinine concentration. The ability of the MELD score to predict outcome after liver transplantation is regularly monitored and was modified to increase its accuracy and improve allocation of donated livers. These modifications include serum sodium concentration as a factor in the model and a reweighting of the MELD components. A separate scoring system pediatric end-stage liver disease (PELD) is used for children (<12 years of age). Transient elastography has also been used to stage cirrhosis and has been shown to be useful in predicting complications such as variceal hemorrhage, ascites development and liver-related death.

TABLE 329-4 Selected Noninvasive Methods of Assessing Hepatic Fibrosis and Cirrhosis

METHOD	PARAMETERS	ADVANCED FIBROSIS	CIRRHOSIS
APRI	AST, platelet count	>1	>1.5 (1–2)
ELF	Age, hyaluronic acid, MMP-3, TIMP-1	>7.7	>9.3
FIB-4	Age, AST, ALT, platelet count	>1.45	>3.25
*Fibro Test	Haptoglobin, α2-macroglobulin, apolipoprotein A1, γGT, total bilirubin	>0.45	>0.63
TE	Measures speed of a shear wave generated by vibration through liver tissue	>7.3 kPa	>15 kPa (9–26.5 kPa)
ARFI	Measures speed of shear wave generated by acoustic radiation force through liver tissue	>1.3 m/s	>1.87 m/s

*Patented models.

Note: The cutpoints presented in the table were mostly derived from patients with chronic hepatitis C. The cutpoints for the non-invasive models and tests presented in the table varies among different liver disease and among patients with the same disease among different populations.

Abbreviations: ALT, alanine aminotransferase; APRI, AST-to-Platelet Ratio; ARFI, acoustic radiation force imaging; AST, aspartate aminotransferase; ELF, Enhanced Liver Fibrosis Panel; γGT, gamma glutamyl-transpeptidase; MMP-3, metalloproteinase-3; TIMP-1, tissue inhibitor of metalloproteinase-1; TE, Transient Elastography.

TABLE 329-5 Child-Pugh Classification of Cirrhosis

FACTOR	POINTS TOWARD TOTAL SCORE			
	UNITS	1	2	3
Serum bilirubin	μmol/L	<34	34–51	>51
	mg/dL	<2.0	2.0–3.0	>3.0
Serum albumin	g/L	>35	30–35	<30
	g/dL	>3.5	3.0–3.5	<3.0
Prothrombin time	seconds prolonged	<4	4–6	>6
	INR[a]	<1.7	1.7–2.3	>2.3
Ascites		None	Easily controlled	Poorly controlled
Hepatic encephalopathy		None	Minimal	Advanced

[a]International normalized ratio.

Note: The Child-Pugh score is calculated by adding the scores for the five factors and can range from 5 to 15. The resulting Child-Pugh class can be A (a score of 5–6), B (7–9), or C (≥10). Decompensation indicates cirrhosis, with a Child-Pugh score of ≥7 (class B). This level has been the accepted criterion for listing a patient for liver transplantation.

The MELD system provides a more objective means of assessing disease severity and has less center-to-center variation than the Child-Pugh score as well as a wider range of values. The MELD and PELD systems are currently used to establish priority listing for liver transplantation in the United States. Convenient MELD and PELD calculators are available via the internet: *(http://optn.transplant.hrsa.gov/resources/MeldPeldCalculator.asp?index=98)*.

NONSPECIFIC ISSUES IN THE MANAGEMENT OF PATIENTS WITH LIVER DISEASE

Specifics on the management of different forms of acute or chronic liver disease are supplied in subsequent chapters, but certain issues are applicable to any patient with liver disease. These issues include advice regarding alcohol use, medication use, vaccination, and surveillance for complications of liver disease. Alcohol should be used sparingly, if at all, by patients with liver disease. Abstinence from alcohol should be encouraged for all patients with alcohol-related liver disease, patients with cirrhosis, and patients receiving interferon-based therapy for hepatitis B and during antiviral therapy of hepatitis C. With regard to vaccinations, all patients with liver disease should receive hepatitis A vaccine, and those with risk factors should receive hepatitis B vaccine as well. Influenza and pneumococcal vaccination should also be encouraged, with adherence to the recommendations of the Centers for Disease Control and Prevention. Patients with liver disease should exercise caution in using any medications other than those that are most necessary. Drug-induced hepatotoxicity can mimic many forms of liver disease and can cause exacerbations of chronic hepatitis and cirrhosis; drugs should be suspected in any situation in which the cause of exacerbation is unknown. Finally, consideration should be given to surveillance for complications of chronic liver disease such as variceal hemorrhage and hepatocellular carcinoma. Cirrhosis warrants upper endoscopy to assess the presence of varices, and the patient should receive chronic therapy with beta blockers or should be offered endoscopic obliteration if large varices are found. Moreover, cirrhosis warrants screening and long-term surveillance for development of hepatocellular carcinoma. While the optimal regimen for such surveillance has not been established, an appropriate approach is US of the liver at 6- to 12-month intervals.

FURTHER READING

Patel K, Bedossa P, Castera L: Diagnosis of liver fibrosis: Present and future. Semin Liver Dis 35:166, 2015.

330 Evaluation of Liver Function

Daniel S. Pratt

There are a number of tests that can be used to evaluate liver function. These tests include biochemical tests, radiologic tests, and pathologic tests.

Serum biochemical tests, also commonly referred to as "liver function tests," can be used to (1) detect the presence of liver disease, (2) distinguish among different types of liver disorders, (3) gauge the extent of known liver damage, and (4) follow the response to treatment. However, serum biochemical tests have shortcomings. They lack sensitivity and specificity; they can be normal in patients with serious liver disease and abnormal in patients with diseases that do not affect the liver. Liver tests rarely suggest a specific diagnosis; rather, they suggest a general category of liver disease, such as hepatocellular or cholestatic, which then further directs the evaluation. The liver carries out thousands of biochemical functions, most of which cannot be easily measured by blood tests. Laboratory tests measure only a limited number of these functions. In fact, many tests, such as the aminotransferases and alkaline phosphatase, do not measure liver function at all. Rather, they detect liver cell damage or interference with bile flow. Thus, no one biochemical test enables the clinician to accurately assess the liver's total functional capacity.

To increase the sensitivity and the specificity of biochemical tests in the detection of liver disease, it is best to use them as a battery. Tests usually employed in clinical practice include the bilirubin, aminotransferases, alkaline phosphatase, albumin, and prothrombin time tests. When more than one of these tests provide abnormal findings or the findings are persistently abnormal on serial determinations, the probability of liver disease is high. When all test results are normal, the probability of missing occult liver disease is low.

Serum Bilirubin (See also Chap. 45) Bilirubin, a breakdown product of the porphyrin ring of heme-containing proteins, is found in the blood in two fractions—conjugated and unconjugated. The unconjugated fraction, also termed the *indirect fraction*, is insoluble in water and is bound to albumin in the blood. The conjugated (direct) bilirubin fraction is water-soluble and can therefore be excreted by the kidney. Normal values of total serum bilirubin are reported between 1 and 1.5 mg/dL with 95% of a normal population falling between 0.2 and 0.9 mg/dL. If the direct-acting fraction is <15% of the total, the bilirubin can be considered to all be indirect. The most frequently reported upper limit of normal for conjugated bilirubin is 0.3 mg/dL.

Elevation of the unconjugated fraction of bilirubin is rarely due to liver disease. An isolated elevation of unconjugated bilirubin is seen primarily in hemolytic disorders and in a number of genetic conditions such as Crigler-Najjar and Gilbert's syndromes **(Chap. 45)**. Isolated unconjugated hyperbilirubinemia (bilirubin elevated but <15% direct) should prompt a workup for hemolysis **(Fig. 330-1)**. In the absence of hemolysis, an isolated, unconjugated hyperbilirubinemia in an otherwise healthy patient can be attributed to Gilbert's syndrome, and no further evaluation is required.

In contrast, conjugated hyperbilirubinemia almost always implies liver or biliary tract disease. The rate-limiting step in bilirubin metabolism is not conjugation of bilirubin, but rather the transport of conjugated bilirubin into the bile canaliculi. Thus, elevation of the conjugated fraction may be seen in any type of liver disease including fulminant liver failure. In most liver diseases, both conjugated and unconjugated fractions of the bilirubin tend to be elevated. Except in the presence of a purely unconjugated hyperbilirubinemia, fractionation of the bilirubin is rarely helpful in determining the cause of jaundice.

Although the degree of elevation of the serum bilirubin has not been critically assessed as a prognostic marker, it is important in a number of conditions. In viral hepatitis, the higher the serum bilirubin, the greater is the hepatocellular damage. Total serum bilirubin correlates with poor outcomes in alcoholic hepatitis. It is also a critical component of the Model for End-Stage Liver Disease (MELD) score, a tool used to estimate survival of patients with end-stage liver disease and assess operative risk of patients with cirrhosis. An elevated total serum bilirubin in patients with drug-induced liver disease indicates more severe injury.

Unconjugated bilirubin always binds to albumin in the serum and is not filtered by the kidney. Therefore, any bilirubin found in the urine is conjugated bilirubin; the presence of bilirubinuria implies the presence of liver disease. A urine dipstick test can theoretically give the same information as fractionation of the serum bilirubin. This test is almost 100% accurate. Phenothiazines may give a false-positive reading with the Ictotest tablet. In patients recovering from jaundice, the urine bilirubin clears prior to the serum bilirubin.

Serum Enzymes The liver contains thousands of enzymes, some of which are also present in the serum in very low concentrations. These enzymes have no known function in the serum and behave like other serum proteins. They are distributed in the plasma and in interstitial fluid and have characteristic half-lives, which are usually measured in days. Very little is known about the catabolism of serum enzymes, although they are probably cleared by cells in the reticuloendothelial system. The elevation of a given enzyme activity in the serum

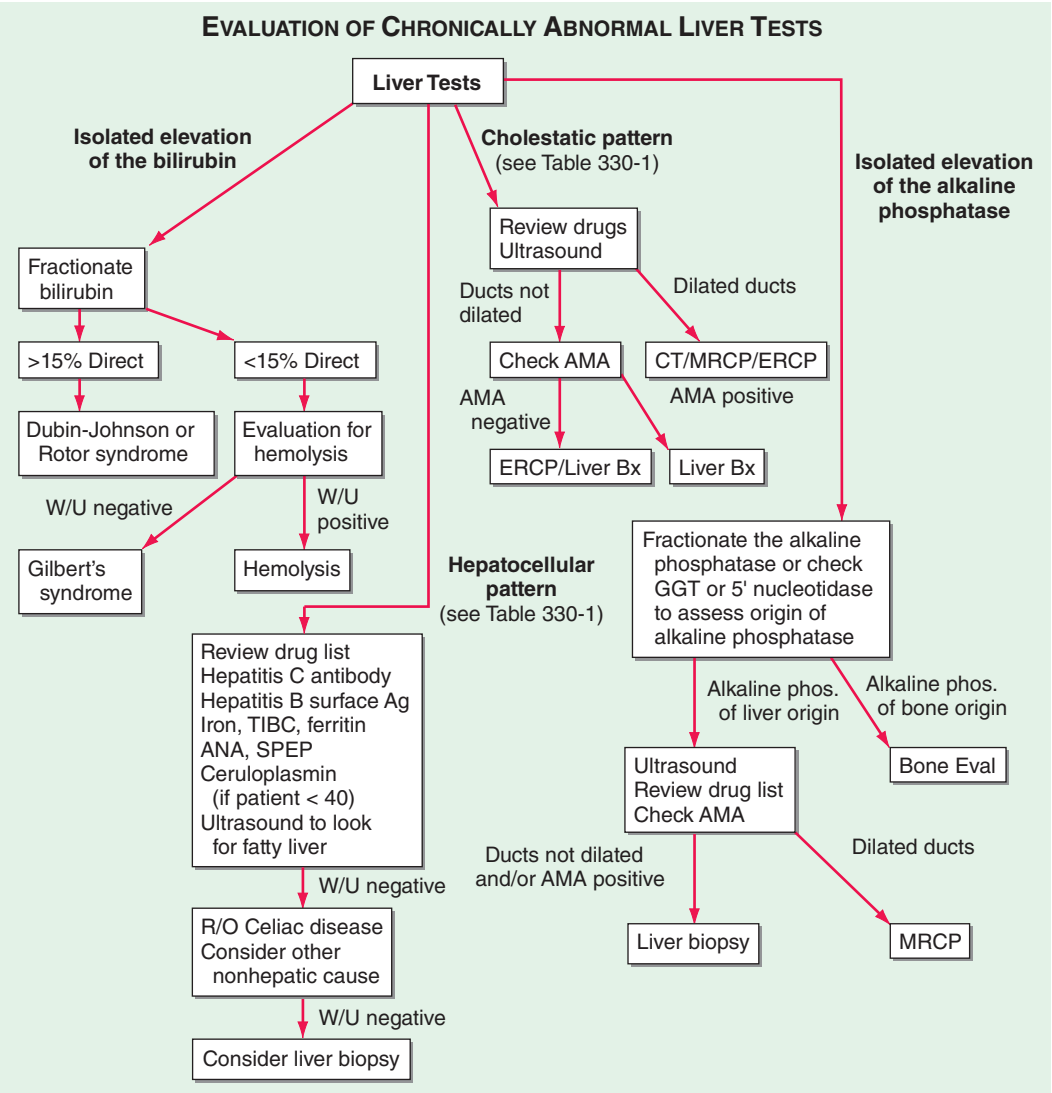


FIGURE 330-1 Algorithm for the evaluation of chronically abnormal liver tests. AMA, antimitochondrial antibody; ANA, antinuclear antibody; Bx, biopsy; CT, computed tomography; ERCP, endoscopic retrograde cholangiopancreatography; GGT, γ glutamyl transpeptidase; MRCP, magnetic resonance cholangiopancreatography; R/O, rule out; SPEP, serum protein electrophoresis; TIBC, total iron-binding capacity; W/U, workup.

is thought to primarily reflect its increased rate of entrance into serum from damaged liver cells.

Serum enzyme tests can be grouped into two categories: (1) enzymes whose elevation in serum reflects damage to hepatocytes and (2) enzymes whose elevation in serum reflects cholestasis.

ENZYMES THAT REFLECT DAMAGE TO HEPATOCYTES The aminotransferases (transaminases) are sensitive indicators of liver cell injury and are most helpful in recognizing acute hepatocellular diseases such as hepatitis. They include aspartate aminotransferase (AST) and alanine aminotransferase (ALT). AST is found in the liver, cardiac muscle, skeletal muscle, kidneys, brain, pancreas, lungs, leukocytes, and erythrocytes in decreasing order of concentration. ALT is found primarily in the liver and is therefore a more specific indicator of liver injury. The aminotransferases are normally present in the serum in low concentrations. These enzymes are released into the blood in greater amounts when there is damage to the liver cell membrane resulting in increased permeability. Liver cell necrosis is not required for the release of the aminotransferases, and there is a poor correlation between the degree of liver cell damage and the level of the aminotransferases. Thus, the absolute elevation of the aminotransferases is of no prognostic significance in acute hepatocellular disorders.

The normal range for aminotransferases varies widely among laboratories, but generally ranges from 10 to 40 IU/L. The interlaboratory variation in normal range is due to technical reasons; no reference standards exist to establish upper limits of normal for ALT and AST. Some have recommended revisions of normal limits of the aminotransferases to adjust for sex and body mass index, but others have noted the potential costs and unclear benefits of implementing this change.

Any type of liver cell injury can cause modest elevations in the serum aminotransferases. Levels of up to 300 IU/L are nonspecific and may be found in any type of liver disorder. Minimal ALT elevations in asymptomatic blood donors rarely indicate severe liver disease; studies have shown that fatty liver disease is the most likely explanation. Striking elevations—that is, aminotransferases >1000 IU/L—occur almost exclusively in disorders associated with extensive hepatocellular injury such as (1) viral hepatitis, (2) ischemic liver injury (prolonged hypotension or acute heart failure), or (3) toxin- or drug-induced liver injury.

The pattern of the aminotransferase elevation can be helpful diagnostically. In most acute hepatocellular disorders, the ALT is higher than or equal to the AST. Whereas the AST:ALT ratio is typically <1 in patients with chronic viral hepatitis and nonalcoholic fatty liver disease, a number of groups have noted that as cirrhosis develops, this ratio rises to >1. An AST:ALT ratio >2:1 is suggestive, whereas a ratio >3:1 is highly suggestive, of alcoholic liver disease. The AST in alcoholic liver disease is rarely >300 IU/L, and the ALT is often normal. A low level of ALT in the serum is due to an alcohol-induced deficiency of pyridoxal phosphate.

The aminotransferases are usually not greatly elevated in obstructive jaundice. One notable exception occurs during the acute phase of biliary obstruction caused by the passage of a gallstone into the common bile duct. In this setting, the aminotransferases can briefly be in the 1000–2000 IU/L range. However, aminotransferase levels decrease quickly, and the biochemical tests rapidly evolve into those typical of cholestasis.

ENZYMES THAT REFLECT CHOLESTASIS The activities of three enzymes—alkaline phosphatase, 5′-nucleotidase, and γ-glutamyl transpeptidase (GGT)—are usually elevated in cholestasis. Alkaline phosphatase and 5′-nucleotidase are found in or near the bile canalicular membrane of hepatocytes, whereas GGT is located in the endoplasmic reticulum and in bile duct epithelial cells. Reflecting its more diffuse localization in the liver, GGT elevation in serum is less specific for cholestasis than are elevations of alkaline phosphatase or 5′-nucleotidase. Some have advocated the use of GGT to identify patients with occult alcohol use. Its lack of specificity makes its use in this setting questionable.

The normal serum alkaline phosphatase consists of many distinct isoenzymes found in the liver, bone, placenta, and, less commonly, in the small intestine. Patients over age 60 can have a mildly elevated alkaline phosphatase (1–1.5 times normal), whereas individuals with blood types O and B can have an elevation of the serum alkaline phosphatase after eating a fatty meal due to the influx of intestinal alkaline phosphatase into the blood. It is also elevated in children and adolescents undergoing rapid bone growth because of bone alkaline phosphatase, and late in normal pregnancies due to the influx of placental alkaline phosphatase.

Elevation of liver-derived alkaline phosphatase is not totally specific for cholestasis, and a less than threefold elevation can be seen in almost any type of liver disease. Alkaline phosphatase elevations greater than four times normal occur primarily in patients with cholestatic liver disorders, infiltrative liver diseases such as cancer and amyloidosis, and bone conditions characterized by rapid bone turnover (e.g., Paget's disease). In bone diseases, the elevation is due to increased amounts of the bone isoenzymes. In liver diseases, the elevation is almost always due to increased amounts of the liver isoenzyme.

If an elevated serum alkaline phosphatase is the only abnormal finding in an apparently healthy person, or if the degree of elevation is higher than expected in the clinical setting, identification of the source of elevated isoenzymes is helpful (Fig. 330-1). This problem can be approached in two ways. First, and most precise, is the fractionation of the alkaline phosphatase by electrophoresis. The second, best substantiated, and most available approach involves the measurement of serum 5′-nucleotidase or GGT. These enzymes are rarely elevated in conditions other than liver disease.

In the absence of jaundice or elevated aminotransferases, an elevated alkaline phosphatase of liver origin often, but not always, suggests early cholestasis and, less often, hepatic infiltration by tumor or granulomata. Other conditions that cause isolated elevations of the alkaline phosphatase include Hodgkin's disease, diabetes, hyperthyroidism, congestive heart failure, amyloidosis, and inflammatory bowel disease.

The level of serum alkaline phosphatase elevation is not helpful in distinguishing between intrahepatic and extrahepatic cholestasis. There is essentially no difference among the values found in obstructive jaundice due to cancer, common duct stone, sclerosing cholangitis, or bile duct stricture. Values are similarly increased in patients with intrahepatic cholestasis due to drug-induced hepatitis, primary biliary cirrhosis, rejection of transplanted livers, and, rarely, alcohol-induced steatohepatitis. Values are also greatly elevated in hepatobiliary disorders seen in patients with AIDS (e.g., AIDS cholangiopathy due to cytomegalovirus or cryptosporidial infection and tuberculosis with hepatic involvement).

TESTS THAT MEASURE BIOSYNTHETIC FUNCTION OF THE LIVER

Serum Albumin Serum albumin is synthesized exclusively by hepatocytes. Serum albumin has a long half-life: 18–20 days, with ~4% degraded per day. Because of this slow turnover, the serum albumin is not a good indicator of acute or mild hepatic dysfunction; only minimal changes in the serum albumin are seen in acute liver conditions such as viral hepatitis, drug-related hepatotoxicity, and obstructive jaundice. In hepatitis, albumin levels <3 g/dL should raise the possibility of chronic liver disease. Hypoalbuminemia is more common in chronic liver disorders such as cirrhosis and usually reflects severe liver damage and decreased albumin synthesis. One exception is the patient with ascites in whom synthesis may be normal or even increased, but levels are low because of the increased volume of distribution. However, hypoalbuminemia is not specific for liver disease and may occur in protein malnutrition of any cause, as well as protein-losing enteropathies, nephrotic syndrome, and chronic infections that are associated with prolonged increases in levels of serum interleukin 1 and/or tumor necrosis factor, cytokines that inhibit albumin synthesis. Serum albumin should not be measured for screening in patients in whom there is no suspicion of liver disease. A general medical clinic study of consecutive patients in whom no indications were present for albumin measurement showed that although 12% of patients had abnormal test results, the finding was of clinical importance in only 0.4%.

Serum Globulins Serum globulins are a group of proteins made up of γ globulins (immunoglobulins) produced by B lymphocytes and α and β globulins produced primarily in hepatocytes. γ globulins are increased in chronic liver disease, such as chronic hepatitis and cirrhosis. In cirrhosis, the increased serum γ globulin concentration is due to the increased synthesis of antibodies, some of which are directed against intestinal bacteria. This occurs because the cirrhotic liver fails to clear bacterial antigens that normally reach the liver through the hepatic circulation.

Increases in the concentration of specific isotypes of γ globulins are often helpful in the recognition of certain chronic liver diseases. Diffuse polyclonal increases in IgG levels are common in autoimmune hepatitis; increases >100% should alert the clinician to this possibility. Increases in the IgM levels are common in primary biliary cirrhosis, whereas increases in the IgA levels occur in alcoholic liver disease.

COAGULATION FACTORS

With the exception of factor VIII, which is produced by vascular endothelial cells, the blood clotting factors are made exclusively in hepatocytes. Their serum half-lives are much shorter than albumin, ranging from 6 h for factor VII to 5 days for fibrinogen. Because of their rapid turnover, measurement of the clotting factors is the single best acute measure of hepatic synthetic function and helpful in both diagnosis and assessing the prognosis of acute parenchymal liver disease. Useful for this purpose is the *serum prothrombin time*, which collectively measures factors II, V, VII, and X. Biosynthesis of factors II, VII, IX, and X depends on vitamin K. The international normalized ratio (INR) is used to express the degree of anticoagulation on warfarin therapy. The INR standardizes prothrombin time measurement according to the characteristics of the thromboplastin reagent used in a particular lab, which is expressed as an International Sensitivity Index (ISI); the ISI is then used in calculating the INR.

The prothrombin time may be elevated in hepatitis and cirrhosis as well as in disorders that lead to vitamin K deficiency such as obstructive jaundice or fat malabsorption of any kind. Marked prolongation of the prothrombin time, >5 s above control and not corrected by parenteral vitamin K administration, is a poor prognostic sign in acute viral hepatitis and other acute and chronic liver diseases. The INR, along with the total serum bilirubin and creatinine, are components of the MELD score, which is used as a measure of hepatic decompensation and to allocate organs for liver transplantation.

OTHER DIAGNOSTIC TESTS

Although tests may direct the physician to a category of liver disease, additional biochemical testing, radiologic testing, and procedures are often necessary to make the proper diagnosis, as shown in Fig. 330-1. The most commonly used ancillary tests are reviewed here, as are the noninvasive tests available for assessing hepatic fibrosis.

Ammonia Ammonia is produced in the body during normal protein metabolism and by intestinal bacteria, primarily those in the colon.

The liver plays a role in the detoxification of ammonia by converting it to urea, which is excreted by the kidneys. Striated muscle also plays a role in detoxification of ammonia, where it is combined with glutamic acid to form glutamine. Patients with advanced liver disease typically have significant muscle wasting, which likely contributes to hyperammonemia. Some physicians use the blood ammonia for detecting encephalopathy or for monitoring hepatic synthetic function, although its use for either of these indications has problems. There is very poor correlation between either the presence or the severity of acute encephalopathy and elevation of blood ammonia; it can be occasionally useful for identifying occult liver disease in patients with mental status changes. There is also a poor correlation of the blood serum ammonia and hepatic function. The ammonia can be elevated in patients with severe portal hypertension and portal blood shunting around the liver even in the presence of normal or near-normal hepatic function. Elevated arterial ammonia levels have been shown to correlate with outcome in fulminant hepatic failure.

Liver Biopsy Percutaneous biopsy of the liver is a safe procedure that can be easily performed at the bedside with local anesthesia and ultrasound guidance. Liver biopsy is of proven value in the following situations: (1) hepatocellular disease of uncertain cause, (2) prolonged hepatitis with the possibility of autoimmune hepatitis, (3) unexplained hepatomegaly, (4) unexplained splenomegaly, (5) hepatic lesions uncharacterized by radiologic imaging, (6) fever of unknown origin, and (7) staging of malignant lymphoma. Liver biopsy is most accurate in disorders causing diffuse changes throughout the liver and is subject to sampling error in focal disorders. Liver biopsy should not be the initial procedure in the diagnosis of cholestasis. The biliary tree should first be assessed for signs of obstruction. Contraindications to performing a percutaneous liver biopsy include significant ascites and prolonged INR. Under these circumstances, the biopsy can be performed via the transjugular approach.

Noninvasive Tests to Detect Hepatic Fibrosis Although liver biopsy is the standard for the assessment of hepatic fibrosis, noninvasive measures of hepatic fibrosis have been developed and show promise. These measures include multiparameter tests aimed at detecting and staging the degree of hepatic fibrosis and imaging techniques. FibroTest (marketed as FibroSure in the United States) is the best evaluated of the multiparameter blood tests. The test incorporates haptoglobin, bilirubin, GGT, apolipoprotein A-I, and α2-macroglobulin and has been found to have high positive and negative predictive values for diagnosing advanced fibrosis in patients with chronic hepatitis C, chronic hepatitis B, and alcoholic liver disease and patients taking methotrexate for psoriasis. Transient elastography (TE), marketed as FibroScan, and magnetic resonance elastography (MRE) both have gained U.S. Food and Drug Administration approval for use in the management of patients with liver disease. TE uses ultrasound waves to measure hepatic stiffness noninvasively. TE has been shown to be accurate for identifying advanced fibrosis in patients with chronic hepatitis C, primary biliary cirrhosis, hemochromatosis, nonalcoholic fatty liver disease, and recurrent chronic hepatitis after liver transplantation. MRE has been found to be superior to TE for staging liver fibrosis in patients with a variety of chronic liver diseases, but requires access to a magnetic resonance imaging scanner.

Ultrasonography Ultrasonography is the first diagnostic test to use in patients whose liver tests suggest cholestasis, to look for the presence of a dilated intrahepatic or extrahepatic biliary tree or to identify gallstones. In addition, it shows space-occupying lesions within the liver, enables the clinician to distinguish between cystic and solid masses, and helps direct percutaneous biopsies. Ultrasound with Doppler imaging can detect the patency of the portal vein, hepatic artery, and hepatic veins and determine the direction of blood flow. This is the first test ordered in patients suspected of having Budd-Chiari syndrome.

USE OF LIVER TESTS

As previously noted, the best way to increase the sensitivity and specificity of laboratory tests in the detection of liver disease is to employ a battery of tests that includes the aminotransferases, alkaline phosphatase, bilirubin, albumin, and prothrombin time along with the judicious use of the other tests described in this chapter. Table 330-1 shows how patterns of liver tests can lead the clinician to a category of disease that will direct further evaluation. However, it is important to remember that no single set of liver tests will necessarily provide a diagnosis. It is often necessary to repeat these tests on several occasions over days to weeks for a diagnostic pattern to emerge. Figure 330-1 is an algorithm for the evaluation of chronically abnormal liver tests.

TABLE 330-1 Liver Test Patterns in Hepatobiliary Disorders

TYPE OF DISORDER	BILIRUBIN	AMINOTRANSFERASES	ALKALINE PHOSPHATASE	ALBUMIN	PROTHROMBIN TIME
Hemolysis/Gilbert's syndrome	Normal to 86 μmol/L (5 mg/dL) 85% due to indirect fractions No bilirubinuria	Normal	Normal	Normal	Normal
Acute hepatocellular necrosis (viral and drug hepatitis, hepatotoxins, acute heart failure)	Both fractions may be elevated Peak usually follows aminotransferases Bilirubinuria	Elevated, often >500 IU, ALT > AST	Normal to <3× normal elevation	Normal	Usually normal. If >5× above control and not corrected by parenteral vitamin K, suggests poor prognosis
Chronic hepatocellular disorders	Both fractions may be elevated Bilirubinuria	Elevated, but usually <300 IU	Normal to <3× normal elevation	Often decreased	Often prolonged Fails to correct with parenteral vitamin K
Alcoholic hepatitis, cirrhosis	Both fractions may be elevated Bilirubinuria	AST:ALT >2 suggests alcoholic hepatitis or cirrhosis	Normal to <3× normal elevation	Often decreased	Often prolonged Fails to correct with parenteral vitamin K
Intra- and extrahepatic cholestasis	Both fractions may be elevated	Normal to moderate elevation	Elevated, often >4× normal elevation	Normal, unless chronic	Normal If prolonged, will correct with parenteral vitamin K
(Obstructive jaundice) Infiltrative diseases (tumor, granulomata); partial bile duct obstruction	Bilirubinuria Usually normal	Rarely >500 IU Normal to slight elevation	Elevated, often >4× normal elevation Fractionate, or confirm liver origin with 5′-nucleotidase or γ glutamyl transpeptidase	Normal	Normal

GLOBAL CONSIDERATIONS

The tests and principles presented in this chapter are applicable worldwide. The causes of liver test abnormalities vary according to region. In developing nations, infectious diseases are more commonly the etiology of abnormal serum liver tests than in developed nations.

ACKNOWLEDGMENT

This chapter represents a revised version of a chapter in previous editions of Harrison's in which Marshall M. Kaplan was a co-author.

FURTHER READING

KAMATH PS, KIM WR: The Model for End-Stage Liver Disease (MELD). Hepatology 45:797, 2007.

KAPLAN M: Alkaline phosphatase. Gastroenterology 62:452, 1972.

MARTÍNEZ SM et al: Noninvasive assessment of liver fibrosis. Hepatology 53:325, 2011.

PRATI D et al: Updated definitions of healthy ranges for serum alanine aminotransferase levels. Ann Intern Med 137:1, 2002.

331 The Hyperbilirubinemias

Allan W. Wolkoff

BILIRUBIN METABOLISM

The details of bilirubin metabolism are presented in Chap. 45. However, the hyperbilirubinemias are best understood in terms of perturbations of specific aspects of bilirubin metabolism and transport, and these will be briefly reviewed here as depicted in Fig. 331-1.

Bilirubin is the end product of heme degradation. Some 70–90% of bilirubin is derived from degradation of the hemoglobin of senescent red blood cells. Bilirubin produced in the periphery is transported to the liver within the plasma, where, due to its insolubility in aqueous solutions, it is tightly bound to albumin. Under normal circumstances, bilirubin is removed from the circulation rapidly and efficiently by hepatocytes. Transfer of bilirubin from blood to bile involves four distinct but interrelated steps (Fig. 331-1).

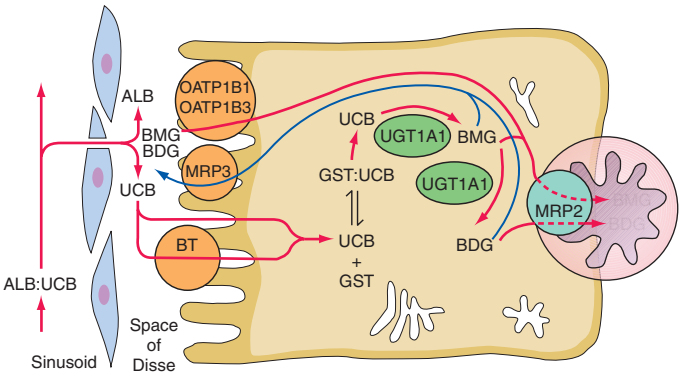


FIGURE 331-1 Hepatocellular bilirubin transport. Albumin-bound bilirubin in sinusoidal blood passes through endothelial cell fenestrae to reach the hepatocyte surface, entering the cell by both facilitated and simple diffusional processes. Within the cell, it is bound to glutathione-S-transferases and conjugated by bilirubin-UDP-glucuronosyltransferase (UGT1A1) to mono- and diglucuronides, which are actively transported across the canalicular membrane into the bile. In addition to this direct excretion of bilirubin glucuronides, a portion are transported into the portal circulation by MRP3 and subjected to reuptake into the hepatocyte by OATP1B1 and OATP1B3. ALB, albumin; BDG, bilirubin diglucuronide; BMG, bilirubin monoglucuronide; BT, proposed bilirubin transporter; GST, glutathione-S-transferase; MRP2 and MRP3, multidrug resistance–associated proteins 2 and 3; OATP1B1 and OATP1B3, organic anion transport proteins 1B1 and 1B3; UCB, unconjugated bilirubin; UGT1A1, bilirubin-UDP-glucuronosyltransferase.

1. *Hepatocellular uptake:* Uptake of bilirubin by the hepatocyte has carrier-mediated kinetics. Although a number of candidate bilirubin transporters have been proposed, the actual transporter remains elusive.
2. *Intracellular binding:* Within the hepatocyte, bilirubin is kept in solution by binding as a nonsubstrate ligand to several of the glutathione-S-transferases, formerly called ligandins.
3. *Conjugation:* Bilirubin is conjugated with one or two glucuronic acid moieties by a specific UDP-glucuronosyltransferase to form bilirubin mono- and diglucuronide, respectively. Conjugation disrupts the internal hydrogen bonding that limits aqueous solubility of bilirubin, and the resulting glucuronide conjugates are highly soluble in water. Conjugation is obligatory for excretion of bilirubin across the bile canalicular membrane into bile. The UDP-glucuronosyltransferases have been classified into gene families based on the degree of homology among the mRNAs for the various isoforms. Those that conjugate bilirubin and certain other substrates have been designated the *UGT1* family. These are expressed from a single gene complex by alternative promoter usage. This gene complex contains multiple substrate-specific first exons, designated A1, A2, etc. (Fig. 331-2), each with its own promoter and each encoding the amino-terminal half of a specific isoform. In addition, there are four common exons (exons 2–5) that encode the shared carboxyl-terminal half of all of the *UGT1* isoforms. The various first exons encode the specific aglycone substrate binding sites for each isoform, while the shared exons encode the binding site for the sugar donor, UDP-glucuronic acid, and the transmembrane domain. Exon A1 and the four common exons, collectively designated as the UGT1A1 gene (Fig. 331-2), encode the physiologically critical enzyme bilirubin-UDP-glucuronosyltransferase (UGT1A1). A functional corollary of the organization of the *UGT1* gene is that a mutation in one of the first exons will affect only a single enzyme isoform. By contrast, a mutation in exons 2–5 will alter all isoforms encoded by the *UGT1* gene complex.
4. *Biliary excretion:* It has been thought until recently that bilirubin mono- and diglucuronides are excreted directly across the canalicular plasma membrane into the bile canaliculus by an ATP-dependent transport process mediated by a canalicular membrane protein called *multidrug resistance–associated protein 2* (MRP2). Mutations of MRP2 result in the Dubin-Johnson syndrome (see below). However, studies in patients with Rotor syndrome (see below) indicate that after formation, a portion of the glucuronides is transported into the portal circulation by a sinusoidal membrane protein called *multidrug resistance–associated protein 3* (MRP3) and is subjected to reuptake into the hepatocyte by the sinusoidal membrane uptake transporters *organic anion transport protein 1B1* (OATP1B1) and OATP1B3.

EXTRAHEPATIC ASPECTS OF BILIRUBIN DISPOSITION

Bilirubin in the Gut Following secretion into bile, conjugated bilirubin reaches the duodenum and passes down the gastrointestinal tract without reabsorption by the intestinal mucosa. An appreciable fraction is converted by bacterial metabolism in the gut to the water-soluble colorless compound urobilinogen. Urobilinogen undergoes enterohepatic cycling. Urobilinogen not taken up by the liver reaches the systemic circulation, from which some is cleared by the kidneys. Unconjugated bilirubin ordinarily does not reach the gut except in neonates or, by ill-defined alternative pathways, in the presence of severe unconjugated hyperbilirubinemia (e.g., Crigler-Najjar syndrome, type I [CN-I]). Unconjugated bilirubin that reaches the gut is partly reabsorbed, amplifying any underlying hyperbilirubinemia.

Renal Excretion of Bilirubin Conjugates Unconjugated bilirubin is not excreted in urine, as it is too tightly bound to albumin for effective glomerular filtration and there is no tubular mechanism for its renal secretion. In contrast, the bilirubin conjugates are readily filtered at the glomerulus and can appear in urine in disorders characterized by increased bilirubin conjugates in the circulation. It should

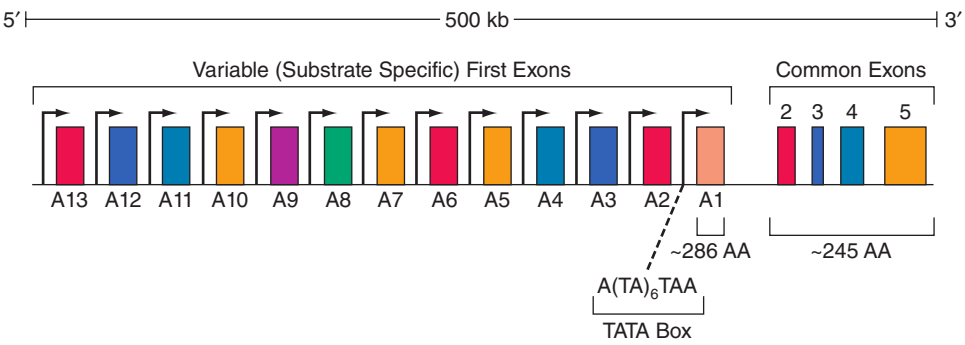


FIGURE 331-2 Structural organization of the human *UGT1* gene complex. This large complex on chromosome 2 contains at least 13 substrate-specific first exons (A1, A2, etc.). Since four of these are pseudogenes, nine UGT1 isoforms with differing substrate specificities are expressed. Each exon 1 has its own promoter and encodes the amino-terminal substrate-specific ~286 amino acids of the various *UGT1*-encoded isoforms, and common exons 2–5 that encode the 245 carboxyl-terminal amino acids common to all of the isoforms. mRNAs for specific isoforms are assembled by splicing a particular first exon such as the bilirubin-specific exon A1 to exons 2 to 5. The resulting message encodes a complete enzyme, in this particular case bilirubin-UDP-glucuronosyltransferase (UGT1A1). Mutations in a first exon affect only a single isoform. Those in exons 2–5 affect all enzymes encoded by the UGT1 complex.

be kept in mind that the kidney can serve as an "overflow valve" for conjugated bilirubin. Consequently, the level of jaundice in individuals with conjugated hyperbilirubinemia can be amplified in the presence of renal failure.

DISORDERS OF BILIRUBIN METABOLISM LEADING TO UNCONJUGATED HYPERBILIRUBINEMIA

INCREASED BILIRUBIN PRODUCTION

Hemolysis Increased destruction of erythrocytes leads to increased bilirubin turnover and unconjugated hyperbilirubinemia; the hyperbilirubinemia is usually modest in the presence of normal liver function. In particular, the bone marrow is only capable of a sustained eightfold increase in erythrocyte production in response to a hemolytic stress. Therefore, hemolysis alone cannot result in a sustained hyperbilirubinemia of more than ~68 μmol/L (4 mg/dL). Higher values imply concomitant hepatic dysfunction. When hemolysis is the only abnormality in an otherwise healthy individual, the result is a purely unconjugated hyperbilirubinemia, with the direct-reacting fraction as measured in a typical clinical laboratory being ≤15% of the total serum bilirubin. In the presence of systemic disease, which may include a degree of hepatic dysfunction, hemolysis may produce a component of conjugated hyperbilirubinemia in addition to an elevated unconjugated bilirubin concentration. Prolonged hemolysis may lead to the precipitation of bilirubin salts within the gallbladder or biliary tree, resulting in the formation of gallstones in which bilirubin, rather than cholesterol, is the major component. Such pigment stones may lead to acute or chronic cholecystitis, biliary obstruction, or any other biliary tract consequence of calculous disease.

Ineffective Erythropoiesis During erythroid maturation, small amounts of hemoglobin may be lost at the time of nuclear extrusion, and a fraction of developing erythroid cells is destroyed within the marrow. These processes normally account for a small proportion of bilirubin that is produced. In various disorders, including thalassemia major, megaloblastic anemias due to folate or vitamin B_{12} deficiency, congenital erythropoietic porphyria, lead poisoning, and various congenital and acquired dyserythropoietic anemias, the fraction of total bilirubin production derived from ineffective erythropoiesis is increased, reaching as much as 70% of the total. This may be sufficient to produce modest degrees of unconjugated hyperbilirubinemia.

Miscellaneous Degradation of the hemoglobin of extravascular collections of erythrocytes, such as those seen in massive tissue infarctions or large hematomas, may lead transiently to unconjugated hyperbilirubinemia.

DECREASED HEPATIC BILIRUBIN CLEARANCE

Decreased Hepatic Uptake Decreased hepatic bilirubin uptake is believed to contribute to the unconjugated hyperbilirubinemia of Gilbert's syndrome (GS), although the molecular basis for this finding remains unclear (see below). Several drugs, including flavaspidic acid, novobiocin, and rifampin, as well as various cholecystographic contrast agents, have been reported to inhibit bilirubin uptake. The resulting unconjugated hyperbilirubinemia resolves with cessation of the medication.

Impaired Conjugation • **PHYSIOLOGIC NEONATAL JAUNDICE** Bilirubin produced by the fetus is cleared by the placenta and eliminated by the maternal liver. Immediately after birth, the neonatal liver must assume responsibility for bilirubin clearance and excretion. However, many hepatic physiologic processes are incompletely developed at birth. Levels of UGT1A1 are low, and alternative excretory pathways allow passage of unconjugated bilirubin into the gut. Since the intestinal flora that convert bilirubin to urobilinogen are also undeveloped, an enterohepatic circulation of unconjugated bilirubin ensues. As a consequence, most neonates develop mild unconjugated hyperbilirubinemia between days 2 and 5 after birth. Peak levels are typically <85–170 μmol/L (5–10 mg/dL) and decline to normal adult concentrations within 2 weeks, as mechanisms required for bilirubin disposition mature. Prematurity, often associated with more profound immaturity of hepatic function and hemolysis, can result in higher levels of unconjugated hyperbilirubinemia. A rapidly rising unconjugated bilirubin concentration, or absolute levels >340 μmol/L (20 mg/dL), puts the infant at risk for bilirubin encephalopathy, or kernicterus. Under these circumstances, bilirubin crosses an immature blood-brain barrier and precipitates in the basal ganglia and other areas of the brain. The consequences range from appreciable neurologic deficits to death. Treatment options include phototherapy, which converts bilirubin into water-soluble photoisomers that are excreted directly into bile, and exchange transfusion. The canalicular mechanisms responsible for bilirubin excretion are also immature at birth, and their maturation may lag behind that of UGT1A1; this can lead to transient conjugated neonatal hyperbilirubinemia, especially in infants with hemolysis.

ACQUIRED CONJUGATION DEFECTS A modest reduction in bilirubin conjugating capacity may be observed in advanced hepatitis or cirrhosis. However, in this setting, conjugation is better preserved than other aspects of bilirubin disposition, such as canalicular excretion. Various drugs, including pregnanediol, novobiocin, chloramphenicol, gentamicin, and atazanavir may produce unconjugated hyperbilirubinemia by inhibiting UGT1A1 activity. Bilirubin conjugation may be inhibited by certain fatty acids that are present in breast milk, but not serum of mothers whose infants have excessive neonatal hyperbilirubinemia (*breast milk jaundice*). Alternatively, there may be increased enterohepatic circulation of bilirubin in these infants. The pathogenesis of breast milk jaundice appears to differ from that of transient familial neonatal hyperbilirubinemia (Lucey-Driscoll syndrome), in which there may be a UGT1A1 inhibitor in maternal serum.

HEREDITARY DEFECTS IN BILIRUBIN CONJUGATION

Three familial disorders characterized by differing degrees of unconjugated hyperbilirubinemia have long been recognized. The defining clinical features of each are described below **(Table 331-1)**. While these disorders have been recognized for decades to reflect differing degrees of deficiency in the ability to conjugate bilirubin, recent advances in the molecular biology of the *UGT1* gene complex have elucidated their interrelationships and clarified previously puzzling features.

TABLE 331-1 Principal Differential Characteristics of Gilbert's and Crigler-Najjar Syndromes

FEATURE	CRIGLER-NAJJAR SYNDROME		GILBERT'S SYNDROME
	TYPE I	TYPE II	
Total serum bilirubin, μmol/L (mg/dL)	310–755 (usually >345) (18–45 [usually >20])	100–430 (usually ≤345) (6–25 [usually ≤20])	Typically ≤70 μmol/L (≤4 mg/dL) in absence of fasting or hemolysis
Routine liver tests	Normal	Normal	Normal
Response to phenobarbital	None	Decreases bilirubin by >25%	Decreases bilirubin to normal
Kernicterus	Usual	Rare	No
Hepatic histology	Normal	Normal	Usually normal; increased lipofuscin pigment in some
Bile characteristics			
Color	Pale or colorless	Pigmented	Normal dark color
Bilirubin fractions	>90% unconjugated	Largest fraction (mean: 57%) monoconjugates	Mainly diconjugates but monoconjugates increased (mean: 23%)
Bilirubin UDP-glucuronosyltransferase activity	Typically absent; traces in some patients	Markedly reduced: 0–10% of normal	Reduced: typically 10–33% of normal
Inheritance (all autosomal)	Recessive	Predominantly recessive	Promoter mutation: recessive Missense mutations: 7 of 8 dominant; 1 reportedly recessive

Crigler-Najjar Syndrome, Type I CN-I is characterized by striking unconjugated hyperbilirubinemia of about 340–765 μmol/L (20–45 mg/dL) that appears in the neonatal period and persists for life. Other conventional hepatic biochemical tests such as serum aminotransferases and alkaline phosphatase are normal, and there is no evidence of hemolysis. Hepatic histology is also essentially normal except for the occasional presence of bile plugs within canaliculi. Bilirubin glucuronides are virtually absent from the bile, and there is no detectable constitutive expression of UGT1A1 activity in hepatic tissue. Neither UGT1A1 activity nor the serum bilirubin concentration responds to administration of phenobarbital or other enzyme inducers. Unconjugated bilirubin accumulates in plasma, from which it is eliminated very slowly by alternative pathways that include direct passage into the bile and small intestine, possibly via bilirubin photoisomers. This accounts for the small amount of urobilinogen found in feces. No bilirubin is found in the urine. First described in 1952, the disorder is rare (estimated prevalence, 0.6–1.0 per million). Many patients are from geographically or socially isolated communities in which consanguinity is common, and pedigree analyses show an autosomal recessive pattern of inheritance. The majority of patients (type IA) exhibit defects in the glucuronide conjugation of a spectrum of substrates in addition to bilirubin, including various drugs and other xenobiotics. These individuals have mutations in one of the common exons (2–5) of the *UGT1* gene (Fig. 331-2). In a smaller subset (type IB), the defect is limited largely to bilirubin conjugation, and the causative mutation is in the bilirubin-specific exon A1. Estrogen glucuronidation is mediated by UGT1A1 and is defective in all CN-I patients. More than 30 different genetic lesions of *UGT1A1* responsible for CN-I have been identified, including deletions, insertions, alterations in intron splice donor and acceptor sites, exon skipping, and point mutations that introduce premature stop codons or alter critical amino acids. Their common feature is that they all encode proteins with absent or, at most, traces of bilirubin-UDP-glucuronosyltransferase enzymatic activity.

Prior to the use of phototherapy, most patients with CN-I died of bilirubin encephalopathy (*kernicterus*) in infancy or early childhood. A few lived as long as early adult life without overt neurologic damage, although more subtle testing usually indicated mild but progressive brain damage. In the absence of liver transplantation, death eventually supervened from late-onset bilirubin encephalopathy, which often followed a nonspecific febrile illness. Although isolated hepatocyte transplantation has been used in a small number of cases of CN-I, early liver transplantation (Chap. 338) remains the best hope to prevent brain injury and death at present. It is anticipated that gene replacement therapy may be an option in the future.

Crigler-Najjar Syndrome, Type II (CN-II) This condition was recognized as a distinct entity in 1962 and is characterized by marked unconjugated hyperbilirubinemia in the absence of abnormalities of other conventional hepatic biochemical tests, hepatic histology, or hemolysis. It differs from CN-I in several specific ways (Table 331-1): (1) Although there is considerable overlap, average bilirubin concentrations are lower in CN-II; (2) accordingly, CN-II is only infrequently associated with kernicterus; (3) bile is deeply colored, and bilirubin glucuronides are present, with a striking, characteristic increase in the proportion of monoglucuronides; (4) UGT1A1 in liver is usually present at reduced levels (typically ≤10% of normal); and (5) while typically detected in infancy, hyperbilirubinemia was not recognized in some cases until later in life and, in one instance, at age 34. As with CN-I, most CN-II cases exhibit abnormalities in the conjugation of other compounds, such as salicylamide and menthol, but in some instances, the defect appears limited to bilirubin. Reduction of serum bilirubin concentrations by >25% in response to enzyme inducers such as phenobarbital distinguishes CN-II from CN-I, although this response may not be elicited in early infancy and often is not accompanied by measurable UGT1A1 induction. Bilirubin concentrations during phenobarbital administration do not return to normal but are typically in the range of 51–86 μmol/L (3–5 mg/dL). Although the incidence of kernicterus in CN-II is low, instances have occurred, not only in infants but also in adolescents and adults, often in the setting of an intercurrent illness, fasting, or another factor that temporarily raises the serum bilirubin concentration above baseline and reduces serum albumin levels. For this reason, phenobarbital therapy is widely recommended, a single bedtime dose often sufficing to maintain clinically safe serum bilirubin concentrations.

Over 100 different mutations in the *UGT1* gene have been identified as causing CN-I or CN-II. It was found that missense mutations are more common in CN-II patients, as would be expected in this less severe phenotype. Their common feature is that they encode for a bilirubin-UDP-glucuronosyltransferase with markedly reduced, but detectable, enzymatic activity. The spectrum of residual enzyme activity explains the spectrum of phenotypic severity of the resulting hyperbilirubinemia. Molecular analysis has established that a large majority of CN-II patients are either homozygotes or compound heterozygotes for CN-II mutations and that individuals carrying one mutated and one entirely normal allele have normal bilirubin concentrations.

Gilbert's Syndrome This syndrome is characterized by mild unconjugated hyperbilirubinemia, normal values for standard hepatic biochemical tests, and normal hepatic histology other than a modest increase of lipofuscin pigment in some patients. Serum bilirubin concentrations are most often <51 μmol/L (<3 mg/dL), although both higher and lower values are frequent. The clinical spectrum of hyperbilirubinemia fades into that of CN-II at serum bilirubin concentrations of 86–136 μmol/L (5–8 mg/dL). At the other end of the scale,

the distinction between mild cases of GS and a normal state is often blurred. Bilirubin concentrations may fluctuate substantially in any given individual, and at least 25% of patients will exhibit temporarily normal values during prolonged follow-up. More elevated values are associated with stress, fatigue, alcohol use, reduced caloric intake, and intercurrent illness, while increased caloric intake or administration of enzyme-inducing agents produces lower bilirubin levels. GS is most often diagnosed at or shortly after puberty or in adult life during routine examinations that include multichannel biochemical analyses. UGT1A1 activity is typically reduced to 10–35% of normal, and bile pigments exhibit a characteristic increase in bilirubin monoglucuronides. Studies of radiobilirubin kinetics indicate that hepatic bilirubin clearance is reduced to an average of one-third of normal. Administration of phenobarbital normalizes both the serum bilirubin concentration and hepatic bilirubin clearance; however, failure of UGT1A1 activity to improve in many such instances suggests the possible coexistence of an additional defect. Compartmental analysis of bilirubin kinetic data suggests that GS patients may have a defect in bilirubin uptake as well as in conjugation, although this has not been shown directly. Defect(s) in the hepatic uptake of other organic anions that at least partially share an uptake mechanism with bilirubin, such as sulfobromophthalein and indocyanine green (ICG), are observed in a minority of patients. The metabolism and transport of bile acids that do not utilize the bilirubin uptake mechanism are normal. The magnitude of changes in the serum bilirubin concentration induced by provocation tests such as 48 hours of fasting or the IV administration of nicotinic acid have been reported to be of help in separating GS patients from normal individuals. Other studies dispute this assertion. Moreover, on theoretical grounds, the results of such studies should provide no more information than simple measurements of the baseline serum bilirubin concentration. Family studies indicate that GS and hereditary hemolytic anemias such as hereditary spherocytosis, glucose-6-phosphate dehydrogenase deficiency, and β-thalassemia trait sort independently. Reports of hemolysis in up to 50% of GS patients are believed to reflect better case finding, since patients with both GS and hemolysis have higher bilirubin concentrations, and are more likely to be jaundiced, than patients with either defect alone.

GS is common, with many series placing its prevalence as high as 8%. Males predominate over females by reported ratios ranging from 1.5:1 to >7:1. However, these ratios may have a large artifactual component since normal males have higher mean bilirubin levels than normal females, but the diagnosis of GS is often based on comparison to normal ranges established in men. The high prevalence of GS in the general population may explain the reported frequency of mild unconjugated hyperbilirubinemia in liver transplant recipients. The disposition of most xenobiotics metabolized by glucuronidation appears to be normal in GS, as is oxidative drug metabolism in the majority of reported studies. The principal exception is the metabolism of the antitumor agent irinotecan (CPT-11), whose active metabolite (SN-38) is glucuronidated specifically by bilirubin-UDP-glucuronosyltransferase. Administration of CPT-11 to patients with GS has resulted in several toxicities, including intractable diarrhea and myelosuppression. Some reports also suggest abnormal disposition of menthol, estradiol benzoate, acetaminophen, tolbutamide, and rifamycin SV. Although some of these studies have been disputed, and there have been no reports of clinical complications from use of these agents in GS, prudence should be exercised in prescribing them, or any agents metabolized primarily by glucuronidation in this condition. It should also be noted that the HIV protease inhibitors indinavir and atazanavir **(Chap. 197)** can inhibit UGT1A1, resulting in hyperbilirubinemia that is most pronounced in patients with preexisting GS.

Most older pedigree studies of GS were consistent with autosomal dominant inheritance with variable expressivity. However, studies of the *UGT1* gene in GS have indicated a variety of molecular genetic bases for the phenotypic picture and several different patterns of inheritance. Studies in Europe and the United States found that nearly all patients had normal coding regions for UGT1A1, but were homozygous for the insertion of an extra TA (i.e., $A[TA]_7TAA$ rather than $A[TA]_6TAA$) in the promoter region of the first exon. This appeared to be necessary, but not sufficient, for clinically expressed GS, since 15% of normal controls were also homozygous for this variant. While normal by standard criteria, these individuals had somewhat higher bilirubin concentrations than the rest of the controls studied. Heterozygotes for this abnormality had bilirubin concentrations identical to those homozygous for the normal $A[TA]_6TAA$ allele. The prevalence of the $A[TA]_7TAA$ allele in a general Western population is 30%, in which case 9% would be homozygotes. This is slightly higher than the prevalence of GS based on purely phenotypic parameters. It was suggested that additional variables, such as mild hemolysis or a defect in bilirubin uptake, might be among the factors enhancing phenotypic expression of the defect.

Phenotypic expression of GS due solely to the $A[TA]_7TAA$ promoter abnormality is inherited as an autosomal recessive trait. A number of CN-II kindreds have been identified in whom there is also an allele containing a normal coding region but the $A[TA]_7TAA$ promoter abnormality. CN-II heterozygotes, who have the $A[TA]_6TAA$ promoter, are phenotypically normal, whereas those with the $A[TA]_7TAA$ promoter express the phenotypic picture of GS. GS in such kindreds may also result from homozygosity for the $A[TA]_7TAA$ promoter abnormality. Seven different missense mutations in the *UGT1* gene that reportedly cause GS with dominant inheritance have been found in Japanese individuals. Another Japanese patient with mild unconjugated hyperbilirubinemia was homozygous for a missense mutation in exon 5. GS in her family appeared to be recessive.

DISORDERS OF BILIRUBIN METABOLISM LEADING TO MIXED OR PREDOMINANTLY CONJUGATED HYPERBILIRUBINEMIA

In hyperbilirubinemia due to acquired liver disease (e.g., acute hepatitis, common bile duct stone), there are usually elevations in the serum concentrations of both conjugated and unconjugated bilirubin. Although biliary tract obstruction or hepatocellular cholestatic injury may present on occasion with a predominantly conjugated hyperbilirubinemia, it is generally not possible to differentiate intrahepatic from extrahepatic causes of jaundice based on the serum levels or relative proportions of unconjugated and conjugated bilirubin. The major reason for determining the amounts of conjugated and unconjugated bilirubin in the serum is for the initial differentiation of hepatic parenchymal and obstructive disorders (mixed conjugated and unconjugated hyperbilirubinemia) from the inheritable and hemolytic disorders discussed above that are associated with unconjugated hyperbilirubinemia.

FAMILIAL DEFECTS IN HEPATIC EXCRETORY FUNCTION

Dubin-Johnson Syndrome (DJS) This benign, relatively rare disorder is characterized by low-grade, predominantly conjugated hyperbilirubinemia **(Table 331-2)**. Total bilirubin concentrations are typically between 34 and 85 μmol/L (2 and 5 mg/dL) but on occasion can be in the normal range or as high as 340–430 μmol/L (20–25 mg/dL) and can fluctuate widely in any given patient. The degree of hyperbilirubinemia may be increased by intercurrent illness, oral contraceptive use, and pregnancy. Because the hyperbilirubinemia is due to a predominant rise in conjugated bilirubin, bilirubinuria is characteristically present. Aside from elevated serum bilirubin levels, other routine laboratory tests are normal. Physical examination is usually normal except for jaundice, although an occasional patient may have hepatosplenomegaly.

Patients with DJS are usually asymptomatic, although some may have vague constitutional symptoms. These latter patients have usually undergone extensive and often unnecessary diagnostic examinations for unexplained jaundice and have high levels of anxiety. In women, the condition may be subclinical until the patient becomes pregnant or receives oral contraceptives, at which time chemical hyperbilirubinemia becomes frank jaundice. Even in these situations, other routine liver function tests, including serum alkaline phosphatase and transaminase activities, are normal.

TABLE 331-2 Principal Differential Characteristics of Inheritable Disorders of Bile Canalicular Function

	DJS	ROTOR	PFIC1	BRIC1	PFIC2	BRIC2	PFIC3
Gene	*ABCCA*	*SLCO1B1/SLCO1B3*	*ATP8B1*	*ATP8B1*	*ABCB11*	*ABCB11*	*ABCB4*
Protein	MRP2	OATP1B1/1B3	FIC1	FIC1	BSEP	BSEP	MDR3
Cholestasis	No	No	Yes	Episodic	Yes	Episodic	Yes
Serum GGT	Normal	Normal	Normal	Normal	Normal	Normal	↑↑
Serum bile acids	Normal	Normal	↑↑	↑↑ during episodes	↑↑	↑↑ during episodes	↑↑
Clinical features	Mild conjugated hyperbilirubinemia; otherwise normal liver function; dark pigment in liver; characteristic pattern of urinary coproporphyrins	Mild conjugated hyperbilirubinemia; otherwise normal liver function; liver without abnormal pigmentation	Severe cholestasis beginning in childhood	Recurrent episodes of cholestasis beginning at any age	Severe cholestasis beginning in childhood	Recurrent episodes of cholestasis beginning at any age	Severe cholestasis beginning in childhood; decreased phospholipids in bile

Abbreviations: BRIC, benign recurrent intrahepatic cholestasis; BSEP, bile salt excretory protein; DJS, Dubin-Johnson syndrome; GGT, γ-glutamyltransferase; MRP2, multidrug resistance–associated protein 2; OATP1A/1B, organic anion transport proteins 1B1 and 1B3; PFIC, progressive familial intrahepatic cholestasis; ↑↑, increased.

A cardinal feature of DJS is the accumulation in the lysosomes of centrilobular hepatocytes of dark, coarsely granular pigment. As a result, the liver may be grossly black in appearance. This pigment is thought to be derived from epinephrine metabolites that are not excreted normally. The pigment may disappear during bouts of viral hepatitis, only to reaccumulate slowly after recovery.

Biliary excretion of a number of anionic compounds is compromised in DJS. These include various cholecystographic agents, as well as sulfobromophthalein (Bromsulphalein, BSP), a synthetic dye formerly used in a test of liver function. In this test, the rate of disappearance of BSP from plasma was determined following bolus IV administration. BSP is conjugated with glutathione in the hepatocyte; the resulting conjugate is normally excreted rapidly into the bile canaliculus. Patients with DJS exhibit characteristic rises in plasma concentrations at 90 min after injection, due to reflux of conjugated BSP into the circulation from the hepatocyte. Dyes such as ICG that are taken up by hepatocytes but are not further metabolized prior to biliary excretion do not show this reflux phenomenon. Continuous BSP infusion studies suggest a reduction in the time to maximum plasma concentration (t_{max}) for biliary excretion. Bile acid disposition, including hepatocellular uptake and biliary excretion, is normal in DJS. These patients have normal serum and biliary bile acid concentrations and do not have pruritus.

By analogy with findings in several mutant rat strains, the selective defect—in biliary excretion of bilirubin conjugates and certain other classes of organic compounds, but not of bile acids—that characterizes DJS in humans was found to reflect defective expression of MRP2, an ATP-dependent canalicular membrane transporter. Several different mutations in the *MRP2* gene produce the Dubin-Johnson phenotype, which has an autosomal recessive pattern of inheritance. Although MRP2 is undoubtedly important in the biliary excretion of conjugated bilirubin, the fact that this pigment is still excreted in the absence of MRP2 suggests that other, as yet uncharacterized, transport proteins may serve in a secondary role in this process.

Patients with DJS also have a diagnostic abnormality in urinary coproporphyrin excretion. There are two naturally occurring coproporphyrin isomers, I and III. Normally, ~75% of the coproporphyrin in urine is isomer III. In urine from DJS patients, total coproporphyrin content is normal, but >80% is isomer I. Heterozygotes for the syndrome show an intermediate pattern. The molecular basis for this phenomenon remains unclear.

Rotor Syndrome This benign, autosomal recessive disorder is clinically similar to DJS (Table 331-2), although it is seen even less frequently. A major phenotypic difference is that the liver in patients with Rotor syndrome has no increased pigmentation and appears totally normal. The only abnormality in routine laboratory tests is an elevation of total serum bilirubin, due to a predominant rise in conjugated bilirubin. This is accompanied by bilirubinuria. Several additional features differentiate Rotor syndrome from DJS. In Rotor syndrome, the gallbladder is usually visualized on oral cholecystography, in contrast to the nonvisualization that is typical of DJS. The pattern of urinary coproporphyrin excretion also differs. The pattern in Rotor syndrome resembles that of many acquired disorders of hepatobiliary function, in which coproporphyrin I, the major coproporphyrin isomer in bile, refluxes from the hepatocyte back into the circulation and is excreted in urine. Thus, total urinary coproporphyrin excretion is substantially increased in Rotor syndrome, in contrast to the normal levels seen in DJS. Although the fraction of coproporphyrin I in urine is elevated, it is usually <70% of the total, compared with ≥80% in DJS. The disorders also can be distinguished by their patterns of BSP excretion. Although clearance of BSP from plasma is delayed in Rotor syndrome, there is no reflux of conjugated BSP back into the circulation as seen in DJS. Kinetic analysis of plasma BSP infusion studies suggests the presence of a defect in intrahepatocellular storage of this compound. This has never been demonstrated directly. Recent studies indicate that the molecular basis of Rotor syndrome results from simultaneous deficiency of the hepatocyte plasma membrane transporters OATP1B1 and OATP1B3. This results in reduced reuptake by these transporters of conjugated bilirubin that has been pumped out of the hepatocyte into the portal circulation by MRP3 (Fig. 331-1).

Benign Recurrent Intrahepatic Cholestasis (BRIC) This rare disorder is characterized by recurrent attacks of pruritus and jaundice. The typical episode begins with mild malaise and elevations in serum aminotransferase levels, followed rapidly by rises in alkaline phosphatase and conjugated bilirubin and onset of jaundice and itching. The first one or two episodes may be misdiagnosed as acute viral hepatitis. The cholestatic episodes, which may begin in childhood or adulthood, can vary in duration from several weeks to months, followed by a complete clinical and biochemical resolution. Intervals between attacks may vary from several months to years. Between episodes, physical examination is normal, as are serum levels of bile acids, bilirubin, transaminases, and alkaline phosphatase. The disorder is familial and has an autosomal recessive pattern of inheritance. BRIC is considered a benign disorder in that it does not lead to cirrhosis or end-stage liver disease. However, the episodes of jaundice and pruritus can be prolonged and debilitating, and some patients have undergone liver transplantation to relieve the intractable and disabling symptoms. Treatment during the cholestatic episodes is symptomatic; there is no specific treatment to prevent or shorten the occurrence of episodes.

A gene termed *FIC1* was recently identified and found to be mutated in patients with BRIC. Curiously, this gene is expressed strongly in the small intestine but only weakly in the liver. The protein encoded by

FIC1 shows little similarity to those that have been shown to play a role in bile canalicular excretion of various compounds. Rather, it appears to be a member of a P-type ATPase family that transports aminophospholipids from the outer to the inner leaflet of a variety of cell membranes. Its relationship to the pathobiology of this disorder remains unclear. A second phenotypically identical form of BRIC, termed BRIC type 2, has been described resulting from mutations in the bile salt excretory protein (BSEP), the protein that is defective in progressive familial intrahepatic cholestasis (PFIC) type 2 (Table 331-2). How some mutations in this protein result in the episodic BRIC phenotype is unknown.

Progressive Familial Intrahepatic Cholestasis (FIC) This name is applied to three phenotypically related syndromes (Table 331-2). PFIC type 1 (Byler disease) presents in early infancy as cholestasis that may be initially episodic. However, in contrast to BRIC, Byler disease progresses to malnutrition, growth retardation, and end-stage liver disease during childhood. This disorder is also a consequence of a *FIC1* mutation. The functional relationship of the FIC1 protein to the pathogenesis of cholestasis in these disorders is unknown. Two other types of PFIC (types 2 and 3) have been described. PFIC type 2 is associated with a mutation in the protein originally named *sister of p-glycoprotein*, now known as *bile salt excretory protein*, which is the major bile canalicular exporter of bile acids. As noted above, some mutations of this protein are associated with BRIC type 2, rather than the progressive FIC type 2 phenotype. Progressive FIC type 3 has been associated with a mutation of MDR3, a protein that is essential for normal hepatocellular excretion of phospholipids across the bile canaliculus. Although all three types of PFIC have similar clinical phenotypes, only type 3 is associated with high serum levels of γ-glutamyltransferase (GGT) activity. In contrast, activity of this enzyme is normal or only mildly elevated in symptomatic BRIC and progressive FIC types 1 and 2. Interestingly, mutations in FIC1 or BSEP are not found in approximately one-third of patients with clinical PFIC and normal GGT. Recent studies have shown that patients with mutations in *NR1H4*, the gene encoding the farnesoid X receptor (FXR), a nuclear hormone receptor activated by bile acids, have a syndrome identical to PFIC2 with absent expression of BSEP. Mutations in tight junction protein 2 (TJP2) have also been associated with severe cholestasis with normal GGT levels, likely due to disruption of tight junctions at the bile canaliculus.

FURTHER READING

Canu G et al: Gilbert and Crigler Najjar syndromes: An update of the UDP-glucuronosyltransferase 1A1 (UGT1A1) gene mutation database. Blood Cells Mol Dis 50:273, 2013.

Erlinger S, Arias IM, Dhumeaux D: Inherited disorders of bilirubin transport and conjugation: New insights into molecular mechanisms and consequences. Gastroenterology 146:1625, 2014.

Gomez-Ospina N et al: Mutations in the nuclear bile acid receptor FXR cause progressive familial intrahepatic cholestasis. Nat Commun 7:10713, 2016.

Hansen TW: Biology of bilirubin photoisomers. Clin Perinatol 43:277, 2016.

Lamola AA: A pharmacologic view of phototherapy. Clin Perinatol 43:259, 2016.

Memon N et al: Inherited disorders of bilirubin clearance. Pediatr Res 79:378, 2016.

Rodrigues C et al: Impact of UGT1A1 gene variants on total bilirubin levels in Gilbert syndrome patients and in healthy subjects. Blood Cells Mol Dis 48:166, 2012.

Sambrotta M et al: Mutations in TJP2 cause progressive cholestatic liver disease. Nat Genet 46:326, 2014.

Soroka CJ, Boyer JL: Biosynthesis and trafficking of the bile salt export pump, BSEP: Therapeutic implications of BSEP mutations. Mol Aspects Med 37:3, 2014.

van de Steeg E et al: Complete OATP1B1 and OATP1B3 deficiency causes human Rotor syndrome by interrupting conjugated bilirubin reuptake into the liver. J Clin Invest 122:519, 2012.

Wolkoff AW: Organic anion uptake by hepatocytes. Compr Physiol 4:1715, 2014.

332 Acute Viral Hepatitis

Jules L. Dienstag

Acute viral hepatitis is a systemic infection affecting the liver predominantly. Almost all cases of acute viral hepatitis are caused by one of five viral agents: hepatitis A virus (HAV), hepatitis B virus (HBV), hepatitis C virus (HCV), the HBV-associated delta agent or hepatitis D virus (HDV), and hepatitis E virus (HEV). All these human hepatitis viruses are RNA viruses, except for hepatitis B, which is a DNA virus but replicates like a retrovirus. Although these agents can be distinguished by their molecular and antigenic properties, all types of viral hepatitis produce clinically similar illnesses. These range from asymptomatic and inapparent to fulminant and fatal acute infections common to all types, on the one hand, and from subclinical persistent infections to rapidly progressive chronic liver disease with cirrhosis and even hepatocellular carcinoma, common to the bloodborne types (HBV, HCV, and HDV), on the other.

VIROLOGY AND ETIOLOGY

Hepatitis A HAV is a nonenveloped 27-nm, heat-, acid-, and ether-resistant RNA virus in the *Hepatovirus* genus of the picornavirus family **(Fig. 332-1)**. Its virion contains four capsid polypeptides, designated VP1–VP4, which are cleaved posttranslationally from the polyprotein product of a 7500-nucleotide genome. Inactivation of viral activity can be achieved by boiling for 1 min, by contact with formaldehyde and chlorine, or by ultraviolet irradiation. Despite nucleotide sequence variation of up to 20% among isolates of HAV, and despite the recognition of four genotypes affecting humans, all strains of this virus are immunologically indistinguishable and belong to one serotype. Human HAV can infect and cause hepatitis in chimpanzees, tamarins (marmosets), and several monkey species. Recently, a hepatotropic *Hepatovirus* related to, and likely to have shared common evolutionary ancestry with, human HAV has been identified in several species of harbor seals, albeit without histologic evidence for liver injury or inflammation. Hepatitis A has an incubation period of ~4 weeks. Its replication is limited to the liver, but the virus is present in the liver, bile, stools, and blood during the late incubation period and acute preicteric/presymptomatic phase of illness. Despite slightly longer persistence of virus in the liver, fecal shedding, viremia, and infectivity diminish rapidly once jaundice becomes apparent. HAV can be cultivated reproducibly in vitro.

Antibodies to HAV (anti-HAV) can be detected during acute illness when serum aminotransferase activity is elevated and fecal HAV shedding is still occurring. This early antibody response is predominantly of the IgM class and persists for several (~3) months, rarely for 6–12 months. During convalescence, however, anti-HAV of the IgG class becomes the predominant antibody **(Fig. 332-2)**. Therefore, the diagnosis of hepatitis A is made during acute illness by demonstrating anti-HAV of the IgM class. After acute illness, anti-HAV of the IgG class remains detectable indefinitely, and patients with serum anti-HAV are immune to reinfection. Neutralizing antibody activity parallels the appearance of anti-HAV, and the IgG anti-HAV present in immune globulin accounts for the protection it affords against HAV infection.

Hepatitis B HBV is a DNA virus with a remarkably compact genomic structure; despite its small, circular, 3200-bp size, HBV DNA codes for four sets of viral products with a complex, multiparticle structure. HBV achieves its genomic economy by relying on an efficient strategy of encoding proteins from four overlapping genes: S, C, P, and X **(Fig. 332-3)**, as detailed below. Once thought to be unique among viruses, HBV is now recognized as one of a family of animal viruses, hepadnaviruses (hepatotropic DNA viruses), and is classified as hepadnavirus type 1. Similar viruses infect certain species of woodchucks, ground and tree squirrels, and Pekin ducks, to mention the most carefully characterized; genetic evidence of ancient HBV-like virus

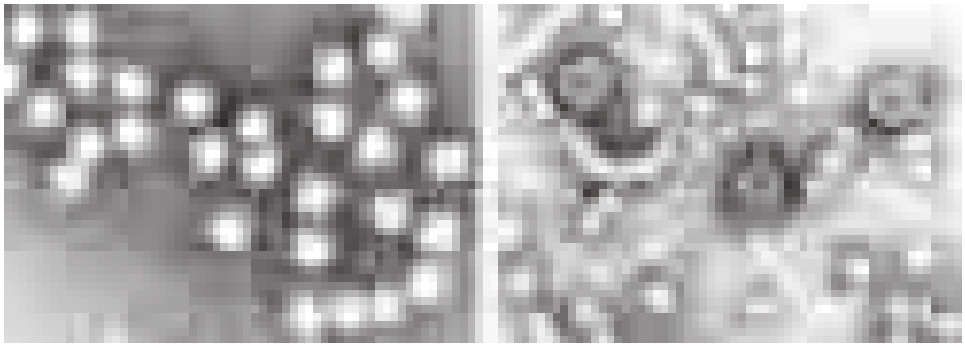


FIGURE 332-1 Electron micrographs of hepatitis A virus particles and serum from a patient with hepatitis B. ***Left:*** 27-nm hepatitis A virus particles purified from stool of a patient with acute hepatitis A and aggregated by antibody to hepatitis A virus. ***Right:*** Concentrated serum from a patient with hepatitis B, demonstrating the 42-nm virions, tubular forms, and spherical 22-nm particles of hepatitis B surface antigen. 132,000×. (Hepatitis D resembles 42-nm virions of hepatitis B but is smaller, 35–37 nm; hepatitis E resembles hepatitis A virus but is slightly larger, 32–34 nm; hepatitis C has been visualized as a 55-nm particle.)

forbears has been found in fossils of ancient birds, and a HBV-like virus has been identified in contemporary fish. Like HBV, all have the same distinctive three morphologic forms, have counterparts to the envelope and nucleocapsid virus antigens of HBV, replicate in the liver but exist in extrahepatic sites, contain their own endogenous DNA polymerase, have partially double-strand and partially single-strand genomes, are associated with acute and chronic hepatitis and hepatocellular carcinoma, and rely on a replicative strategy unique among DNA viruses but typical of retroviruses. Entry of HBV into hepatocytes is mediated by binding to the sodium taurocholate cotransporting polypeptide receptor. Instead of DNA replication directly from a DNA template, hepadnaviruses rely on reverse transcription (effected by the DNA polymerase) of minus-strand DNA from a "pregenomic" RNA intermediate. Then, plus-strand DNA is transcribed from the minus-strand DNA template by the DNA-dependent DNA polymerase and converted in the hepatocyte nucleus to a covalently closed circular DNA, which serves as a template for messenger RNA and pregenomic RNA. Viral proteins are translated by the messenger RNA, and the proteins and genome are packaged into virions and secreted from the hepatocyte. Although HBV is difficult to cultivate in vitro in the conventional sense from clinical material, several cell lines have been transfected with HBV DNA. Such transfected cells support in vitro replication of the intact virus and its component proteins.

VIRAL PROTEINS AND PARTICLES Of the three particulate forms of HBV (Table 332-1), the most numerous are the 22-nm particles, which appear as spherical or long filamentous forms; these are antigenically indistinguishable from the outer surface or envelope protein of HBV and are thought to represent excess viral envelope protein. Outnumbered in serum by a factor of 100 or 1000 to 1 compared with the spheres and tubules are large, 42-nm, double-shelled spherical particles, which represent the intact hepatitis B virion (Fig. 332-1). The envelope protein expressed on the outer surface of the virion and on the smaller spherical and tubular structures is referred to as *hepatitis B surface antigen* (HBsAg). The concentration of HBsAg and virus particles in the blood may reach 500 μg/mL and 10 trillion particles per milliliter, respectively. The envelope protein, HBsAg, is the product of the S gene of HBV.

Envelope HBsAg subdeterminants include a common group-reactive antigen, *a*, shared by all HBsAg isolates and one of several subtype-specific antigens—*d* or *y*, *w* or *r*—as well as other specificities. Hepatitis B isolates fall into one of at least 8 subtypes and 10 genotypes (A–J). Geographic distribution of genotypes and subtypes varies; genotypes A (corresponding to subtype *adw*) and D (*ayw*) predominate in the United States and Europe, whereas genotypes B (*adw*) and C (*adr*) predominate in Asia. Clinical course and outcome are independent of subtype, but genotype B appears to be associated with less rapidly progressive liver disease and cirrhosis and a lower likelihood, or delayed appearance, of hepatocellular carcinoma than genotype C or D. Patients with genotype A are more likely to clear circulating viremia and achieve *hepatitis B e antigen* (HBeAg) and HBsAg seroconversion, both spontaneously and in response to antiviral therapy. In addition, "precore" mutations are favored by certain genotypes (see below).

Upstream of the S gene are the pre-S genes (Fig. 332-3), which code for pre-S gene products, including receptors on the HBV surface for polymerized human serum albumin and for hepatocyte membrane proteins. The pre-S region actually consists of both pre-S1 and pre-S2. Depending on where translation is initiated, three potential HBsAg gene products are synthesized. The protein product of the S gene is HBsAg (*major protein*), the product of the S region plus the adjacent pre-S2 region is the *middle protein*, and the product of the pre-S1 plus pre-S2 plus S regions is the *large protein*. Compared with the smaller spherical and tubular particles of HBV, complete 42-nm virions are enriched in the large protein. Both pre-S proteins and their respective antibodies can be detected during HBV infection, and the period of pre-S antigenemia appears to coincide with other markers of virus replication,

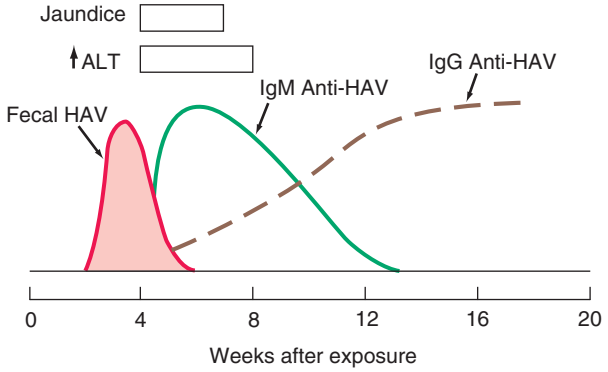


FIGURE 332-2 Scheme of typical clinical and laboratory features of hepatitis A virus (HAV). ALT, alanine aminotransferase.

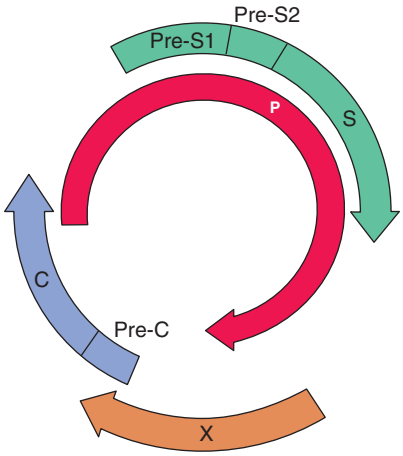


FIGURE 332-3 Compact genomic structure of hepatitis B virus (HBV). This structure, with overlapping genes, permits HBV to code for multiple proteins. The S gene codes for the "major" envelope protein, HBsAg. Pre-S1 and pre-S2, upstream of S, combine with S to code for two larger proteins, "middle" protein, the product of pre-S2 + S, and "large" protein, the product of pre-S1 + pre-S2 + S. The largest gene, P, codes for DNA polymerase. The C gene codes for two nucleocapsid proteins, HBeAg, a soluble, secreted protein (initiation from the pre-C region of the gene), and HBcAg, the intracellular core protein (initiation after pre-C). The X gene codes for HBxAg, which can transactivate the transcription of cellular and viral genes; its clinical relevance is not known, but it may contribute to carcinogenesis by binding to p53.

TABLE 332-1 Nomenclature and Features of Hepatitis Viruses

HEPATITIS TYPE	VIRUS PARTICLE, nm	MORPHOLOGY	GENOME[a]	CLASSIFICATION	ANTIGEN(s)	ANTIBODIES	REMARKS
HAV	27	Icosahedral nonenveloped	7.5-kb RNA, linear, ss, +	Hepatovirus	HAV	Anti-HAV	Early fecal shedding Diagnosis: IgM anti-HAV Previous infection: IgG anti-HAV
HBV	42	Double-shelled virion (surface and core) spherical	3.2-kb DNA, circular, ss/ds	Hepadnavirus	HBsAg HBcAg HBeAg	Anti-HBs Anti-HBc Anti-HBe	Bloodborne virus; carrier state Acute diagnosis: HBsAg, IgM anti-HBc Chronic diagnosis: IgG anti-HBc, HBsAg Markers of replication: HBeAg, HBV DNA Liver, lymphocytes, other organs
	27	Nucleocapsid core			HBcAg HBeAg	Anti-HBc Anti-HBe	Nucleocapsid contains DNA and DNA polymerase; present in hepatocyte nucleus; HBcAg does not circulate; HBeAg (soluble, nonparticulate) and HBV DNA circulate—correlate with infectivity and complete virions
	22	Spherical and filamentous; represents excess virus coat material			HBsAg	Anti-HBs	HBsAg detectable in >95% of patients with acute hepatitis B; found in serum, body fluids, hepatocyte cytoplasm; anti-HBs appears following infection—protective antibody
HCV	Approx. 50–80	Enveloped	9.4-kb RNA, linear, ss, +	Hepacivirus	HCV core antigen	Anti-HCV	Bloodborne agent, formerly labeled non-A, non-B hepatitis Acute diagnosis: anti-HCV, HCV RNA Chronic diagnosis: anti-HCV, HCV RNA; cytoplasmic location in hepatocytes
HDV	35–37	Enveloped hybrid particle with HBsAg coat and HDV core	1.7-kb RNA, circular, ss, –	Resembles viroids and plant satellite viruses (genus Deltavirus)	HBsAg HDAg	Anti-HBs Anti-HDV	Defective RNA virus, requires helper function of HBV (hepadnaviruses); HDV antigen (HDAg) present in hepatocyte nucleus Diagnosis: anti-HDV, HDV RNA; HBV/HDV co-infection—IgM anti-HBc and anti-HDV; HDV superinfection—IgG anti-HBc and anti-HDV
HEV	32–34	Nonenveloped icosahedral	7.6-kb RNA, linear, ss, +	Hepevirus	HEV antigen	Anti-HEV	Agent of enterically transmitted hepatitis; rare in the United States; occurs in Asia, Mediterranean countries, Central America Diagnosis: IgM/IgG anti-HEV (assays not routinely available); virus in stool, bile, hepatocyte cytoplasm

[a]ss, single-strand; ss/ds, partially single-strand, partially double-strand; –, minus-strand; +, plus-strand.

Note: See text for abbreviations.

as detailed below; however, pre-S proteins have little clinical relevance and are not included in routine serologic testing repertoires.

The intact 42-nm virion contains a 27-nm nucleocapsid core particle. Nucleocapsid proteins are coded for by the C gene. The antigen expressed on the surface of the nucleocapsid core is *hepatitis B core antigen* (HBcAg), and its corresponding antibody is anti-HBc. A third HBV antigen is HBeAg, a soluble, nonparticulate, nucleocapsid protein that is immunologically distinct from intact HBcAg but is a product of the same C gene. The C gene has two initiation codons: a precore and a core region (Fig. 332-3). If translation is initiated at the precore region, the protein product is HBeAg, which has a signal peptide that binds it to the smooth endoplasmic reticulum, the secretory apparatus of the cell, leading to its secretion into the circulation. If translation begins at the core region, HBcAg is the protein product; it has no signal peptide and is not secreted, but it assembles into nucleocapsid particles, which bind to and incorporate RNA, and which, ultimately, contain HBV DNA. Also packaged within the nucleocapsid core is a DNA polymerase, which directs replication and repair of HBV DNA. When packaging within viral proteins is complete, synthesis of the incomplete plus strand stops; this accounts for the single-strand gap and for differences in the size of the gap. HBcAg particles remain in the hepatocyte, where they are readily detectable by immunohistochemical staining and are exported after encapsidation by an envelope of HBsAg. Therefore, naked core particles do not circulate in the serum. The secreted nucleocapsid protein, HBeAg, provides a convenient, readily detectable, qualitative marker of HBV replication and relative infectivity.

HBsAg-positive serum containing HBeAg is more likely to be highly infectious and to be associated with the presence of hepatitis B virions (and detectable HBV DNA, see below) than HBeAg-negative or anti-HBe-positive serum. For example, HBsAg-positive mothers who are HBeAg-positive almost invariably (>90%) transmit hepatitis B infection to their offspring, whereas HBsAg-positive mothers with anti-HBe rarely (10–15%) infect their offspring.

Early during the course of acute hepatitis B, HBeAg appears transiently; its disappearance may be a harbinger of clinical improvement and resolution of infection. Persistence of HBeAg in serum beyond the first 3 months of acute infection may be predictive of the development of chronic infection, and the presence of HBeAg during chronic hepatitis B tends to be associated with ongoing viral replication, infectivity, and inflammatory liver injury (except during the early decades after perinatally acquired HBV infection; see below).

The third and largest of the HBV genes, the P gene (Fig. 332-3), codes for HBV DNA polymerase; as noted above, this enzyme has both DNA-dependent DNA polymerase and RNA-dependent reverse transcriptase activities. The fourth gene, X, codes for a small, nonparticulate protein, *hepatitis B x antigen* (HBxAg), that is capable of transactivating the transcription of both viral and cellular genes (Fig. 332-3). In the cytoplasm, HBxAg effects calcium release (possibly from mitochondria), which activates signal-transduction pathways that lead to stimulation of HBV reverse transcription and HBV DNA replication. Such transactivation may enhance the replication of HBV, leading to the clinical association observed between the expression of HBxAg and antibodies

to it in patients with severe chronic hepatitis and hepatocellular carcinoma. The transactivating activity can enhance the transcription and replication of other viruses besides HBV, such as HIV. Cellular processes transactivated by X include the human interferon-γ gene and class I major histocompatibility genes; potentially, these effects could contribute to enhanced susceptibility of HBV-infected hepatocytes to cytolytic T cells. The expression of X can also induce programmed cell death (apoptosis). The clinical relevance of HBxAg is limited, however, and testing for it is not part of routine clinical practice.

SEROLOGIC AND VIROLOGIC MARKERS After a person is infected with HBV, the first virologic marker detectable in serum within 1–12 weeks, usually between 8 and 12 weeks, is HBsAg **(Fig. 332-4)**. Circulating HBsAg precedes elevations of serum aminotransferase activity and clinical symptoms by 2–6 weeks and remains detectable during the entire icteric or symptomatic phase of acute hepatitis B and beyond. In typical cases, HBsAg becomes undetectable 1–2 months after the onset of jaundice and rarely persists beyond 6 months. After HBsAg disappears, antibody to HBsAg (anti-HBs) becomes detectable in serum and remains detectable indefinitely thereafter. Because HBcAg is intracellular and, when in the serum, sequestered within an HBsAg coat, naked core particles do not circulate in serum, and therefore HBcAg is not detectable routinely in the serum of patients with HBV infection. By contrast, anti-HBc is readily demonstrable in serum, beginning within the first 1–2 weeks after the appearance of HBsAg and preceding detectable levels of anti-HBs by weeks to months. Because variability exists in the time of appearance of anti-HBs after HBV infection, occasionally a gap of several weeks or longer may separate the disappearance of HBsAg and the appearance of anti-HBs. During this "gap" or "window" period, anti-HBc may represent the only serologic evidence of current or recent HBV infection, and blood containing anti-HBc in the absence of HBsAg and anti-HBs has been implicated in transfusion-associated hepatitis B. In part because the sensitivity of immunoassays for HBsAg and anti-HBs has increased, however, this window period is rarely encountered. In some persons, years after HBV infection, anti-HBc may persist in the circulation longer than anti-HBs. Therefore, isolated anti-HBc does not necessarily indicate active virus replication; most instances of isolated anti-HBc represent hepatitis B infection in the remote past. Rarely, however, isolated anti-HBc represents low-level hepatitis B viremia, with HBsAg below the detection threshold, and, occasionally, isolated anti-HBc represents a cross-reacting or false-positive immunologic specificity. Recent and remote HBV infections can be distinguished by determination of the immunoglobulin class of anti-HBc. Anti-HBc of the IgM class (IgM anti-HBc) predominates during the first 6 months after acute infection, whereas IgG anti-HBc is the predominant class of anti-HBc beyond 6 months. Therefore, patients with current or recent acute hepatitis B, including those in the anti-HBc window, have IgM anti-HBc in their serum. In patients who have recovered from hepatitis B in the remote past as well as those with chronic HBV infection, anti-HBc is predominantly of the IgG class. Infrequently, in ≤1–5% of patients with acute HBV infection, levels of HBsAg are too low to be detected; in such cases, the presence of IgM anti-HBc establishes the diagnosis of acute hepatitis B. When isolated anti-HBc occurs in the rare patient with chronic hepatitis B whose HBsAg level is below the sensitivity threshold of contemporary immunoassays (a low-level carrier), anti-HBc is of the IgG class. Generally, in persons who have recovered from hepatitis B, anti-HBs and anti-HBc persist indefinitely.

The temporal association between the appearance of anti-HBs and resolution of HBV infection as well as the observation that persons with anti-HBs in serum are protected against reinfection with HBV suggests that *anti-HBs is the protective antibody*. Therefore, strategies for prevention of HBV infection are based on providing susceptible persons with circulating anti-HBs (see below). Occasionally, in ~10% of patients with chronic hepatitis B, low-level, low-affinity anti-HBs can be detected. This antibody is directed against a subtype determinant different from that represented by the patient's HBsAg; its presence is thought to reflect the stimulation of a related clone of antibody-forming cells, but it has no clinical relevance and does not signal imminent clearance of hepatitis B. These patients with HBsAg and such nonneutralizing anti-HBs should be categorized as having chronic HBV infection.

The other readily detectable serologic marker of HBV infection, HBeAg, appears concurrently with or shortly after HBsAg. Its appearance coincides temporally with high levels of virus replication and reflects the presence of circulating intact virions and detectable HBV DNA (with the notable exception of patients with precore mutations who cannot synthesize HBeAg—see "Molecular Variants"). Pre-S1 and pre-S2 proteins are also expressed during periods of peak replication, but assays for these gene products are not routinely available. In self-limited HBV infections, HBeAg becomes undetectable shortly after peak elevations in aminotransferase activity, before the disappearance of HBsAg, and anti-HBe then becomes detectable, coinciding with a period of relatively lower infectivity (Fig. 332-4). Because markers of HBV replication appear transiently during acute infection, testing for such markers is of little clinical utility in typical cases of acute HBV infection. In contrast, markers of HBV replication provide valuable information in patients with protracted infections.

Departing from the pattern typical of acute HBV infections, in chronic HBV infection, HBsAg remains detectable beyond 6 months, anti-HBc is primarily of the IgG class, and anti-HBs is either undetectable or detectable at low levels (see "Laboratory Features") **(Fig. 332-5)**.

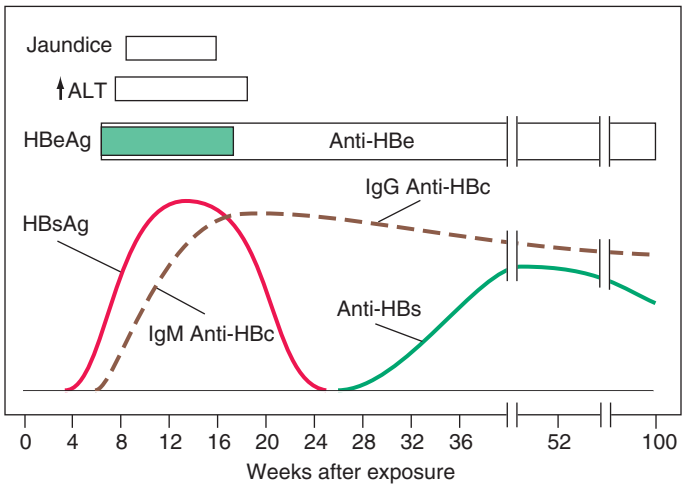


FIGURE 332-4 Scheme of typical clinical and laboratory features of acute hepatitis B. ALT, alanine aminotransferase.

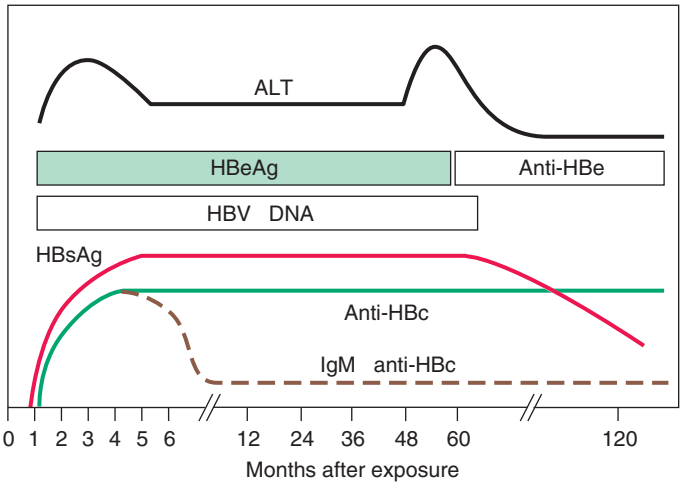


FIGURE 332-5 Scheme of typical laboratory features of wild-type chronic hepatitis B. HBeAg and hepatitis B virus (HBV) DNA can be detected in serum during the relatively *replicative phase* of chronic infection, which is associated with infectivity and liver injury. Seroconversion from the replicative phase to the relatively *nonreplicative phase* occurs at a rate of ~10% per year and is heralded by an acute hepatitis–like elevation of alanine aminotransferase (ALT) activity; during the nonreplicative phase, infectivity and liver injury are limited. In HBeAg-negative chronic hepatitis B associated with mutations in the precore region of the HBV genome, replicative chronic hepatitis B occurs in the absence of HBeAg.

During early chronic HBV infection, HBV DNA can be detected both in serum and in hepatocyte nuclei, where it is present in free or episomal form. This relatively highly *replicative stage* of HBV infection is the time of maximal infectivity and liver injury; HBeAg is a qualitative marker and HBV DNA a quantitative marker of this replicative phase, during which all three forms of HBV circulate, including intact virions. Over time, the relatively replicative phase of chronic HBV infection gives way to a relatively *nonreplicative phase*. This occurs at a rate of ~10% per year and is accompanied by seroconversion from HBeAg to anti-HBe. In many cases, this seroconversion coincides with a transient, usually mild, acute hepatitis-like elevation in aminotransferase activity, believed to reflect cell-mediated immune clearance of virus-infected hepatocytes. In the nonreplicative phase of chronic infection, when HBV DNA is demonstrable in hepatocyte nuclei, it tends to be integrated into the host genome. In this phase, only spherical and tubular forms of HBV, *not intact virions*, circulate, and liver injury tends to subside. Most such patients would be characterized as *inactive HBV carriers*. In reality, the designations *replicative* and *nonreplicative* are only relative; even in the so-called nonreplicative phase, HBV replication can be detected at levels of approximately $\leq 10^3$ virions/mL with highly sensitive amplification probes such as the polymerase chain reaction (PCR); below this replication threshold, liver injury and infectivity of HBV are limited to negligible. Still, the distinctions are pathophysiologically and clinically meaningful. Occasionally, nonreplicative HBV infection converts back to replicative infection. Such spontaneous reactivations are accompanied by reexpression of HBeAg and HBV DNA, and sometimes of IgM anti-HBc, as well as by exacerbations of liver injury. Because high-titer IgM anti-HBc can reappear during acute exacerbations of chronic hepatitis B, relying on IgM anti-HBc versus IgG anti-HBc to distinguish between acute and chronic hepatitis B infection, respectively, may not always be reliable; in such cases, patient history and additional follow-up monitoring over time are invaluable in helping to distinguish de novo acute hepatitis B infection from acute exacerbation of chronic hepatitis B infection.

MOLECULAR VARIANTS Variation occurs throughout the HBV genome, and clinical isolates of HBV that do not express typical viral proteins have been attributed to mutations in individual or even multiple gene locations. For example, variants have been described that lack nucleocapsid proteins (commonly), envelope proteins (very rarely), or both. Two categories of naturally occurring HBV variants have attracted the most attention. One of these was identified initially in Mediterranean countries among patients with severe chronic HBV infection and detectable HBV DNA, but with anti-HBe instead of HBeAg. These patients were found to be infected with an HBV mutant that contained an alteration in the precore region rendering the virus incapable of encoding HBeAg. Although several potential mutation sites exist in the pre-C region, the region of the C gene necessary for the expression of HBeAg (see "Virology and Etiology"), the most commonly encountered in such patients is a single base substitution, from G to A in the second to last codon of the pre-C gene at nucleotide 1896. This substitution results in the replacement of the TGG tryptophan codon by a stop codon (TAG), which prevents the translation of HBeAg. Another mutation, in the core-promoter region, prevents transcription of the coding region for HBeAg and yields an HBeAg-negative phenotype. Patients with such mutations in the precore region and who are unable to secrete HBeAg may have severe liver disease that progresses more rapidly to cirrhosis, or alternatively, they are identified clinically later in the course of the natural history of chronic hepatitis B, when the disease is more advanced. Both "wild-type" HBV and precore-mutant HBV can coexist in the same patient, or mutant HBV may arise late during wild-type HBV infection. In addition, clusters of fulminant hepatitis B in Israel and Japan were attributed to common-source infection with a precore mutant. Fulminant hepatitis B in North America and western Europe, however, occurs in patients infected with wild-type HBV, in the absence of precore mutants, and both precore mutants and other mutations throughout the HBV genome occur commonly, even in patients with typical, self-limited, milder forms of HBV infection. HBeAg-negative chronic hepatitis with mutations in the precore region is now the most frequently encountered form of hepatitis B in Mediterranean countries and in Europe. In the United States, where HBV genotype A (less prone to G1896A mutation) is prevalent, precore-mutant HBV is much less common; however, as a result of immigration from Asia and Europe, the proportion of HBeAg-negative hepatitis B–infected individuals has increased in the United States, and they now represent ~30–40% of patients with chronic hepatitis B. Characteristic of such HBeAg-negative chronic hepatitis B are lower levels of HBV DNA (usually $\leq 10^5$ IU/mL) and one of several patterns of aminotransferase activity—persistent elevations, periodic fluctuations above the normal range, and periodic fluctuations between the normal and elevated range.

The second important category of HBV mutants consists of *escape mutants*, in which a single amino acid substitution, from glycine to arginine, occurs at position 145 of the immunodominant *a* determinant common to all HBsAg subtypes. This HBsAg alteration leads to a critical conformational change that results in a loss of neutralizing activity by anti-HBs. This specific HBV/*a* mutant has been observed in two situations, active and passive immunization, in which humoral immunologic pressure may favor evolutionary change ("escape") in the virus—in a small number of hepatitis B vaccine recipients who acquired HBV infection despite the prior appearance of neutralizing anti-HBs and in HBV-infected liver transplant recipients treated with a high-potency human monoclonal anti-HBs preparation. Although such mutants have not been recognized frequently, their existence raises a concern that may complicate vaccination strategies and serologic diagnosis.

Different types of mutations emerge during antiviral therapy of chronic hepatitis B with nucleoside analogues; such "YMDD" and similar mutations in the polymerase motif of HBV are described in Chap. 334.

EXTRAHEPATIC SITES Hepatitis B antigens and HBV DNA have been identified in extrahepatic sites, including the lymph nodes, bone marrow, circulating lymphocytes, spleen, and pancreas. Although the virus does not appear to be associated with tissue injury in any of these extrahepatic sites, its presence in these "remote" reservoirs has been invoked (but is not necessary) to explain the recurrence of HBV infection after orthotopic liver transplantation. The clinical relevance of such extrahepatic HBV is limited.

Hepatitis D

The delta hepatitis agent, or HDV, the only member of the genus *Deltavirus*, is a defective RNA virus that co-infects with and requires the helper function of HBV (or other hepadnaviruses) for its replication and expression. Slightly smaller than HBV, HDV is a formalin-sensitive, 35- to 37-nm virus with a hybrid structure. Its nucleocapsid expresses HDV antigen (HDAg), which bears no antigenic homology with any of the HBV antigens, and contains the virus genome. The HDV core is "encapsidated" by an outer envelope of HBsAg, indistinguishable from that of HBV except in its relative compositions of major, middle, and large HBsAg component proteins. The genome is a small, 1700-nucleotide, circular, single-strand RNA of negative polarity that is nonhomologous with HBV DNA (except for a small area of the polymerase gene) but that has features and the rolling circle model of replication common to genomes of plant satellite viruses or viroids. HDV RNA contains many areas of internal complementarity; therefore, it can fold on itself by internal base pairing to form an unusual, very stable, rodlike structure that contains a very stable, self-cleaving and self-ligating ribozyme. HDV RNA requires host RNA polymerase II for its replication in the hepatocyte nucleus via RNA-directed RNA synthesis by transcription of genomic RNA to a complementary antigenomic (plus strand) RNA; the antigenomic RNA, in turn, serves as a template for subsequent genomic RNA synthesis effected by host RNA polymerase I. HDV RNA has only one open reading frame, and HDAg, a product of the antigenomic strand, is the only known HDV protein; HDAg exists in two forms: a small, 195-amino-acid species, which plays a role in facilitating HDV RNA replication, and a large, 214-amino-acid species, which appears to suppress replication but is required for assembly of the antigen into virions. HDV antigens have been shown to bind directly to RNA polymerase II,

resulting in stimulation of transcription. Although complete hepatitis D virions and liver injury require the cooperative helper function of HBV, intracellular replication of HDV RNA can occur without HBV. Genomic heterogeneity among HDV isolates has been described; however, pathophysiologic and clinical consequences of this genetic diversity have not been recognized. The clinical spectrum of hepatitis D is common to all eight genotypes identified, the predominant of which is genotype 1.

HDV can either infect a person simultaneously with HBV (*co-infection*) or superinfect a person already infected with HBV (*superinfection*); when HDV infection is transmitted from a donor with one HBsAg subtype to an HBsAg-positive recipient with a different subtype, HDV assumes the HBsAg subtype of the recipient, rather than the donor. Because HDV relies absolutely on HBV, the duration of HDV infection is determined by the duration of (and cannot outlast) HBV infection. HDV replication tends to suppress HBV replication; therefore, patients with hepatitis D tend to have lower levels of HBV replication. HDV antigen is expressed primarily in hepatocyte nuclei and is occasionally detectable in serum. During acute HDV infection, anti-HDV of the IgM class predominates, and 30–40 days may elapse after symptoms appear before anti-HDV can be detected. In self-limited infection, anti-HDV is low-titer and transient, rarely remaining detectable beyond the clearance of HBsAg and HDV antigen. In chronic HDV infection, anti-HDV circulates in high titer, and both IgM and IgG anti-HDV can be detected. HDV antigen in the liver and HDV RNA in serum and liver can be detected during HDV replication.

Hepatitis C Hepatitis C virus, which, before its identification was labeled "non-A, non-B hepatitis," is a linear, single-strand, positive-sense, 9600-nucleotide RNA virus, the genome of which is similar in organization to that of flaviviruses and pestiviruses; HCV is the only member of the genus *Hepacivirus* in the family Flaviviridae. The HCV genome contains a single, large open reading frame (ORF) (gene) that codes for a virus polyprotein of ~3000 amino acids, which is cleaved after translation to yield 10 viral proteins. The 5′ end of the genome consists of an untranslated region (containing an internal ribosomal entry site [IRES]) adjacent to the genes for three structural proteins, the nucleocapsid core protein, C, and two envelope glycoproteins, E1 and E2. The 5′ untranslated region and core gene are highly conserved among genotypes, but the envelope proteins are coded for by the hypervariable region, which varies from isolate to isolate and may allow the virus to evade host immunologic containment directed at accessible virus-envelope proteins. The 3′ end of the genome also includes an untranslated region and contains the genes for seven nonstructural (NS) proteins: p7, NS2, NS3, NS4A, NS4B, NS5A, and NS5B. p7 is a membrane ion channel protein necessary for efficient assembly and release of HCV. The NS2 cysteine protease cleaves NS3 from NS2, and the NS3-4A serine protease cleaves all the downstream proteins from the polyprotein. Important NS proteins involved in virus replication include the NS3 helicase; NS3-4A serine protease; the multifunctional membrane-associated phosphoprotein NS5A, an essential component of the viral replication membranous web (along with NS4B); and the NS5B RNA-dependent RNA polymerase **(Fig. 332-6)**. Because HCV does not replicate via a DNA intermediate, it does not integrate into the host genome. Because HCV tends to circulate in relatively low titer, 10^3–10^7 virions/mL, visualization of the 50- to 80-nm virus particles remains difficult. Still, the replication rate of HCV is very high, 10^{12} virions per day; its half-life is 2.7 h. The chimpanzee is a helpful but cumbersome animal model. Although a robust, reproducible, small animal model is lacking, HCV replication has been documented in an immunodeficient mouse model containing explants of human liver and in transgenic mouse and rat models. Although in vitro replication is difficult, replicons in hepatocellular carcinoma–derived cell lines support replication of genetically manipulated, truncated, or full-length HCV RNA (but not intact virions); infectious pseudotyped retroviral HCV particles have been shown to yield functioning envelope proteins. In 2005, complete replication of HCV and intact 55-nm virions were described in cell culture systems. HCV entry into the hepatocyte occurs via the nonliver-specific CD81 receptor and the liver-specific tight junction protein claudin-1. A growing list of additional host receptors to which HCV binds on cell entry includes occludin, low-density lipoprotein receptors, glycosaminoglycans, scavenger receptor B1, and epidermal growth factor receptor, among others. Relying on the same assembly and secretion pathway as low-density and very-low-density lipoproteins, HCV is a lipoviroparticle and masquerades as a lipoprotein, which may limit its visibility to the adaptive immune system and explain its ability to evade immune containment and clearance. After viral entry and uncoating, translation is initiated by the IRES on the endoplasmic reticulum membrane, and the HCV polyprotein is cleaved during translation and posttranslationally by host cellular proteases as well as HCV NS2-3 and NS3-4A proteases. Host cofactors involved in HCV replication include cyclophilin A, which binds to NS5A and yields conformational changes required for viral replication, and liver-specific host microRNA miR-122.

At least six distinct major genotypes (and a minor genotype 7), as well as >50 subtypes within genotypes, of HCV have been identified by nucleotide sequencing. Genotypes differ from one another in sequence homology by ≥30%, and subtypes differ by ~20%. Because divergence of HCV isolates within a genotype or subtype and within the same host may vary insufficiently to define a distinct genotype, these intragenotypic differences are referred to as *quasispecies* and differ in sequence homology by only a few percent. The genotypic and quasispecies diversity of HCV, resulting from its high mutation rate, interferes with effective humoral immunity. Neutralizing antibodies to HCV have been demonstrated, but they tend to be short-lived, and HCV infection does not induce lasting immunity against reinfection with different virus isolates or even the same virus isolate. Thus, neither *heterologous* nor *homologous* immunity appears to develop commonly after acute HCV infection. Some HCV genotypes are distributed worldwide, whereas others are more geographically confined (see "Epidemiology and Global Features"). In addition, differences exist among genotypes in responsiveness to antiviral therapy but not in pathogenicity or clinical progression (except for genotype 3, in which hepatic steatosis and clinical progression are more likely).

Currently available, third-generation immunoassays, which incorporate proteins from the core, NS3, and NS5 regions, detect anti-HCV antibodies during acute infection. The most sensitive indicator of HCV infection is the presence of HCV RNA, which requires molecular amplification by PCR or transcription-mediated amplification (TMA)

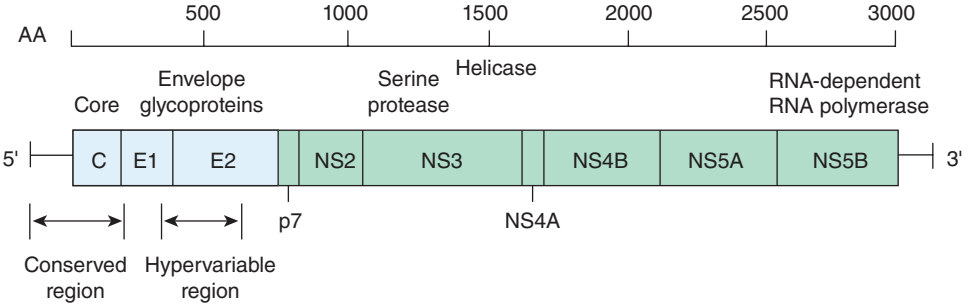


FIGURE 332-6 Organization of the hepatitis C virus genome and its associated, 3000-amino-acid (AA) proteins. The three structural genes at the 5′ end are the core region, C, which codes for the nucleocapsid, and the envelope regions, E1 and E2, which code for envelope glycoproteins. The 5′ untranslated region and the C region are highly conserved among isolates, whereas the envelope domain E2 contains the hypervariable region. At the 3′ end are seven nonstructural (NS) regions—p7, a membrane protein adjacent to the structural proteins that appears to function as an ion channel; NS2, which codes for a cysteine protease; NS3, which codes for a serine protease and an RNA helicase; NS4 and NS4B; NS5A, a multifunctional membrane-associated phosphoprotein, an essential component of the viral replication membranous web; and NS5B, which codes for an RNA-dependent RNA polymerase. After translation of the entire polyprotein, individual proteins are cleaved by both host and viral proteases.

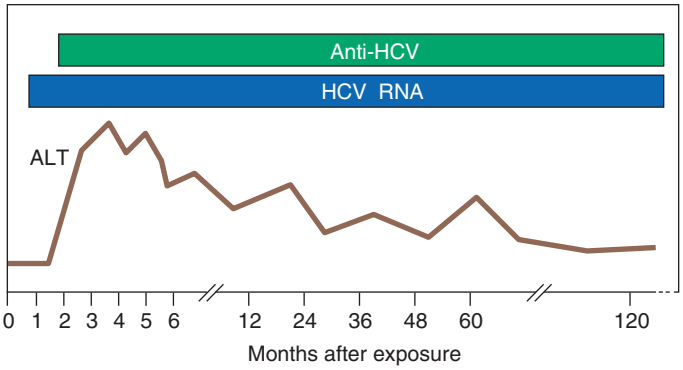


FIGURE 332-7 Scheme of typical laboratory features during acute hepatitis C progressing to chronicity. Hepatitis C virus (HCV) RNA is the first detectable event, preceding alanine aminotransferase (ALT) elevation and the appearance of anti-HCV.

(Fig. 332-7). To allow standardization of the quantification of HCV RNA among laboratories and commercial assays, HCV RNA is reported as international units (IUs) per milliliter; quantitative assays with a broad dynamic range are available that allow detection of HCV RNA with a sensitivity as low as 5 IU/mL. HCV RNA can be detected within a few days of exposure to HCV—well before the appearance of anti-HCV—and tends to persist for the duration of HCV infection. Application of sensitive molecular probes for HCV RNA has revealed the presence of replicative HCV in peripheral blood lymphocytes of infected persons; however, as is the case for HBV in lymphocytes, the clinical relevance of HCV lymphocyte infection is not known.

Hepatitis E Previously labeled *epidemic* or *enterically transmitted non-A, non-B hepatitis*, HEV is an enterically transmitted virus that causes clinically apparent hepatitis primarily in India, Asia, Africa, and Central America; in those geographic areas, HEV is the most common cause of acute hepatitis; one-third of the global population appears to have been infected. This agent, with epidemiologic features resembling those of hepatitis A, is a 27- to 34-nm, nonenveloped, heat-stable, HAV-like virus with a 7200-nucleotide, single-strand, positive-sense RNA genome. HEV has three overlapping ORFs (genes), the largest of which, *ORF1*, encodes nonstructural proteins involved in virus replication (viral replicase). A middle-sized gene, *ORF2*, encodes the nucleocapsid protein, the major structural protein, and the smallest, *ORF3*, encodes a small structural protein involved in virus particle secretion. All HEV isolates appear to belong to a single serotype, despite genomic heterogeneity of up to 25% and the existence of five genotypes, only four of which have been detected in humans; genotypes 1 and 2 (common in developing countries) appear to be more virulent, whereas genotypes 3 (the most common in the United States and Europe) and 4 (seen in China) are more attenuated and account for subclinical infections. Contributing to the perpetuation of this virus are animal reservoirs, most notably in swine but also in camels, deer, rats, and rabbits, among others. No genomic or antigenic homology, however, exists between HEV and HAV or other picornaviruses; and HEV, although resembling caliciviruses, is sufficiently distinct from any known agent to merit its own classification as a unique genus, *Hepevirus*, within the family Hepeviridae. The virus has been detected in stool, bile, and liver, and is excreted in the stool during the late incubation period. Both IgM anti-HEV during early acute infection and IgG anti-HEV predominating after the first 3 months can be detected. The presence of HEV RNA in serum and stool accompanies acute infection; viremia resolves as clinical-biochemical recovery ensues, while HEV RNA in stool may outlast viremia by several weeks. Currently, availability and reliability of serologic/virologic testing for HEV infection is limited—and not FDA-approved or licensed—but can be done in specialized laboratories (e.g., the Centers for Disease Control and Prevention).

PATHOGENESIS

Under ordinary circumstances, none of the hepatitis viruses is known to be directly cytopathic to hepatocytes. Evidence suggests that the clinical manifestations and outcomes after acute liver injury associated with viral hepatitis are determined by the immunologic responses of the host. Among the viral hepatitides, the immunopathogenesis of hepatitis B and C has been studied most extensively.

Hepatitis B For HBV, the existence of inactive hepatitis B carriers with normal liver histology and function suggests that the virus is not directly cytopathic. The fact that patients with defects in cellular immune competence are more likely to remain chronically infected rather than to clear HBV supports the role of cellular immune responses in the pathogenesis of hepatitis B–related liver injury. The model that has the most experimental support involves cytolytic T cells sensitized specifically to recognize host and hepatitis B viral antigens on the liver cell surface. Nucleocapsid proteins (HBcAg and possibly HBeAg), present on the cell membrane in minute quantities, are the viral target antigens that, with host antigens, invite cytolytic T cells to destroy HBV-infected hepatocytes. Differences in the robustness and broad polyclonality of CD8+ cytolytic T cell responsiveness; in the level of HBV-specific helper CD4+ T cells; in attenuation, depletion, and exhaustion of virus-specific T cells; in viral T cell epitope escape mutations that allow the virus to evade T cell containment; and in the elaboration of antiviral cytokines by T cells have been invoked to explain differences in outcomes between those who recover after acute hepatitis and those who progress to chronic hepatitis, or between those with mild and those with severe (fulminant) acute HBV infection.

Although a robust cytolytic T cell response occurs and eliminates virus-infected liver cells during acute hepatitis B, >90% of HBV DNA has been found in experimentally infected chimpanzees to disappear from the liver and blood before maximal T cell infiltration of the liver and before most of the biochemical and histologic evidence of liver injury. This observation suggests that components of the innate immune system and inflammatory cytokines, independent of cytopathic antiviral mechanisms, participate in the early immune response to HBV infection; this effect has been shown to represent elimination of HBV replicative intermediates from the cytoplasm and covalently closed circular viral DNA from the nucleus of infected hepatocytes. In turn, the innate immune response to HBV infection is mediated largely by natural killer (NK) cell cytotoxicity, activated by immunosuppressive cytokines (e.g., interleukin [IL] 10 and transforming growth factor [TGF] β), reduced signals from inhibitory receptor expression (e.g., major histocompatibility complex), or increased signals from activating receptor expression on infected hepatocytes. In addition, NK cells reduce helper CD4+ cells, which results in reduced CD8+ cells and exhaustion of the virus-specific T cell response to HBV infection. Ultimately, HBV-HLA–specific cytolytic T cell responses of the adaptive immune system are felt to be responsible for recovery from HBV infection.

Debate continues over the relative importance of viral and host factors in the pathogenesis of HBV-associated liver injury and its outcome. As noted above, precore genetic mutants of HBV have been associated with the more severe outcomes of HBV infection (severe chronic and fulminant hepatitis), suggesting that, under certain circumstances, relative pathogenicity is a property of the virus, not the host. The facts that concomitant HDV and HBV infections are associated with more severe liver injury than HBV infection alone and that cells transfected in vitro with the gene for HDV antigen express HDV antigen and then become necrotic in the absence of any immunologic influences are also consistent with a viral effect on pathogenicity. Similarly, in patients who undergo liver transplantation for end-stage chronic hepatitis B, occasionally, rapidly progressive liver injury appears in the new liver. This clinical pattern is associated with an unusual histologic pattern in the new liver, *fibrosing cholestatic hepatitis*, which, ultrastructurally, appears to represent a choking of the cell with overwhelming quantities of HBsAg. This observation suggests that, under the influence of the potent immunosuppressive agents required to prevent allograft rejection, HBV may have a direct cytopathic effect on liver cells, independent of the immune system.

Although the precise mechanism of liver injury in HBV infection remains elusive, studies of nucleocapsid proteins have shed light on

the profound immunologic tolerance to HBV of babies born to mothers with highly replicative (HBeAg-positive), chronic HBV infection. In HBeAg-expressing transgenic mice, in utero exposure to HBeAg, which is sufficiently small to traverse the placenta, induces T cell tolerance to both nucleocapsid proteins. This, in turn, may explain why, when infection occurs so early in life, immunologic clearance does not occur, and protracted, lifelong infection ensues. An alternative explanation proposed to explain why robust liver injury does not accompany neonatal HBV infection but predisposes to chronic infection is defective priming of HBV-specific T cells during in utero exposure to HBV.

An important distinction should be drawn between HBV infection acquired at birth, common in endemic areas, such as East Asia, and infection acquired in adulthood, common in the West. Infection in the neonatal period is associated with the acquisition of what appears to be a high level of immunologic tolerance to HBV and absence of an acute hepatitis illness, but the almost invariable establishment of chronic, often lifelong infection. Neonatally acquired HBV infection can culminate decades later in cirrhosis and hepatocellular carcinoma (see "Complications and Sequelae"). In contrast, when HBV infection is acquired during adolescence or early adulthood, the host immune response to HBV-infected hepatocytes tends to be robust, an acute hepatitis-like illness is the rule, and failure to recover is the exception. After adulthood-acquired infection, chronicity is uncommon, and the risk of hepatocellular carcinoma is very low. Based on these observations, some authorities categorize HBV infection into an "immunotolerant" phase, an "immunoreactive" phase, and an "inactive" phase. This somewhat simplistic formulation does not apply at all to the typical adult in the West with self-limited acute hepatitis B, in whom no period of immunologic tolerance occurs. Even among those with neonatally acquired HBV infection, in whom immunologic tolerance appears to be established definitively, immunologic responses to HBV infection have been demonstrated, and intermittent bursts of hepatic necroinflammatory activity punctuate the early decades of life during which liver injury appears to be quiescent (labeled by some as the "immunotolerant" phase; however, it more accurately is a period of dissociation between high-level HBV replication and a paucity of inflammatory liver injury). In addition, even when clinically apparent liver injury and progressive fibrosis emerge during later decades (the so-called immunoreactive, or immunointolerant, phase), the level of immunologic tolerance to HBV remains substantial. More accurately, in patients with neonatally acquired HBV infection, a dynamic equilibrium exists between tolerance and intolerance, the outcome of which determines the clinical expression of chronic infection. Persons infected as neonates tend to have a relatively higher level of immunologic tolerance (high replication, low necroinflammatory activity) during the early decades of life and a relatively lower level (but only rarely a loss) of tolerance (and necroinflammatory activity reflecting the level of virus replication) in the later decades of life.

Hepatitis C Cell-mediated immune responses and elaboration by T cells of antiviral cytokines contribute to the multicellular innate and adaptive immune responses involved in the containment of infection and pathogenesis of liver injury associated with hepatitis C. The fact that HCV is so efficient in evading these immune mechanisms is a testament to its highly evolved ability to disrupt host immune responses at multiple levels. After exposure to HCV, the host cell identifies viral product motifs (pattern recognition receptors) that distinguish the virus from "self," resulting in the elaboration of interferons and other cytokines that result in activation of innate and adaptive immune responses. Intrahepatic HLA class I–restricted cytolytic T cells directed at nucleocapsid, envelope, and nonstructural viral protein antigens have been demonstrated in patients with chronic hepatitis C; however, such virus-specific cytolytic T cell responses do not correlate adequately with the degree of liver injury or with recovery. Yet, a consensus has emerged supporting a role in the pathogenesis of HCV-associated liver injury of virus-activated CD4+ helper T cells that stimulate, via the cytokines they elaborate, HCV-specific CD8+ cytotoxic T cells. These responses appear to be more robust (higher in number, more diverse in viral antigen specificity, more functionally effective, and more long lasting) in those who recover from HCV infection than in those who have chronic infection. Contributing to chronic infection are a CD4+ proliferative defect that results in rapid contraction of CD4+ responses, mutations in CD8+ T cell–targeted viral epitopes that allow HCV to escape immune-mediated clearance, and upregulation of inhibitory receptors on functionally impaired, exhausted T cells. Although attention has focused on adaptive immunity, HCV proteins have been shown to interfere with innate immunity by resulting in blocking of type 1 interferon responses and inhibition of interferon signaling and effector molecules in the interferon signaling cascade. Several HLA alleles have been linked with self-limited hepatitis C, the most convincing of which is the CC haplotype of the *IL28B* gene, which codes for interferon λ3, a component of innate immune antiviral defense. The *IL28B* association is even stronger when combined with HLA class II *DQB1*03:01*. The link between non-CC *IL28B* polymorphisms and failure to clear HCV infection has been explained by a chromosome 19q13.13 frameshift variant upstream of *IL28B*, the ΔG polymorphism of which creates an ORF in a novel interferon gene (*IFN-λ4*) associated with impaired HCV clearance. Also shown to contribute to limiting HCV infection are NK cells of the innate immune system that function when HLA class I molecules required for successful adaptive immunity are underexpressed. Both peripheral cytotoxicity and intrahepatic NK cell cytotoxicity are dysfunctional in persistent HCV infection. Adding to the complexity of the immune response, HCV core, NS4B, and NS5B have been shown to suppress the immunoregulatory nuclear factor (NF)-κB pathway, resulting in reduced antiapoptotic proteins and a resultant increased vulnerability to tumor necrosis factor (TNF) α–mediated cell death. Patients with hepatitis C and unfavorable (non-CC, associated with reduced HCV clearance) *IL28B* alleles have been shown to have depressed NK cell/innate immune function. Of note, the emergence of substantial viral quasispecies diversity and HCV sequence variation allow the virus to evade attempts by the host to contain HCV infection by both humoral and cellular immunity.

Finally, cross-reactivity between viral antigens (HCV NS3 and NS5A) and host autoantigens (cytochrome P450 2D6) has been invoked to explain the association between hepatitis C and a subset of patients with autoimmune hepatitis and antibodies to liver-kidney microsomal (LKM) antigen (anti-LKM) **(Chap. 334)**.

EXTRAHEPATIC MANIFESTATIONS

Immune complex–mediated tissue damage appears to play a pathogenetic role in the extrahepatic manifestations of acute hepatitis B. The occasional prodromal serum sickness–like syndrome observed in acute hepatitis B appears to be related to the deposition in tissue blood vessel walls of HBsAg-anti-HBs circulating immune complexes, leading to activation of the complement system and depressed serum complement levels.

In patients with chronic hepatitis B, other types of immune-complex disease may be seen. Glomerulonephritis with the nephrotic syndrome is observed occasionally; HBsAg, immunoglobulin, and C3 deposition has been found in the glomerular basement membrane. Whereas generalized vasculitis (polyarteritis nodosa) develops in considerably <1% of patients with chronic HBV infection, 20–30% of patients with polyarteritis nodosa have HBsAg in serum **(Chap. 356)**. In these patients, the affected small- and medium-size arterioles contain HBsAg, immunoglobulins, and complement components. Another extrahepatic manifestation of viral hepatitis, essential mixed cryoglobulinemia (EMC), was reported initially to be associated with hepatitis B. The disorder is characterized clinically by arthritis, cutaneous vasculitis (palpable purpura), and, occasionally, glomerulonephritis and serologically by the presence of circulating cryoprecipitable immune complexes of more than one immunoglobulin class **(Chaps. 308 and 356)**. Many patients with this syndrome have chronic liver disease, but the association with HBV infection is limited; instead, a substantial proportion has chronic HCV infection, with circulating immune complexes containing HCV RNA. Immune-complex glomerulonephritis is another recognized extrahepatic manifestation of chronic hepatitis C. Immune-complex disorders have been linked, albeit rarely, with both hepatitis A and E. In hepatitis E,

rare neurologic complications have been postulated to result from both immunologic mechanisms and/or direct CNS infection with the virus.

PATHOLOGY

The typical morphologic lesions of all types of viral hepatitis are similar and consist of panlobular infiltration with mononuclear cells, hepatic cell necrosis, hyperplasia of Kupffer cells, and variable degrees of cholestasis. Hepatic cell regeneration is present, as evidenced by numerous mitotic figures, multinucleated cells, and "rosette" or "pseudoacinar" formation. The mononuclear infiltration consists primarily of small lymphocytes, although plasma cells and eosinophils occasionally are present. Liver cell damage consists of hepatic cell degeneration and necrosis, cell dropout, ballooning of cells, and acidophilic degeneration of hepatocytes (forming so-called Councilman or apoptotic bodies). Large hepatocytes with a ground-glass appearance of the cytoplasm may be seen in chronic but not in acute HBV infection; these cells contain HBsAg and can be identified histochemically with orcein or aldehyde fuchsin. In uncomplicated viral hepatitis, the reticulin framework is preserved.

In hepatitis C, the histologic lesion is often remarkable for a relative paucity of inflammation, a marked increase in activation of sinusoidal lining cells, lymphoid aggregates, the presence of fat (more frequent in genotype 3 and linked to increased fibrosis), and, occasionally, bile duct lesions in which biliary epithelial cells appear to be piled up without interruption of the basement membrane. Occasionally, microvesicular steatosis occurs in hepatitis D. In hepatitis E, a common histologic feature is marked cholestasis. A cholestatic variant of slowly resolving acute hepatitis A also has been described.

A more severe histologic lesion, *bridging hepatic necrosis*, also termed *subacute* or *confluent necrosis* or *interface hepatitis*, is observed occasionally in acute hepatitis. "Bridging" between lobules results from large areas of hepatic cell dropout, with collapse of the reticulin framework. Characteristically, the bridge consists of condensed reticulum, inflammatory debris, and degenerating liver cells that span adjacent portal areas, portal to central veins, or central vein to central vein. This lesion had been thought to have prognostic significance; in many of the originally described patients with this lesion, a subacute course terminated in death within several weeks to months, or severe chronic hepatitis and cirrhosis developed; however, the association between bridging necrosis and a poor prognosis in patients with acute hepatitis has not been upheld. Therefore, although demonstration of this lesion in patients with chronic hepatitis has prognostic significance **(Chap. 334)**, its demonstration during acute hepatitis is less meaningful, and liver biopsies to identify this lesion are no longer undertaken routinely in patients with acute hepatitis. In *massive hepatic necrosis* (fulminant hepatitis, "acute yellow atrophy"), the striking feature at postmortem examination is the finding of a small, shrunken, soft liver. Histologic examination reveals massive necrosis and dropout of liver cells of most lobules with extensive collapse and condensation of the reticulin framework. When histologic documentation is required in the management of fulminant or very severe hepatitis, a biopsy can be done by the angiographically guided transjugular route, which permits the performance of this invasive procedure in the presence of severe coagulopathy.

Immunohistochemical and electron-microscopic studies have localized HBsAg to the cytoplasm and plasma membrane of infected liver cells. In contrast, HBcAg predominates in the nucleus, but, occasionally, scant amounts are also seen in the cytoplasm and on the cell membrane. HDV antigen is localized to the hepatocyte nucleus, whereas HAV, HCV, and HEV antigens are localized to the cytoplasm.

EPIDEMIOLOGY AND GLOBAL FEATURES

Before the availability of serologic tests for hepatitis viruses, all viral hepatitis cases were labeled either as "infectious" or "serum" hepatitis. Modes of transmission overlap, however, and *a clear distinction among the different types of viral hepatitis cannot be made solely on the basis of clinical or epidemiologic features* **(Table 332-2)**. The most accurate means to distinguish the various types of viral hepatitis involves specific serologic testing.

Hepatitis A *This agent is transmitted almost exclusively by the fecal-oral route*. Person-to-person spread of HAV is enhanced by poor personal hygiene and overcrowding; large outbreaks as well as sporadic cases have been traced to contaminated food, water, milk, frozen raspberries and strawberries, green onions imported from Mexico, and shellfish (e.g., scallops imported from the Philippines used to make sushi, the culprit identified in a 2016 Hawaiian outbreak). Intrafamily and intrainstitutional spreads are also common. Early epidemiologic observations supported a predilection for hepatitis A to occur in late fall and early winter. In temperate zones, epidemic waves have been recorded every 5–20 years as new segments of nonimmune population appeared; however, in developed countries, the incidence of hepatitis A has been declining, presumably as a function of improved sanitation, and these cyclic patterns are no longer observed. No HAV carrier state has been identified after acute hepatitis A; perpetuation of the virus in nature depends presumably on nonepidemic, inapparent subclinical infection, ingestion of contaminated food or water in, or imported from, endemic areas, and/or contamination linked to environmental reservoirs.

In the general population, anti-HAV, a marker for previous HAV infection, increases in prevalence as a function of increasing age and of decreasing socioeconomic status. In the 1970s, serologic evidence of prior hepatitis A infection occurred in ~40% of urban populations in the United States, most of whose members never recalled having had a symptomatic case of hepatitis. In subsequent decades, however, the prevalence of anti-HAV has been declining in the United States. In developing countries, exposure, infection, and subsequent immunity are almost universal in childhood. As the frequency of subclinical childhood infections declines in developed countries, a susceptible cohort of adults emerges. Hepatitis A tends to be more symptomatic in adults; therefore, paradoxically, as the frequency of HAV infection declines, the likelihood of clinically apparent, even severe, HAV illnesses increases in the susceptible adult population. Travel to endemic areas is a common source of infection for adults from nonendemic areas. More recently recognized epidemiologic foci of HAV infection include child care centers, neonatal intensive care units, promiscuous men who have sex with men, injection drug users, and unvaccinated close contacts of newly arrived international adopted children, most of whom emanate from countries with intermediate-to-high hepatitis A endemicity. Although hepatitis A is rarely bloodborne, several outbreaks have been recognized in recipients of clotting-factor concentrates. In the United States, the introduction of hepatitis A vaccination programs among children from high-incidence states has resulted in a >70% reduction in the annual incidence of new HAV infections and has shifted the burden of new infections from children to adults. In the 2007–2012 U.S. Public Health Service National Health and Nutrition Examination Survey (NHANES), the prevalence of anti-HAV in the U.S. population aged ≥20 years had declined to 24.2% from the 29.5% measured in NHANES 1999–2006. While universal childhood vaccination accounted for a high prevalence of vaccine-induced immunity in children aged 2–19 years, the lowest age-specific prevalence of anti-HAV (16.1–17.6%) occurred in adults in the fourth and fifth decades, respectively (aged 30–49 years). This is a subgroup of the population who remain susceptible to acute hepatitis A acquired during travel to endemic areas and from contaminated foods, especially those imported from endemic countries.

Hepatitis B Percutaneous inoculation has long been recognized as a major route of hepatitis B transmission, but the outmoded designation "serum hepatitis" is an inaccurate label for the epidemiologic spectrum of HBV infection. As detailed below, most of the hepatitis transmitted by blood transfusion is not caused by HBV; moreover, in approximately two-thirds of patients with acute type B hepatitis, no history of an identifiable percutaneous exposure can be elicited. We now recognize that many cases of hepatitis B result from less obvious modes of nonpercutaneous or covert percutaneous transmission. HBsAg has been identified in almost every body fluid from infected persons, and at least some of these body fluids—most notably semen and saliva—are infectious, albeit less so than serum, when administered percutaneously or nonpercutaneously to experimental animals. Among the nonpercutaneous

TABLE 332-2 Clinical and Epidemiologic Features of Viral Hepatitis

FEATURE	HAV	HBV	HCV	HDV	HEV
Incubation (days)	15–45, mean 30	30–180, mean 60–90	15–160, mean 50	30–180, mean 60–90	14–60, mean 40
Onset	Acute	Insidious or acute	Insidious or acute	Insidious or acute	Acute
Age preference	Children, young adults	Young adults (sexual and percutaneous), babies, toddlers	Any age, but more common in adults	Any age (similar to HBV)	Epidemic cases: young adults (20–40 years); sporadic cases: older adults (>60)
Transmission					
Fecal-oral	+++	–	–	–	+++
Percutaneous	Unusual	+++	+++	+++	–
Perinatal	–	+++	±[a]	+	–
Sexual	±	++	±[a]	++	–
Clinical					
Severity	Mild	Occasionally severe	Moderate	Occasionally severe	Mild
Fulminant	0.1%	0.1–1%	0.1%	5–20%[b]	1–2%[e]
Progression to chronicity	None	Occasional (1–10%) (90% of neonates)	Common (85%)	Common[d]	None[f]
Carrier	None	0.1–30%[c]	1.5–3.2%	Variable[g]	None
Cancer	None	+ (neonatal infection)	+	±	None
Prognosis	Excellent	Worse with age, debility	Moderate	Acute, good Chronic, poor	Good
Prophylaxis	Ig, inactivated vaccine	HBIG, recombinant vaccine	None	HBV vaccine (none for HBV carriers)	Vaccine
Therapy	None	Interferon Lamivudine Adefovir Pegylated interferon[h] Entecavir[h] Telbivudine Tenofovir[h]	Pegylated interferon ribavirin telaprevir,[i] boceprevir,[i] simeprevir, sofosbuvir, ledipasvir, paritaprevir/ritonavir ombitasvir, dasabuvir daclatasvir, velpatasvir, grazoprevir, elbasvir	Pegylated interferon ±	None[j]

[a]Primarily with HIV co-infection and high-level viremia in index case; more likely in persons with multiple sex partners or sexually transmitted diseases; risk ~5%. [b]Up to 5% in acute HBV/HDV co-infection; up to 20% in HDV superinfection of chronic HBV infection.[c] Varies considerably throughout the world and in subpopulations within countries; see text. [d]In acute HBV/HDV co-infection, the frequency of chronicity is the same as that for HBV; in HDV superinfection, chronicity is invariable. [e]10–20% in pregnant women. [f]Except as observed in immunosuppressed liver allograft recipients or other immunosuppressed hosts. [g]Common in Mediterranean countries; rare in North America and western Europe. [h]First-line agents. [i]No longer recommended. [j]Anecdotal reports and retrospective studies suggest that pegylated interferon and/or ribavirin are effective in treating chronic hepatitis E, observed in immunocompromised persons; ribavirin monotherapy has been used successfully in acute, severe hepatitis E.

Abbreviation: HBIG, hepatitis B immunoglobulin. See text for other abbreviations.

modes of HBV transmission, oral ingestion has been documented as a potential but inefficient route of exposure. By contrast, the two nonpercutaneous routes considered to have the greatest impact are intimate (especially sexual) contact and perinatal transmission.

In sub-Saharan Africa, intimate contact among toddlers is considered instrumental in contributing to the maintenance of the high frequency of hepatitis B in the population. Perinatal transmission occurs primarily in infants born to mothers with chronic hepatitis B or (rarely) mothers with acute hepatitis B during the third trimester of pregnancy or during the early postpartum period. Perinatal transmission is uncommon in North America and western Europe but occurs with great frequency and is the most important mode of HBV perpetuation in East Asia and developing countries. Although the precise mode of perinatal transmission is unknown, and although ~10% of infections may be acquired in utero, epidemiologic evidence suggests that most infections occur approximately at the time of delivery and are not related to breast-feeding (which is not contraindicated in women with hepatitis B). The likelihood of perinatal transmission of HBV correlates with the presence of HBeAg and high-level viral replication; 90% of HBeAg-positive mothers but only 10–15% of anti-HBe-positive mothers transmit HBV infection to their offspring. In most cases, acute infection in the neonate is clinically asymptomatic, but the child is very likely to remain chronically infected.

The >350–400 million persons with chronic HBV infection in the world constitute the main reservoir of hepatitis B in human beings. Whereas serum HBsAg is infrequent (0.1–0.5%) in normal populations in the United States and western Europe, a prevalence of up to 5–20% has been found in East Asia and in some tropical countries; in persons with Down's syndrome, lepromatous leprosy, leukemia, Hodgkin's disease, or polyarteritis nodosa; in patients with chronic renal disease on hemodialysis; and in injection drug users.

Other groups with high rates of HBV infection include spouses of acutely infected persons; sexually promiscuous persons (especially promiscuous men who have sex with men); health care workers exposed to blood; persons who require repeated transfusions especially with pooled blood-product concentrates (e.g., hemophiliacs); residents and staff of custodial institutions for the developmentally handicapped; prisoners; and, to a lesser extent, family members of chronically infected patients. In volunteer blood donors, the prevalence of anti-HBs, a reflection of previous HBV infection, ranges from 5% to 10%, but the prevalence is higher in lower socioeconomic strata, older age groups, and persons—including those mentioned above—exposed to blood products. Because of highly sensitive virologic screening of donor blood, the risk of acquiring HBV infection from a blood transfusion is 1 in 230,000.

Prevalence of infection, modes of transmission, and human behavior conspire to mold geographically different epidemiologic patterns of HBV infection. In East Asia and Africa, hepatitis B, a disease of the newborn and young children, is perpetuated by a cycle of maternal-neonatal spread. In North America and western Europe, hepatitis B is primarily a disease of adolescence and early adulthood, the time of life when intimate sexual contact and recreational and occupational percutaneous exposures tend to occur. To some degree, however, this dichotomy between high-prevalence and low-prevalence geographic regions has been minimized by immigration from high-prevalence to low-prevalence areas. For example, in the United States, NHANES data from 2007 to 2012 revealed an overall prevalence of current HBV infection (detectable HBsAg) of 0.3%; however, the prevalence in Asian persons,

TABLE 332-3 High-Risk Populations for Whom HBV Infection Screening Is Recommended

Persons born in countries/regions with a high (≥8%) and intermediate (≥2%) prevalence of HBV infection including immigrants and adopted children and including persons born in the United States who were not vaccinated as infants and whose parents emigrated from areas of high HBV endemicity
Household and sexual contacts of persons with hepatitis B
Babies born to HBsAg-positive mothers
Persons who have used injection drugs
Persons with multiple sexual contacts or a history of sexually transmitted disease
Men who have sex with men
Inmates of correctional facilities
Persons with elevated alanine or aspartate aminotransferase levels
Blood/plasma/organ/tissue/semen donors
Persons with HCV or HIV infection
Hemodialysis patients
Pregnant women
Persons who are the source of blood or body fluids that would be an indication for postexposure prophylaxis (e.g., needlestick, mucosal exposure, sexual assault)
Persons who require immunosuppressive or cytotoxic therapy (including anti–tumor necrosis factor α therapy for rheumatologic or inflammatory bowel disorders)

93% of whom were foreign-born, was 10-fold higher, 3.1%, representing 50% of the U.S. national disease burden. The introduction of hepatitis B vaccine in the early 1980s and adoption of universal childhood vaccination policies in many countries resulted in a dramatic, ~90% decline in the incidence of new HBV infections in those countries as well as in the dire consequences of chronic infection, including hepatocellular carcinoma. In the United States, as demonstrated in NHANES 2007–2012, following the 1991 implementation of universal childhood vaccination, HBsAg seropositivity had declined in children aged 6–19 years to as low as 0.03%, an ~85% reduction. Populations and groups for whom HBV infection screening is recommended are listed in **Table 332-3**.

Hepatitis D Infection with HDV has a worldwide distribution, but two epidemiologic patterns exist. In Mediterranean countries (northern Africa, southern Europe, the Middle East), HDV infection is endemic among those with hepatitis B, and the disease is transmitted predominantly by nonpercutaneous means, especially close personal contact. In nonendemic areas, such as the United States (where hepatitis D is rare among persons with chronic hepatitis B) and northern Europe, HDV infection is confined to persons exposed frequently to blood and blood products, primarily injection drug users, (especially so in HIV-infected injection drug users) and hemophiliacs. In the United States, the prevalence of HDV infection in the national population is 0.02% (NHANES 1999–2012); however, among HBsAg-positive persons, the prevalence of HDV infection is highest in injection drug users (11%) and hemophiliacs (19%). HDV infection can be introduced into a population through drug users or by migration of persons from endemic to nonendemic areas. Thus, patterns of population migration and human behavior facilitating percutaneous contact play important roles in the introduction and amplification of HDV infection. Occasionally, the migrating epidemiology of hepatitis D is expressed in explosive outbreaks of severe hepatitis, such as those that have occurred in remote South American villages (e.g., "Lábrea fever" in the Amazon basin) as well as in urban centers in the United States. Ultimately, such outbreaks of hepatitis D—either of co-infections with acute hepatitis B or of superinfections in those already infected with HBV—may blur the distinctions between endemic and nonendemic areas. On a global scale, HDV infection declined at the end of the 1990s. Even in Italy, an HDV-endemic area, public health measures introduced to control HBV infection (e.g., mass hepatitis B vaccination) resulted during the 1990s in a 1.5%/year reduction in the prevalence of HDV infection. Still, the frequency of HDV infection during the first decade of the twenty-first century has not fallen below levels reached during the 1990s; the reservoir has been sustained by survivors infected during 1970–1980 and recent immigrants from still-endemic (e.g., Eastern Europe and Central Asia) to less-endemic countries.

Hepatitis C Routine screening of blood donors for HBsAg and the elimination of commercial blood sources in the early 1970s reduced the frequency of, but did not eliminate, transfusion-associated hepatitis. During the 1970s, the likelihood of acquiring hepatitis after transfusion of voluntarily donated, HBsAg-screened blood was ~10% per patient (up to 0.9% per unit transfused); 90–95% of these cases were classified, based on serologic exclusion of hepatitis A and B, as "non-A, non-B" hepatitis. For patients requiring transfusion of pooled products, such as clotting factor concentrates, the risk was even higher, up to 20–30%.

During the 1980s, voluntary self-exclusion of blood donors with risk factors for AIDS and then the introduction of donor screening for anti-HIV reduced further the likelihood of transfusion-associated hepatitis to <5%. During the late 1980s and early 1990s, the introduction first of "surrogate" screening tests for non-A, non-B hepatitis (alanine aminotransferase [ALT] and anti-HBc, both shown to identify blood donors with a higher likelihood of transmitting non-A, non-B hepatitis to recipients) and, subsequently, after the discovery of HCV, first-generation immunoassays for anti-HCV reduced the frequency of transfusion-associated hepatitis even further. A prospective analysis of transfusion-associated hepatitis conducted between 1986 and 1990 showed that the frequency of transfusion-associated hepatitis at one urban university hospital fell from a baseline of 3.8% per patient (0.45% per unit transfused) to 1.5% per patient (0.19% per unit) after the introduction of surrogate testing and to 0.6% per patient (0.03% per unit) after the introduction of first-generation anti-HCV assays. The introduction of second-generation anti-HCV assays reduced the frequency of transfusion-associated hepatitis C to almost imperceptible levels—1 in 100,000—and these gains were reinforced by the application of third-generation anti-HCV assays and of automated PCR testing of donated blood for HCV RNA, which has resulted in a reduction in the risk of transfusion-associated HCV infection to 1 in 2.3 million transfusions.

In addition to being transmitted by transfusion, hepatitis C can be transmitted by other percutaneous routes, such as injection drug use. In addition, this virus can be transmitted by occupational exposure to blood, and the likelihood of infection is increased in hemodialysis units. Although the frequency of transfusion-associated hepatitis C fell as a result of blood-donor screening, the overall frequency of hepatitis C remained the same until the early 1990s, when the overall frequency of reported cases fell by 80%, in parallel with a reduction in the number of new cases in injection drug users. After the exclusion of anti-HCV-positive plasma units from the donor pool, rare, sporadic instances have occurred of hepatitis C among recipients of immunoglobulin preparations for intravenous (but not intramuscular) use.

Serologic evidence for HCV infection occurs in 90% of patients with a history of transfusion-associated hepatitis (almost all occurring before 1992, when second-generation HCV screening tests were introduced); hemophiliacs and others treated with clotting factors; injection drug users; 60–70% of patients with sporadic "non-A, non-B" hepatitis who lack identifiable risk factors; 0.5% of volunteer blood donors; and, in the NHANES survey conducted in the United States between 1999 and 2002, 1.6% of the general population in the United States, which translates into 4.1 million persons (3.2 million with viremia), the majority of whom are unaware of their infections. Moreover, such population surveys do not include higher-risk groups such as incarcerated persons, homeless persons, and active injection drug users, indicating that the actual prevalence is even higher. Comparable frequencies of HCV infection occur in most countries around the world, with 170 million persons infected worldwide, but extraordinarily high prevalences of HCV infection occur in certain countries such as Egypt, where >20% of the population (as high as 50% in persons born prior to 1960) in some cities is infected. The high frequency in Egypt is attributable to contaminated equipment used for medical procedures and unsafe injection practices in the 1950s to 1980s (during a campaign to eradicate schistosomiasis with intravenous tartar emetic). In the United States,

African Americans and Mexican Americans have higher frequencies of HCV infection than whites. Data from NHANES showed that between 1988 and 1994, 30- to 40-year-old men had the highest prevalence of HCV infection; however, in a survey conducted between 1999 and 2002, the peak age decile had shifted to those age 40–49 years; an increase in hepatitis C–related mortality has paralleled this secular trend, increasing since 1995 predominantly in the 45- to 65-year age group. Thus, despite an 80% reduction in new HCV infections during the 1990s, the prevalence of HCV infection in the population was sustained by an aging cohort that had acquired their infections three to four decades earlier, during the 1960s and 1970s, as a result predominantly of self-inoculation with recreational drugs. Retrospective phylogenetic mapping of >45,000 HCV genotype 1a isolates revealed that the hepatitis C epidemic emerged in the United States between 1940 and 1965, peaking in 1950 and aligning temporally with the post-World-War-II expansion of medical procedures (including re-use of glass syringes). Thus, HCV was amplified iatrogenically not only in Egypt but also in the United States; in the United States, the seeds sewn by medical procedures in the 1950s were reaped in the 1960s and 1970s among transfusion recipients and injection drug users, even those whose drug use was confined to brief adolescent experimentation.

In NHANES 2003–2010, the prevalence of HCV infection (HCV RNA reactivity) in the United States had actually fallen to 1% (2.7 million persons) from 1.3% (3.2 million) the decade before (NHANES 1999–2002), attributable to deaths among the HCV-infected population. As death resulting from HIV infection fell after 1999, age-adjusted mortality associated with HCV infection surpassed that of HIV infection in 2007; >70% of HCV-associated deaths occurred in the "baby boomer" cohort born between 1945 and 1965. By 2012, HCV mortality had surpassed deaths from HIV, tuberculosis, hepatitis B, and 57 other notifiable infectious diseases (i.e., *all* infectious diseases) reported to the Centers for Disease Control and Prevention. In NHANES 1999–2002, compared to the 1.6% prevalence of HCV infection in the population at large, the prevalence in the 1945–1965 birth cohort was 3.2%, representing three-quarters of all infected persons. Therefore, in 2012, the Centers for Disease Control and Prevention recommended that all persons born between 1945 and 1965 be screened for hepatitis C, without ascertainment of risk, a recommendation shown to be cost-effective and predicted to identify 800,000 infected persons. Because of the availability of highly effective antiviral therapy, such screening would have the potential to avert 200,000 cases of cirrhosis and 47,000 cases of hepatocellular carcinoma and to prevent 120,000 hepatitis-related deaths; with the availability of the new generation of direct-acting antivirals (efficacy >95%, see **Chap 334**), screening baby boomers and treating those with hepatitis C have been predicted to reduce the HCV-associated disease burden by 50–70% through 2050.

Hepatitis C accounts for 40% of chronic liver disease, is the most frequent indication for liver transplantation, and is estimated to account for 8000–10,000 deaths per year in the United States. The distribution of HCV genotypes varies in different parts of the world. Worldwide, genotype 1 is the most common. In the United States, genotype 1 accounts for 70% of HCV infections, whereas genotypes 2 and 3 account for the remaining 30%; among African Americans, the frequency of genotype 1 is even higher (i.e., 90%). Genotype 4 predominates in Egypt; genotype 5 is localized to South Africa, genotype 6 to Hong Kong, and genotype 7 to Central Africa. Most asymptomatic blood donors found to have anti-HCV and ~20–30% of persons with reported cases of acute hepatitis C do not fall into a recognized risk group; however, many such blood donors do recall risk-associated behaviors when questioned carefully.

As a bloodborne infection, HCV potentially can be transmitted sexually and perinatally; however, both of these modes of transmission are inefficient for hepatitis C. Although 10–15% of patients with acute hepatitis C report having potential sexual sources of infection, most studies have failed to identify sexual transmission of this agent. The chances of sexual and perinatal transmission have been estimated to be ~5% but shown in a prospective study to be only 1% between monogamous sexual partners, well below comparable rates for HIV and HBV infections. Moreover, sexual transmission appears to be confined to such subgroups as persons with multiple sexual partners and sexually transmitted diseases; for example, isolated clusters of sexually transmitted HCV infection have been reported in HIV-infected men who have sex with men. Breast-feeding does not increase the risk of HCV infection between an infected mother and her infant. Infection of health workers is not dramatically higher than among the general population; however, health workers are more likely to acquire HCV infection through accidental needle punctures, the efficiency of which is ~3%. Infection of household contacts is rare as well.

TABLE 332-4 High-Risk Populations for Whom HCV-Infection Screening is Recommended

Persons born between 1945 and 1965
Persons who have ever used injection drugs
Persons with HIV infection
Hemophiliacs treated with clotting factor concentrates prior to 1987
Persons who have ever undergone long-term hemodialysis
Persons with unexplained elevations of aminotransferase levels
Transfusion or transplantation recipients prior to July 1992
Recipients of blood or organs from a donor found to be positive for hepatitis C
Children born to women with hepatitis C
Health care, public safety, and emergency medical personnel following needle injury or mucosal exposure to HCV-contaminated blood
Sexual partners of persons with hepatitis C infection

Besides persons born between 1945 and 1965, other groups with an increased frequency of HCV infection are listed in **Table 332-4**. In immunosuppressed individuals, levels of anti-HCV may be undetectable, and a diagnosis may require testing for HCV RNA. Although new acute cases of hepatitis C are rare outside of the injection-drug using community, newly diagnosed cases are common among otherwise healthy persons who experimented briefly with injection drugs, as noted above, three or four decades earlier. Such instances usually remain unrecognized for years, until unearthed by laboratory screening for routine medical examinations, insurance applications, and attempted blood donation. Although, overall, the annual incidence of new HCV infections has continued to fall, the rate of new infections has been increasing since 2002, amplified by the recent epidemic of opioid use, in a new cohort of young injection drug users, age 15–24 years (accounting for more than two-thirds of all acute cases), who, unlike older cohorts, had not learned to take precautions to prevent bloodborne infections.

Hepatitis E This type of hepatitis, identified in India, Asia, Africa, the Middle East, and Central America, resembles hepatitis A in its primarily enteric mode of spread. The commonly recognized cases occur after contamination of water supplies such as after monsoon flooding, but sporadic, isolated cases occur. An epidemiologic feature that distinguishes HEV from other enteric agents is the rarity of secondary person-to-person spread from infected persons to their close contacts. Large waterborne outbreaks in endemic areas are linked to genotypes 1 and 2, arise in populations that are immune to HAV, favor young adults, and account for antibody prevalences of 30–80%. In nonendemic areas of the world, such as the United States, clinically apparent acute hepatitis E is extremely rare; however, during the 1988–1994 NHANES survey conducted by the U.S. Public Health Service, the prevalence of anti-HEV was 21%, reflecting subclinical infections, infection with genotypes 3 and 4, predominantly in older males (>60 years). In nonendemic areas, HEV accounts hardly at all for cases of sporadic (labeled "autochthonous" or indigenous) hepatitis; however, cases imported from endemic areas have been found in the United States. Evidence supports a zoonotic reservoir for HEV primarily in swine, which may account for the mostly subclinical infections in nonendemic areas. A previously unrecognized high distribution of HEV infection, linked to pork-product ingestion, has been discovered in western Europe (e.g., in Germany, an estimated annual incidence of 300,000 cases and a 17% prevalence of anti-HEV among adults; in France, a 22% prevalence of anti-HEV in healthy blood donors).

CLINICAL AND LABORATORY FEATURES

Symptoms and Signs Acute viral hepatitis occurs after an incubation period that varies according to the responsible agent. Generally, incubation periods for hepatitis A range from 15 to 45 days (mean, 4 weeks), for hepatitis B and D from 30 to 180 days (mean, 8–12 weeks), for hepatitis C from 15 to 160 days (mean, 7 weeks), and for hepatitis E from 14 to 60 days (mean, 5–6 weeks). The *prodromal symptoms* of acute viral hepatitis are systemic and quite variable. Constitutional symptoms of anorexia, nausea and vomiting, fatigue, malaise, arthralgias, myalgias, headache, photophobia, pharyngitis, cough, and coryza may precede the onset of jaundice by 1–2 weeks. The nausea, vomiting, and anorexia are frequently associated with alterations in olfaction and taste. A low-grade fever between 38° and 39°C (100°–102°F) is more often present in hepatitis A and E than in hepatitis B or C, except when hepatitis B is heralded by a serum sickness–like syndrome; rarely, a fever of 39.5°–40°C (103°–104°F) may accompany the constitutional symptoms. Dark urine and clay-colored stools may be noticed by the patient from 1–5 days before the onset of clinical jaundice.

With the onset of *clinical jaundice*, the constitutional prodromal symptoms usually diminish, but in some patients, mild weight loss (2.5–5 kg) is common and may continue during the entire icteric phase. The liver becomes enlarged and tender and may be associated with right upper quadrant pain and discomfort. Infrequently, patients present with a cholestatic picture, suggesting extrahepatic biliary obstruction. Splenomegaly and cervical adenopathy are present in 10–20% of patients with acute hepatitis. Rarely, a few spider angiomas appear during the icteric phase and disappear during convalescence. During the *recovery phase*, constitutional symptoms disappear, but usually some liver enlargement and abnormalities in liver biochemical tests are still evident. The duration of the posticteric phase is variable, ranging from 2 to 12 weeks, and is usually more prolonged in acute hepatitis B and C. Complete clinical and biochemical recovery is to be expected 1–2 months after all cases of hepatitis A and E and 3–4 months after the onset of jaundice in three-quarters of uncomplicated, self-limited cases of hepatitis B and C (among healthy adults, acute hepatitis B is self-limited in 95–99%, whereas hepatitis C is self-limited in only ~15–20%). In the remainder, biochemical recovery may be delayed. A substantial proportion of patients with viral hepatitis never become icteric.

Infection with HDV can occur in the presence of acute or chronic HBV infection; the duration of HBV infection determines the duration of HDV infection. When acute HDV and HBV infections occur simultaneously, clinical and biochemical features may be indistinguishable from those of HBV infection alone, although occasionally they are more severe. As opposed to patients with *acute* HBV infection, patients with *chronic* HBV infection can support HDV replication indefinitely, as when acute HDV infection occurs in the presence of a nonresolving acute HBV infection or, more commonly, when acute hepatitis D is superimposed on underlying chronic hepatitis B. In such cases, the HDV superinfection appears as a clinical exacerbation or an episode resembling acute viral hepatitis in someone already chronically infected with HBV. Superinfection with HDV in a patient with chronic hepatitis B often leads to clinical deterioration (see below).

In addition to superinfections with other hepatitis agents, acute hepatitis-like clinical events in persons with chronic hepatitis B may accompany spontaneous HBeAg to anti-HBe seroconversion or spontaneous reactivation (i.e., reversion from relatively nonreplicative to replicative infection). Such reactivations can occur as well in therapeutically immunosuppressed patients with chronic HBV infection when cytotoxic/immunosuppressive drugs are withdrawn; in these cases, restoration of immune competence is thought to allow resumption of previously checked cell-mediated immune cytolysis of HBV-infected hepatocytes. Occasionally, acute clinical exacerbations of chronic hepatitis B may represent the emergence of a precore mutant (see "Virology and Etiology"), and the subsequent course in such patients may be characterized by periodic exacerbations. Cytotoxic chemotherapy can lead to reactivation of chronic hepatitis C as well, and anti-TNF-α therapy can lead to reactivation of both hepatitis B and C.

Laboratory Features The serum aminotransferases aspartate aminotransferase (AST) and ALT (previously designated SGOT and SGPT) increase to a variable degree during the prodromal phase of acute viral hepatitis and precede the rise in bilirubin level (Figs. 332-2 and 332-4). The level of these enzymes, however, does not correlate well with the degree of liver cell damage. Peak levels vary from ~400 to ~4000 IU or more; these levels are usually reached at the time the patient is clinically icteric and diminish progressively during the recovery phase of acute hepatitis. The diagnosis of anicteric hepatitis is based on clinical features and on aminotransferase elevations.

Jaundice is usually visible in the sclera or skin when the serum bilirubin value is >43 µmol/L (2.5 mg/dL). When jaundice appears, the serum bilirubin typically rises to levels ranging from 85 to 340 µmol/L (5–20 mg/dL). The serum bilirubin may continue to rise despite falling serum aminotransferase levels. In most instances, the total bilirubin is equally divided between the conjugated and unconjugated fractions. Bilirubin levels >340 µmol/L (20 mg/dL) extending and persisting late into the course of viral hepatitis are more likely to be associated with severe disease. In certain patients with underlying hemolytic anemia, however, such as glucose-6-phosphate dehydrogenase deficiency and sickle cell anemia, a high serum bilirubin level is common, resulting from superimposed hemolysis. In such patients, bilirubin levels >513 µmol/L (30 mg/dL) have been observed and are not necessarily associated with a poor prognosis.

Neutropenia and lymphopenia are transient and are followed by a relative lymphocytosis. Atypical lymphocytes (varying between 2 and 20%) are common during the acute phase. Measurement of the prothrombin time (PT) is important in patients with acute viral hepatitis, because a prolonged value may reflect a severe hepatic synthetic defect, signify extensive hepatocellular necrosis, and indicate a worse prognosis. Occasionally, a prolonged PT may occur with only mild increases in the serum bilirubin and aminotransferase levels. Prolonged nausea and vomiting, inadequate carbohydrate intake, and poor hepatic glycogen reserves may contribute to hypoglycemia noted occasionally in patients with severe viral hepatitis. Serum alkaline phosphatase may be normal or only mildly elevated, whereas a fall in serum albumin is uncommon in uncomplicated acute viral hepatitis. In some patients, mild and transient steatorrhea has been noted, as well as slight microscopic hematuria and minimal proteinuria.

A diffuse but mild elevation of the γ globulin fraction is common during acute viral hepatitis. Serum IgG and IgM levels are elevated in about one-third of patients during the acute phase of viral hepatitis, but the serum IgM level is elevated more characteristically during acute hepatitis A. During the acute phase of viral hepatitis, antibodies to smooth muscle and other cell constituents may be present, and low titers of rheumatoid factor, nuclear antibody, and heterophile antibody can also be found occasionally. In hepatitis C and D, antibodies to LKM may occur; however, the species of LKM antibodies in the two types of hepatitis are different from each other as well as from the LKM antibody species characteristic of autoimmune hepatitis type 2 (Chap. 334). The autoantibodies in viral hepatitis are nonspecific and can also be associated with other viral and systemic diseases. In contrast, virus-specific antibodies, which appear during and after hepatitis virus infection, are serologic markers of diagnostic importance.

As described above, serologic tests are available routinely with which to establish a diagnosis of hepatitis A, B, D, and C. Tests for fecal or serum HAV are not routinely available. Therefore, a diagnosis of hepatitis A is based on detection of IgM anti-HAV during acute illness (Fig. 332-2). Rheumatoid factor can give rise to false-positive results in this test.

A diagnosis of HBV infection can usually be made by detection of HBsAg in serum. Infrequently, levels of HBsAg are too low to be detected during acute HBV infection, even with contemporary, highly sensitive immunoassays. In such cases, the diagnosis can be established by the presence of IgM anti-HBc.

The titer of HBsAg bears little relation to the severity of clinical disease. Indeed, an inverse correlation exists between the serum concentration of HBsAg and the degree of liver cell damage. For example, titers are highest in immunosuppressed patients, lower in patients with chronic liver disease (but higher in mild chronic than in severe chronic

hepatitis), and very low in patients with acute fulminant hepatitis. These observations suggest that in hepatitis B the degree of liver cell damage and the clinical course are related to variations in the patient's immune response to HBV rather than to the amount of circulating HBsAg. In immunocompetent persons, however, a correlation exists between markers of HBV *replication* and liver injury (see below).

Another important serologic marker in patients with hepatitis B is HBeAg. Its principal clinical usefulness is as an indicator of relative infectivity. Because HBeAg is invariably present during early acute hepatitis B, HBeAg testing is indicated primarily in chronic infection.

In patients with hepatitis B surface antigenemia of unknown duration (e.g., blood donors found to be HBsAg-positive) testing for IgM anti-HBc may be useful to distinguish between acute or recent infection (IgM anti-HBc-positive) and chronic HBV infection (IgM anti-HBc-negative, IgG anti-HBc-positive). A false-positive test for IgM anti-HBc may be encountered in patients with high-titer rheumatoid factor. Also, IgM anti-HBc may be reexpressed during acute reactivation of chronic hepatitis B.

Anti-HBs is rarely detectable in the presence of HBsAg in patients with *acute* hepatitis B, but 10–20% of persons with *chronic* HBV infection may harbor low-level anti-HBs. This antibody is directed not against the common group determinant, *a*, but against the heterotypic subtype determinant (e.g., HBsAg of subtype *ad* with anti-HBs of subtype *y*). In most cases, this serologic pattern cannot be attributed to infection with two different HBV subtypes but, instead, is thought (based on the clonal selection theory of antibody diversity) to reflect the stimulation of a related clone of antibody-forming cells and is not a harbinger of imminent HBsAg clearance. When such antibody is detected, its presence is of no recognized clinical significance (see "Virology and Etiology").

After immunization with hepatitis B vaccine, which consists of HBsAg alone, anti-HBs is the only serologic marker to appear. The commonly encountered serologic patterns of hepatitis B and their interpretations are summarized in **Table 332-5**. Tests for the detection of HBV DNA in liver and serum are now available. Like HBeAg, serum HBV DNA is an indicator of HBV replication, but tests for HBV DNA are more sensitive and quantitative. First-generation hybridization assays for HBV DNA had a sensitivity of 10^5–10^6 virions/mL, a relative threshold below which infectivity and liver injury are limited and HBeAg is usually undetectable. Currently, testing for HBV DNA has shifted from insensitive hybridization assays to amplification assays (e.g., the PCR-based assay, which can detect as few as 10 or 100 virions/mL); among the commercially available PCR assays, the most useful are those with the highest sensitivity (5–10 IU/mL) and the largest dynamic range (10^0–10^9 IU/mL). With increased sensitivity, amplification assays remain reactive well below the current 10^3–10^4 IU/mL threshold for infectivity and liver injury. These markers are useful in following the course of HBV replication in patients with chronic hepatitis B receiving antiviral chemotherapy **(Chap. 334)**. Except for the early decades of life after perinatally acquired HBV infection (see above), in immunocompetent adults with chronic hepatitis B, a general correlation exists between the level of HBV replication, as reflected by the level of serum HBV DNA, and the degree of liver injury. High-serum HBV DNA levels, increased expression of viral antigens, and necroinflammatory activity in the liver go hand in hand unless immunosuppression interferes with cytolytic T cell responses to virus-infected cells; reduction of HBV replication with antiviral drugs tends to be accompanied by an improvement in liver histology. Among patients with chronic hepatitis B, high levels of HBV DNA increase the risk of cirrhosis, hepatic decompensation, and hepatocellular carcinoma (see "Complications and Sequelae").

In patients with hepatitis C, an episodic pattern of aminotransferase elevation is common. A specific serologic diagnosis of hepatitis C can be made by demonstrating the presence in serum of anti-HCV. When contemporary immunoassays are used, anti-HCV can be detected in acute hepatitis C during the initial phase of elevated aminotransferase activity and remains detectable after recovery (rare) and during chronic infection (common). Nonspecificity can confound immunoassays for anti-HCV, especially in persons with a low prior probability of infection, such as volunteer blood donors, or in persons with circulating rheumatoid factor, which can bind nonspecifically to assay reagents; testing for HCV RNA can be used in such settings to distinguish between true-positive and false-positive anti-HCV determinations. Assays for HCV RNA are the most sensitive tests for HCV infection and represent the "gold standard" in establishing a diagnosis of hepatitis C. HCV RNA can be detected even before acute elevation of aminotransferase activity and before the appearance of anti-HCV in patients with acute hepatitis C. In addition, HCV RNA remains detectable indefinitely, continuously in most but intermittently in some, in patients with chronic hepatitis C (detectable as well in some persons with normal liver tests, i.e., inactive carriers). In the very small minority of patients with hepatitis C who lack anti-HCV, a diagnosis can be supported by detection of HCV RNA. If all these tests are negative and the patient has a well-characterized case of hepatitis after percutaneous exposure to blood or blood products, a diagnosis of hepatitis caused by an unidentified agent can be entertained.

Amplification techniques are required to detect HCV RNA. Currently, such target amplification (i.e., synthesis of multiple copies of the viral genome) is achieved by PCR, in which the viral RNA is reverse transcribed to complementary DNA and then amplified by repeated cycles of DNA synthesis. Quantitative PCR assays provide a measurement of relative "viral load"; current PCR assays have a sensitivity of 10 (lower limit of detection)-25 (lower limit of quantitation) IU/mL and a wide dynamic range (10–10^7 IU/mL). Determination of HCV RNA level is not a reliable marker of disease severity or prognosis but is helpful in predicting relative responsiveness to antiviral therapy. The same is true for determinations of HCV genotype **(Chap. 334)**. Of course, HCV RNA monitoring during and after antiviral therapy is the *sine qua non* for determining on-treatment and durable responsiveness.

A proportion of patients with hepatitis C have isolated anti-HBc in their blood, a reflection of a common risk in certain populations of exposure to multiple bloodborne hepatitis agents. The anti-HBc in such cases is almost

TABLE 332-5 Commonly Encountered Serologic Patterns of Hepatitis B Infection

HBsAg	ANTI-HBs	ANTI-HBc	HBeAg	ANTI-HBe	INTERPRETATION
+	–	IgM	+	–	Acute hepatitis B, high infectivity[a]
+	–	IgG	+	–	Chronic hepatitis B, high infectivity
+	–	IgG	–	+	1. Late acute or chronic hepatitis B, low infectivity 2. HBeAg-negative ("precore-mutant") hepatitis B (chronic or, rarely, acute)
+	+	+	+/–	+/–	1. HBsAg of one subtype and heterotypic anti-HBs (common) 2. Process of seroconversion from HBsAg to anti-HBs (rare)
–	–	IgM	+/–	+/–	1. Acute hepatitis B[a] 2. Anti-HBc "window"
–	–	IgG	–	+/–	1. Low-level hepatitis B carrier 2. Hepatitis B in remote past
–	+	IgG	–	+/–	Recovery from hepatitis B
–	+	–	–	–	1. Immunization with HBsAg (after vaccination) 2. Hepatitis B in the remote past (?) 3. False-positive

[a]IgM anti-HBc may reappear during acute reactivation of chronic hepatitis B.

Note: See text for abbreviations.

invariably of the IgG class and usually represents HBV infection in the remote past (HBV DNA undetectable); it rarely represents current HBV infection with low-level virus carriage. Detectable anti-HCV in the absence of HCV RNA signifies spontaneous or therapeutically induced recovery from ("cured") hepatitis C.

The presence of HDV infection can be identified by demonstrating intrahepatic HDV antigen or, more practically, an anti-HDV seroconversion (a rise in titer of anti-HDV or de novo appearance of anti-HDV). Circulating HDV antigen, also diagnostic of acute infection, is detectable only briefly, if at all. Because anti-HDV is often undetectable once HBsAg disappears, retrospective serodiagnosis of acute self-limited, simultaneous HBV and HDV infection is difficult. Early diagnosis of acute infection may be hampered by a delay of up to 30–40 days in the appearance of anti-HDV.

When a patient presents with acute hepatitis and has HBsAg and anti-HDV in serum, determination of the class of anti-HBc is helpful in establishing the relationship between infection with HBV and HDV. Although IgM anti-HBc does not distinguish *absolutely* between acute and chronic HBV infection, its presence is a reliable indicator of recent infection and its absence a reliable indicator of infection in the remote past. In simultaneous acute HBV and HDV infections, IgM anti-HBc will be detectable, whereas in acute HDV infection superimposed on chronic HBV infection, anti-HBc will be of the IgG class. Assays for HDV RNA, available in specialized laboratories and yet to be standardized, can be used to confirm HDV infection and to monitor treatment during chronic infection.

The serologic/virologic course of events during acute hepatitis E is entirely analogous to that of acute hepatitis A, with brief fecal shedding of virus and viremia and an early IgM anti-HEV response that predominates during approximately the first 3 months, but is eclipsed thereafter by long-lasting IgG anti-HEV. Diagnostic tests of varying reliability for hepatitis E are commercially available but used routinely primarily outside the United States; in the United States, diagnostic serologic/virologic assays can be performed at the Centers for Disease Control and Prevention or other specialized reference laboratories.

Liver biopsy is rarely necessary or indicated in acute viral hepatitis, except when the diagnosis is questionable or when clinical evidence suggests a diagnosis of chronic hepatitis.

A diagnostic algorithm can be applied in the evaluation of cases of acute viral hepatitis. A patient with acute hepatitis should undergo four serologic tests: HBsAg, IgM anti-HAV, IgM anti-HBc, and anti-HCV **(Table 332-6)**. The presence of HBsAg, with or without IgM anti-HBc, represents HBV infection. If IgM anti-HBc is present, the HBV infection is considered acute; if IgM anti-HBc is absent, the HBV infection is considered chronic. A diagnosis of acute hepatitis B can be made in the absence of HBsAg when IgM anti-HBc is detectable. A diagnosis of acute hepatitis A is based on the presence of IgM anti-HAV. If IgM anti-HAV coexists with HBsAg, a diagnosis of simultaneous HAV and HBV infections can be made; if IgM anti-HBc (with or without HBsAg) is detectable, the patient has simultaneous acute hepatitis A and B, and if IgM anti-HBc is undetectable, the patient has acute hepatitis A superimposed on chronic HBV infection. The presence of anti-HCV supports a diagnosis of acute hepatitis C. Occasionally, testing for HCV RNA or repeat anti-HCV testing later during the illness is necessary to establish the diagnosis. Absence of all serologic markers is consistent with a diagnosis of "non-A, non-B, non-C" hepatitis (no other proven human hepatitis viruses have been identified), if the epidemiologic setting is appropriate.

In patients with chronic hepatitis, initial testing should consist of HBsAg and anti-HCV. Anti-HCV supports and HCV RNA testing establishes the diagnosis of chronic hepatitis C. If a serologic diagnosis of chronic hepatitis B is made, testing for HBeAg and anti-HBe is indicated to evaluate relative infectivity. Testing for HBV DNA in such patients provides a more quantitative and sensitive measure of the level of virus replication, and therefore is very helpful during antiviral therapy **(Chap. 334)**. In patients with chronic hepatitis B and normal aminotransferase activity in the absence of HBeAg, serial testing over time is often required to distinguish between inactive carriage and HBeAg-negative chronic hepatitis B with fluctuating virologic and necroinflammatory activity. In persons with hepatitis B, testing for anti-HDV is useful in those with severe and fulminant disease, with severe chronic disease, with chronic hepatitis B and acute hepatitis-like exacerbations, with frequent percutaneous exposures, and from areas where HDV infection is endemic.

TABLE 332-6 Simplified Diagnostic Approach in Patients Presenting with Acute Hepatitis

SEROLOGIC TESTS OF PATIENT'S SERUM				
HBsAg	IgM ANTI-HAV	IgM ANTI-HBc	ANTI-HCV	DIAGNOSTIC INTERPRETATION
+	–	+	–	Acute hepatitis B
+	–	–	–	Chronic hepatitis B
+	+	–	–	Acute hepatitis A superimposed on chronic hepatitis B
+	+	+	–	Acute hepatitis A and B
–	+	–	–	Acute hepatitis A
–	+	+	–	Acute hepatitis A and B (HBsAg below detection threshold)
–	–	+	–	Acute hepatitis B (HBsAg below detection threshold)
–	–	–	+	Acute hepatitis C

Note: See text for abbreviations.

PROGNOSIS

Virtually all previously healthy patients with hepatitis A recover completely with no clinical sequelae. Similarly, in acute hepatitis B, 95–99% of previously healthy adults have a favorable course and recover completely. Certain clinical and laboratory features, however, suggest a more complicated and protracted course. Patients of advanced age and with serious underlying medical disorders may have a prolonged course and are more likely to experience severe hepatitis. Initial presenting features such as ascites, peripheral edema, and symptoms of hepatic encephalopathy suggest a poorer prognosis. In addition, a prolonged PT, low serum albumin level, hypoglycemia, and very high serum bilirubin values suggest severe hepatocellular disease. Patients with these clinical and laboratory features deserve prompt hospital admission. The case fatality rate in hepatitis A and B is very low (~0.1%) but is increased by advanced age and underlying debilitating disorders. Among patients ill enough to be hospitalized for acute hepatitis B, the fatality rate is 1%. Hepatitis C is less severe during the acute phase than hepatitis B and is more likely to be anicteric; fatalities are rare, but the precise case fatality rate is not known. In outbreaks of waterborne hepatitis E in India and Asia, the case fatality rate is 1–2% and up to 10–20% in pregnant women. Contributing to fulminant hepatitis E in endemic countries (but only very rarely or not at all in nonendemic countries) are instances of acute hepatitis E superimposed on underlying chronic liver disease ("acute-on-chronic" liver disease). Patients with simultaneous acute hepatitis B and D do not necessarily experience a higher mortality rate than do patients with acute hepatitis B alone; however, in several outbreaks of acute simultaneous HBV and HDV infection among injection drug users, the case fatality rate was ~5%. When HDV superinfection occurs in a person with chronic hepatitis B, the likelihood of fulminant hepatitis and death is increased substantially. Although the case fatality rate for hepatitis D is not known definitively, in outbreaks of severe HDV superinfection in isolated populations with a high hepatitis B carrier rate, a mortality rate >20% has been recorded.

COMPLICATIONS AND SEQUELAE

A small proportion of patients with hepatitis A experience *relapsing hepatitis* weeks to months after apparent recovery from acute hepatitis. Relapses are characterized by recurrence of symptoms, aminotransferase elevations, occasional jaundice, and

fecal excretion of HAV. Another unusual variant of acute hepatitis A is *cholestatic hepatitis*, characterized by protracted cholestatic jaundice and pruritus. Rarely, liver test abnormalities persist for many months, even up to 1 year. Even when these complications occur, hepatitis A remains self-limited and does not progress to chronic liver disease. During the prodromal phase of acute hepatitis B, a serum sickness–like syndrome characterized by arthralgia or arthritis, rash, angioedema, and, rarely, hematuria and proteinuria may develop in 5–10% of patients. This syndrome occurs before the onset of clinical jaundice, and these patients are often diagnosed erroneously as having rheumatologic diseases. The diagnosis can be established by measuring serum aminotransferase levels, which are almost invariably elevated, and serum HBsAg. As noted above, EMC is an immune-complex disease that can complicate chronic hepatitis C and is part of a spectrum of B cell lymphoproliferative disorders, which, in rare instances, can evolve to B cell lymphoma **(Chap. 104)**. Attention has been drawn as well to associations between hepatitis C and such cutaneous disorders as porphyria cutanea tarda and lichen planus. A mechanism for these associations is unknown. Finally, related to the reliance of HCV on lipoprotein secretion and assembly pathways and on interactions of HCV with glucose metabolism, HCV infection may be complicated by hepatic steatosis, hypercholesterolemia, insulin resistance (and other manifestations of the metabolic syndrome), and type 2 diabetes mellitus; both hepatic steatosis and insulin resistance appear to accelerate hepatic fibrosis and blunt responsiveness to interferon-based antiviral therapy **(Chap. 334)**.

The most feared complication of viral hepatitis is *fulminant hepatitis* (massive hepatic necrosis); fortunately, this is a rare event. Fulminant hepatitis is seen primarily in hepatitis B, D, and E, but rare fulminant cases of hepatitis A occur primarily in older adults and in persons with underlying chronic liver disease, including, according to some reports, chronic hepatitis B and C. Hepatitis B accounts for >50% of fulminant cases of viral hepatitis, a sizable proportion of which are associated with HDV infection and another proportion with underlying chronic hepatitis C. Fulminant hepatitis is hardly ever seen in hepatitis C, but hepatitis E, as noted above, can be complicated by fatal fulminant hepatitis in 1–2% of all cases and in up to 20% of cases in pregnant women. Patients usually present with signs and symptoms of encephalopathy that may evolve to deep coma. The liver is usually small and the PT excessively prolonged. The combination of rapidly shrinking liver size, rapidly rising bilirubin level, and marked prolongation of the PT, even as aminotransferase levels fall, together with clinical signs of confusion, disorientation, somnolence, ascites, and edema, indicates that the patient has hepatic failure with encephalopathy. Cerebral edema is common; brainstem compression, gastrointestinal bleeding, sepsis, respiratory failure, cardiovascular collapse, and renal failure are terminal events. The mortality rate is exceedingly high (>80% in patients with deep coma), but patients who survive may have a complete biochemical and histologic recovery. If a donor liver can be located in time, liver transplantation may be lifesaving in patients with fulminant hepatitis **(Chap. 338)**.

Documenting the disappearance of HBsAg after apparent clinical recovery from acute hepatitis B is particularly important. Before laboratory methods were available to distinguish between acute hepatitis and acute hepatitis-like exacerbations (*spontaneous reactivations*) of chronic hepatitis B, observations suggested that ~10% of previously healthy patients remained HBsAg-positive for >6 months after the onset of clinically apparent acute hepatitis B. One-half of these persons cleared the antigen from their circulations during the next several years, but the other 5% remained chronically HBsAg-positive. More recent observations suggest that the true rate of chronic infection after clinically apparent acute hepatitis B is as low as 1% in normal, immunocompetent, young adults. Earlier, higher estimates may have been confounded by inadvertent inclusion of acute exacerbations in chronically infected patients; these patients, chronically HBsAg-positive before exacerbation, were unlikely to seroconvert to HBsAg-negative thereafter. Whether the rate of chronicity is 10% or 1%, such patients have IgG anti-HBc in serum; anti-HBs is either undetected or detected at low titer against the opposite subtype specificity of the antigen (see "Laboratory Features"). These patients may (1) be inactive carriers; (2) have low-grade, mild chronic hepatitis; or (3) have moderate to severe chronic hepatitis with or without cirrhosis. The likelihood of remaining chronically infected after acute HBV infection is especially high among neonates, persons with Down's syndrome, chronically hemodialyzed patients, and immunosuppressed patients, including persons with HIV infection.

Chronic hepatitis is an important late complication of acute hepatitis B occurring in a small proportion of patients with acute disease but more common in those who present with chronic infection without having experienced an acute illness, as occurs typically after neonatal infection or after infection in an immunosuppressed host **(Chap. 334)**. The following clinical and laboratory features suggest progression of acute hepatitis to chronic hepatitis: (1) lack of complete resolution of clinical symptoms of anorexia, weight loss, fatigue, and the persistence of hepatomegaly; (2) the presence of bridging/interface or multilobular hepatic necrosis on liver biopsy during protracted, severe acute viral hepatitis; (3) failure of the serum aminotransferase, bilirubin, and globulin levels to return to normal within 6–12 months after the acute illness; and (4) the persistence of HBeAg for >3 months or HBsAg for >6 months after acute hepatitis.

Although acute hepatitis D infection does not increase the likelihood of chronicity of simultaneous acute hepatitis B, hepatitis D has the potential for contributing to the severity of chronic hepatitis B. Hepatitis D superinfection can transform inactive or mild chronic hepatitis B into severe, progressive chronic hepatitis and cirrhosis; it also can accelerate the course of chronic hepatitis B. Some HDV superinfections in patients with chronic hepatitis B lead to fulminant hepatitis. As defined in longitudinal studies over three decades, the annual rate of cirrhosis in patients with chronic hepatitis D is 4%. Although HDV and HBV infections are associated with severe liver disease, mild hepatitis and even inactive carriage have been identified in some patients, and the disease may become indolent beyond the early years of infection.

After acute HCV infection, the likelihood of remaining chronically *infected* approaches 85–90%. Although many patients with chronic hepatitis C have no symptoms, cirrhosis may develop in as many as 20% within 10–20 years of acute illness; in some series of cases reported by referral centers, cirrhosis has been reported in as many as 50% of patients with chronic hepatitis C. Among cirrhotic patients with chronic hepatitis C, the annual risk of hepatic decompensation is ~4%. Although chronic hepatitis C accounts for at least 40% of cases of chronic liver disease and of patients undergoing liver transplantation for end-stage liver disease in the United States and Europe, in the majority of patients with chronic hepatitis C, morbidity and mortality are limited during the initial 20 years after the onset of infection. Progression of chronic hepatitis C may be influenced by advanced age of acquisition, long duration of infection, immunosuppression, coexisting excessive alcohol use, concomitant hepatic steatosis, other hepatitis virus infection, or HIV co-infection. In fact, instances of severe and rapidly progressive chronic hepatitis B and C are being recognized with increasing frequency in patients with HIV infection **(Chap. 197)**. In contrast, neither HAV nor HEV causes chronic liver disease in immunocompetent hosts; however, cases of chronic hepatitis E (including cirrhosis and end-stage liver disease) have been observed in immunosuppressed organ-transplant recipients, persons receiving cytotoxic chemotherapy, and persons with HIV infection. Among patients with chronic hepatitis (e.g., caused by hepatitis B or C, alcohol, etc.) in endemic countries, hepatitis E has been reported as the cause of acute-on-chronic liver failure; however, in most experiences among patients from nonendemic countries, HEV has not been found to contribute to hepatic decompensation in patients with chronic hepatitis.

Persons with chronic hepatitis B, particularly those infected in infancy or early childhood and especially those with HBeAg and/or high-level HBV DNA, have an enhanced risk of hepatocellular carcinoma. The risks of cirrhosis and hepatocellular carcinoma increase with the level of HBV replication. The annual rate of hepatocellular carcinoma in patients with chronic hepatitis D and cirrhosis is ~3%. The risk of hepatocellular carcinoma is increased as well in patients with chronic hepatitis C, almost exclusively in patients with cirrhosis, and almost always after at least several decades, usually after three decades

of disease **(Chap. 78)**. Among such cirrhotic patients with chronic hepatitis C, the annual risk of hepatocellular carcinoma is ~1–4%.

Rare complications of viral hepatitis include pancreatitis, myocarditis, atypical pneumonia, aplastic anemia, transverse myelitis, and peripheral neuropathy. In children, hepatitis B may present rarely with anicteric hepatitis, a nonpruritic papular rash of the face, buttocks, and limbs, and lymphadenopathy (papular acrodermatitis of childhood or Gianotti-Crosti syndrome).

Rarely, autoimmune hepatitis **(Chap. 334)** can be triggered by a bout of otherwise self-limited acute hepatitis, as reported after acute hepatitis A, B, and C.

DIFFERENTIAL DIAGNOSIS

Viral diseases such as infectious mononucleosis; those due to cytomegalovirus, herpes simplex, and coxsackieviruses; and toxoplasmosis may share certain clinical features with viral hepatitis and cause elevations in serum aminotransferase and, less commonly, in serum bilirubin levels. Tests such as the differential heterophile and serologic tests for these agents may be helpful in the differential diagnosis if HBsAg, anti-HBc, IgM anti-HAV, and anti-HCV determinations are negative. Aminotransferase elevations can accompany almost any systemic viral infection; other rare causes of liver injury confused with viral hepatitis are infections with *Leptospira*, *Candida*, *Brucella*, *Mycobacteria*, and *Pneumocystis*. A complete drug history is particularly important because many drugs and certain anesthetic agents can produce a picture of either acute hepatitis or cholestasis **(Chap. 333)**. Equally important is a past history of unexplained "repeated episodes" of acute hepatitis. This history should alert the physician to the possibility that the underlying disorder is chronic hepatitis, for example autoimmune hepatitis **(Chap. 334)**. Alcoholic hepatitis must also be considered, but usually the serum aminotransferase levels are not as markedly elevated, and other stigmata of alcoholism may be present. The finding on liver biopsy of fatty infiltration, a neutrophilic inflammatory reaction, and "alcoholic hyaline" would be consistent with alcohol-induced rather than viral liver injury. Because acute hepatitis may present with right upper quadrant abdominal pain, nausea and vomiting, fever, and icterus, it is often confused with acute cholecystitis, common duct stone, or ascending cholangitis. Patients with acute viral hepatitis may tolerate surgery poorly; therefore, it is important to exclude this diagnosis, and in confusing cases, a percutaneous liver biopsy may be necessary before laparotomy. Viral hepatitis in the elderly is often misdiagnosed as obstructive jaundice resulting from a common duct stone or carcinoma of the pancreas. Because acute hepatitis in the elderly may be quite severe and the operative mortality high, a thorough evaluation including biochemical tests, radiographic studies of the biliary tree, and even liver biopsy may be necessary to exclude primary parenchymal liver disease. Another clinical constellation that may mimic acute hepatitis is right ventricular failure with passive hepatic congestion or hypoperfusion syndromes, such as those associated with shock, severe hypotension, and severe left ventricular failure. Also included in this general category is any disorder that interferes with venous return to the heart, such as right atrial myxoma, constrictive pericarditis, hepatic vein occlusion (Budd-Chiari syndrome), or venoocclusive disease. Clinical features are usually sufficient to distinguish among these vascular disorders and viral hepatitis. Acute fatty liver of pregnancy, cholestasis of pregnancy, eclampsia, and the HELLP (*h*emolysis, *e*levated *l*iver tests, and *l*ow *p*latelets) syndrome can be confused with viral hepatitis during pregnancy. Very rarely, malignancies metastatic to the liver can mimic acute or even fulminant viral hepatitis. Occasionally, genetic or metabolic liver disorders (e.g., Wilson's disease, α_1 antitrypsin deficiency) and nonalcoholic fatty liver disease are confused with acute viral hepatitis.

TREATMENT

Acute Viral Hepatitis

Most persons with acute hepatitis (especially hepatitis A, B, and E) recover spontaneously and do not require specific antiviral therapy. In hepatitis B, among previously healthy adults who present with clinically apparent acute hepatitis, recovery occurs in ~99%; therefore, antiviral therapy is not likely to improve the rate of recovery and is not required. In rare instances of severe acute hepatitis B, treatment with a nucleoside analogue at oral doses used to treat chronic hepatitis B **(Chap. 334)** has been attempted successfully. Although clinical trials have not been done to establish the efficacy or duration of this approach, most authorities would recommend institution of antiviral therapy with a nucleoside analogue (entecavir or tenofovir, the most potent and least resistance-prone agents) for severe, but not mild–moderate, acute hepatitis B. Treatment should continue until 3 months after HBsAg seroconversion or 6 months after HBeAg seroconversion.

In typical cases of acute hepatitis C, recovery is rare (~15–20% in most experiences), progression to chronic hepatitis is the rule, and small clinical trials during the era of interferon-based regimens suggested that antiviral therapy with courses (usually 24 weeks) of standard or pegylated interferon α monotherapy reduced the rate of chronicity considerably by inducing sustained responses in 30–70% of patients (according to a meta-analyses of published studies) and in up to 98% in a small German multicenter study (treatment initiated an average of 3 months after infection). In the current interferon-free therapy era, as of 2016, six different all-oral, brief-duration (most lasting 12 weeks), very well-tolerated, highly effective (sustained virologic response rates exceeding 90–95%) combination regimens (of polymerase inhibitors, protease inhibitors, and/or NS5A inhibitors) are available to treat patients with chronic hepatitis C **(see Chap. 334)**; the same regimens are available and recommended to treat patients with acute hepatitis C. Although the duration of therapy for acute hepatitis C has not been determined definitively, in a study of 20 patients, acute hepatitis C resolved after treatment lasting only 6 weeks. In 2016, the European Association for the Study of the Liver (EASL) recommended 8 weeks of treatment for acute hepatitis C with a genotype-appropriate **(see Chap. 334)** direct-acting antiviral regimen consisting of sofosbuvir plus one of the three approved NS5A inhibitors without ribavirin (12 weeks for patients with acute hepatitis C and either a baseline HCV RNA level >1 million IU/mL or HIV co-infection).

Because spontaneous recovery can occur and because most cases of acute hepatitis C are not clinically severe or rapidly progressive, delaying antiviral therapy of acute hepatitis C for at least 12–16 weeks and even up to 6 months (after which recovery is unlikely) is a recommended approach. Patients with jaundice, those with HCV genotype 1, women, earlier age of infection, lower level of HCV RNA, HBV co-infection, and absence of current injection-drug use are more likely to recover from acute hepatitis C, as are persons who have genetic markers associated with spontaneous recovery (*IL28B* CC haplotype). Because of the marked reduction over the past three decades in the frequency of acute hepatitis C, opportunities to identify and treat patients with acute hepatitis C are rare, except in two population subsets: (1) In health workers who sustain hepatitis C–contaminated needle sticks (occupational accidents), monitoring for ALT elevations and the presence of HCV RNA identify acute hepatitis C in ~3%, and this group should be treated; (2) in injection-drug users, the risk of acute hepatitis C has been on the rise, and the epidemic of opioid use has contributed to an amplification of HCV infection among drug users. Such patients are candidates for antiviral therapy, and efforts to combine antiviral therapy with drug-rehabilitation therapy have been very successful.

Notwithstanding these specific therapeutic considerations, in most cases of typical acute viral hepatitis, specific treatment generally is not necessary. Although hospitalization may be required for clinically severe illness, most patients do not require hospital care. Forced and prolonged bed rest is not essential for full recovery, but many patients will feel better with restricted physical activity. A high-calorie diet is desirable, and because many patients may experience nausea late in the day, the major caloric intake is best tolerated in the morning. Intravenous feeding is necessary in the acute stage if the patient has persistent vomiting and cannot maintain oral intake. Drugs capable of producing adverse reactions such as

cholestasis and drugs metabolized by the liver should be avoided. If severe pruritus is present, the use of the bile salt-sequestering resin cholestyramine is helpful. Glucocorticoid therapy has no value in acute viral hepatitis, even in severe cases, and may be deleterious, even increasing the risk of chronicity (e.g., of acute hepatitis B).

Physical isolation of patients with hepatitis to a single room and bathroom is rarely necessary except in the case of fecal incontinence for hepatitis A and E or uncontrolled, voluminous bleeding for hepatitis B (with or without concomitant hepatitis D) and C. Because most patients hospitalized with hepatitis A excrete little, if any, HAV, the likelihood of HAV transmission from these patients during their hospitalization is low. Therefore, burdensome *enteric precautions are no longer recommended*. Although gloves should be worn when the bed pans or fecal material of patients with hepatitis A are handled, these precautions do not represent a departure from sensible procedure and contemporary universal precautions for all hospitalized patients. For patients with hepatitis B and C, emphasis should be placed on blood precautions (i.e., avoiding direct, ungloved hand contact with blood and other body fluids). Enteric precautions are unnecessary. The importance of simple hygienic precautions such as hand washing cannot be overemphasized. Universal precautions that have been adopted for all patients apply to patients with viral hepatitis. Hospitalized patients may be discharged following substantial symptomatic improvement, a significant downward trend in the serum aminotransferase and bilirubin values, and a return to normal of the PT. Mild aminotransferase elevations should not be considered contraindications to the gradual resumption of normal activity.

In *fulminant hepatitis*, the goal of therapy is to support the patient by maintenance of fluid balance, support of circulation and respiration, control of bleeding, correction of hypoglycemia, and treatment of other complications of the comatose state in anticipation of liver regeneration and repair. Protein intake should be restricted, and oral lactulose administered. Glucocorticoid therapy has been shown in controlled trials to be ineffective. Likewise, exchange transfusion, plasmapheresis, human cross-circulation, porcine liver cross-perfusion, hemoperfusion, and extracorporeal liver-assist devices have not been proven to enhance survival. Meticulous intensive care that includes prophylactic antibiotic coverage is the one factor that appears to improve survival. Orthotopic liver transplantation is resorted to with increasing frequency, with excellent results, in patients with fulminant hepatitis **(Chap. 338)**. Fulminant hepatitis C is very rare; however, in fulminant hepatitis B, oral antiviral therapy has been used successfully, as reported anecdotally. In clinically severe hepatitis E (with jaundice and coagulopathy), successful therapy with ribavirin (600 mg twice daily, 15 mg/kg) has been reported anecdotally. Unfortunately, when fulminant hepatitis E occurs in pregnant women (as it does in up to 20% of pregnant women with acute hepatitis E), ribavirin, which is teratogenic, is contraindicated.

PROPHYLAXIS

Because application of therapy for acute viral hepatitis is limited and because antiviral therapy for chronic viral hepatitis is cumbersome, costly, and not effective in all patients **(Chap. 334)**, emphasis is placed on prevention through immunization. The prophylactic approach differs for each of the types of viral hepatitis. In the past, immunoprophylaxis relied exclusively on passive immunization with antibody-containing globulin preparations purified by cold ethanol fractionation from the plasma of hundreds of normal donors. Currently, for hepatitis A, B, and E, active immunization with vaccines is the preferable approach to prevention.

Hepatitis A Both passive immunization with IG and active immunization with killed vaccines are available. All preparations of IG contain anti-HAV concentrations sufficient to be protective. When administered before exposure or during the early incubation period, IG is effective in preventing clinically apparent hepatitis A. For postexposure prophylaxis of intimate contacts (household, sexual, institutional) of persons with hepatitis A, the administration of 0.02 mL/kg is recommended as early after exposure as possible; it may be effective even when administered as late as 2 weeks after exposure. Prophylaxis is not necessary for those who have already received hepatitis A vaccine, for casual contacts (office, factory, school, or hospital), for most elderly persons, who are very likely to be immune, or for those known to have anti-HAV in their serum. In day care centers, recognition of hepatitis A in children or staff should provide a stimulus for immunoprophylaxis in the center and in the children's family members. By the time most common-source outbreaks of hepatitis A are recognized, it is usually too late in the incubation period for IG to be effective; however, prophylaxis may limit the frequency of secondary cases. For travelers to tropical countries, developing countries, and other areas outside standard tourist routes, IG prophylaxis had been recommended before a vaccine became available. When such travel lasted <3 months, 0.02 mL/kg was given; for longer travel or residence in these areas, a dose of 0.06 mL/kg every 4–6 months was recommended. Administration of plasma-derived globulin is safe; all contemporary lots of IG are subjected to viral inactivation steps and must be free of HCV RNA as determined by PCR testing. Administration of IM lots of IG has not been associated with transmission of HBV, HCV, or HIV.

Formalin-inactivated vaccines made from strains of HAV attenuated in tissue culture have been shown to be safe, immunogenic, and effective in preventing hepatitis A. Hepatitis A vaccines are approved for use in persons who are at least 1 year old and appear to provide adequate protection beginning 4 weeks after a primary inoculation. If it can be given within 4 weeks of an expected exposure, such as by travel to an endemic area, hepatitis A vaccine is the preferred approach to *preexposure* immunoprophylaxis. If travel is more imminent, IG (0.02 mL/kg) should be administered at a different injection site, along with the first dose of vaccine. Because vaccination provides long-lasting protection (protective levels of anti-HAV should last 20 years after vaccination), persons whose risk will be sustained (e.g., frequent travelers or those remaining in endemic areas for prolonged periods) should be vaccinated, and vaccine should supplant the need for repeated IG injections. Shortly after its introduction, hepatitis A vaccine was recommended for children living in communities with a high incidence of HAV infection; in 1999, this recommendation was extended to include all children living in states, counties, and communities with high rates of HAV infection. As of 2006, the Advisory Committee on Immunization Practices of the U.S. Public Health Service recommended *routine hepatitis A vaccination of all children*. Other groups considered being at increased risk for HAV infection and who are candidates for hepatitis A vaccination include military personnel, populations with cyclic outbreaks of hepatitis A (e.g., Alaskan natives), employees of day care centers, primate handlers, laboratory workers exposed to hepatitis A or fecal specimens, and patients with chronic liver disease. Because of an increased risk of fulminant hepatitis A—observed in some experiences but not confirmed in others—among patients with chronic hepatitis C, patients with chronic hepatitis C are candidates for hepatitis A vaccination, as are persons with chronic hepatitis B. Other populations whose recognized risk of hepatitis A is increased should be vaccinated, including men who have sex with men, injection drug users, persons with clotting disorders who require frequent administration of clotting-factor concentrates, persons traveling from the United States to countries with high or intermediate hepatitis A endemicity, postexposure prophylaxis for contacts of persons with hepatitis A, and household members and other close contacts of adopted children arriving from countries with high and moderate hepatitis A endemicity. Recommendations for dose and frequency differ for the two approved vaccine preparations **(Table 332-7)**; all injections are IM. Hepatitis A vaccine has been reported to be effective in preventing secondary household and day care center–associated cases of acute hepatitis A. Because the vaccine provides long-lasting protection and is simpler to use, in 2006, the Immunization Practices Advisory Committee of the U.S. Public Health Service favored hepatitis A vaccine to IG for postexposure prophylaxis of healthy persons age 2–40 years; for younger or older persons, for immunosuppressed patients, and for patients with chronic liver disease, IG should continue to be used. In the United States, reported mortality resulting from hepatitis A

TABLE 332-7 Hepatitis A Vaccination Schedules

AGE, YEARS	NO. OF DOSES	DOSE	SCHEDULE, MONTHS
HAVRIX (GlaxoSmithKline)[a]			
1–18	2	720 ELU[b] (0.5 mL)	0, 6–12
≥19	2	1440 ELU (1 mL)	0, 6–12
VAQTA (Merck)			
1–18	2	25 units (0.5 mL)	0, 6–18
≥19	2	50 units (1 mL)	0, 6–18

[a]A combination of this hepatitis A vaccine and hepatitis B vaccine, TWINRIX, is licensed for simultaneous protection against both of these viruses among adults (age ≥18 years). Each 1-mL dose contains 720 ELU of hepatitis A vaccine and 20 μg of hepatitis B vaccine. These doses are recommended at months 0, 1, and 6. [b]Enzyme-linked immunoassay units.

declined in parallel with hepatitis A vaccine–associated reductions in the annual incidence of new infections.

Hepatitis B Until 1982, prevention of hepatitis B was based on *passive* immunoprophylaxis either with standard immunoglobulin, containing modest levels of anti-HBs, or hepatitis B immunoglobulin (HBIG), containing high-titer anti-HBs. The efficacy of standard IG has never been established and remains questionable; even the efficacy of HBIG, demonstrated in several clinical trials, has been challenged, and its contribution appears to be in reducing the frequency of clinical *illness*, not in preventing *infection*. The first vaccine for *active* immunization, introduced in 1982, was prepared from purified, noninfectious 22-nm spherical HBsAg particles derived from the plasma of healthy HBsAg carriers. In 1987, the plasma-derived vaccine was supplanted by a genetically engineered vaccine derived from recombinant yeast. The latter vaccine consists of HBsAg particles that are nonglycosylated but are otherwise indistinguishable from natural HBsAg; two recombinant vaccines are licensed for use in the United States. Current recommendations can be divided into those for preexposure and postexposure prophylaxis.

For *preexposure* prophylaxis against hepatitis B in settings of frequent exposure (health workers exposed to blood; first-responder public safety workers; hemodialysis patients and staff; residents and staff of custodial institutions for the developmentally handicapped; injection drug users; inmates of long-term correctional facilities; persons with multiple sexual partners or who have had a sexually transmitted disease; men who have sex with men; persons such as hemophiliacs who require long-term, high-volume therapy with blood derivatives; household and sexual contacts of persons with chronic HBV infection; persons living in or traveling extensively in endemic areas; unvaccinated children aged <18; unvaccinated children who are Alaskan natives, Pacific Islanders, or residents in households of first-generation immigrants from endemic countries; persons born in countries with a prevalence of HBV infection ≥2%; patients with chronic liver disease; persons <age 60 with diabetes mellitus [those ≥60 at the discretion of their physicians]; persons with end-stage renal disease; and persons with HIV infection), three IM (deltoid, not gluteal) injections of hepatitis B vaccine are recommended at 0, 1, and 6 months (other, optional schedules are summarized in Table 332-8). Pregnancy is *not* a contraindication to vaccination. In areas of low HBV endemicity such as the United States, despite the availability of safe and effective hepatitis B vaccines, a strategy of vaccinating persons in high-risk groups was not effective. The incidence of new hepatitis B cases continued to increase in the United States after the introduction of vaccines; <10% of all targeted persons in high-risk groups were actually vaccinated, and ~30% of persons with sporadic acute hepatitis B did not fall into any high-risk-group category. Therefore, to have an impact on the frequency of HBV infection in an area of low endemicity such as the United States, universal hepatitis B vaccination in childhood has been recommended. For unvaccinated children born after the implementation of universal infant vaccination, vaccination during early adolescence, at age 11–12 years, was recommended, and this recommendation has been extended to include all unvaccinated children age 0–19 years. In HBV-hyperendemic areas (e.g., Asia), universal vaccination of children has resulted in a marked (~70–90%) 30-year decline in complications of hepatitis B, including liver-related mortality and hepatocellular carcinoma.

The two available recombinant hepatitis B vaccines are comparable, one containing 10 μg of HBsAg (Recombivax-HB) and the other containing 20 μg of HBsAg (Engerix-B), and recommended doses for each injection vary for the two preparations (Table 332-8). Combinations of hepatitis B vaccine with other childhood vaccines are available as well (Table 332-8).

For unvaccinated persons sustaining an exposure to HBV, *postexposure* prophylaxis with a combination of HBIG (for rapid achievement of high-titer circulating anti-HBs) and hepatitis B vaccine (for achievement of long-lasting immunity as well as its apparent efficacy in attenuating clinical illness after exposure) is recommended. For *perinatal* exposure of infants born to HBsAg-positive mothers, a single dose of HBIG, 0.5 mL, should be administered IM in the thigh *immediately after birth*, followed by a complete course of three injections of recombinant hepatitis B vaccine (see doses above) to be started within the first 12 h of life. For those experiencing a direct percutaneous inoculation or transmucosal exposure to HBsAg-positive blood or body fluids (e.g., accidental *needle stick*, other mucosal penetration, or ingestion), a single IM dose of HBIG, 0.06 mL/kg, administered as soon after exposure as possible, is followed by a complete course of hepatitis B vaccine to begin within the first week. For pregnant mothers with high-level HBV DNA (>2×10^5 IU/mL), adding antiviral nucleoside analogues (e.g., pregnancy class B tenofovir, see Chap 334) during the third trimester of pregnancy reduces perinatal transmission even further. For persons exposed by *sexual* contact to a patient with acute hepatitis B, a single IM dose of HBIG, 0.06 mL/kg, should be given within 14 days of exposure,

TABLE 332-8 Preexposure Hepatitis B Vaccination Schedules

TARGET GROUP	NO. OF DOSES	DOSE	SCHEDULE, MONTHS
RECOMBIVAX-HB (Merck)[a]			
Infants, children (<1–10 years)	3	5 μg (0.5 mL)	0, 1–2, 4–6
Adolescents (11–19 years)	3 or 4	5 μg (0.5 mL)	0–2, 1–4, 4–6 *or* 0, 12, 24 *or* 0, 1, 2, 12
	Or		
Adults (≥20 years)	2	10 μg (1 mL)	0, 4–6 (age 11–15)
Hemodialysis patients[b]	3	10 μg (1 mL)	0–2, 1–4, 4–6
<20 years	3	5 μg (0.5 mL)	0, 1, 6
≥20 years	3	40 μg (4 mL)	0, 1, 6
ENGERIX-B (GlaxoSmithKline)[c]			
Infants, children (<1–10 years)	3 or 4	10 μg (0.5 mL)	0, 1–2, 4–6 *or* 0, 1, 2, 12
Adolescents (10–19 years)	3 or 4	10 μg (0.5 mL)	0, 1–2, 4–6 *or* 0, 12, 24 *or* 0, 1, 2, 12
Adults (≥20 years)	3 or 4	20 μg (1 mL)	0–2, 1–4, 4–6 or 0, 1, 2, 12
Hemodialysis patients[b]			
<20 years	4	10 μg (0.5 mL)	0, 1, 2, 6
≥20 years	4	40 μg (2 mL)	0, 1, 2, 6

[a]This manufacturer produces a licensed combination of hepatitis B vaccine and vaccines against *Haemophilus influenzae* type b and *Neisseria meningitides*, Comvax, for use in infants and young children. Please consult product insert for dose and schedule. [b]This group also includes other immunocompromised persons. [c]This manufacturer produces two licensed combination hepatitis B vaccines: (1) Twinrix, recombinant hepatitis B vaccine plus inactivated hepatitis A vaccine, is licensed for simultaneous protection against both of these viruses among adults (age ≥18 years). Each 1-mL dose contains 720 ELU of hepatitis A vaccine and 20 μg of hepatitis B vaccine. These doses are recommended at months 0, 1, and 6. (2) Pediarix, recombinant hepatitis B vaccine plus diphtheria and tetanus toxoid, pertussis, and inactivated poliovirus, is licensed for use in infants and young children. Please consult product insert for doses and schedules.

to be followed by a complete course of hepatitis B vaccine. When both HBIG and hepatitis B vaccine are recommended, they may be given at the same time but at separate sites. Testing adults for anti-HBs after a course of vaccine is advisable to document the acquisition of immunity, but, because hepatitis B vaccine immunogenicity is nearly universal in infants, postvaccination anti-HBs testing of children is not recommended.

The precise duration of protection afforded by hepatitis B vaccine is unknown; however, ~80–90% of immunocompetent adult vaccinees retain protective levels of anti-HBs for at least 5 years, and 60–80% for 10 years, and protective antibody has been documented to last for at least two decades after vaccination in infancy. Thereafter and even after anti-HBs becomes undetectable, protection persists against clinical hepatitis B, hepatitis B surface antigenemia, and chronic HBV infection. Currently, *booster* immunizations are not recommended routinely, except in immunosuppressed persons who have lost detectable anti-HBs or immunocompetent persons who sustain percutaneous HBsAg-positive inoculations after losing detectable antibody. Specifically, for hemodialysis patients, annual anti-HBs testing is recommended after vaccination; booster doses are recommended when anti-HBs levels fall to <10 mIU/mL. As noted above, for persons at risk of both hepatitis A and B, a combined vaccine is available containing 720 enzyme-linked immunoassay units (ELUs) of inactivated HAV and 20 μg of recombinant HBsAg (at 0, 1, and 6 months).

Hepatitis D Infection with hepatitis D can be prevented by vaccinating susceptible persons with hepatitis B vaccine. No product is available for immunoprophylaxis to prevent HDV superinfection in persons with chronic HBV infection; for them, avoidance of percutaneous exposures and limitation of intimate contact with persons who have HDV infection are recommended.

Hepatitis C IG is ineffective in preventing hepatitis C and is no longer recommended for postexposure prophylaxis in cases of perinatal, needle stick, or sexual exposure. Although prototype vaccines that induce antibodies to HCV envelope proteins have been developed, currently, hepatitis C vaccination is not feasible practically. Genotype and quasispecies viral heterogeneity, as well as rapid evasion of neutralizing antibodies by this rapidly mutating virus, conspire to render HCV a difficult target for immunoprophylaxis with a vaccine. Prevention of transfusion-associated hepatitis C has been accomplished by the following successively introduced measures: exclusion of commercial blood donors and reliance on a volunteer blood supply; screening donor blood with surrogate markers such as ALT (no longer recommended) and anti-HBc, markers that identify segments of the blood donor population with an increased risk of bloodborne infections; exclusion of blood donors in high-risk groups for AIDS and the introduction of anti-HIV screening tests; and progressively sensitive serologic and virologic screening tests for HCV infection.

In the absence of active or passive immunization, prevention of hepatitis C includes behavior changes and precautions to limit exposures to infected persons. Recommendations designed to identify patients with clinically inapparent hepatitis as candidates for medical management have as a secondary benefit the identification of persons whose contacts could be at risk of becoming infected. A so-called look-back program has been recommended to identify persons who were transfused before 1992 with blood from a donor found subsequently to have hepatitis C. In addition, anti-HCV testing is recommended for persons born between 1945 and 1965, anyone who received a blood transfusion or a transplanted organ before the introduction of second-generation screening tests in 1992, those who ever used injection drugs (or took other illicit drugs by noninjection routes), chronically hemodialyzed patients, persons with clotting disorders who received clotting factors made before 1987 from pooled blood products, persons with elevated aminotransferase levels, health workers exposed to HCV-positive blood or contaminated needles, recipients of blood or organs from a donor found to be positive for hepatitis C, persons with HIV infection, health care and public safety personnel following a needle stick or other nonpercutaneous exposure to HCV-infected material, sexual partners of persons with hepatitis C, and children born to HCV-positive mothers (Table 332-4).

For stable, monogamous sexual partners, sexual transmission of hepatitis C is unlikely, and sexual barrier precautions are not recommended. For persons with multiple sexual partners or with sexually transmitted diseases, the risk of sexual transmission of hepatitis C is increased, and barrier precautions (latex condoms) are recommended. A person with hepatitis C should avoid sharing such items as razors, toothbrushes, and nail clippers with sexual partners and family members. No special precautions are recommended for babies born to mothers with hepatitis C, and breast-feeding does not have to be restricted.

Hepatitis E Whether IG prevents hepatitis E remains undetermined. Safe and effective recombinant genotype 1 vaccines, which protect against other genotypes as well, have been developed and are available in endemic areas but not in the United States. Protection provided by the Chinese hepatitis E vaccine is long-lasting, documented in a clinical trial up to 4.5 years.

FURTHER READING

Buckley GJ, Strom BL (eds): *Eliminating the Public Health Problem of Hepatitis B and C in the United States: Phase One Report*. Washington DC, National Academies Press, 2016. Available from *http://www.nap.edu/23407*. Accessed September 22, 2016.

Centers for Disease Control and Prevention: Recommendations for the identification of chronic hepatitis C virus infection among persons born during 1945–1965. MMWR 61(RR-4):1, 2012.

Debing Y et al: Update on hepatitis E virology: Implications for clinical practice. J Hepatol 65:200, 2016.

Denniston MM et al: Chronic hepatitis C virus infection in the United States, National Health and Nutrition Examination Survey 2003–2010. Ann Intern Med 160:293, 2014.

Douam F et al: The mechanism of HCV entry into host cells. Prog Mol Biol Transl Sci 129:63, 2015.

Joy JB et al: The spread of hepatitis C virus genotype 1a in North America: A retrospective phylogenetic study. Lancet Infect Dis 16:698, 2016.

Lin H-H et al: Changing hepatitis D virus epidemiology in a hepatitis B virus endemic area with a national vaccination program. Hepatology 61:1870, 2016.

Pan CQ et al: Tenofovir to prevent hepatitis B transmission in mothers with high viral load. N Engl J Med 374:2324, 2016.

Roberts H et al: Prevalence of chronic hepatitis B virus (HBV) infection in U.S. households: National Health and Nutrition Examination Survey (NHANES), 1988–2012. Hepatology 63:388, 2016.

Trépo C et al: Hepatitis B virus infection. Lancet 384:2053, 2014.

333 Toxic and Drug-Induced Hepatitis

William M. Lee, Jules L. Dienstag

Liver injury is a possible consequence of ingestion of any xenobiotic, including industrial toxins, pharmacologic agents, and complementary and alternative medications (CAMs). Among patients with acute liver failure, drug-induced liver injury (DILI) is the most common cause, and evidence for hepatotoxicity detected during clinical trials for drug development is the most common reason for failure of compounds to reach approval status. DILI requires careful history taking to identify unrecognized exposure to chemicals used in work or at home, drugs taken by prescription or bought over the counter, and herbal or dietary supplement medicines. Hepatotoxic drugs can injure the hepatocyte directly, for example, via a free-radical or metabolic intermediate that causes peroxidation of membrane lipids and that results in liver cell injury. Alternatively, a drug or its metabolite may activate components

of the innate or adaptive immune system, stimulate apoptotic pathways, or initiate damage to bile excretory pathways **(Fig. 333-1)**. Interference with bile canalicular pumps can allow endogenous bile acids, which can injure the liver, to accumulate. Such secondary injury, in turn, may lead to necrosis of hepatocytes; injure bile ducts, producing cholestasis; or block pathways of lipid movement, inhibit protein synthesis, or impair mitochondrial oxidation of fatty acids, resulting in lactic acidosis and intracellular triglyceride accumulation (expressed histologically as microvesicular steatosis). In other instances, drug metabolites sensitize hepatocytes to toxic cytokines. The differences observed between susceptible and nonsusceptible drug recipients may be attributable to human leukocyte antigens (HLA) haplotypes that determine binding of drug-related haptens on the cell surface as well as to polymorphisms in elaboration of competing, protective cytokines, as has been suggested for acetaminophen hepatotoxicity (see below). Immune mechanisms may include cytotoxic lymphocytes or antibody-mediated cellular cytotoxicity. In addition, a role has been shown for activation of nuclear transporters, such as the constitutive androstane receptor (CAR) or, more recently, the pregnane X receptor (PXR), in the induction of drug hepatotoxicity.

DRUG METABOLISM

Most drugs, which are water-insoluble, undergo a series of metabolic steps, culminating in a water-soluble form appropriate for renal or biliary excretion. This process begins with oxidation or methylation mediated initially by the microsomal mixed-function oxygenases, cytochrome P450 (phase I reaction), followed by glucuronidation or sulfation (phase II reaction) or inactivation by glutathione. Most drug hepatotoxicity is mediated by a phase I toxic metabolite, but glutathione depletion, precluding inactivation of harmful compounds by glutathione S-transferase, can contribute as well.

LIVER INJURY CAUSED BY DRUGS

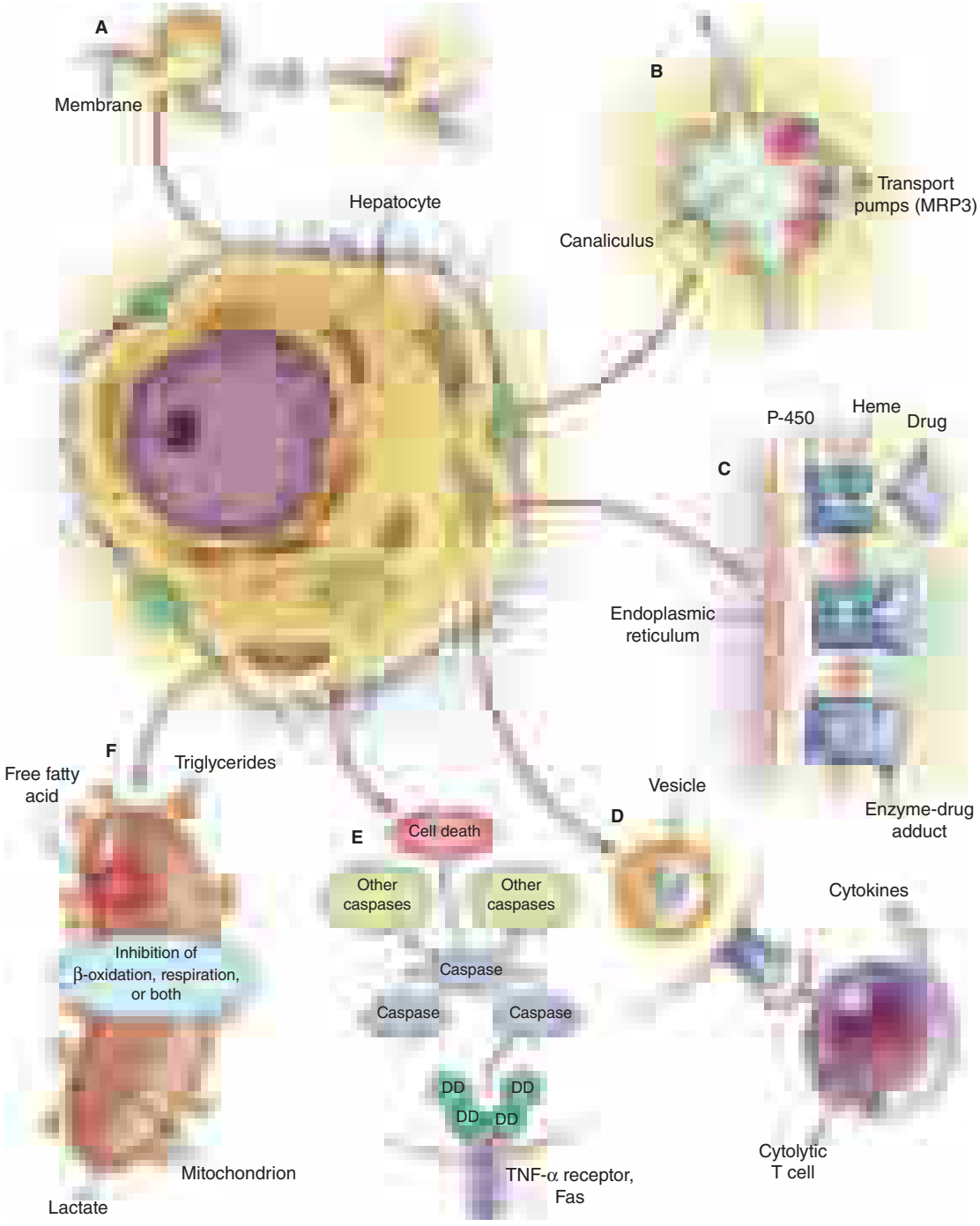
In general, two major types of chemical hepatotoxicity have been recognized: (1) direct toxic and (2) idiosyncratic. As shown in **Table 333-1**, direct toxic hepatitis occurs with predictable regularity in individuals exposed to the offending agent and is dose-dependent. The latent period between exposure and liver injury is usually short (often several hours), although clinical manifestations may be delayed for 24–48 h. Agents producing toxic hepatitis are generally systemic poisons or are converted in the liver to toxic metabolites. The direct hepatotoxins result in morphologic abnormalities that are reasonably characteristic and reproducible for each toxin. For example, carbon tetrachloride and trichloroethylene characteristically produce a centrilobular zonal necrosis, whereas yellow phosphorus poisoning typically results in periportal injury. The hepatotoxic octapeptides of *Amanita phalloides* usually produce massive hepatic necrosis; the lethal dose of the toxin is ~10 mg, the amount found in a single deathcap mushroom. Liver injury, which is often only one facet of the toxicity produced by the direct hepatotoxins, may go unrecognized until jaundice appears.

In idiosyncratic drug reactions, the occurrence of hepatitis is usually infrequent (1 in 10^3–10^5 patients) and unpredictable; the response is not as clearly dose-dependent as is injury associated with direct hepatotoxins, and liver injury may occur at any time during or shortly after exposure to the drug. That said, recent data suggest that most agents causing idiosyncratic toxicity are given at a daily dose exceeding 100 mg, suggesting a role for dose—drugs with low potency must be given in higher doses that engender greater chances for "off-target" effects. Adding to the difficulty of predicting or identifying idiosyncratic drug hepatotoxicity is the occurrence of mild, transient, nonprogressive serum aminotransferase elevations that resolve with continued drug use. Such "adaptation," the mechanism of which is unknown, is well recognized for drugs such as isoniazid (INH), valproate, phenytoin, and HMG-CoA reductase inhibitors (statins). Extrahepatic manifestations of hypersensitivity, such as rash, arthralgias, fever, leukocytosis, and eosinophilia, occur in about one-quarter of patients with idiosyncratic hepatotoxic drug reactions but are characteristic for certain drugs and not others. Both primary immunologic injury and direct hepatotoxicity related to idiosyncratic differences in generation of toxic metabolites have been invoked to explain idiosyncratic drug reactions. The most current data appear to implicate the adaptive immune system responding to the formation of immune stimulatory compounds resulting from phase I metabolic activation of the offending drug. Differences in host susceptibility may result from varying kinetics of toxic metabolite generation and genetic polymorphisms in downstream drug-metabolizing pathways or cytokine activation; in addition, certain HLA haplotypes have been associated with hepatotoxicity of certain drugs such as amoxicillin-clavulanate and flucloxacillin. Occasionally, however, the clinical features of an allergic reaction (prominent tissue eosinophilia, autoantibodies, etc.) are difficult to ignore and suggest activation of IgE pathways. A few instances of drug hepatotoxicity are observed to be associated with autoantibodies, including a class of antibodies to liver-kidney microsomes, anti-LKM2, directed against a cytochrome P450 enzyme.

Idiosyncratic reactions lead to a morphologic pattern that is more variable than those produced by direct toxins; a single agent is often capable of causing a variety of lesions, although certain patterns tend to predominate. Depending on the agent involved, idiosyncratic hepatitis may result in a clinical and morphologic picture indistinguishable from that of viral hepatitis (e.g., INH or ciprofloxacin). So-called hepatocellular injury is the most common form, featuring spotty necrosis in the liver lobule with a predominantly lymphocytic infiltrate resembling that observed in acute hepatitis A, B, or C. Drug-induced cholestasis ranges from mild to increasingly severe: (1) bland cholestasis with limited hepatocellular injury (e.g., estrogens, 17,α-substituted androgens); (2) inflammatory cholestasis (e.g., amoxicillin-clavulanic acid [the most frequently implicated antibiotic among cases of DILI], oxacillin, erythromycin estolate); (3) sclerosing cholangitis (e.g., after intrahepatic infusion of the chemotherapeutic agent floxuridine for hepatic metastases from a primary colonic carcinoma); and (4) disappearance of bile ducts, "ductopenic" cholestasis, similar to that observed in chronic rejection **(Chap. 338)** following liver transplantation (e.g., carbamazepine, levofloxacin). Cholestasis may result from binding of drugs to canalicular membrane transporters, accumulation of toxic bile acids resulting from canalicular pump failure, or genetic defects in canalicular transporter proteins. Clinically, the distinction between a hepatocellular and a cholestatic reaction is indicated by the R value, the ratio of alanine aminotransferase (ALT) to alkaline phosphatase values, both expressed as multiples of the upper limit of normal. An R value of >5.0 is associated with hepatocellular injury, R <2.0 with cholestatic injury, and R between 2.0 and 5.0 with mixed hepatocellular-cholestatic injury.

Morphologic alterations may also include bridging hepatic necrosis (e.g., methyldopa) or, infrequently, hepatic granulomas (e.g., sulfonamides). Some drugs result in macrovesicular or microvesicular steatosis or steatohepatitis, which, in some cases, has been linked to mitochondrial dysfunction and lipid peroxidation. Severe hepatotoxicity associated with steatohepatitis, most likely a result of mitochondrial toxicity, is being recognized with increasing frequency among patients receiving antiretroviral therapy with reverse transcriptase inhibitors for HIV infection (e.g., zidovudine, didanosine), although many of these drugs have been withdrawn because of such hepatotoxicity **(Chap. 197)**. Generally, such mitochondrial hepatotoxicity of these antiretroviral agents is reversible, but dramatic, nonreversible hepatotoxicity associated with mitochondrial injury (inhibition of DNA polymerase γ) was the cause of acute liver failure encountered during early clinical trials of now-abandoned fialuridine, a fluorinated pyrimidine analogue with potent antiviral activity against hepatitis B virus. Another potential target for idiosyncratic drug hepatotoxicity is sinusoidal lining cells; when these are injured, such as by high-dose chemotherapeutic agents (e.g., cyclophosphamide, melphalan, busulfan) administered prior to bone marrow transplantation, venoocclusive disease can result. Nodular regenerative hyperplasia, a subtle form of portal hypertension, may also result from vascular injury to portal venous endothelium following systemic chemotherapy, such as with oxaliplatin, as part of adjuvant treatment for colon cancer.

Not all adverse hepatic drug reactions can be classified as either toxic or idiosyncratic. For example, oral contraceptives, which



A. Rupture of cell membrane.
B. Injury of bile canaliculus (disruption of transport pumps).
C. P-450-drug covalent binding (drug adducts).
D. Drug adducts targeted by CTLs/cytokines.
E. Activation of apoptotic pathway by TNFα/Fas.
F. Inhibition of mitochondrial function.

FIGURE 333-1 Potential mechanisms of drug-induced liver injury. The normal hepatocyte may be affected adversely by drugs through (***A***) disruption of intracellular calcium homeostasis that leads to the disassembly of actin fibrils at the surface of the hepatocyte, resulting in blebbing of the cell membrane, rupture, and cell lysis; (***B***) disruption of actin filaments next to the canaliculus (the specialized portion of the cell responsible for bile excretion), leading to loss of villous processes and interruption of transport pumps such as multidrug resistance–associated protein 3 (MRP3), which, in turn, prevents the excretion of bilirubin and other organic compounds; (***C***) covalent binding of the heme-containing cytochrome P450 enzyme to the drug, thus creating nonfunctioning adducts; (***D***) migration of these enzyme-drug adducts to the cell surface in vesicles to serve as target immunogens for cytolytic attack by T cells, stimulating an immune response involving cytolytic T cells and cytokines; (***E***) activation of apoptotic pathways by tumor necrosis factor α (TNF-α) receptor or Fas (DD denotes death domain), triggering the cascade of intercellular caspases, resulting in programmed cell death; or (***F***) inhibition of mitochondrial function by a dual effect on both β-oxidation and the respiratory-chain enzymes, leading to failure of free fatty acid metabolism, a lack of aerobic respiration, and accumulation of lactate and reactive oxygen species (which may disrupt mitochondrial DNA). Toxic metabolites excreted in bile may damage bile-duct epithelium (not shown). CTLs, cytolytic T lymphocytes. *(Reproduced from WM Lee: Drug-induced hepatotoxicity. N Engl J Med 349:474, 2003, with permission.)*

combine estrogenic and progestational compounds, may result in impairment of liver tests and, occasionally, jaundice; however, they do not produce necrosis or fatty change, manifestations of hypersensitivity are generally absent, and susceptibility to the development of oral contraceptive–induced cholestasis appears to be genetically determined. Such estrogen-induced cholestasis is more common in women with cholestasis of pregnancy, a disorder linked to genetic defects in multidrug resistance–associated canalicular transporter proteins. A rare

TABLE 333-1 Some Features of Toxic and Drug-Induced Hepatic Injury

	DIRECT TOXIC EFFECT[a]		IDIOSYNCRATIC[a]			OTHER[a]
FEATURES	CARBON TETRACHLORIDE	ACETAMINOPHEN	AMOXICILLIN-CLAVULANATE	ISONIAZID	CIPROFLOXACIN	ESTROGENS/ ANDROGENIC STEROIDS
Predictable and dose-related toxicity	+	+	0	0	0	+
Latent period	Short	Short	Delayed onset	Variable	May be short	Variable
Arthralgia, fever, rash, eosinophilia	0	0	0	0	0	0
Liver morphology	Necrosis, fatty infiltration	Centrilobular necrosis	Mixed hepatocellular/ cholestatic	Hepatocellular injury resembling viral hepatitis	Hepatocellular injury resembling viral hepatitis	Cholestasis *without* portal inflammation

[a]The drugs listed are typical examples.

complication of oral contraceptive therapy is hepatic sinusoidal dilatation localized to the periportal zone of the liver lobule.

Any idiosyncratic reaction that occurs in <1:10,000 recipients will go unrecognized in most clinical trials, which involve only several thousand recipients. The U.S. Food and Drug Administration (FDA) and pharmaceutical companies have learned to look for even subtle indications of serious toxicity and monitor regularly the number of trial subjects in whom any aminotransferase elevations develop, as a possible surrogate for more serious toxicity. Even more valid as a predictor of severe hepatotoxicity is the occurrence of jaundice in patients enrolled in a clinical drug trial, so-called "Hy's Law," named after Hyman Zimmerman, one of the pioneers of the field of drug hepatotoxicity. He recognized that, if jaundice occurred during a phase III trial, more serious liver injury was likely, with a 10:1 ratio between cases of jaundice and liver failure—10 patients with jaundice to 1 patient with acute liver failure. Thus, the finding of such Hy's Law cases during drug development often portends failure of approval, particularly if any of the subjects sustains a bad outcome. Troglitazone, a peroxisome proliferator-activated receptor γ agonist, was the first in its class of thiazolidinedione insulin-sensitizing agents. Although in retrospect, Hy's Law cases of jaundice had occurred during phase III trials, no instances of liver failure were recognized until well after the drug was introduced, emphasizing the importance of postmarketing surveillance in identifying toxic drugs and in leading to their withdrawal from use. Fortunately, such hepatotoxicity is not characteristic of the second-generation thiazolidinedione insulin-sensitizing agents rosiglitazone and pioglitazone; in clinical trials, the frequency of aminotransferase elevations in patients treated with these medications did not differ from that in placebo recipients, and isolated reports of liver injury among recipients are extremely rare.

Proving that an episode of liver injury is caused by a drug is difficult in many cases. DILI is nearly always a presumptive diagnosis, and many other disorders produce a similar clinicopathologic picture. Thus, causality may be difficult to establish and requires several separate supportive assessment variables to lead to a high level of certainty, including temporal association (time of onset, time to resolution), clinical-biochemical features, type of injury (hepatocellular versus cholestatic), extrahepatic features, likelihood that a given agent is to blame based on its past record, and exclusion of other potential causes. Scoring systems such as the Roussel-Uclaf Causality Assessment Method (RUCAM) yield residual uncertainty and have not been adopted widely. Currently, the U.S. Drug-Induced Liver Injury Network (DILIN) relies on a structured expert opinion process requiring detailed data on each case and a comprehensive review by three experts who arrive at a consensus on a five-degree scale of likelihood (definite, highly likely, probable, possible, unlikely); however, this approach is not practical for routine clinical application.

Generally, drug hepatotoxicity is not more frequent in persons with underlying chronic liver disease, although the severity of the outcome may be amplified. Reported exceptions include hepatotoxicity of aspirin, methotrexate, INH (only in certain experiences), antiretroviral therapy for HIV infection, and certain drugs such as conditioning regimens for bone marrow transplantation in the presence of hepatitis C.

TREATMENT

Toxic and Drug-Induced Hepatic Disease

Treatment is largely supportive, except in acetaminophen hepatotoxicity (for which *N*-acetylcysteine is effective, see below). In patients with fulminant hepatitis resulting from drug hepatotoxicity, liver transplantation may be lifesaving **(Chap. 338)**. Withdrawal of the suspected agent is indicated at the first sign of an adverse reaction. A number of studies have suggested that lethal outcomes follow continued use of an agent in the face of symptoms and signs of liver injury. In the case of the direct toxins, liver involvement should not divert attention from renal or other organ involvement, which may also threaten survival. A number of agents are used occasionally but are of questionable value: glucocorticoids for drug hepatotoxicity with allergic features, silibinin for hepatotoxic mushroom poisoning, and ursodeoxycholic acid for cholestatic drug hepatotoxicity have not been shown to be effective and cannot be recommended. A double-blind, randomized controlled trial of the use of *N*-acetylcysteine for non-acetaminophen acute liver failure, including cases of DILI demonstrated benefit particularly for patients with early-stage hepatic encephalopathy; however, the drug is not FDA-approved for this indication.

In **Table 333-2**, several classes of chemical agents are listed together with examples of the pattern of liver injury they produce. Certain drugs appear to be responsible for the development of chronic as well as acute hepatic injury. For example, nitrofurantoin, minocycline, hydralazine, and methyldopa have been associated with moderate to severe chronic hepatitis with autoimmune features. Methotrexate, tamoxifen, and amiodarone have been implicated in the development of cirrhosis. Portal hypertension in the absence of cirrhosis may result from alterations in hepatic architecture produced by vitamin A or arsenic intoxication, industrial exposure to polyvinyl chloride, or administration of thorium dioxide. The latter three agents have also been associated with angiosarcoma of the liver. Oral contraceptives have been implicated in the development of hepatic adenoma and, rarely, hepatocellular carcinoma and hepatic vein occlusion (Budd-Chiari syndrome). Another unusual lesion, peliosis hepatis (blood cysts of the liver), has been observed in some patients treated with anabolic or contraceptive steroids. The existence of these hepatic disorders expands the spectrum of liver injury induced by chemical agents and emphasizes the need for a thorough drug history in all patients with liver dysfunction. The comprehensive, authoritative LiverTox website, which contains up-to-date information on drug-induced liver injury, is available as a valuable reference through the National Institutes of Health and the National Library of Medicine (*www.livertox.nih.gov*).

The following are patterns of adverse hepatic reactions for some prototypic agents.

ACETAMINOPHEN HEPATOTOXICITY (DIRECT TOXIN)

Acetaminophen represents the most prevalent cause of acute liver failure in the Western world; up to 72% of patients with acetaminophen hepatotoxicity in Scandinavia—somewhat lower frequencies in the

TABLE 333-2 Principal Alterations of Hepatic Morphology Produced by Some Commonly Used Drugs and Chemicals[a]

PRINCIPAL MORPHOLOGIC CHANGE	CLASS OF AGENT	EXAMPLE
Cholestasis	Anabolic steroid	Methyl testosterone, many other body-building supplements
	Antibiotic	Erythromycin estolate, nitrofurantoin, rifampin, amoxicillin-clavulanic acid, oxacillin
	Anticonvulsant	Carbamazepine
	Antidepressant	Duloxetine, mirtazapine, tricyclic antidepressants
	Anti-inflammatory	Sulindac
	Antiplatelet	Clopidogrel
	Antihypertensive	Irbesartan, fosinopril
	Antithyroid	Methimazole
	Calcium channel blocker	Nifedipine, verapamil
	Immunosuppressive	Cyclosporine
	Lipid-lowering	Ezetimibe
	Oncotherapeutic	Anabolic steroids, busulfan, tamoxifen, irinotecan, cytarabine, temozolomide
	Oral contraceptive	Norethynodrel with mestranol
	Oral hypoglycemic	Chlorpropamide
	Tranquilizer	Chlorpromazine[b]
Fatty liver	Antiarrhythmic	Amiodarone
	Antibiotic	Tetracycline (high-dose, IV)
	Anticonvulsant	Valproic acid
	Antiviral	Dideoxynucleosides (e.g., zidovudine), protease inhibitors (e.g., indinavir, ritonavir)
	Oncotherapeutic	Asparaginase, methotrexate, tamoxifen
Hepatitis	Anesthetic	Halothane, fluothane
	Antiandrogen	Flutamide
	Antibiotic	Isoniazid,[c] rifampicin, nitrofurantoin, telithromycin, minocycline,[d] pyrazinamide, trovafloxacin[e]
	Anticonvulsant	Phenytoin, carbamazepine, valproic acid, phenobarbital
	Antidepressant	Iproniazid, amitriptyline, trazodone, venlafaxine, fluoxetine, paroxetine, duloxetine, sertraline, nefazodone[e]
	Antifungal	Ketoconazole, fluconazole, itraconazole
	Antihypertensive	Methyldopa,[c] captopril, enalapril, lisinopril, losartan
	Anti-inflammatory	Ibuprofen, indomethacin, diclofenac, sulindac, bromfenac
	Antipsychotic	Risperidone
	Antiviral	Zidovudine, didanosine, stavudine, nevirapine, ritonavir, indinavir, tipranavir, zalcitabine
	Calcium channel blocker	Nifedipine, verapamil, diltiazem
	Cholinesterase inhibitor	Tacrine
	Diuretic	Chlorothiazide
	Laxative	Oxyphenisatin[c,e]
	Norepinephrine-reuptake inhibitor	Atomoxetine
	Oral hypoglycemic	Troglitazone,[e] acarbose
Mixed hepatitis/cholestatic	Antibiotic	Amoxicillin-clavulanic acid, trimethoprim-sulfamethoxazole
	Antibacterial	Clindamycin
	Antifungal	Terbinafine
	Antihistamine	Cyproheptadine
	Immunosuppressive	Azathioprine
	Lipid-lowering	Nicotinic acid, lovastatin, ezetimibe
Toxic (necrosis)	Analgesic	Acetaminophen
	Hydrocarbon	Carbon tetrachloride
	Metal	Yellow phosphorus
	Mushroom	*Amanita phalloides*
	Solvent	Dimethylformamide
Granulomas	Antiarrhythmic	Quinidine, diltiazem
	Antibiotic	Sulfonamides
	Anticonvulsant	Carbamazepine
	Anti-inflammatory	Phenylbutazone
	Xanthine oxidase inhibitor	Allopurinol
Vascular injury	Chemotherapeutic	Oxaliplatin, melphalan

[a]Several agents cause more than one type of liver lesion and appear under more than one category. [b]Rarely associated with primary biliary cirrhosis–like lesion. [c]Occasionally associated with chronic hepatitis or bridging hepatic necrosis or cirrhosis. [d]Associated with an autoimmune hepatitis–like syndrome. [e]Withdrawn from use because of severe hepatotoxicity.

United Kingdom and the United States—progress to encephalopathy and coagulopathy. Acetaminophen causes dose-related centrilobular hepatic necrosis after single-time-point ingestions, as intentional self-harm, or over extended periods, as unintentional overdoses, when multiple drug preparations or inappropriate drug amounts are used daily for several days, for example, for relief of pain or fever. In these instances, 8 g/d, twice the daily recommended maximum dose, over several days can readily lead to liver failure. Use of opioid-acetaminophen combinations appears to be particularly harmful, because habituation to the opioid may occur with a gradual increase in opioid-acetaminophen combination dosing over days or weeks. A single dose of 10–15 g, occasionally less, may produce clinical evidence of liver injury. Fatal fulminant disease is usually (although not invariably) associated with ingestion of ≥25 g. Blood levels of acetaminophen correlate with severity of hepatic injury (levels >300 µg/mL 4 h after ingestion are predictive of the development of severe damage; levels <150 µg/mL suggest that hepatic injury is highly unlikely). Nausea, vomiting, diarrhea, abdominal pain, and shock are early manifestations occurring 4–12 h after ingestion. Then 24–48 h later, when these features are abating, hepatic injury becomes apparent. Maximal abnormalities and hepatic failure are evident 3–5 days after ingestion, and aminotransferase levels exceeding 10,000 IU/L are not uncommon (i.e., levels far exceeding those in patients with viral hepatitis). Renal failure and myocardial injury may be present. Whether or not a clear history of overdose can be elicited, clinical suspicion of acetaminophen hepatotoxicity should be raised by the presence of the extremely high aminotransferase levels in association with low bilirubin levels that are characteristic of this hyperacute injury. This biochemical signature should trigger further questioning of the subject if possible; however, denial or altered mentation may confound diagnostic efforts. In this setting, a presumptive diagnosis is reasonable, and the proven antidote, *N*-acetylcysteine—both safe and presumed to be effective even when injury has already begun to evolve—should be instituted.

Acetaminophen is metabolized predominantly by a phase II reaction to innocuous sulfate and glucuronide metabolites; however, a small proportion of acetaminophen is metabolized by a phase I reaction to a hepatotoxic metabolite formed from the parent compound by cytochrome P450 CYP2E1. This metabolite, *N*-acetyl-p-benzoquinone-imine (NAPQI), is detoxified by binding to "hepatoprotective" glutathione to become harmless, water-soluble mercapturic acid, which undergoes renal excretion. When excessive amounts of NAPQI are formed, or when glutathione levels are low, glutathione levels are depleted and overwhelmed, permitting covalent binding to nucleophilic hepatocyte macromolecules forming acetaminophen-protein "adducts." These adducts, which can be measured in serum by high-performance liquid chromatography, hold promise as diagnostic markers of acetaminophen hepatotoxicity, and a point-of-care assay for acetaminophen-Cys adducts is under development. The binding of acetaminophen to hepatocyte macromolecules is believed to lead to hepatocyte necrosis; the precise sequence and mechanism are unknown. Hepatic injury may be potentiated by prior administration of alcohol, phenobarbital, INH, or other drugs; by conditions that stimulate the mixed-function oxidase system; or by conditions such as starvation (including inability to maintain oral intake during severe febrile illnesses) that reduce hepatic glutathione levels. Alcohol induces cytochrome P450 CYP2E1; consequently, increased levels of the toxic metabolite NAPQI may be produced in chronic alcoholics after acetaminophen ingestion, but the role of alcohol in potentiating acute acetaminophen injury is still debated. Alcohol also suppresses hepatic glutathione production. Therefore, in chronic alcoholics, the toxic dose of acetaminophen may be as low as 2 g, and alcoholic patients should be warned specifically about the dangers of even standard doses of this commonly used drug. In a 2006 study, aminotransferase elevations were identified in 31–44% of normal subjects treated for 14 days with the maximal recommended dose of acetaminophen, 4 g daily (administered alone or as part of an acetaminophen-opioid combination); because these changes were transient and never associated with bilirubin elevation, the clinical relevance of these findings remains to be determined. Although underlying hepatitis C virus (HCV) infection was found to be associated with an increased risk of acute liver injury in patients hospitalized for acetaminophen overdose, generally, in patients with nonalcoholic liver disease, acetaminophen taken in recommended doses is well tolerated. Acetaminophen use in cirrhotic patients has not been associated with hepatic decompensation. On the other hand, because of the link between acetaminophen use and liver injury, and because of the limited safety margin between safe and toxic doses, the FDA has recommended that the daily dose of acetaminophen be reduced from 4 g to 3 g (even lower for persons with chronic alcohol use), that all acetaminophen-containing products be labeled prominently as containing acetaminophen, and that the potential for liver injury be prominent in the packaging of acetaminophen and acetaminophen-containing products. Within opioid combination products, the limit for the acetaminophen component has been lowered to 325 mg per tablet.

TREATMENT

Acetaminophen Overdosage

Treatment includes gastric lavage, supportive measures, and oral administration of activated charcoal or cholestyramine to prevent absorption of residual drug. Neither charcoal nor cholestyramine appears to be effective if given >30 min after acetaminophen ingestion; if they are used, the stomach lavage should be done before other agents are administered orally. The chances of possible, probable, and high-risk hepatotoxicity can be derived from a nomogram plot (Fig. 333-2), readily available in emergency departments, as a function of measuring acetaminophen plasma levels 8 h after ingestion. In patients with high acetaminophen blood levels (>200 µg/mL measured at 4 h or >100 µg/mL at 8 h after ingestion), the administration of *N*-acetylcysteine reduces the severity of hepatic necrosis. This agent provides sulfhydryl donor groups to replete glutathione, which is required to render harmless toxic metabolites that would otherwise bind covalently via sulfhydryl linkages to cell proteins, resulting in the formation of drug metabolite-protein adducts. Therapy should be begun within 8 h of ingestion but may be at least partially effective when given as late as 24–36 h after overdose.

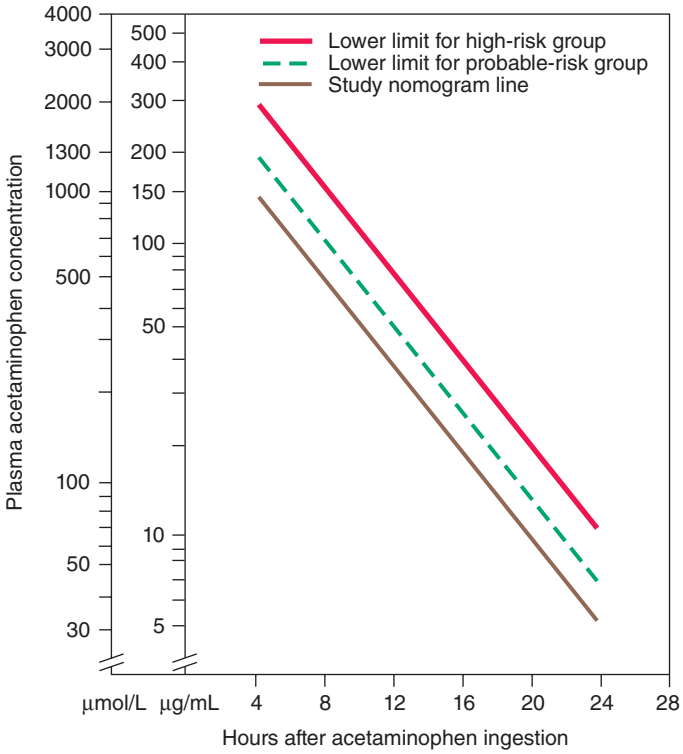


FIGURE 333-2 Nomogram to define risk of acetaminophen hepatotoxicity according to initial plasma acetaminophen concentration. *(After BH Rumack, H Matthew: Pediatrics 55:871, 1975.)*

Later administration of sulfhydryl compounds is of uncertain value. Routine use of *N*-acetylcysteine has substantially reduced the occurrence of fatal acetaminophen hepatotoxicity. *N*-acetylcysteine may be given orally but is more commonly used as an IV solution, with a loading dose of 140 mg/kg over 1 h, followed by 70 mg/kg every 4 h for 15–20 doses. Whenever a patient with potential acetaminophen hepatotoxicity is encountered, a local poison control center should be contacted. Treatment can be stopped when plasma acetaminophen levels indicate that the risk of liver damage is low. If signs of hepatic failure (e.g., progressive jaundice, coagulopathy, confusion) occur despite *N*-acetylcysteine therapy for acetaminophen hepatotoxicity, liver transplantation may be the only option. Early arterial blood lactate levels among such patients with acute liver failure may distinguish patients highly likely to require liver transplantation (lactate levels >3.5 mmol/L) from those likely to survive without liver replacement. Acute renal injury occurs in nearly 75% of patients with severe acetaminophen injury but is virtually always self-limited.

Survivors of acute acetaminophen overdose rarely, if ever, have ongoing liver injury or sequelae.

ISONIAZID HEPATOTOXICITY (TOXIC AND IDIOSYNCRATIC REACTION)

INH remains central to most antituberculous prophylactic and therapeutic regimens, despite its long-standing recognition as a hepatotoxin. In 10% of patients treated with INH, elevated serum aminotransferase levels develop during the first few weeks of therapy; however, these elevations in most cases are self-limited, mild (values for ALT <200 IU/L), and resolve despite continued drug use. This adaptive response allows continuation of the agent if symptoms and progressive enzyme elevations do not follow the initial elevations. Acute hepatocellular drug-induced liver injury secondary to INH is evident with a variable latency period up to 6 months and is more frequent in alcoholics and patients taking certain other medications, such as barbiturates, rifampin, and pyrazinamide. If the clinical threshold of encephalopathy is reached, severe hepatic injury is likely to be fatal or to require liver transplantation. Liver biopsy reveals morphologic changes similar to those of viral hepatitis or bridging hepatic necrosis. Substantial liver injury appears to be age-related, increasing substantially after age 35; the highest frequency is in patients over age 50, and the lowest is in patients under the age of 20. Even for patients >50 years of age monitored carefully during therapy, hepatotoxicity occurs in only ~2%, well below the risk estimate derived from earlier experiences. Fever, rash, eosinophilia, and other manifestations of drug allergy are distinctly unusual. Antibodies to INH have been detected in INH recipients, but a link to causality of liver injury remains unclear. A clinical picture resembling chronic hepatitis has been observed in a few patients. Many public health programs that require INH prophylaxis for a positive tuberculin skin test or blood test (Quantiferon or T-Spot) include monthly monitoring of aminotransferase levels, although this practice has been called into question. Even more effective in limiting serious outcomes may be encouraging patients to be alert for symptoms such as nausea, fatigue, or jaundice, because most fatalities occur in the setting of continued INH use despite clinically apparent illness.

SODIUM VALPROATE HEPATOTOXICITY (TOXIC AND IDIOSYNCRATIC REACTION)

Sodium valproate, an anticonvulsant useful in the treatment of petit mal and other seizure disorders, has been associated with the development of severe hepatic toxicity and, rarely, fatalities, predominantly in children but also in adults. Among children listed as candidates for liver transplantation, valproate is the most common antiepileptic drug implicated. Asymptomatic elevations of serum aminotransferase levels have been recognized in as many as 45% of treated patients. These "adaptive" changes, however, appear to have no clinical importance, because major hepatotoxicity is not seen in the majority of patients despite continuation of drug therapy. In the rare patients in whom jaundice, encephalopathy, and evidence of hepatic failure are found, examination of liver tissue reveals microvesicular fat and bridging hepatic necrosis, predominantly in the centrilobular zone. Bile duct injury may also be apparent. Most likely, sodium valproate is not directly hepatotoxic, but its metabolite, 4-pentenoic acid, may be responsible for hepatic injury. Valproate hepatotoxicity is more common in persons with mitochondrial enzyme deficiencies and may be ameliorated by IV administration of carnitine, which valproate therapy can deplete. Valproate toxicity has been linked to HLA haplotypes (*DR4* and *B*1502*) and to mutations in mitochondrial DNA polymerase gamma 1.

NITROFURANTOIN HEPATOTOXICITY (IDIOSYNCRATIC REACTION)

This commonly used antibiotic for urinary tract infections may cause an acute hepatitis leading to fatal outcome or, more frequently, chronic hepatitis of varying severity but indistinguishable from autoimmune hepatitis. These two scenarios may reflect the frequent use and reuse of the drug for treatment of recurrent cystitis in women. Although most toxic agents manifest injury within 6 months of first ingestion, nitrofurantoin may have a longer latency period, in part perhaps because of its intermittent, recurrent use. Autoantibodies to nuclear components, smooth muscle, and mitochondria are seen and may subside after resolution of infection; however, glucocorticoid or other immunosuppressive medication may be necessary to resolve the autoimmune injury, and cirrhosis may be seen in cases that are not recognized quickly. Interstitial pulmonary fibrosis presenting as chronic cough and dyspnea may be present and resolve slowly with medication withdrawal. Histologic findings are identical to those of autoimmune hepatitis. A similar disease pattern can be observed with minocycline that is used repeatedly for the treatment of acne in teenagers as well as with hydralazine and alpha methyldopa.

AMOXICILLIN-CLAVULANATE HEPATOTOXICITY (IDIOSYNCRATIC MIXED REACTION)

Currently, the most common agent implicated as causing drug-induced liver injury in the United States and in Europe is amoxicillin-clavulanate (most frequent brand name: Augmentin). This medication causes a very specific syndrome of mixed or primarily cholestatic injury. Because hepatotoxicity may follow amoxicillin-clavulanate therapy after a relatively long latency period, the liver injury may begin to manifest at the time of drug withdrawal or after the drug has been withdrawn. The high prevalence of hepatotoxicity reflects in part the very frequent use of this drug for respiratory tract infections, including community-acquired pneumonia. The mechanism of hepatotoxicity is unclear, but the liver injury is thought to be caused by amoxicillin toxicity that is potentiated in some way by clavulanate, which itself appears not to be toxic. Symptoms include nausea, anorexia, fatigue, and jaundice—which may be prolonged—with pruritus. Rash is quite uncommon. On occasion, amoxicillin-clavulanate, like other cholestatic hepatotoxic drugs, causes permanent injury to small bile ducts, leading to the so-called "vanishing bile duct syndrome." In vanishing bile duct syndrome, initially, liver injury is minimal except for severe cholestasis; however, over time, histologic evidence of bile duct abnormalities is replaced by a paucity and eventual absence of discernible ducts on subsequent biopsies.

PHENYTOIN HEPATOTOXICITY (IDIOSYNCRATIC REACTION)

Phenytoin, formerly diphenylhydantoin, a mainstay in the treatment of seizure disorders, has been associated in rare instances with the development of severe hepatitis-like liver injury leading to fulminant hepatic failure. In many patients, the hepatitis is associated with striking fever, lymphadenopathy, rash (Stevens-Johnson syndrome or exfoliative dermatitis), leukocytosis, and eosinophilia, suggesting an immunologically mediated hypersensitivity mechanism. Despite these observations, evidence suggests that metabolic idiosyncrasy may be responsible for hepatic injury. In the liver, phenytoin is converted by cytochrome P450 to metabolites that include the highly reactive electrophilic arene oxides. These metabolites are normally metabolized further by epoxide hydrolases. A defect (genetic or acquired) in epoxide hydrolase activity could permit covalent binding of arene oxides to hepatic macromolecules, thereby leading to hepatic injury. Hepatic injury is usually manifest within the first 2 months after beginning

phenytoin therapy. With the exception of an abundance of eosinophils in the liver, the clinical, biochemical, and histologic picture resembles that of viral hepatitis. In rare instances, bile duct injury may be the salient feature of phenytoin hepatotoxicity, with striking features of intrahepatic cholestasis. Asymptomatic elevations of aminotransferase and alkaline phosphatase levels have been observed in a sizable proportion of patients receiving long-term phenytoin therapy. These liver changes are believed by some authorities to represent the potent hepatic enzyme-inducing properties of phenytoin and are accompanied histologically by swelling of hepatocytes in the absence of necroinflammatory activity or evidence of chronic liver disease.

AMIODARONE HEPATOTOXICITY (TOXIC AND IDIOSYNCRATIC REACTION)

Therapy with this potent antiarrhythmic drug is accompanied in 15–50% of patients by modest elevations of serum aminotransferase levels that may remain stable or diminish despite continuation of the drug. Such abnormalities may appear days to many months after beginning therapy. A proportion of those with elevated aminotransferase levels have detectable hepatomegaly, and clinically important liver disease develops in <5% of patients. Features that represent a direct effect of the drug on the liver and that are common to the majority of long-term recipients are ultrastructural phospholipidosis, unaccompanied by clinical liver disease, and interference with hepatic mixed-function oxidase metabolism of other drugs. The cationic amphiphilic drug and its major metabolite desethylamiodarone accumulate in hepatocyte lysosomes and mitochondria and in bile duct epithelium. The relatively common elevations in aminotransferase levels are also considered a predictable, dose-dependent, direct hepatotoxic effect. On the other hand, in the rare patient with clinically apparent, symptomatic liver disease, liver injury resembling that seen in alcoholic liver disease is observed. The so-called pseudoalcoholic liver injury can range from steatosis, to alcoholic hepatitis–like neutrophilic infiltration and Mallory's hyaline, to cirrhosis. Electron-microscopic demonstration of phospholipid-laden lysosomal lamellar bodies can help to distinguish amiodarone hepatotoxicity from typical alcoholic hepatitis. This category of liver injury appears to be a metabolic idiosyncrasy that allows hepatotoxic metabolites to be generated. Rarely, an acute idiosyncratic hepatocellular injury resembling viral hepatitis or cholestatic hepatitis occurs. Hepatic granulomas have occasionally been observed. Because amiodarone has a long half-life, liver injury may persist for months after the drug is stopped.

ERYTHROMYCIN HEPATOTOXICITY (CHOLESTATIC IDIOSYNCRATIC REACTION)

The most important adverse effect associated with erythromycin, more common in children than adults, is the infrequent occurrence of a cholestatic reaction. Although most of these reactions have been associated with erythromycin estolate, other erythromycins may also be responsible. The reaction usually begins during the first 2 or 3 weeks of therapy and includes nausea, vomiting, fever, right upper quadrant abdominal pain, jaundice, leukocytosis, and moderately elevated aminotransferase and alkaline phosphatase levels. The clinical picture can resemble acute cholecystitis or bacterial cholangitis. Liver biopsy reveals variable cholestasis; portal inflammation comprising lymphocytes, polymorphonuclear leukocytes, and eosinophils; and scattered foci of hepatocyte necrosis. Symptoms and laboratory findings usually subside within a few days of drug withdrawal, and evidence of chronic liver disease has not been found on follow-up. The precise mechanism remains ill-defined.

ORAL CONTRACEPTIVE HEPATOTOXICITY (CHOLESTATIC REACTION)

The administration of oral contraceptive combinations of estrogenic and progestational steroids leads to intrahepatic cholestasis with pruritus and jaundice in a small proportion of patients weeks to months after taking these agents. Especially susceptible seem to be patients with recurrent idiopathic jaundice of pregnancy, severe pruritus of pregnancy, or a family history of these disorders. With the exception of liver biochemical tests, laboratory studies are normal, and extrahepatic manifestations of hypersensitivity are absent. Liver biopsy reveals cholestasis with bile plugs in dilated canaliculi and striking bilirubin staining of liver cells. In contrast to chlorpromazine-induced cholestasis, portal inflammation is absent. The lesion is reversible on withdrawal of the agent. The two steroid components appear to act synergistically on hepatic function, although the estrogen may be primarily responsible. Oral contraceptives are contraindicated in patients with a history of recurrent jaundice of pregnancy. Primarily benign, but rarely malignant, neoplasms of the liver, hepatic vein occlusion, and peripheral sinusoidal dilatation have also been associated with oral contraceptive therapy. Focal nodular hyperplasia of the liver is not more frequent among users of oral contraceptives.

ANABOLIC STEROIDS (CHOLESTATIC REACTION)

The most common form of liver injury caused by CAMs is the profound cholestasis associated with anabolic steroids used by body builders. Unregulated agents sold in gyms and health food stores as diet supplements, which are taken by athletes to improve their performance, may contain anabolic steroids. In a young male, jaundice that is accompanied by a cholestatic, rather than a hepatitic, laboratory profile almost invariably will turn out to be caused by the use of one of a variety of androgen congeners. Such agents have the potential to injure bile transport pumps and to cause intense cholestasis; the time to onset is variable, and resolution, which is the rule, may require many weeks to months. Initially, anorexia, nausea, and malaise may occur, followed by pruritus in some but not all patients. Serum aminotransferase levels are usually <100 IU/L and serum alkaline phosphatase levels are generally moderately elevated with bilirubin levels frequently exceeding 342 µmol/L (20 mg/dL). Examination of liver tissue reveals cholestasis without substantial inflammation or necrosis. Anabolic steroids have also been used by prescription to treat bone marrow failure. In this setting, hepatic centrizonal sinusoidal dilatation and peliosis hepatis have been reported in rare patients, as have hepatic adenomas and hepatocellular carcinoma.

TRIMETHOPRIM-SULFAMETHOXAZOLE HEPATOTOXICITY (IDIOSYNCRATIC REACTION)

This antibiotic combination is used routinely for urinary tract infections in immunocompetent persons and for prophylaxis against and therapy of *Pneumocystis carinii* pneumonia in immunosuppressed persons (transplant recipients, patients with AIDS). With its increasing use, its occasional hepatotoxicity is being recognized with growing frequency. Its likelihood is unpredictable, but when it occurs, trimethoprim-sulfamethoxazole hepatotoxicity follows a relatively uniform latency period of several weeks and is often accompanied by eosinophilia, rash, and other features of a hypersensitivity reaction. Biochemically and histologically, acute hepatocellular necrosis predominates, but cholestatic features are quite frequent. Occasionally, cholestasis without necrosis occurs, and, very rarely, a severe cholangiolytic pattern of liver injury is observed. In most cases, liver injury is self-limited, but rare fatalities have been recorded. The hepatotoxicity is attributable to the sulfamethoxazole component of the drug and is similar in features to that seen with other sulfonamides; tissue eosinophilia and granulomas may be seen. The risk of trimethoprim-sulfamethoxazole hepatotoxicity is increased in persons with HIV infection.

HMG-COA REDUCTASE INHIBITORS (STATINS) (IDIOSYNCRATIC MIXED HEPATOCELLULAR AND CHOLESTATIC REACTION)

Between 1 and 2% of patients taking lovastatin, simvastatin, pravastatin, fluvastatin, or one of the newer statin drugs for the treatment of hypercholesterolemia experience asymptomatic, reversible elevations (>threefold) of aminotransferase activity. Acute hepatitis-like histologic changes, centrilobular necrosis, and centrilobular cholestasis have been described in a very small number of cases. In a larger proportion, minor aminotransferase elevations appear during the first several weeks of therapy. Careful laboratory monitoring can distinguish between patients with minor, transitory changes, who may continue therapy and those with more profound and sustained abnormalities, who should discontinue therapy. Because clinically meaningful aminotransferase elevations are so rare after statin use and do not differ in meta-analyses from the frequency of such laboratory abnormalities in placebo recipients, a

panel of liver experts recommended to the National Lipid Association's Safety Task Force that liver test monitoring was not necessary in patients treated with statins and that statin therapy need not be discontinued in patients found to have asymptomatic isolated aminotransferase elevations during therapy. Statin hepatotoxicity is not increased in patients with chronic hepatitis C, hepatic steatosis, or other underlying liver diseases, and statins can be used safely in these patients.

TOTAL PARENTERAL NUTRITION (STEATOSIS, CHOLESTASIS)

Total parenteral nutrition (TPN) is often complicated by cholestatic hepatitis attributable to steatosis, cholestasis, or gallstones (or gallbladder sludge). Steatosis or steatohepatitis may result from the excess carbohydrate calories in these nutritional supplements and is the predominant form of TPN-associated liver disorder in adults. The frequency of this complication has been reduced substantially by the introduction of balanced TPN formulas that rely on lipid as an alternative caloric source. Cholestasis and cholelithiasis, caused by the absence of stimulation of bile flow and secretion resulting from the lack of oral intake, is the predominant form of TPN-associated liver disease in infants, especially in premature neonates. Often, cholestasis in such neonates is multifactorial, contributed to by other factors such as sepsis, hypoxemia, and hypotension; occasionally, TPN-induced cholestasis in neonates culminates in chronic liver disease and liver failure. When TPN-associated liver test abnormalities occur in adults, balancing the TPN formula with more lipid is the intervention of first recourse. In infants with TPN-associated cholestasis, the addition of oral feeding may ameliorate the problem. Therapeutic interventions suggested, but not shown, to be of proven benefit, include cholecystokinin, ursodeoxycholic acid, *S* adenosyl methionine, and taurine.

ALTERNATIVE AND COMPLEMENTARY MEDICINES (IDIOSYNCRATIC HEPATITIS, STEATOSIS)

Herbal medications that are of scientifically unproven efficacy and that lack prospective safety oversight by regulatory agencies account currently for more than 20% of drug-induced liver injury in the United States. Besides anabolic steroids, the most common category of dietary or herbal products is weight loss agents. Included among the herbal remedies associated with toxic hepatitis are Jin Bu Huan, xiao-chai-hu-tang, germander, chaparral, senna, mistletoe, skullcap, gentian, comfrey (containing pyrrolizidine alkaloids), ma huang, bee pollen, valerian root, pennyroyal oil, kava, celandine, Impila (*Callilepis laureola*), LipoKinetix, Hydroxycut, herbal nutritional supplements, and herbal teas containing *Camellia sinensis* (green tea extract). Well characterized are the acute hepatitis-like histologic lesions following Jin Bu Huan use: focal hepatocellular necrosis, mixed mononuclear portal tract infiltration, coagulative necrosis, apoptotic hepatocyte degeneration, tissue eosinophilia, and microvesicular steatosis. Megadoses of vitamin A can injure the liver, as can pyrrolizidine alkaloids, which often contaminate Chinese herbal preparations and can cause a venoocclusive injury leading to sinusoidal hepatic vein obstruction. Because some alternative medicines induce toxicity via active metabolites, alcohol and drugs that stimulate cytochrome P450 enzymes may enhance the toxicity of some of these products. Conversely, some alternative medicines also stimulate cytochrome P450 and may result in or amplify the toxicity of recognized drug hepatotoxins. Given the widespread use of such poorly defined herbal preparations, hepatotoxicity is likely to be encountered with increasing frequency; therefore, a drug history in patients with acute and chronic liver disease should include use of "alternative medicines" and other nonprescription preparations sold in so-called health food stores.

HIGHLY ACTIVE ANTIRETROVIRAL THERAPY (HAART) FOR HIV INFECTION (MITOCHONDRIAL TOXIC, IDIOSYNCRATIC, STEATOSIS; HEPATOCELLULAR, CHOLESTATIC, AND MIXED)

The recognition of drug hepatotoxicity in persons with HIV infection is complicated in this population by the many alternative causes of liver injury (chronic viral hepatitis, fatty infiltration, infiltrative disorders, mycobacterial infection, etc.), but drug hepatotoxicity associated with HAART is an emerging and common type of liver injury in HIV-infected persons **(Chap. 197)**. Although no one antiviral agent is recognized as a potent hepatotoxin, combination regimens including reverse transcriptase and protease inhibitors cause hepatotoxicity in ~10% of treated patients. Implicated most frequently are combinations including nucleoside analogue reverse transcriptase inhibitors zidovudine, didanosine, and, to a lesser extent, stavudine; protease inhibitors ritonavir and indinavir (and amprenavir when used together with ritonavir), as well as tipranavir; and nonnucleoside reverse transcriptase inhibitors nevirapine and, to a lesser extent, efavirenz. These drugs cause predominantly hepatocellular injury but cholestatic injury as well, and prolonged (>6 months) use of reverse transcriptase inhibitors has been associated with mitochondrial injury, steatosis, and lactic acidosis. Indirect hyperbilirubinemia, resulting from direct inhibition of bilirubin-conjugating activity by UDP-glucuronosyltransferase, usually without elevation of aminotransferase or alkaline phosphatase activities, occurs in ~10% of patients treated with the protease inhibitor indinavir. Distinguishing the impact of HAART hepatotoxicity in patients with HIV and hepatitis virus co-infection is made challenging by the following: (1) both chronic hepatitis B and hepatitis C can affect the natural history of HIV infection and the response to HAART, and (2) HAART can have an impact on chronic viral hepatitis. For example, immunologic reconstitution with HAART can result in immunologically mediated liver-cell injury in patients with chronic hepatitis B co-infection if treatment with an antiviral agent for hepatitis B (e.g., the nucleoside analogue lamivudine) is withdrawn or if nucleoside analogue resistance emerges. Infection with HIV, especially with low CD4+ T cell counts, has been reported to increase the rate of hepatic fibrosis associated with chronic hepatitis C, and HAART therapy can increase levels of serum aminotransferases and HCV RNA in patients with hepatitis C co-infection. Didanosine or stavudine should not be used with ribavirin in patients with HIV/HCV co-infection because of an increased risk of severe mitochondrial toxicity and lactic acidosis.

Acknowledgment

Kurt J. Isselbacher, MD, contributed to this chapter in previous editions of Harrison's.

FURTHER READING

Björnsson ES, Hoofnagle JL: Categorization of drugs implicated in causing liver injury: Critical assessment based upon published case reports. Hepatology 63:590, 2016.

Chalasani N et al: Features and outcomes of 899 patients with drug-induced liver injury: The DILIN prospective study. Gastroenterology 148:1340, 2015.

Chen M et al: A model to predict severity of drug-induced liver injury in humans. Hepatology 64:931, 2016.

Fontana RJ et al: Idiosyncratic drug-induced liver injury is associated with substantial morbidity and mortality within 6 months from onset. Gastroenterology 147:96, 2014.

Goldberg DS et al: Population-representative incidence of drug-induced acute liver failure based on an analysis of an integrated health care system. Gastroenterology 148:1353, 2015.

Hayashi PH et al: Under-reporting and poor adherence to monitoring guidelines for severe cases of isoniazid hepatotoxicity. Clin Gastroenterol Hepatol 13:1676, 2015.

Hoofnagle JH et al: LiverTox: A website on drug-induced liver injury. Hepatology 57:873, 2013. Available from *www.livertox.nih.gov*. Accessed September 26, 2016.

Kaplowitz N, Deleve LD (eds): *Drug-Induced Liver Disease*, 3rd ed. London, Elsevier/Academic Press, 2013.

Lee WM et al: Intravenous N-acetylcysteine improves transplant-free survival in early stage non-acetaminophen acute liver failure. Gastroenterology 137:856, 2009.

Medina-Caliz I et al: Definition and risk factors for chronicity following acute idiosyncratic drug-induced liver injury. J Hepatol 65:532, 2016.

334 Chronic Hepatitis

Jules L. Dienstag

Chronic hepatitis represents a series of liver disorders of varying causes and severity in which hepatic inflammation and necrosis continue for at least 6 months. Milder forms are nonprogressive or only slowly progressive, while more severe forms may be associated with scarring and architectural reorganization, which, when advanced, lead ultimately to cirrhosis. Several categories of chronic hepatitis have been recognized. These include chronic viral hepatitis, drug-induced chronic hepatitis **(Chap. 333)**, and autoimmune chronic hepatitis. In many cases, clinical and laboratory features are insufficient to allow assignment into one of these three categories; these "idiopathic" cases are also believed to represent autoimmune chronic hepatitis. Finally, clinical and laboratory features of chronic hepatitis are observed occasionally in patients with such hereditary/metabolic disorders as Wilson's disease (copper overload), α_1 antitrypsin deficiency **(Chaps. 337 and 408)**, and nonalcoholic fatty liver disease **(Chap. 336)** and even occasionally in patients with alcoholic liver injury **(Chap. 335)**. Although all types of chronic hepatitis share certain clinical, laboratory, and histopathologic features, chronic viral and chronic autoimmune hepatitis are sufficiently distinct to merit separate discussions. **For discussion of acute hepatitis, see Chap. 332.**

CLASSIFICATION OF CHRONIC HEPATITIS

Common to all forms of chronic hepatitis are histopathologic distinctions based on localization and extent of liver injury. These vary from the milder forms, previously labeled *chronic persistent hepatitis* and *chronic lobular hepatitis*, to the more severe form, formerly called *chronic active hepatitis*. When first defined, these designations were believed to have prognostic implications, which were not corroborated by subsequent observations. Categorization of chronic hepatitis based primarily on histopathologic features has been replaced by a more informative classification based on a combination of clinical, serologic, and histologic variables. Classification of chronic hepatitis is based on (1) its *cause*; (2) its histologic activity, or *grade*; and (3) its degree of progression based on level of fibrosis, or *stage*. Thus, neither clinical features alone nor histologic features—requiring liver biopsy or noninvasive markers of fibrosis—alone are sufficient to characterize and distinguish among the several categories of chronic hepatitis.

CLASSIFICATION BY CAUSE

Clinical and serologic features allow the establishment of a diagnosis of *chronic viral hepatitis*, caused by hepatitis B, hepatitis B plus D, or hepatitis C; *autoimmune hepatitis*, including several subcategories, I and II, based on serologic distinctions; *drug-associated chronic hepatitis*; and a category of unknown cause, or *cryptogenic chronic hepatitis* **(Table 334-1)**. These are addressed in more detail below.

CLASSIFICATION BY GRADE

Grade, a histologic assessment of necroinflammatory activity, is based on examination of the liver biopsy. An assessment of important histologic features includes the degree of *periportal necrosis* and the disruption of the limiting plate of periportal hepatocytes by inflammatory cells (so-called *piecemeal necrosis* or *interface hepatitis*); the degree of confluent necrosis that links or forms bridges between vascular structures—between portal tract and portal tract or even more important bridges between portal tract and central vein—referred to as *bridging necrosis*; the degree of hepatocyte degeneration and focal necrosis within the lobule; and the degree of *portal inflammation*. Several scoring systems that take these histologic features into account have been devised, and the most popular are the histologic activity index (HAI), used commonly in the United States, and the METAVIR score, used in Europe **(Table 334-2)**. Based on the presence and degree of these features of histologic activity, chronic hepatitis can be graded as mild, moderate, or severe.

CLASSIFICATION BY STAGE

The stage of chronic hepatitis, which reflects the level of progression of the disease, is based on the degree of hepatic fibrosis. When fibrosis is so extensive that fibrous septa surround parenchymal nodules and alter the normal architecture of the liver lobule, the histologic lesion is defined as *cirrhosis*. Staging is based on the degree of fibrosis as categorized on a numerical scale 0–6 (HAI) or 0–4 (METAVIR) (Table 334-2). Several noninvasive approaches have been introduced to provide approximations of hepatic histologic stage, including serum biomarkers of fibrosis and imaging determinations of liver elasticity.

CHRONIC VIRAL HEPATITIS

Both the enterically transmitted forms of viral hepatitis, hepatitis A and E, are self-limited and do not cause chronic hepatitis (rare reports notwithstanding in which acute hepatitis A serves as a trigger for the onset of autoimmune hepatitis in genetically susceptible patients or in which hepatitis E **(Chap. 332)** can cause chronic liver disease in immunosuppressed hosts, for example, after liver transplantation). In contrast, the

TABLE 334-1 Clinical and Laboratory Features of Chronic Hepatitis

TYPE OF HEPATITIS	DIAGNOSTIC TEST(s)	AUTOANTIBODIES	THERAPY
Chronic hepatitis B	HBsAg, IgG anti-HBc, HBeAg, HBV DNA	Uncommon	IFN-α, PEG IFN-α Oral agents: First-line: entecavir, tenofovir Second-line: lamivudine, adefovir, telbivudine
Chronic hepatitis C	Anti-HCV, HCV RNA	Anti-LKM1[a]	PEG IFN-α plus ribavirin[b] Direct-acting oral agents: sofosbuvir, ledipasvir, velpatasvir ritonavir-boosted paritaprevir, ombitasvir, dasabavir elbasvir, grazoprevir daclatasvir, simeprevir
Chronic hepatitis D	Anti-HDV, HDV RNA, HBsAg, IgG anti-HBc	Anti-LKM3	IFN-α, PEG IFN-α[c]
Autoimmune hepatitis	ANA[d] (homogeneous), anti-LKM1 (±) Hyperglobulinemia	ANA, anti-LKM1 anti-SLA[e]	Prednisone, azathioprine
Drug-associated	—	Uncommon	Withdraw drug
Cryptogenic	All negative	None	Prednisone (?), azathioprine (?)

[a]Antibodies to liver-kidney microsomes type 1 (autoimmune hepatitis type II and some cases of hepatitis C). [b]Supplanted in almost all cases by combinations of the direct-acting antiviral agents listed (see *www.hcvguidelines.org*). [c]Early clinical trials suggested benefit of IFN-α therapy; PEG IFN-α is as effective, if not more so, and has supplanted standard IFN-α. [d]Antinuclear antibody (autoimmune hepatitis type I). [e]Antibodies to soluble liver antigen (autoimmune hepatitis type III).

Abbreviations: HBc, hepatitis B core; HBeAg, hepatitis B e antigen; HBsAg, hepatitis B surface antigen; HBV, hepatitis B virus; HCV, hepatitis C virus; HDV, hepatitis D virus; IFN-α, interferon α; IgG, immunoglobulin G; LKM, liver-kidney microsome; PEG IFN-α, pegylated interferon α; SLA, soluble liver antigen.

TABLE 334-2 Histologic Grading and Staging of Chronic Hepatitis

		HISTOLOGIC ACTIVITY INDEX (HAI)[a]		METAVIR[b]	
HISTOLOGIC FEATURE		**SEVERITY**	**SCORE**	**SEVERITY**	**SCORE**
Necroinflammatory Activity (grade)					
Periportal necrosis, including piecemeal necrosis and/or bridging necrosis (BN)		None	0	None	0
		Mild	1	Mild	1
		Mild/moderate	2	Moderate	2
		Moderate	3	Severe	3
		Severe	4		
				Bridging necrosis	Yes
					No
Intralobular necrosis	Confluent	—None	0	None or mild	0
		—Focal	1	Moderate	1
		—Zone 3 some	2	Severe	2
		—Zone 3 most	3		
		—Zone 3 + BN few	4		
		—Zone 3 + BN multiple	5		
		—Panacinar/multiacinar	6		
	Focal	—None	0		
		—≤1 focus/10× field	1		
		—2–4 foci/10× field	2		
		—5–10 foci/10× field	3		
		—>10 foci/10× field	4		
Portal Inflammation		None	0		
		Mild	1		
		Moderate	2		
		Moderate/marked	3		
		Marked	4		
		Total	0–18	A0–A3[c]	
Fibrosis (stage)					
None			0	F0	
Portal fibrosis—some			1	F1	
Portal fibrosis—most			2	F1	
Bridging fibrosis—few			3	F2	
Bridging fibrosis—many			4	F3	
Incomplete cirrhosis			5	F4	
Cirrhosis			6	F4	
		Total	6	4	

[a]Ishak K, Baptista A, Bianchi L, et al: Histologic grading and staging of chronic hepatitis. J Hepatol 22:696, 1995. [b]Bedossa P, Poynard T, French METAVIR Cooperative Study Group: An algorithm for grading activity in chronic hepatitis C. Hepatology 24:289, 1996. [c]Necroinflammatory grade: A0 = none; A1 = mild; A2 = moderate; A3 = severe.

entire clinicopathologic spectrum of chronic hepatitis occurs in patients with chronic viral hepatitis B and C as well as in patients with chronic hepatitis D superimposed on chronic hepatitis B.

CHRONIC HEPATITIS B

The likelihood of chronicity after acute hepatitis B varies as a function of age. Infection at birth is associated with clinically silent acute infection but a 90% chance of chronic infection, whereas infection in young adulthood in immunocompetent persons is typically associated with clinically apparent acute hepatitis but a risk of chronicity of only ~1%. Most cases of chronic hepatitis B among adults, however, occur in patients who never had a recognized episode of clinically apparent acute viral hepatitis. The degree of liver injury (grade) in patients with chronic hepatitis B is variable, ranging from none in inactive carriers to mild to moderate to severe. Among adults with chronic hepatitis B, histologic features are of prognostic importance. In one long-term study of patients with chronic hepatitis B, investigators found a 5-year survival rate of 97% for patients with mild chronic hepatitis, 86% for patients with moderate to severe chronic hepatitis, and only 55% for patients with chronic hepatitis and postnecrotic cirrhosis. The 15-year survival in these cohorts was 77%, 66%, and 40%, respectively. On the other hand, more recent observations do not allow us to be so sanguine about the prognosis in patients with mild chronic hepatitis; among such patients followed for 1–13 years, progression to more severe chronic hepatitis and cirrhosis has been observed in more than a quarter of cases.

More important to consider than histology alone in patients with chronic hepatitis B is the degree of hepatitis B virus (HBV) replication. As reviewed in **Chap. 332**, chronic HBV infection can occur in the presence or absence of serum hepatitis B e antigen (HBeAg), and generally, for both HBeAg-reactive and HBeAg-negative chronic hepatitis B, the level of HBV DNA correlates with the level of liver injury and risk of progression. In *HBeAg-reactive chronic hepatitis B*, two phases have been recognized based on the relative level of HBV replication. The relatively *replicative phase* is characterized by the presence in the serum of HBeAg and HBV DNA levels well in excess of 10^3–10^4 IU/mL, sometimes exceeding 10^9 IU/mL; by the presence in the liver of detectable intrahepatocyte nucleocapsid antigens (primarily hepatitis B core antigen [HBcAg]); by high infectivity; and by accompanying liver injury. In contrast, the relatively *nonreplicative phase* is characterized by the absence of the conventional serum marker of HBV replication (HBeAg), the appearance of anti-HBe, levels of HBV DNA below a threshold of ~10^3 IU/mL, the absence of intrahepatocytic HBcAg, limited infectivity, and minimal liver injury. Patients in the relatively replicative phase tend to have more severe chronic hepatitis, whereas those in the relatively nonreplicative phase tend to have minimal or mild chronic hepatitis or to be inactive hepatitis B carriers. The likelihood in a patient with HBeAg-reactive chronic hepatitis B of converting spontaneously from relatively replicative to nonreplicative infection is ~10% per year. Distinctions in HBV replication and in histologic category, however, do not always coincide. In patients with HBeAg-reactive chronic HBV infection, especially when acquired at birth or in early childhood, as recognized commonly in Asian countries, a dichotomy is common between very high levels of HBV replication during the early decades of life (when the level of apparent host immunologic tolerance of HBV is relatively high) and negligible levels of liver injury. Yet despite the relatively immediate, apparently benign nature of liver disease for many decades in this population, in the middle decades, activation of liver injury emerges as what appears to be the relative tolerance of the host to HBV declines, and these patients with childhood-acquired HBV infection are ultimately at increased risk later in life of cirrhosis, hepatocellular carcinoma (HCC) **(Chap. 78)**, and liver-related death. **A discussion of the pathogenesis of liver injury in patients with chronic hepatitis B appears in Chap. 332.**

HBeAg-negative chronic hepatitis B (i.e., chronic HBV infection with active virus replication, readily detectable HBV DNA but without HBeAg [anti-HBe-reactive]), is more common than HBeAg-reactive chronic hepatitis B in Mediterranean and European countries and in Asia (and, correspondingly, in HBV genotypes other than A). Compared to patients with HBeAg-reactive chronic hepatitis B, patients with HBeAg-negative chronic hepatitis B have HBV DNA levels several orders of magnitude lower (no more than 10^5–10^6 IU/mL) than those observed in the HBeAg-reactive subset. Most such cases represent precore or core-promoter mutations acquired

late in the natural history of the disease (mostly early-life onset; age range 40–55 years, older than that for HBeAg-reactive chronic hepatitis B); these mutations prevent translation of HBeAg from the precore component of the HBV genome (precore mutants) or are characterized by down-regulated transcription of precore mRNA (core-promoter mutants; **Chap. 332**). Although their levels of HBV DNA tend to be lower than among patients with HBeAg-reactive chronic hepatitis B, patients with HBeAg-negative chronic hepatitis B can have progressive liver injury (complicated by cirrhosis and HCC) and experience episodic reactivation of liver disease reflected in fluctuating levels of aminotransferase activity ("flares"). The biochemical and histologic activity of HBeAg-negative disease tends to correlate closely with levels of HBV replication, unlike the case mentioned above of Asian patients with HBeAg-reactive chronic hepatitis B during the early decades of their HBV infection. Worth reiterating, the level of HBV replication is the most important risk factor for the ultimate development of cirrhosis and HCC in both HBeAg-reactive (beyond the early decades of "relatively nonreplicative" infection) and HBeAg-negative patients. Although levels of HBV DNA are lower and more readily suppressed by therapy to undetectable levels in HBeAg-negative (compared to HBeAg-reactive) chronic hepatitis B, achieving sustained responses that permit discontinuation of antiviral therapy is less likely in HBeAg-negative patients (see below). Inactive carriers are patients with circulating hepatitis B surface antigen (HBsAg), normal serum aminotransferase levels, undetectable HBeAg, and levels of HBV DNA that are either undetectable or present at a threshold of $\leq 10^3$ IU/mL. This serologic profile occurs not only in inactive carriers but also in patients with HBeAg-negative chronic hepatitis B during periods of relative inactivity; distinguishing between the two requires sequential biochemical and virologic monitoring over many months.

The spectrum of *clinical features* of chronic hepatitis B is broad, ranging from asymptomatic infection to debilitating disease or even end-stage, fatal hepatic failure. As noted above, the onset of the disease tends to be insidious in most patients, with the exception of the very few in whom chronic disease follows failure of resolution of clinically apparent acute hepatitis B. **The clinical and laboratory features associated with progression from acute to chronic hepatitis B are discussed in Chap. 332.**

Fatigue is a common symptom, and persistent or intermittent *jaundice* is a common feature in severe or advanced cases. Intermittent deepening of jaundice and recurrence of malaise and anorexia, as well as worsening fatigue, are reminiscent of acute hepatitis; such exacerbations may occur spontaneously, often coinciding with evidence of virologic reactivation; may lead to progressive liver injury; and, when superimposed on well-established cirrhosis, may cause hepatic decompensation. Complications of cirrhosis occur in end-stage chronic hepatitis and include ascites, edema, bleeding gastroesophageal varices, hepatic encephalopathy, coagulopathy, and hypersplenism. Occasionally, these complications bring the patient to initial clinical attention. Extrahepatic complications of chronic hepatitis B, similar to those seen during the prodromal phase of acute hepatitis B, are associated with tissue deposition of circulating hepatitis B antigen–antibody immune complexes. These include arthralgias and arthritis, which are common, and the more rare purpuric cutaneous lesions (leukocytoclastic vasculitis), immune-complex glomerulonephritis, and generalized vasculitis (polyarteritis nodosa) **(Chaps. 332 and 356)**.

Laboratory features of chronic hepatitis B do not distinguish adequately between histologically mild and severe hepatitis. Aminotransferase elevations tend to be modest for chronic hepatitis B but may fluctuate in the range of 100–1000 units. As is true for acute viral hepatitis B, alanine aminotransferase (ALT) tends to be more elevated than aspartate aminotransferase (AST); however, once cirrhosis is established, AST tends to exceed ALT. Levels of alkaline phosphatase activity tend to be normal or only marginally elevated. In severe cases, moderate elevations in serum bilirubin (51.3–171 μmol/L [3–10 mg/dL]) occur. Hypoalbuminemia and prolongation of the prothrombin time occur in severe or end-stage cases. Hyperglobulinemia and detectable circulating autoantibodies are distinctly absent in chronic hepatitis B (in contrast to autoimmune hepatitis). **Viral markers of chronic HBV infection are discussed in Chap. 332.**

TREATMENT

Chronic Hepatitis B

Although progression to cirrhosis is more likely in severe than in mild or moderate chronic hepatitis B, all forms of chronic hepatitis B can be progressive, and progression occurs primarily in patients with active HBV replication. Moreover, in populations of patients with chronic hepatitis B who are at risk for HCC **(Chap. 78)**, the risk is highest for those with continued, high-level HBV replication and lower for persons in whom initially high-level HBV DNA falls spontaneously over time. Therefore, management of chronic hepatitis B is directed at suppressing the level of virus replication. Although clinical trials tend to focus on clinical endpoints achieved over 1–2 years (e.g., suppression of HBV DNA to undetectable levels, loss of HBeAg/HBsAg, improvement in histology, normalization of ALT), these short-term gains translate into reductions in the risk of clinical progression, hepatic decompensation, HCC, liver transplantation, and death; regression of cirrhosis and of esophageal varices have been documented to follow long-term pharmacologic suppression of HBV replication. In addition, restoration of impaired HBV-specific T-cell function has been shown following successful suppression of HBV replication with antiviral therapy. To date, seven drugs have been approved for treatment of chronic hepatitis B: injectable interferon (IFN) α and pegylated interferon (long-acting IFN bound to polyethylene glycol, PEG [PEG IFN]) and the oral agents lamivudine, adefovir dipivoxil, entecavir, telbivudine, and tenofovir disoproxil fumarate (TDF).

Antiviral therapy for hepatitis B has evolved rapidly since the mid-1990s, as has the sensitivity of tests for HBV DNA. When IFN and lamivudine were evaluated in clinical trials, HBV DNA was measured by insensitive hybridization assays with detection thresholds of 10^5–10^6 virions/mL; when adefovir, entecavir, telbivudine, tenofovir, and PEG IFN were studied in clinical trials, HBV DNA was measured by sensitive amplification assays (polymerase chain reaction [PCR]) with detection thresholds of 10^1–10^3 viral copies/mL or IU/mL. Recognition of these distinctions is helpful when comparing results of clinical trials that established the efficacy of these therapies (reviewed below in chronological order of publication of these efficacy trials).

INTERFERON

IFN-α was the first approved therapy (1992) for chronic hepatitis B. Although it is no longer used to treat hepatitis B, standard IFN is important historically, having provided important lessons about antiviral therapy in general. For immunocompetent adults with HBeAg-reactive chronic hepatitis B (who tend to have high-level HBV DNA [$>10^5$–10^6 virions/mL] and histologic evidence of chronic hepatitis on liver biopsy), a 16-week course of IFN given subcutaneously at a daily dose of 5 million units, or three times a week at a dose of 10 million units, resulted in a loss of HBeAg and hybridization-detectable HBV DNA (i.e., a reduction to levels below 10^5–10^6 virions/mL) in ~30% of patients, with a concomitant improvement in liver histology. Seroconversion from HBeAg to anti-HBe occurred in ~20%, and, in early trials, ~8% lost HBsAg. Successful IFN therapy and seroconversion were often accompanied by an acute hepatitis-like elevation in aminotransferase activity, postulated to result from enhanced cytolytic T cell clearance of HBV-infected hepatocytes. Relapse after successful therapy was rare (1 or 2%). The likelihood of responding to IFN was higher in patients with lower levels of HBV DNA and substantial elevations of ALT. Although children can respond as well as adults, IFN therapy was not effective in very young children infected at birth. Similarly, IFN therapy was not effective in immunosuppressed persons, Asian patients with neonatal acquisition of infection and minimal-to-mild ALT elevations, or patients with decompensated chronic hepatitis B (in whom such therapy was actually detrimental, sometimes precipitating decompensation, often associated with severe adverse effects). Among patients with HBeAg loss during therapy, long-term follow-up

demonstrated that 80% experienced eventual loss of HBsAg (i.e., all serologic markers of infection, and normalization of ALT over a 9-year posttreatment period). In addition, improved long-term and complication-free survival as well as a reduction in the frequency of HCC were documented among IFN responders, supporting the conclusion that successful antiviral therapy improves the natural history of chronic hepatitis B.

Initial trials of brief-duration IFN therapy in patients with *HBeAg-negative chronic hepatitis B* were disappointing, suppressing HBV replication transiently during therapy but almost never resulting in sustained antiviral responses. In subsequent IFN trials among patients with HBeAg-negative chronic hepatitis B, however, more protracted courses, lasting up to 1.5 years, were reported to result in sustained remissions documented to last for several years, with suppressed HBV DNA and aminotransferase activity, in ~20%.

Complications of IFN therapy include systemic "flu-like" symptoms; marrow suppression; emotional lability (irritability, depression, anxiety); autoimmune reactions (especially autoimmune thyroiditis); and miscellaneous side effects such as alopecia, rashes, diarrhea, and numbness and tingling of the extremities. With the possible exception of autoimmune thyroiditis, all these side effects are reversible upon dose lowering or cessation of therapy.

Although no longer competitive with the newer generation of antivirals, IFN did represent the first successful antiviral approach and set a standard against which to measure subsequent drugs in the achievement of durable virologic, serologic, biochemical, and histologic responses; consolidation of virologic and biochemical benefit in the ensuing years after therapy; and improvement in the natural history of chronic hepatitis B. Standard IFN has been supplanted by long-acting PEG IFN (see below), and IFN nonresponders are now treated with one of the newer oral nucleoside analogues.

LAMIVUDINE

The first of the nucleoside analogues to be approved (in 1998) for hepatitis B, the dideoxynucleoside lamivudine inhibits reverse transcriptase activity of both HIV and HBV and is an effective agent for patients with chronic hepatitis B. Although generally superseded by newer, more potent, less resistance-prone agents, lamivudine is still used in regions of the world where newer agents are not yet available or affordable. In clinical trials among patients with HBeAg-reactive chronic hepatitis B, lamivudine therapy at daily doses of 100 mg for 48–52 weeks suppressed HBV DNA by a median of ~5.5 $\log_{10}$ copies/mL and to undetectable levels, as measured by PCR amplification assays, in ~40% of patients. Therapy was associated with HBeAg loss in 32–33%, HBeAg seroconversion (i.e., conversion from HBeAg-reactive to anti-HBe-reactive) in 16–21%, normalization of ALT in 40–75%, improvement in histology in 50–60%, retardation in hepatic fibrosis in 20–30%, and prevention of progression to cirrhosis. HBeAg responses occur even in patients resistant to IFN (e.g., those with high-level HBV DNA) or who failed in the past to respond to it. As is true for IFN therapy of chronic hepatitis B, patients with near-normal ALT activity tend not to experience HBeAg responses (despite suppression of HBV DNA), whereas those with ALT levels exceeding 5 × the upper limit of normal can expect 1-year HBeAg seroconversion rates of 50–60%. Generally, HBeAg seroconversions are confined to patients who achieve suppression of HBV DNA to $<10^4$ copies/mL (equivalent to ~10^3 IU/mL). Lamivudine-associated HBeAg responses are accompanied by a delayed posttreatment HBsAg seroconversion rate comparable to that seen after IFN-induced HBeAg responses. Among Western patients who undergo HBeAg responses during a year-long course of therapy and in whom the response is sustained for 4–6 months after cessation of therapy, the response is durable thereafter in the vast majority (>80%); therefore, the achievement of an HBeAg response represents a viable stopping point in therapy. Reduced durability has been reported in Asian patients; therefore, to support the durability of HBeAg responses, patients should receive a period of consolidation therapy of ≥6 months in Western patients and ≥1 year in Asian patients after HBeAg seroconversion (see treatment guidelines below). Close posttreatment monitoring is necessary to identify HBV reactivation promptly and to resume therapy. If HBeAg is unaffected by lamivudine therapy, the current approach is to continue therapy until an HBeAg response occurs, but long-term therapy may be required to suppress HBV replication and, in turn, limit liver injury; HBeAg seroconversions can increase to a level of 50% after 5 years of therapy. Histologic improvement continues to accrue with therapy beyond the first year; after a cumulative course of 3 years of lamivudine therapy, necroinflammatory activity is reduced in the majority of patients, and even cirrhosis has been shown to regress to precirrhotic stages in as many as three-quarters of patients.

Losses of HBsAg have been few during the first year of lamivudine therapy, and this observation had been cited as an advantage of IFN-based over lamivudine therapy; however, in head-to-head comparisons between standard IFN and lamivudine monotherapy, HBsAg losses were rare in both groups. Trials in which lamivudine and IFN were administered in combination failed to show a benefit of combination therapy over lamivudine monotherapy for either treatment-naïve patients or prior IFN nonresponders.

In patients with *HBeAg-negative chronic hepatitis B* (i.e., in those with precore and core-promoter HBV mutations), 1 year of lamivudine therapy results in HBV DNA suppression and normalization of ALT in three-quarters of patients and in histologic improvement in approximately two-thirds. Therapy has been shown to suppress HBV DNA by ~4.5 $\log_{10}$ copies/mL (baseline HBV DNA levels are lower than in patients with HBeAg-reactive hepatitis B) and to undetectable levels in ~70%, as measured by sensitive PCR amplification assays. Lacking HBeAg at the outset, patients with HBeAg-negative chronic hepatitis B cannot achieve an HBeAg response—a stopping point in HBeAg-reactive patients; almost invariably, when therapy is discontinued, reactivation is the rule. Therefore, these patients require long-term therapy; with successive years, the proportion with suppressed HBV DNA and normal ALT increases.

Clinical and laboratory side effects of lamivudine are negligible and indistinguishable from those observed in placebo recipients. Still, lamivudine doses should be reduced in patients with reduced creatinine clearance. During lamivudine therapy, transient ALT elevations, resembling those seen during IFN therapy and during spontaneous HBeAg-to-anti-HBe seroconversions, occur in one-fourth of patients. These ALT elevations may result from restored cytolytic T cell activation permitted by suppression of HBV replication. Similar ALT elevations, however, occurred at an identical frequency in placebo recipients; however, ALT elevations associated with HBeAg seroconversion in clinical trials were confined to lamivudine-treated patients. When therapy is stopped after a year of therapy, two- to threefold ALT elevations occur in 20–30% of lamivudine-treated patients, representing renewed liver-cell injury as HBV replication returns. Although these posttreatment flares are almost always transient and mild, rare severe exacerbations, especially in cirrhotic patients, have been observed, mandating close and careful clinical and virologic monitoring after discontinuation of treatment. Many authorities caution against discontinuing therapy in patients with cirrhosis, in whom posttreatment flares could precipitate decompensation.

Long-term monotherapy with lamivudine is associated with methionine-to-valine (M204V) or methionine-to-isoleucine (M204I) mutations, primarily at amino acid 204 in the tyrosine-methionine-aspartate-aspartate (YMDD) motif of the C domain of HBV DNA polymerase, analogous to mutations that occur in HIV-infected patients treated with this drug. During a year of therapy, YMDD mutations occur in 15–30% of patients; the frequency increases with each year of therapy, reaching 70% at year 5. Ultimately, patients with YMDD mutants experience degradation of clinical, biochemical, and histologic responses; therefore, if treatment is begun with lamivudine monotherapy, the emergence of lamivudine resistance, reflected clinically by a breakthrough from suppressed levels of HBV DNA and ALT, is managed by adding another antiviral to which YMDD variants are sensitive (e.g., adefovir, tenofovir; see below).

Currently, although lamivudine is very safe and still used widely in other parts of the world, in the United States and Europe, lamivudine has been eclipsed by more potent antivirals that have superior resistance profiles (see below); it is no longer recommended as first-line therapy. Still, as the first successful oral antiviral agent for use in hepatitis B, lamivudine provided proof of principle that polymerase inhibitors can achieve virologic, serologic, biochemical, and histologic benefits. In addition, lamivudine has been shown to be effective in the treatment of patients with decompensated hepatitis B (for whom IFN is contraindicated), in some of whom decompensation can be reversed. Moreover, among patients with cirrhosis or advanced fibrosis, lamivudine has been shown to be effective in reducing the risk of progression to hepatic decompensation and, marginally, the risk of HCC. In the half decade following the introduction in the United States of lamivudine therapy for hepatitis B, referral of patients with HBV-associated end-stage liver disease for liver transplantation was reduced by ~30%, supporting further the beneficial impact of oral antiviral therapy on the natural history of chronic hepatitis B.

Because lamivudine monotherapy can result universally in the rapid emergence of YMDD variants in persons with HIV infection, patients with chronic hepatitis B should be tested for anti-HIV prior to therapy; if HIV infection is identified, lamivudine monotherapy at the HBV daily dose of 100 mg is contraindicated. These patients should be treated for both HIV and HBV with an HIV drug regimen that includes or is supplemented by at least two drugs active against HBV; antiretroviral therapy (ART) often contains two drugs with antiviral activity against HBV (e.g., tenofovir and emtricitabine), but if lamivudine is part of the regimen, the daily dose should be 300 mg **(Chap. 197)**. The safety of lamivudine during pregnancy has not been established; however, the drug is not teratogenic in rodents and has been used safely in pregnant women with HIV infection and with HBV infection. Administration of lamivudine during the last months of pregnancy to mothers with high-level hepatitis B viremia ($\geq 10^8$ IU/mL) can reduce the likelihood of perinatal transmission of hepatitis B.

ADEFOVIR DIPIVOXIL

At an oral daily dose of 10 mg, the acyclic nucleotide analogue adefovir dipivoxil, the prodrug of adefovir (approved for hepatitis B in 2002), reduces HBV DNA by ~3.5–4 $\log_{10}$ copies/mL and is equally effective in treatment-naïve patients and prior IFN nonresponders. In HBeAg-reactive chronic hepatitis B, a 48-week course of adefovir dipivoxil was shown to achieve histologic improvement (and reduce the progression of fibrosis) and normalization of ALT in just over one-half of patients, HBeAg seroconversion in 12%, HBeAg loss in 23%, and suppression to an undetectable level of HBV DNA in 13–21%, as measured by PCR. Similar to IFN and lamivudine, adefovir dipivoxil is more likely to achieve an HBeAg response in patients with high baseline ALT; among adefovir-treated patients with ALT level >5 × the upper limit of normal, HBeAg seroconversions occurred in 25%. The durability of adefovir-induced HBeAg responses is high (91% in one study); therefore, HBeAg response can be relied upon as a stopping point for adefovir therapy, after a period of consolidation therapy, as outlined above. Although data on the impact of additional therapy beyond 1 year are limited, biochemical, serologic, and virologic outcomes improve progressively as therapy is continued.

In patients with *HBeAg-negative chronic hepatitis B*, a 48-week course of 10 mg/d of adefovir dipivoxil resulted in histologic improvement in two-thirds, normalization of ALT in three-fourths, and suppression of HBV DNA to PCR-undetectable levels in one-half to two-thirds. As was true for lamivudine, because HBeAg responses—a potential stopping point—cannot be achieved in this group, reactivation is the rule when adefovir therapy is discontinued, and indefinite, long-term therapy is required. Treatment beyond the first year consolidates the gain of the first year; after 5 years of therapy, improvement in hepatic inflammation and regression of fibrosis were observed in three-fourths of patients, ALT was normal in 70%, and HBV DNA was undetectable in almost 70%. In one study, stopping adefovir after 5 years was followed by sustained suppression of HBV DNA and ALT, but most HBeAg-negative patients are treated indefinitely unless HBsAg loss, albeit very rare, is achieved.

Adefovir contains a flexible acyclic linker instead of the L-nucleoside ring of lamivudine, avoiding steric hindrance by mutated amino acids. In addition, the molecular structure of phosphorylated adefovir is very similar to that of its natural substrate; therefore, mutations to adefovir would also affect binding of the natural substrate, dATP. Hypothetically, these are among the reasons that resistance to adefovir dipivoxil is much less likely than resistance to lamivudine; no resistance was encountered in 1 year of clinical trial therapy. In subsequent years, however, adefovir resistance begins to emerge (asparagine to threonine at amino acid 236 [N236T] and alanine to valine or threonine at amino acid 181 [A181V/T], primarily), occurring in 2.5% after 2 years, but in 29% after 5 years of therapy (reported in HBeAg-negative patients). Among patients co-infected with HBV and HIV and who have normal CD4+ T cell counts, adefovir dipivoxil is effective in suppressing HBV dramatically (by 5 $\log_{10}$ in one study). Moreover, adefovir dipivoxil is effective in lamivudine-resistant, YMDD-mutant HBV and can be used when such lamivudine-induced variants emerge. When lamivudine resistance occurs, adding adefovir (i.e., maintaining lamivudine to preempt the emergence of adefovir resistance) is superior to switching to adefovir. Almost invariably, patients with adefovir-induced HBV mutations respond to lamivudine (or newer agents, such as entecavir, see below). When, in the past, adefovir had been evaluated as therapy for HIV infection, doses of 60–120 mg were required to suppress HIV, and, at these doses, the drug was nephrotoxic. Even at 30 mg/d, creatinine elevations of 44 µmol/L (0.5 mg/dL) occurred in 10% of patients; however, at the HBV-effective dose of 10 mg, such creatinine elevations are rarely encountered. If any nephrotoxicity does occur, it rarely appears before 6–8 months of therapy. Although renal tubular injury is a rare potential side effect, and although creatinine monitoring is recommended during treatment, the therapeutic index of adefovir dipivoxil is high, and the nephrotoxicity observed in clinical trials at higher doses was reversible. For patients with underlying renal disease, frequency of administration of adefovir dipivoxil should be reduced to every 48 h for creatinine clearances of 30–49 mL/min; to every 72 h for creatinine clearances of 10–29 mL/min; and to once a week, following dialysis, for patients undergoing hemodialysis. Adefovir dipivoxil is very well tolerated, and ALT elevations during and after withdrawal of therapy are similar to those observed and described above in clinical trials of lamivudine. An advantage of adefovir is its relatively favorable resistance profile; however, it is not as potent as the other approved oral agents, it does not suppress HBV DNA as rapidly or as uniformly as the others, it is the least likely of all agents to result in HBeAg seroconversion, and 20–50% of patients fail to suppress HBV DNA by 2 $\log_{10}$ ("primary nonresponders"). For these reasons, adefovir, which has been supplanted in both treatment-naïve and lamivudine-resistant patients by the more potent, less resistance-prone nucleotide analogue tenofovir (see below), is no longer recommended as first-line therapy.

PEGYLATED IFN

After long-acting PEG IFN was shown to be effective in the treatment of hepatitis C (see below), this more convenient drug was evaluated in the treatment of chronic hepatitis B. Once-a-week PEG IFN is more effective than the more frequently administered, standard IFN, and several large-scale trials of PEG IFN versus oral nucleoside analogues were conducted among patients with HBeAg-reactive and HBeAg-negative chronic hepatitis B.

In HBeAg-reactive chronic hepatitis B, two large-scale studies were done. In one study, PEG IFN-α 2b (100 µg weekly for 32 weeks, then 50 µg weekly for another 20 weeks for a total of 52 weeks) was evaluated against a comparison arm of combination PEG IFN with oral lamivudine in 307 subjects. The other study involved PEG IFN-α 2a (180 µg weekly for 48 weeks) in 814 primarily Asian patients, three-fourths of whom had ALT ≥2 × the upper limit of normal, with

comparison arms of lamivudine monotherapy and combination PEG IFN plus lamivudine. At the end of therapy (48–52 weeks) in the PEG IFN monotherapy arms, HBeAg loss occurred in ~30%, HBeAg seroconversion in 22–27%, undetectable HBV DNA (<400 copies/mL by PCR) in 10–25%, normal ALT in 34–39%, and a mean reduction in HBV DNA of 2 $\log_{10}$ copies/mL (PEG IFN-α 2b) to 4.5 $\log_{10}$ copies/mL (PEG IFN-α 2a). Six months after completing PEG IFN monotherapy in these trials, HBeAg losses were present in ~35%, HBeAg seroconversion in ~30%, undetectable HBV DNA in 7–14%, normal ALT in 32–41%, and a mean reduction in HBV DNA of 2–2.4 $\log_{10}$ copies/mL. Although the combination of PEG IFN and lamivudine was superior at the end of therapy in one or more serologic, virologic, or biochemical outcomes, neither the combination arm (in both studies) nor the lamivudine monotherapy arm (in the PEG IFN-α 2a trial) demonstrated any benefit compared to the PEG IFN monotherapy arms 6 months after therapy. Moreover, HBsAg seroconversion occurred in 3–7% of PEG IFN recipients (with or without lamivudine); some of these seroconversions were identified by the end of therapy, but many were identified during the post-treatment follow-up period. The likelihood of HBeAg loss in PEG IFN-treated HBeAg-reactive patients is associated with HBV genotype A > B > C > D (shown for PEG IFN-α2b but not for α-2a). PEG IFN-α 2a was approved in the US for hepatitis B in 2005; PEG IFN-α 2b, not approved for hepatitis B in the US, is used in other countries.

Based on these results, some authorities concluded that PEG IFN monotherapy should be the first-line therapy of choice in HBeAg-reactive chronic hepatitis B; however, this conclusion has been challenged. Although a finite, 1-year course of PEG IFN results in a higher rate of sustained response (6 months after treatment) than is achieved with oral nucleoside/nucleotide analogue therapy, the comparison is confounded by the fact that oral agents are not discontinued at the end of 1 year. Instead, taken orally and free of side effects, therapy with oral agents is extended indefinitely or until after the occurrence of an HBeAg response. The rate of HBeAg responses after 2 years of oral-agent nucleoside analogue therapy is at least as high as, if not higher than, that achieved with PEG IFN after 1 year; favoring oral agents is the absence of injections, difficult-to-tolerate side effects, and laboratory monitoring as well as lower direct and indirect medical care costs and inconvenience. The association of HBsAg responses with PEG IFN therapy occurs in such a small proportion of patients that subjecting everyone to PEG IFN for the marginal gain of HBsAg responses during or immediately after therapy in such a very small minority is questionable. Moreover, HBsAg responses occur in a comparable proportion of patients treated with early-generation nucleoside/nucleotide analogues in the years *after* therapy, and, with the newer, more potent nucleoside analogues, the frequency of HBsAg loss during the first year of therapy equals that of PEG IFN and is exceeded during year 2 and beyond (see below). Of course, resistance is not an issue during PEG IFN therapy, but the risk of resistance is much lower with new agents (≤1% up to 3–8 years in previously treatment-naïve, entecavir-treated and 0% of tenofovir-treated patients; see below). Finally, the level of HBV DNA inhibition that can be achieved with the newer agents, and even with lamivudine, exceeds that which can be achieved with PEG IFN, in some cases by several orders of magnitude.

In HBeAg-negative chronic hepatitis B, a trial of PEG IFN-α 2a (180 μg weekly for 48 weeks versus comparison arms of lamivudine monotherapy and of combination therapy) in 564 patients showed that PEG IFN monotherapy resulted at the end of therapy in suppression of HBV DNA by a mean of 4.1 $\log_{10}$ copies/mL, undetectable HBV DNA (<400 copies/mL by PCR) in 63%, normal ALT in 38%, and loss of HBsAg in 4%. Although lamivudine monotherapy and combination lamivudine–PEG IFN therapy were both superior to PEG IFN at the end of therapy, no advantage of lamivudine monotherapy or combination therapy was apparent over PEG IFN monotherapy 6 months after therapy—suppression of HBV DNA by a mean of 2.3 $\log_{10}$ copies/mL, undetectable HBV DNA in 19%, and normal ALT in 59%. In subjects involved in this trial followed for up to 5 years, among the two-thirds followed who had been treated initially with PEG IFN, 17% maintained HBV DNA suppression to <400 copies/mL, but ALT remained normal in only 22%; HBsAg loss increased gradually to 12%. Among the half followed who had been treated initially with lamivudine monotherapy, HBV DNA remained <400 copies/mL in 7% and ALT normal in 16%; by year 5, 3.5% had lost HBsAg. As was the case for standard IFN therapy in HBeAg-negative patients, only a small proportion maintained responsiveness after completion of PEG IFN therapy, raising questions about the relative value of a finite period of PEG IFN, versus a longer course with a potent, low-resistance oral nucleoside analogue in these patients. Moreover, the value of PEG IFN for HBeAg-negative chronic hepatitis B has not been confirmed. In the only other controlled clinical trial of PEG IFN for HBeAg-negative chronic hepatitis B, the hepatitis C regimen of PEG IFN plus ribavirin was compared to PEG IFN monotherapy. In this trial, HBV DNA suppression (<400 copies/mL) occurred in only 7.5% of the two groups combined, and no study subject lost HBsAg.

In patients treated with PEG IFN, HBeAg and HBsAg responses have been associated with *IL28B* genotype CC, the favorable genotype identified in trials of PEG IFN for chronic hepatitis C. Also, reductions in quantitative HBsAg levels have been shown to correlate with and to be predictive of responsiveness to PEG IFN in chronic hepatitis B. If HBsAg levels fail to fall within the first 12–24 weeks or to reach <20,000 IU/mL by week 24, PEG IFN therapy is unlikely to be effective and should be discontinued. (Similar observations of HBsAg levels in oral-agent-treated patients are of interest, but of limited clinical relevance, given the very high likelihood of virologic responses during such therapy.)

ENTECAVIR

Entecavir, an oral cyclopentyl guanosine analogue polymerase inhibitor (approved 2005), appears to be the most potent of the HBV antivirals and is just as well tolerated as lamivudine. In a 709-subject clinical trial among HBeAg-reactive patients, oral entecavir, 0.5 mg daily, was compared to lamivudine, 100 mg daily. At 48 weeks, entecavir was superior to lamivudine in suppression of HBV DNA (mean 6.9 vs 5.5 $\log_{10}$ copies/mL), percentage with undetectable HBV DNA (<300 copies/mL by PCR; 67% vs 36%), histologic improvement (≥2-point improvement in necroinflammatory HAI score; 72% vs 62%), and normal ALT (68% vs 60%). The two treatments were indistinguishable in percentage with HBeAg loss (22% vs 20%) and seroconversion (21% vs 18%). Among patients treated with entecavir for 96 weeks, HBV DNA was undetectable cumulatively in 80% (vs 39% for lamivudine), and HBeAg seroconversions had occurred in 31% (vs 26% for lamivudine). After 3–6 years of entecavir, HBeAg seroconversions have been observed in 39–44% and HBsAg loss in 5–6%. Similarly, in a 638-subject clinical trial among HBeAg-negative patients, at week 48, oral entecavir, 0.5 mg daily, was superior to lamivudine, 100 mg daily, in suppression of HBV DNA (mean 5.0 vs 4.5 $\log_{10}$ copies/mL) and in percentage with undetectable HBV DNA (90% vs 72%), histologic improvement (70% vs 61%), and normal ALT (78% vs 71%). No resistance mutations were encountered in previously treatment-naïve, entecavir-treated patients during 96 weeks of therapy, and in a cohort of subjects treated for up to 6 years, resistance emerged in only 1.2%. Entecavir-induced HBeAg seroconversions are as durable as those achieved with other antivirals. Its high barrier to resistance coupled with its high potency renders entecavir a first-line drug for patients with chronic hepatitis B.

Entecavir is also effective against lamivudine-resistant HBV infection. In a trial of 286 lamivudine-resistant patients, entecavir, at a higher daily dose of 1 mg, was superior to lamivudine, as measured at week 48, in achieving suppression of HBV DNA (mean 5.1 vs 0.48 $\log_{10}$ copies/mL), undetectable HBV DNA (72% vs 19%), normal ALT (61% versus 15%), HBeAg loss (10% vs 3%), and HBeAg seroconversion (8% vs 3%). In this population of lamivudine-experienced patients, however, entecavir resistance emerged in 7% at 48 weeks. Although entecavir resistance requires both a YMDD mutation and a second mutation at one of several other sites

(e.g., T184A, S202G/I, or M250V), resistance to entecavir in lamivudine-resistant chronic hepatitis B has been recorded to increase progressively to 43% at 4 years and 57% at 6 years; therefore, entecavir is not as attractive a choice (and is not recommended, despite its approval for this indication) as adefovir or tenofovir for patients with lamivudine-resistant hepatitis B.

In clinical trials, entecavir had an excellent safety profile. In addition, on-treatment and posttreatment ALT flares are relatively uncommon and relatively mild in entecavir-treated patients. Doses should be reduced for patients with reduced creatinine clearance. Entecavir does have low-level antiviral activity against HIV and cannot be used as monotherapy to treat HBV infection in HIV/HBV co-infected persons.

TELBIVUDINE

Telbivudine, a cytosine analogue (approved 2006), is similar in efficacy to entecavir but slightly less potent in suppressing HBV DNA (a slightly less profound median 6.4 $\log_{10}$ reduction in HBeAg-reactive disease and a similar 5.2 $\log_{10}$ reduction in HBeAg-negative disease). In its registration trial, telbivudine at an oral daily dose of 600 mg suppressed HBV DNA to <300 copies/mL in 60% of HBeAg-positive and 88% of HBeAg-negative patients, reduced ALT to normal in 77% of HBeAg-positive and 74% of HBeAg-negative patients, and improved histology in 65% of HBeAg-positive and 67% of HBeAg-negative patients. Although resistance to telbivudine (M204I, not M204V, mutations) was less frequent than resistance to lamivudine at the end of 1 year, resistance mutations after 2 years of treatment occurred in up to 22%. Generally well tolerated, telbivudine has been associated with a low frequency of asymptomatic creatine kinase elevations and with a very low frequency of peripheral neuropathy; frequency of administration should be reduced for patients with impaired creatinine clearance. Its excellent potency notwithstanding, the inferior resistance and safety profile of telbivudine has limited its appeal; telbivudine is neither recommended as first-line therapy nor widely used.

TENOFOVIR

TDF, an acyclic nucleotide analogue and potent antiretroviral agent used to treat HIV infection (approved for hepatitis B in 2008), is similar to adefovir but more potent in suppressing HBV DNA and inducing HBeAg responses; it is highly active against both wild-type and lamivudine-resistant HBV and active in patients whose response to adefovir is slow and/or limited. At an oral once-daily dose of 300 mg for 48 weeks, tenofovir suppressed HBV DNA by 6.2 $\log_{10}$ (to undetectable levels [<400 copies/mL] in 76%) in HBeAg-positive patients and by 4.6 $\log_{10}$ (to undetectable levels in 93%) in HBeAg-negative patients; reduced ALT to normal in 68% of HBeAg-positive and 76% of HBeAg-negative patients; and improved histology in 74% of HBeAg-positive and 72% of HBeAg-negative patients. In HBeAg-positive patients, HBeAg seroconversions occurred in 21% by the end of year 1, 27% by year 2, 34% by year 3, and 40% by year 5 of tenofovir treatment; HBsAg loss occurred in 3% by the end of year 1 and 6% at year 2, and 8% by year 5. After 5 years of tenofovir therapy, 87% of patients experienced histologic improvement, including reduction in fibrosis score (51%) and regression of cirrhosis (71%). The 5-year safety (negligible renal toxicity, in 1%, and mild reduction in bone density, in ~0.5%) and resistance profiles (none recorded through 8 years) of tenofovir are very favorable as well; therefore, tenofovir has supplanted adefovir both as first-line therapy for chronic hepatitis B and as add-on therapy for lamivudine-resistant chronic hepatitis B. Studies of tenofovir and entecavir reviewed in 2015 showed no difference in long-term risks of renal and bone toxicity; however, among patients treated with tenofovir, instances of acute renal failure and of low blood phosphate levels have been reported. Thus, in patients receiving tenofovir, monitoring bone density is not recommended, but periodic (at least annual) monitoring for renal injury is (serum creatinine and phosphate, urine glucose and protein). Frequency of tenofovir administration should be reduced for patients with impaired creatinine clearance.

A comparison of the six antiviral therapies in current use appears in Table 334-3; their relative potencies in suppressing HBV DNA are shown in Fig. 334-1.

COMBINATION THERAPY

Although the combination of lamivudine and PEG IFN suppresses HBV DNA more profoundly during therapy than does monotherapy with either drug alone (and is much less likely to be associated with lamivudine resistance), this combination used for a year is no better than a year of PEG IFN in achieving sustained responses. To date, combinations of oral nucleoside/nucleotide agents have not achieved an enhancement in virologic, serologic, or biochemical efficacy over that achieved by the more potent of the combined drugs given individually. In a 2-year trial of combination entecavir and tenofovir versus entecavir monotherapy, for a small subgroup of patients with very high HBV DNA levels ($\geq 10^8$ IU/mL), a reduction in HBV DNA to <50 IU/mL was higher in the combination group (79% vs 62%); however, no differences in HBeAg responses or any other endpoint were observed between the combination-therapy and monotherapy groups, even in the high-HBV DNA subgroup. On the other hand, combining agents that are not cross-resistant (e.g., lamivudine or entecavir with adefovir or tenofovir) has the potential to reduce the risk or perhaps even to preempt entirely the emergence of drug resistance. In the future, the treatment paradigm may shift from the current approach of sequential monotherapy to preemptive combination therapy, perhaps not for all patients but for subsets (e.g., patients with very high levels of HBV DNA, immunosuppressed patients); however, designing and executing clinical trials that demonstrate superior efficacy and resistance profile of combination therapy over monotherapy with entecavir or tenofovir will remain challenging. Whereas, initially, in clinical studies of adefovir as rescue therapy for lamividune resistance, *adding* adefovir to lamivudine (combination therapy) was considered a better strategy than replacing lamivudine with adefovir monotherapy, according to the 2016 treatment recommendations of the American Association for the Study of Liver Diseases (AASLD), data to support adding or switching agents are insufficient. Therefore, while sound virologic principles would favor adding as opposed to switching, according to current recommendations involving the more potent first-line agents, entecavir for tenofovir resistance and tenofovir for entecavir resistance, either strategy is acceptable. For patients who already have acquired multidrug resistance (to both nucleoside analogues [lamivudine, entecavir, telbivudine] and nucleotide analogues [adefovir, tenofovir]), treatment with a combination of entecavir and tenofovir has been shown to be highly effective in suppressive HBV DNA and overcoming drug resistance.

NOVEL ANTIVIRALS AND STRATEGIES

In addition to the seven approved antiviral drugs for hepatitis B, emtricitabine, a fluorinated cytosine analogue very similar to lamivudine in structure, efficacy, and resistance profile, offers no advantage over lamivudine. A combination of emtricitabine and tenofovir is approved for the treatment of HIV infection and is an appealing combination therapy for hepatitis B, especially for lamivudine-resistant disease; however, neither emtricitabine nor the combination is approved for hepatitis B. Several initially promising antiviral agents have been abandoned because of toxicity (e.g., clevudine, which was linked to myopathy during its clinical development). As noted above, the current formulation of tenofovir, TDF, has been associated with renal toxicity and loss of bone density, especially in patients with HIV infection, less so in patients with HBV infection. A new formulation, tenofovir alafenamide (TAF), is a prodrug of tenofovir that is metabolized to the active agent in its target organ (the liver for HBV infection); such targeting permits higher dose delivery to the liver with markedly reduced systemic exposure. Studies in patients with chronic hepatitis B treated with 25 mg of TAF or 300 mg of TDF demonstrate comparable virologic efficacy as well as less reduction in bone mineral density and estimated glomerular filtration rate for TAF. Based on its better renal and bone safety profile than TDF, TAF has been approved for HBV infection

TABLE 334-3 Comparison of Pegylated Interferon (PEG IFN), Lamivudine, Adefovir, Entecavir, Telbivudine, and Tenofovir Therapy for Chronic Hepatitis B[a]

FEATURE	PEG IFN[b]	LAMIVUDINE	ADEFOVIR	ENTECAVIR	TELBIVUDINE	TENOFOVIR
Route of administration	Subcutaneous injection	Oral	Oral	Oral	Oral	Oral
Duration of therapy[c]	48–52 weeks	≥52 weeks	≥48 weeks	≥48 weeks	≥52 weeks	≥48 weeks
Tolerability	Poorly tolerated	Well tolerated	Well tolerated; creatinine monitoring recommended	Well tolerated	Well tolerated	Well tolerated; creatinine monitoring recommended
HBeAg seroconversion						
1 yr Rx	18–20%	16–21%	12%	21%	22%	21%
>1 yr Rx	NA	up to 50% @ 5 yrs	43% @ 3 yrs[d]	31% @ 2 yrs 44% @ 6 yrs	30% @ 2 yrs	40% @ 5 yrs
Log_{10} HBV DNA reduction (mean copies/mL)						
HBeAg-reactive	4.5	5.5	median 3.5–5	6.9	6.4	6.2
HBeAg-negative	4.1	4.4–4.7	median 3.5–3.9	5.0	5.2	4.6
HBV DNA PCR negative (<300–400 copies/mL; <1000 copies/mL for adefovir) at end of yr 1						
HBeAg-reactive	10–25%	36–44%	13–21%	67% (91% @ 4 yrs)	60%	76%
HBeAg-negative	63%	60–73%	48–77%	90%	88%	93%
ALT normalization at end of yr 1						
HBeAg-reactive	39%	41–75%	48–61%	68%	77%	68%
HBeAg-negative	34–38%	62–79%	48–77%	78%	74%	76%
HBsAg loss yr 1	3–4%	≤1%	0%	2%	<1%	3%
>yr 1	12% 5 yr after 1 yr of Rx	No data	5% at yr 5	6% at yr 6	No data	8% at yr 5
Histologic improvement (≥2 point reduction in HAI) at yr 1						
HBeAg-reactive	38% 6 months after	49–62%	53–68%	72%	65%	74%
HBeAg-negative	48% 6 months after	61–66%	64%	70%	67%	72%
Viral resistance	None	15–30% @ 1 yr 70% @ 5 yrs	None @ 1 yr 29% @ 5 yrs	≤1% @ 1 yr[e] 1.2% @ 6 yrs[e]	Up to 5% @ yr 1 Up to 22% @ yr 2	0% @ yr 1 0% through yr 8
Pregnancy category	C	C[f]	C	C	B	B
Cost (US$) for 1 yr	~$18,000	~$2,500	~$6,500	~$8,700[g]	~$6,000	~$6,000

[a]Generally, these comparisons are based on data on each drug tested individually versus placebo in registration clinical trials; because, with rare exception, these comparisons are not based on head-to-head testing of these drugs, relative advantages and disadvantages should be interpreted cautiously. [b]Although standard interferon α administered daily or three times a week is approved as therapy for chronic hepatitis B, it has been supplanted by PEG IFN, which is administered once a week and is more effective. Standard interferon has no advantages over PEG IFN. [c]Duration of therapy in clinical efficacy trials; use in clinical practice may vary. [d]Because of a computer-generated randomization error that resulted in misallocation of drug versus placebo during the second year of clinical trial treatment, the frequency of HBeAg seroconversion beyond the first year is an estimate (Kaplan-Meier analysis) based on the small subset in whom adefovir was administered correctly. [e]7% during a year of therapy (43% at year 4) in lamivudine-resistant patients. [f]Despite its Class C designation, lamivudine has an extensive pregnancy safety record in women with HIV/AIDS. [g]Approximately $17,400 for lamivudine-refractory patients.

Abbreviations: ALT, alanine aminotransferase; HAI, histologic activity index; HBeAg, hepatitis B e antigen; HBsAg, hepatitis B surface antigen; HBV, hepatitis B virus; NA, not applicable; PEG IFN, pegylated interferon; PCR, polymerase chain reaction; Rx, therapy; yr, year.

and provides an alternative to TDF in patients with TDF-associated elevations in serum creatinine and/or reductions in serum phosphorus. Direct-acting antivirals (DAAs) have been very successful in the management of chronic hepatitis B; however, most patients require long-duration, usually indefinite, therapy. Ideally, an approach to achieving "cure" (eradication of HBV infection) with finite-duration therapy would be welcome. Currently, innovative approaches being investigated include viral entry inhibitors, nucleocapsid assembly inhibitors, HBV secretion (HBsAg release) inhibitors, immunomodulators (e.g., toll receptor agonists, T-cell vaccines, programmed cell death [PD-1] blockade, reconstitution of innate and adaptive immune responses, HBV mRNA recognition and activation of innate immune signaling by retinoic acid-inducible gene-I [RIG-I]), covalently closed circular (ccc) DNA silencing/inhibition/cleavage, RNA interference, and HBx inhibitors. While data supporting several of these unconventional approaches have begun to appear, none has been shown to "cure" hepatitis B, and none is likely to be competitive, unless it can be shown to go beyond current antivirals in achieving recovery (HBsAg seroconversion) from HBV infection. Finally, initial emphasis in the development of antiviral therapy for hepatitis B was placed on monotherapy; whether combination regimens will yield additive or synergistic efficacy remains to be determined.

TREATMENT RECOMMENDATIONS

Several learned societies and groups of expert physicians have issued treatment recommendations for patients with chronic hepatitis B; the most authoritative and updated (and free of financial support by pharmaceutical companies) are those of the AASLD and of the European Association for the Study of the Liver (EASL). Although the recommendations differ slightly, a consensus has emerged on most of the important points **(Table 334-4)**. No treatment is recommended or available for inactive "nonreplicative" hepatitis B carriers (undetectable HBeAg with normal ALT and HBV

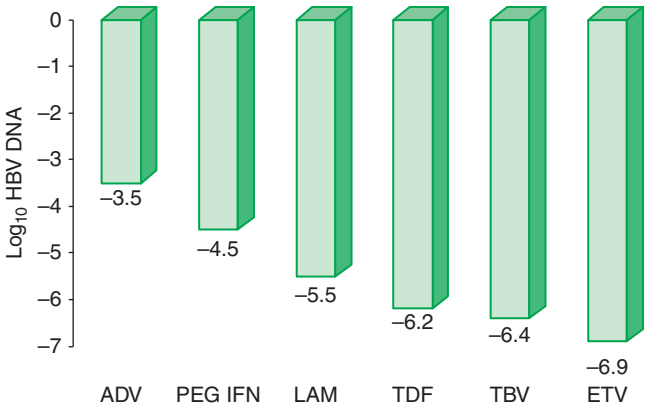


FIGURE 334-1 Relative potency of antiviral drugs for hepatitis B, as reflected by median $\log_{10}$ HBV DNA reduction in HBeAg-positive chronic hepatitis B. These data are from individual reports of large, randomized controlled registration trials that were the basis for approval of the drugs. In most instances, these data do not represent direct comparisons among the drugs, because study populations were different, baseline patient variables were not always uniform, and the sensitivity and dynamic range of the HBV DNA assays used in the trials varied. ADV, adefovir dipivoxil; ETV, entecavir; LAM, lamivudine; PEG IFN, pegylated interferon α2a; TBV, telbivudine; TDF, tenofovir disoproxil fumarate.

DNA ≤10^3 IU/mL documented serially over time). In patients with detectable HBeAg and HBV DNA levels >2 × 10^4 IU/mL, treatment is recommended by the AASLD for those with ALT levels above 2 × the upper limit of normal. (The EASL recommends treatment in HBeAg-positive patients for HBV DNA levels >2 × 10^3 IU/mL and ALT above the upper limit of normal.) For HBeAg-positive patients with ALT ≤2 × the upper limit of normal, in whom sustained responses are not likely and who would require multiyear therapy, antiviral therapy is not recommended currently. This pattern is common during the early decades of life among Asian patients infected at birth; even in this group, therapy would be considered for those >40 years of age, ALT persistently at the high end of the twofold range, and/or with a family history of HCC, especially if the liver biopsy shows moderate to severe necroinflammatory activity or fibrosis. In this group, when, eventually, ALT becomes elevated later in life, antiviral therapy should be instituted. For patients with HBeAg-negative chronic hepatitis B, ALT >2 × the upper limit of normal (above the upper limit of normal according to EASL), and HBV DNA >2 × 10^3 IU/mL, antiviral therapy is recommended. If HBV DNA is >2 × 10^3 IU/mL and ALT is 1 to >2 × the upper limit of normal, liver biopsy should be considered to help in arriving at a decision to treat if substantial liver injury is present (treatment in this subset would be recommended according to EASL guidelines, because ALT is elevated). Per current AASLD recommendations, antiviral treatment with oral agents can be stopped after HBeAg seroconversion in noncirrhotics, and the suggested period of consolidation therapy is 12 months with close monitoring for recurrent viremia (monthly × 6, then every 3 months for the rest of a year) after cessation of therapy. For patients with HBeAg-negative chronic hepatitis, the current recommendation with oral agents is for indefinite therapy; although sufficient data are lacking, stopping therapy in this group can be considered after HBsAg loss.

For patients with compensated cirrhosis, because antiviral therapy has been shown to retard clinical progression, treatment is recommended regardless of HBeAg status and ALT as long as HBV DNA is detectable at >2 × 10^3 IU/mL (detectable at any level according to the EASL); monitoring without therapy is recommended for those with HBV DNA <2 × 10^3 IU/mL, unless ALT is elevated. For patients with decompensated cirrhosis, treatment is recommended regardless of serologic and biochemical status, as long as HBV DNA is detectable. Patients with decompensated cirrhosis should be evaluated as candidates for liver transplantation.

Among the seven available drugs for hepatitis B, PEG IFN has supplanted standard IFN, entecavir has supplanted lamivudine, and tenofovir has supplanted adefovir. PEG IFN, entecavir, or tenofovir is recommended as first-line therapy (Table 334-3). PEG

TABLE 334-4 Recommendations for Treatment of Chronic Hepatitis B[a]

HBeAg STATUS	CLINICAL	HBV DNA (IU/mL)	ALT	RECOMMENDATION
HBeAg-reactive	[b]	>2 × 10^4	≤2 × ULN[c,d]	No treatment; monitor. In patients >40, with family history of hepatocellular carcinoma, and/or ALT persistently at the high end of the twofold range, liver biopsy may help in decision to treat
	Chronic hepatitis	>2 × 10^4[d]	>2 × ULN[d]	Treat[e]
	Cirrhosis compensated	>2 × 10^3	< or > ULN	Treat[e] with oral agents, not PEG IFN
	Cirrhosis decompensated	<2 × 10^3	>ULN	Consider treatment[f]
		Detectable	< or > ULN	Treat[e] with oral agents[g], not PEG IFN; refer for liver transplantation
		Undetectable	< or > ULN	Observe; refer for liver transplantation
HBeAg-negative	[b]	≤2 × 10^3	≤ULN	Inactive carrier; treatment not necessary
	Chronic hepatitis	>10^3	1 to >2 × ULN[d]	Consider liver biopsy; treat[h] if biopsy shows moderate to severe inflammation or fibrosis
	Chronic hepatitis	>10^4	>2 × ULN[d]	Treat[h,i]
	Cirrhosis compensated	>2 × 10^3	< or > ULN	Treat[e] with oral agents, not PEG IFN
		<2 × 10^3	>ULN	Consider treatment[f]
	Cirrhosis decompensated	Detectable	< or > ULN	Treat[h] with oral agents[g], not PEG IFN; refer for liver transplantation
		Undetectable	< or > ULN	Observe; refer for liver transplantation

[a]Based on practice guidelines of the American Association for the Study of Liver Diseases (AASLD). Except as indicated in footnotes, these guidelines are similar to those issued by the European Association for the Study of the Liver (EASL). [b]Liver disease tends to be mild or inactive clinically; most such patients do not undergo liver biopsy. [c]This pattern is common during early decades of life in Asian patients infected at birth. [d]According to the EASL guidelines, treat if HBV DNA is >2 × 10^3 IU/mL and ALT >ULN. [e]One of the potent oral drugs with a high barrier to resistance (entecavir or tenofovir) or PEG IFN can be used as first-line therapy (see text). These oral agents, but not PEG IFN, should be used for interferon-refractory/intolerant and immunocompromised patients. PEG IFN is administered weekly by subcutaneous injection for a year; the oral agents are administered daily for at least a year and continued indefinitely or until at least 6 months after HBeAg seroconversion. [f]According to EASL guidelines, patients with compensated cirrhosis and detectable HBV DNA at any level, even with normal ALT, are candidates for therapy. Most authorities would treat indefinitely, even in HBeAg-positive disease after HBeAg seroconversion. [g]Because the emergence of resistance can lead to loss of antiviral benefit and further deterioration in decompensated cirrhosis, a low-resistance regimen is recommended—entecavir or tenofovir monotherapy or combination therapy with the more resistance-prone lamivudine (or telbivudine) plus adefovir. Therapy should be instituted urgently. [h]Because HBeAg seroconversion is not an option, the goal of therapy is to suppress HBV DNA and maintain a normal ALT. PEG IFN is administered by subcutaneous injection weekly for a year; caution is warranted in relying on a 6-month posttreatment interval to define a sustained response, because the majority of such responses are lost thereafter. Oral agents, entecavir or tenofovir, are administered daily, usually indefinitely or until, as very rarely occurs, virologic and biochemical responses are accompanied by HBsAg seroconversion. [i]For older patients and those with advanced fibrosis, consider lowering the HBV DNA threshold to >2 × 10^3 IU/mL.

Abbreviations: AASLD, American Association for the Study of Liver Diseases; ALT, alanine aminotransferase; EASL, European Association for the Study of the Liver; HBeAg, hepatitis B e antigen; HBsAg, hepatitis B surface antigen; HBV, hepatitis B virus; PEG IFN, pegylated interferon; ULN, upper limit of normal.

IFN requires finite-duration therapy, achieves the highest rate of HBeAg responses after a year of therapy, and does not support viral mutations, but it requires subcutaneous injections and is associated with inconvenience, more intensive clinical and laboratory monitoring, and intolerability. Oral nucleoside analogues require long-term therapy in most patients, and when used alone, lamivudine and telbivudine foster the emergence of viral mutations, adefovir somewhat less so, and entecavir (except in lamivudine-experienced patients) and tenofovir rarely at all. Oral agents do not require injections or cumbersome laboratory monitoring, are very well tolerated, lead to improved histology in 50–90% of patients, suppress HBV DNA more profoundly than PEG IFN, and are effective even in patients who fail to respond to IFN-based therapy. Although oral agents are less likely to result in HBeAg responses during the first year of therapy, as compared to PEG IFN, treatment with oral agents tends to be extended beyond the first year and, by the end of the second year, yields HBeAg responses (and even HBsAg responses) comparable in frequency to those achieved after 1 year of PEG IFN (and without the associated side effects) **(Table 334-5)**. In a 2016 systematic review of 1716 patients involved in 25 clinical trials, responses after oral-agent therapy were found to be durable. Among patients with HBeAg-reactive chronic hepatitis B, the pooled rates of durable HBeAg seroconversions maintained after cessation of nucleoside/nucleotide analogue therapy (including all the oral agents) were 92% and 88% at posttreatment months 12 and 24, respectively, unaffected by the duration of post-HBeAg-response consolidation therapy (>6 months in all studies evaluated); the pooled rate of durable biochemical remission after therapy in this population was 76%. Even for HBeAg-negative chronic hepatitis B, for which most authorities recommend indefinite therapy, pooled rates of virologic remissions maintained after cessation of oral-agent therapy were 44%, 31%, and 30% at posttreatment months 12, 24, and 36, and the pooled rate of durable biochemical remission in this population was 57%.

TABLE 334-5 Pegylated Interferon Versus Oral Nucleoside Analogues for the Treatment of Chronic Hepatitis B

	PEG IFN	NUCLEOSIDE ANALOGUES
Administration	Weekly injection	Daily, orally
Tolerability	Poorly tolerated, intensive monitoring	Well tolerated, limited monitoring
Duration of therapy	Finite 48 weeks	≥1 year, indefinite in most patients
Maximum mean HBV DNA suppression	4.5 $\log_{10}$	6.9 $\log_{10}$
Effective in high-level HBV DNA ($\geq 10^9$ IU/mL)	No	Yes
HBeAg seroconversion		
During 1 year of therapy	~30%	~20%
During >1 year of therapy	Not applicable	30% (year 2) to up to 50% (year 5)
HBeAg-negative posttreatment HBV DNA suppression	17% @ 5 years	7% @ 4 years (lamivudine)
HBsAg loss		
During 1 year of therapy	3–4%	0–3%
During >1 year of therapy	Not applicable	3–8% @ 5 years of therapy
After 1 year of therapy–HBeAg-negative	12% @ 5 years	3.5% @ 5 years
Antiviral resistance	None	Lamivudine: ~30% year 1, ~70% year 5 Adefovir: 0% year 1, ~30% year 5 Telbivudine: up to 4% year 1, 22% year 2 Entecavir: ≤1.2% through year 6 Tenofovir: 0% through year 8
Use in cirrhosis, transplantation, immunosuppressed	No	Yes
Cost, 1 year of therapy	++++	+ to ++

Abbreviations: HBV, hepatitis B virus; HBeAg, hepatitis B e antigen; HBsAg, hepatitis B surface antigen; PEG IFN, pegylated interferon.

Although adefovir and tenofovir are safe, renal monitoring (e.g., serum creatinine and phosphate, urine glucose and protein) is recommended. Substantial experience with lamivudine during pregnancy (see above) has identified no teratogenicity; although widely used during pregnancy, lamivudine remains classified as pregnancy category C. Although IFNs do not appear to cause congenital anomalies, these have antiproliferative properties and should be avoided during pregnancy. Adefovir during pregnancy has not been associated with birth defects; however, the risk of spontaneous abortion may be increased, and adefovir is categorized as pregnancy category C. Data on the safety of entecavir during pregnancy have not been published (pregnancy category C). Sufficient data in animals and limited data in humans suggest that telbivudine and tenofovir (both pregnancy category B) can be used safely during pregnancy; however, telbivudine is not an acceptable first-line drug. In general, then, except for lamivudine and tenofovir, and until additional data become available, the other antivirals for hepatitis B should be avoided or used with extreme caution during pregnancy.

For children aged 2 to <18 with HBeAg-reactive hepatitis B (most children will be HBeAg-reactive; no studies have been done in children with HBeAg-negative chronic hepatitis B), treatment is recommended if HBV DNA is detectable and ALT levels are elevated, but not if ALT levels are normal. Each of the available drugs, except telbivudine, is approved for different childhood age groups (standard IFN α-2b age ≥1 year; PEG IFN α-2a age ≥5 years [approved for hepatitis C, not B, but can be used in hepatitis B]; lamivudine and entecavir age ≥2 years; adefovir and tenofovir age ≥12 years). Package inserts should be consulted for childhood doses.

As noted above, some physicians prefer to begin with PEG IFN, while other physicians and patients prefer oral agents as first-line therapy. For patients with decompensated cirrhosis, the emergence of resistance can result in further deterioration and loss of antiviral effectiveness. Therefore, in this patient subset, the threshold for relying on therapy with a very favorable resistance profile (e.g., entecavir or tenofovir) or on combination therapy is low. PEG IFN should not be used in patients with compensated or decompensated cirrhosis.

For patients with end-stage chronic hepatitis B who undergo liver transplantation, reinfection of the new liver is almost universal in the absence of antiviral therapy. The majority of patients become high-level viremic carriers with minimal liver injury. Before the availability of antiviral therapy, an unpredictable proportion experienced severe hepatitis B–related liver injury, sometimes a fulminant-like hepatitis and sometimes a rapid recapitulation of the original severe chronic hepatitis B **(Chap. 332)**. Currently, however, prevention of recurrent hepatitis B after liver transplantation has been achieved definitively by *combining* hepatitis B immune globulin with one of the low-resistance oral nucleoside (entecavir) or nucleotide analogues (tenofovir) **(Chap. 338)**; preliminary data suggest that the newer, more potent, and less resistance-prone oral agents may be used instead of hepatitis B immune globulin for posttransplantation therapy. In patients documented at the time of liver transplantation to have undetectable HBV DNA in serum and cccDNA in the liver (i.e., with low risk for recurrence of HBV infection), a preliminary clinical trial suggested that, after patients received 5 years of combined therapy, both hepatitis B immune globulin and oral-agent therapy can be withdrawn sequentially (over two 6-month periods) with a success rate, as monitored over a median of 6 years postwithdrawal, of 90% and an anti-HBs seroconversion rate of 60% (some with transient reappearance of HBV DNA and/or HBsAg).

Patients with HBV-HIV co-infection can have progressive HBV-associated liver disease and, occasionally, a severe exacerbation of hepatitis B resulting from immunologic reconstitution following

ART. Lamivudine should never be used as monotherapy in patients with HBV-HIV infection because HIV resistance emerges rapidly to both viruses. Adefovir has been used successfully to treat chronic hepatitis B in HBV-HIV co-infected patients but is no longer considered a first-line agent for HBV. Entecavir has low-level activity against HIV and can result in selection of HIV resistance; therefore, it should be avoided in HBV-HIV co-infection. Tenofovir and the combination of tenofovir and emtricitabine in one pill are approved therapies for HIV and represent excellent choices for treating HBV infection in HBV-HIV co-infected patients. Generally, even for HBV-HIV co-infected patients who do not yet meet treatment criteria for HIV infection, treating for both HBV and HIV is recommended.

Patients with chronic hepatitis B who undergo cytotoxic chemotherapy for treatment of malignancies as well as patients treated with immunosuppressive, anticytokine, or antitumor necrosis factor therapies (the risk varies, from highest [e.g., B-cell-depleting agents, anthracycline derivatives, moderate/high-dose corticosteroids for ≥4 weeks] to moderate [e.g., tumor necrosis factor alpha inhibitors, cytokine or integrin inhibitors, tyrosine kinase inhibitors, low-dose corticosteroids for ≥4 weeks], to lowest [e.g., immunosuppressive agents like methotrexate and azathioprine, intraarticular corticosteroids, any dose of corticosteroids for ≤1 week]) experience enhanced HBV replication and viral expression on hepatocyte membranes during chemotherapy coupled with suppression of cellular immunity. When chemotherapy is withdrawn, such patients are at risk for reactivation of hepatitis B, often severe and occasionally fatal. Such rebound reactivation represents restoration of cytolytic T cell function against a target organ enriched in HBV expression. Preemptive treatment with the first of the oral HBV antivirals, lamivudine, prior to the initiation of chemotherapy was shown to reduce the risk of such reactivation substantially; treating *after* reactivation has occurred is less effective. The newer, more potent oral antiviral agents, entecavir and tenofovir, which are even more effective in preventing hepatitis B reactivation and with a lower risk of antiviral drug resistance, are preferred. The optimal duration of antiviral therapy after completion of chemotherapy is not known, but a suggested approach is 6 months (12 months for B-cell-depleting agents) for inactive hepatitis B carriers and longer-duration therapy in patients with baseline HBV DNA levels $>2 \times 10^3$ IU/mL, until standard clinical endpoints are met (Table 334-4). Such chemotherapy-associated reactivation of hepatitis B is common (4–68%, median 25%, in a meta-analysis) in persons with ongoing HBV infection (HBsAg-reactive); however, such reactivation can occur albeit less commonly in persons who have cleared HBsAg, but express anti-HBc (moderate risk, <10%) and rarely (<5%) even in persons with serologic evidence of recovery from HBV infection (anti-HBs-reactive, anti-HBc-reactive). Therefore, most authorities (e.g., Centers for Disease Control and Prevention; AASLD; American Gastroenterological Association; EASL) recommend HBsAg and anti-HBc (± anti-HBs) screening of all patients undergoing such chemotherapy and preemptive antiviral prophylaxis for HBsAg-reactive persons and close on-therapy monitoring of anti-HBc-reactive/anti-HBs-reactive persons with treatment if and when reactivation occurs.

CHRONIC HEPATITIS D (DELTA HEPATITIS)

Chronic hepatitis D virus (HDV) may follow acute co-infection with HBV but at a rate no higher than the rate of chronicity of acute hepatitis B. That is, although HDV co-infection can increase the severity of acute hepatitis B, HDV does not increase the likelihood of progression to chronic hepatitis B. When, however, HDV superinfection occurs in a person who is already chronically infected with HBV, long-term HDV infection is the rule, and a worsening of the liver disease is the expected consequence. Except for severity, chronic hepatitis B plus D has similar clinical and laboratory features to those seen in chronic hepatitis B alone. Relatively severe and progressive chronic hepatitis, with or without cirrhosis, is the rule, and mild chronic hepatitis is the exception. Occasionally, however, mild hepatitis or even, rarely, inactive carriage occurs in patients with chronic hepatitis B plus D, and the disease may become indolent after several years of infection. A distinguishing serologic feature of chronic hepatitis D is the presence in the circulation of antibodies to liver-kidney microsomes (anti-LKM); however, the anti-LKM seen in hepatitis D, anti-LKM3, are directed against uridine diphosphate glucuronosyltransferase and are distinct from anti-LKM1 seen in patients with autoimmune hepatitis and in a subset of patients with chronic hepatitis C (see below). **The clinical and laboratory features of chronic HDV infection are summarized in Chap. 332.**

TREATMENT

Chronic Hepatitis D

Management is not well defined, and the host cellular RNA polymerase upon which HDV replication depends cannot be targeted by conventional antiviral agents. Glucocorticoids are ineffective and are not used. Preliminary experimental trials of IFN-α suggested that conventional doses and durations of therapy lower levels of HDV RNA and aminotransferase activity only transiently during treatment but have no impact on the natural history of the disease. In contrast, high-dose IFN-α (9 million units three times a week) for 12 months was reported to be associated with a sustained loss of HDV replication and clinical improvement in up to 50% of patients. Moreover, in anecdotal reports, the beneficial impact of treatment has been observed to persist for 15 years and to be associated with a reduction in grade of hepatic necrosis and inflammation, reversion of advanced fibrosis (improved stage), and clearance of HDV RNA in some patients. A suggested approach to therapy has been high-dose, long-term IFN for at least a year and, in responders, extension of therapy until HDV RNA and HBsAg clearance; however, extension of therapy to a second year provided no advantage, and sustained responses after completion of therapy have been rare. PEG IFN has also been shown to be more effective in the treatment of chronic hepatitis D (e.g., after 48 weeks of therapy, associated with undetectable HDV RNA, durable for at least 24 posttreatment weeks, in a quarter to a half of patients) and is a more convenient replacement for standard IFN; however, loss of virologic responses (reappearance of HDV RNA) was observed during long-term (median 4.5-year) monitoring in over half of initial, 24-week-post-treatment responders. Even extending PEG IFN therapy for 5 years and driving treatment doses up to 270 μg weekly (of PEG IFN- α2a), as reported in a small trial among 13 patients, while achieving serologic, virologic, histologic, biochemical, and clinical improvement, yielded sustained virologic responses (SVRs) in only 3 patients (58–246 weeks of posttreatment observation). None of the nucleoside analogue antiviral agents for hepatitis B is effective in hepatitis D, and adding oral nucleoside agents to PEG IFN is no more effective than PEG IFN monotherapy. While recommended, PEG IFN therapy is far from satisfactory. Preliminary trials have been performed with an oral prenylation inhibitor, lonafarnib, and with an inhibitor of HBV/HDV viral entry into hepatocytes, myrcludex B. Prenylation, the posttranslational covalent addition of the prenyl lipid farnesyl to large HDV antigen, is required for this HDV protein to interact and form secreted viral particles with HBsAg. In 14 patients treated twice daily for 28 days with 100 or 200 mg of lonafarnib, HDV RNA fell by 0.73 $\log_{10}$ IU/mL and 1.54 $\log_{10}$ IU/mL, respectively, before rebounding after completion of therapy. Hepatitis B virus entry into hepatocytes requires the binding of the myristolated N-terminal pre-S1 peptide of large HBsAg to sodium taurocholate co-transporting peptide, the functional receptor for HBV into hepatocytes. The application of myrcludex B, a synthetic homologous myristolated lipopeptide that competes for binding with HBsAg, was reported in a study of 24 patients (with a baseline mean of 4.1–4.2 $\log_{10}$ copies/mL of HDV RNA) randomized to 24 weeks of treatment with myrcludex B (2 mg daily subcutaneously) as monotherapy or combined with PEG IFN compared to PEG IFN alone. A reduction in HDV RNA occurred in all three groups, by 1.67 $\log_{10}$ copies/mL (in two of eight patients RNA became undetectable), 2.59 $\log_{10}$ copies/mL (in five of eight patients RNA became undetectable), and 2.17 $\log_{10}$ copies/mL

(in two of eight patients RNA became undetectable), respectively. No change occurred, however, in the level of HBsAg, which would have been expected. In these two exploratory brief-duration trials, sustained responses were not achieved, and toxicities were encountered (e.g., intermittent vomiting and weight loss [lonafarnib] and transient amylase and lipase elevations [myrcludex B]); however, from these proof-of-principle trials, potentially, more definitive and larger-scale studies will follow.

In patients with end-stage liver disease secondary to chronic hepatitis D, liver transplantation has been effective. If hepatitis D recurs in the new liver without the expression of hepatitis B (an unusual serologic profile in immunocompetent persons but common in transplant patients), liver injury is limited. In fact, the outcome of transplantation for chronic hepatitis D is superior to that for chronic hepatitis B; in such patients, combination hepatitis B immune globulin and nucleoside analogue therapy for hepatitis B is indicated **(Chap. 338)**.

CHRONIC HEPATITIS C

Regardless of the epidemiologic mode of acquisition of hepatitis C virus (HCV) infection, chronic hepatitis follows acute hepatitis C in 50–70% of cases; chronic infection is common even in those with a return to normal in aminotransferase levels after acute hepatitis C, adding up to an 85% likelihood of chronic HCV infection after acute hepatitis C. Few clues had emerged to explain host differences associated with chronic infection until recently, when variation in a single nucleotide polymorphism (SNP) on chromosome 19, *IL28B* (which codes for IFN-λ3), was identified that distinguished between responders and nonresponders to IFN-based antiviral therapy (see below). The same variants correlated with spontaneous resolution after acute infection: 53% in genotype C/C, 30% in genotype C/T, but only 23% in genotype T/T. The association with HCV clearance after acute infection is even stronger when *IL28B* haplotype is combined with haplotype G/G of a SNP near human leukocyte antigen (HLA) Class II *DBQ1*03:01*.

In patients with chronic hepatitis C followed for 20 years, progression to cirrhosis occurs in about 20–25%. Such is the case even for patients with relatively clinically mild chronic hepatitis, including those without symptoms, with only modest elevations of aminotransferase activity, and with mild chronic hepatitis on liver biopsy. Even in cohorts of well compensated patients with chronic hepatitis C referred for clinical research trials (no complications of chronic liver disease and with normal hepatic synthetic function), the prevalence of cirrhosis may be as high as 50%. Most cases of hepatitis C are identified initially in asymptomatic patients who have no history of acute hepatitis C (e.g., those discovered while attempting to donate blood, while undergoing lab testing as part of an application for life insurance, or as a result of routine laboratory tests). The source of HCV infection in many of these cases is not defined, although a long-forgotten percutaneous exposure (e.g., injection drug use) in the remote past can be elicited in a substantial proportion and probably accounts for most infections; most of these infections were acquired in the 1960s and 1970s, coming to clinical attention decades later.

Approximately one-third of patients with chronic hepatitis C have normal or near-normal aminotransferase activity; although one-third to one-half of these patients have chronic hepatitis on liver biopsy, the grade of liver injury and stage of fibrosis tend to be mild in the vast majority. In some cases, more severe liver injury has been reported—even, rarely, cirrhosis, most likely the result of previous histologic activity. Among patients with persistent normal aminotransferase activity sustained over ≥5–10 years, histologic progression has been shown to be rare; however, approximately one-fourth of patients with normal aminotransferase activity experience subsequent aminotransferase elevations, and histologic injury can be progressive once abnormal biochemical activity resumes. Therefore, continued clinical monitoring and antiviral therapy are indicated, even for patients with normal aminotransferase activity.

Despite this substantial rate of progression of chronic hepatitis C, and despite the fact that liver failure can result from end-stage chronic hepatitis C, the long-term prognosis over 1–2 decades for chronic hepatitis C in a majority of patients is relatively benign. Mortality >10–20 years among patients with transfusion-associated chronic hepatitis C has been shown not to differ from mortality in a matched population of transfused patients in whom hepatitis C did not develop. Although death in the hepatitis group is more likely to result from liver failure, and although hepatic decompensation may occur in ~15% of such patients over the course of a decade, the majority (almost 60%) of patients remain asymptomatic and well compensated, with no clinical sequelae of chronic liver disease. Overall, chronic hepatitis C tends to be very slowly and insidiously progressive, if at all, in the vast majority of patients, whereas in approximately one-fourth of cases, chronic hepatitis C will progress eventually to end-stage cirrhosis. In fact, because HCV infection is so prevalent, and because a proportion of patients progress inexorably to end-stage liver disease, hepatitis C is the most frequent indication for liver transplantation **(Chap. 338)**. In the United States, hepatitis C accounts for up to 40% of all chronic liver disease; as of 2007, mortality caused by hepatitis C surpassed that associated with HIV/AIDS, and as of 2012, reported deaths caused by hepatitis C surpassed those associated with all other notifiable infectious diseases (HIV, tuberculosis, hepatitis B, and 57 other infectious diseases). Moreover, because the prevalence of HCV infection is so much higher in the "baby boomer" cohort born between 1945 and 1965, three-quarters of the mortality associated with hepatitis C occurs in this age cohort. Referral bias may account for the more severe outcomes described in cohorts of patients reported from tertiary care centers (20-year progression of ≥20%) versus the more benign outcomes in cohorts of patients monitored from initial blood-product-associated acute hepatitis or identified in community settings (20-year progression of only 4–7%). Still unexplained, however, are the wide ranges in reported progression to cirrhosis, from 2% over 17 years in a population of Irish women with hepatitis C infection acquired from contaminated anti-D immune globulin to 30% over ≤11 years in recipients of contaminated intravenous immune globulin.

Progression of liver disease in patients with chronic hepatitis C has been reported to be more likely in patients with older age, longer duration of infection, advanced histologic stage and grade, more complex HCV quasispecies diversity, increased hepatic iron, concomitant other liver disorders (alcoholic liver disease, chronic hepatitis B, hemochromatosis, α_1 antitrypsin deficiency, and steatohepatitis), HIV infection, and obesity. Among these variables, however, duration of infection appears to be one of the most important, and some of the others probably reflect disease duration to some extent (e.g., quasispecies diversity, hepatic iron accumulation). No other epidemiologic or clinical features of chronic hepatitis C (e.g., severity of acute hepatitis, level of aminotransferase activity, level of HCV RNA, presence or absence of jaundice during acute hepatitis) are predictive of eventual outcome. Despite the relatively benign nature of chronic hepatitis C over time in many patients, cirrhosis following chronic hepatitis C has been associated with the late development, after several decades, of HCC **(Chap. 78)**; the annual rate of HCC in cirrhotic patients with hepatitis C is 1–4%, occurring primarily in patients who have had HCV infection for 30 years or more.

Perhaps the best prognostic indicator in chronic hepatitis C is liver histology; the rate of hepatic fibrosis may be slow, moderate, or rapid. Patients with mild necrosis and inflammation as well as those with limited fibrosis have an excellent prognosis and limited progression to cirrhosis. In contrast, among patients with moderate to severe necroinflammatory activity or fibrosis, including septal or bridging fibrosis, progression to cirrhosis is highly likely over the course of 10–20 years. The pace of fibrosis progression may be accelerated by such factors as concomitant HIV infection, other causes of liver disease, excessive alcohol use, and hepatic steatosis. Among patients with compensated cirrhosis associated with hepatitis C, the 10-year survival rate is close to 80%; mortality occurs at a rate of 2–6% per year; decompensation at a rate of 4–5% per year; and, as noted above, HCC at a rate of 1–4% per year. Estimates of the natural history of chronic hepatitis C have been made, based on data available on the prevalence of HCV infection in the U.S. population and on the rate of disease progression. Weighted primarily by the concentration of chronic hepatitis C in the baby boomer generation, the peak prevalence was estimated to have

occurred in 2015. The calculated frequency of cirrhosis in U.S. patients with hepatitis C was 5% in 1990, 25% in 2010, and is projected to be 37% in 2020. Estimated peak mortality has been predicted to occur in 2032. **A discussion of the pathogenesis of liver injury in patients with chronic hepatitis C appears in Chap. 332.**

Clinical features of chronic hepatitis C are similar to those described above for chronic hepatitis B. Generally, fatigue is the most common symptom; jaundice is rare. Immune complex–mediated extrahepatic complications of chronic hepatitis C are less common than in chronic hepatitis B (despite the fact that assays for immune complexes are often positive in patients with chronic hepatitis C), with the exception of essential mixed cryoglobulinemia **(Chap. 332)**, which is linked to cutaneous vasculitis and membranoproliferative glomerulonephritis as well as lymphoproliferative disorders such as B-cell lymphoma and unexplained monoclonal gammopathy. In addition, chronic hepatitis C has been associated with extrahepatic complications unrelated to immune-complex injury. These include Sjögren's syndrome, lichen planus, porphyria cutanea tarda, type 2 diabetes mellitus, and the metabolic syndrome (including insulin resistance and steatohepatitis).

Laboratory features of chronic hepatitis C are similar to those in patients with chronic hepatitis B, but aminotransferase levels tend to fluctuate more (the characteristic episodic pattern of aminotransferase activity) and to be lower, especially in patients with long-standing disease. An interesting and occasionally confusing finding in patients with chronic hepatitis C is the presence of autoantibodies. Rarely, patients with autoimmune hepatitis (see below) and hyperglobulinemia have false-positive immunoassays for anti-HCV. On the other hand, some patients with serologically confirmable chronic hepatitis C have circulating anti-LKM. These antibodies are anti-LKM1, as seen in patients with autoimmune hepatitis type 2 (see below), and are directed against a 33-amino-acid sequence of cytochrome P450 IID6. The occurrence of anti-LKM1 in some patients with chronic hepatitis C may result from the partial sequence homology between the epitope recognized by anti-LKM1 and two segments of the HCV polyprotein. In addition, the presence of this autoantibody in some patients with chronic hepatitis C suggests that autoimmunity may be playing a role in the pathogenesis of chronic hepatitis C.

Histopathologic features of chronic hepatitis C, especially those that distinguish hepatitis C from hepatitis B, are described in Chap. 332.

TREATMENT

Chronic Hepatitis C

Therapy for chronic hepatitis C has evolved substantially in the 25 years since IFN-α was introduced for this indication in 1991. The therapeutic armamentarium grew to include PEG IFN with ribavirin and, then, in 2011, the introduction of the first protease inhibitors, telaprevir and boceprevir, used in combination with PEG IFN and ribavirin in patients with HCV genotype 1. The field of antiviral therapy for hepatitis C was transformed beginning in 2013, with the approval of the first nucleoside analogue, sofosbuvir. As of 2016, no fewer than six, all-oral, highly effective (>95%), low-resistance, well tolerated, short-duration (usually 12 weeks) combination regimens of DAA drugs are available. The remarkable historical evolution of antiviral therapy for hepatitis C is instructive.

THE INTERFEON ERA (1991–2011)

IFN-based therapy has been supplanted by DAA agents introduced in the second decade of the twenty-first century; however, many important lessons about antiviral therapy for chronic hepatitis C were learned from the experience with IFN-based treatment, and many of the limitations of—and disparities in responsiveness to—IFN-based therapy have been overcome by current-generation DAA treatments. When first approved, IFN-α was administered via subcutaneous injection three times a week for 6 months but achieved an SVR **(Fig. 334-2)** (defined then as a reduction of HCV RNA to undetectable levels by PCR when measured ≥24 weeks after completion of therapy) <10%. Doubling the duration of therapy—but

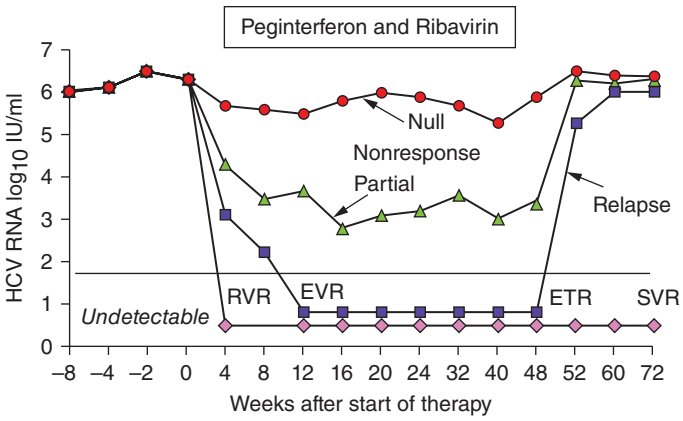


FIGURE 334-2 Classification of virologic responses based on outcomes during and after a 48-week course of pegylated interferon (PEG IFN) plus ribavirin antiviral therapy in patients with hepatitis C, genotype 1 or 4 (for genotype 2 or 3, the course would be 24 weeks). Nonresponders can be classified as null responders (hepatitis C virus [HCV] RNA reduction of <2 $\log_{10}$ IU/mL) or partial responders (HCV RNA reduction ≥2 $\log_{10}$ IU/mL but not suppressed to undetectable) by week 24 of therapy. In responders, HCV RNA can become undetectable, as shown with sensitive amplification assays, within 4 weeks (RVR, rapid virologic response); can be reduced by ≥2 $\log_{10}$ IU/mL within 12 weeks (early virologic response, EVR; if HCV RNA is undetectable at 12 weeks, the designation is "complete" EVR); or at the end of therapy, 48 weeks (ETR, end-treatment response). In responders, if HCV RNA remains undetectable for 24 weeks after ETR, week 72, the patient has a sustained virologic response (SVR), but if HCV RNA becomes detectable again, the patient is considered to have relapsed. The posttreatment week-24 SVR (SVR_{24}) has been supplanted by an SVR at week 12 (SVR_{12}), which has been shown to be equivalent to an SVR_{24}. In patients treated with DAA therapy, RVR and EVR milestones are largely irrelevant, being met by almost all patients. *(Reproduced with permission, courtesy of Marc G. Ghany, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health and the American Association for the Study of Liver Diseases. Hepatology 49:1335, 2009.)*

not increasing the dose or changing IFN preparations—increased the SVR rate to ~20%, and addition to the regimen of daily ribavirin, an oral guanosine nucleoside, increased the SVR rate to 40%. When used alone, ribavirin is ineffective and does not reduce HCV RNA levels appreciably, but ribavirin enhances the efficacy of IFN by reducing the likelihood of virologic relapse after the achievement of an end-treatment response (Fig. 334-2) (response measured during, and maintained to the end of, treatment). Proposed mechanisms to explain the role of ribavirin include subtle direct reduction of HCV replication, inhibition of host inosine monophosphate dehydrogenase activity (and associated depletion of guanosine pools), immune modulation, induction of virologic mutational catastrophe, and enhancement of IFN-stimulated gene expression. Ribavirin, despite its poorly understood mechanism of action, retains a modest role in supporting DAA agents as well (see below). IFN therapy results in activation of the JAK-STAT signal transduction pathway, which culminates in the intracellular elaboration of genes and their protein products that have antiviral properties. Hepatitis C proteins inhibit JAK-STAT signaling at several steps along the pathway, and exogenous IFN restores expression of IFN-stimulated genes and their antiviral effects.

Treatment with the combination of PEG IFN and ribavirin increased responsiveness (frequency of SVR) to as high as 55% overall—to >40% in genotypes 1 and 4, and to >80% in genotypes 2 and 3. Even in the absence of biochemical and virologic responses, histologic improvement occurred in approximately three-fourths of all treated patients. In chronic hepatitis C, ALT levels fall precipitously during therapy, and up to 90% of virologic responses are achieved within the first 12 weeks of therapy; responses thereafter are rare. Most relapses occur within the first 12 weeks after treatment; therefore, an SVR at week 12 posttreatment (SVR_{12}) is roughly equivalent to a 24-week SVR, and SVR_{12} has become the new standard. SVRs are very durable; normal ALT, improved histology, and absence of HCV RNA in serum and liver have been documented a decade after

successful therapy, and "relapses" 2 years after sustained responses are almost unheard of. Thus, an SVR to antiviral therapy of chronic hepatitis C is tantamount to a cure, which is followed by marked improvements in liver-disease outcomes (see below).

Patient variables that correlate with sustained virologic responsiveness to IFN-based therapy include favorable genotype (genotypes 2 and 3 as opposed to genotypes 1 and 4; genotype 1b as opposed to genotype 1a); low baseline HCV RNA level (<800,000 IU/mL), low HCV quasispecies diversity, and histologically mild hepatitis and minimal fibrosis, especially absence of cirrhosis; immunocompetence, low liver iron levels, age <40; female gender; and absence of obesity, insulin resistance, type 2 diabetes mellitus, and hepatic steatosis. High levels of HCV RNA, more histologically advanced liver disease, and high HCV quasispecies diversity all go hand in hand with advanced duration of infection and reduced IFN responsiveness. Also associated with poor responses to IFN-based therapy are African-American ethnicity (contributed to, but not explained entirely by, a higher proportion with genotype 1, slower early treatment viral kinetics, impaired HCV-specific immunity, and host genetic differences in *IL28B* alleles, described below), Latino ethnicity, and poor treatment adherence (<80% of IFN and ribavirin doses and <80% of prescribed duration of therapy). Ironically, patients whose disease was least likely to progress were the ones *most* likely to respond to IFN and vice versa. For patients treated with combination IFN-ribavirin, therapy for those with genotype 1 usually required a full 48 weeks with SVRs in the range of 40–45%, whereas in those with genotypes 2 and 3, a 24-week course of therapy sufficed with SVRs in the range of 80% (although refined tailoring of treatment duration could be indicated based on rapidity of response or associated cofactors, see below).

Genetic changes in the virus may explain differences in treatment responsiveness in some patients (e.g., among patients with genotype 1b, responsiveness to IFN is enhanced in those with amino-acid-substitution mutations in the nonstructural protein 5A gene). As described above in the discussion of spontaneous recovery from acute hepatitis C, IFN gene variants discovered in genome-wide association studies were shown to have a substantial impact on responsiveness of patients with genotype 1 to antiviral therapy. In studies of patients treated with PEG IFN and ribavirin, variants of the *IL28B* SNP that code for IFN-λ3 (a type III IFN, the receptors for which are more discretely distributed than IFN-α receptors and more concentrated in hepatocytes) correlate significantly with responsiveness. Patients homozygous for the C allele at this locus have the highest frequency of achieving an SVR (~80%), those homozygous for the T allele at this locus are least likely to achieve an SVR (~25%), and those heterozygous at this locus (C/T) have an intermediate level of responsiveness (SVRs in ~35%).

Side effects of IFN therapy are described in the section on treatment of chronic hepatitis B. The most pronounced side effect of ribavirin therapy is hemolysis—an expected reduction in hemoglobin of up to 2–3 g or in hematocrit of 5–10% but also a small, unpredictable proportion with profound, brisk hemolysis, resulting in symptomatic anemia; therefore, close monitoring of blood counts is crucial, and ribavirin should be avoided in patients with anemia or hemoglobinopathies; in patients with coronary artery disease or cerebrovascular disease, in whom anemia can precipitate an ischemic event; in patients with renal insufficiency (the drug is excreted renally); and in pregnancy (the drug is teratogenic, mandating scrupulous use of efficient contraception during, and for several months after, therapy in women of child-bearing age [because of their antiproliferative properties, IFNs also are contraindicated during pregnancy]). When symptomatic anemia occurs, ribavirin dose reductions or addition of erythropoietin to boost red blood cell levels may be required; erythropoietin was shown to improve patients' quality of life but not the likelihood of achieving an SVR. If ribavirin was stopped during therapy, SVR rates fell, but responsiveness could be maintained as long as ribavirin was not stopped and the total ribavirin dose exceeded 60% of the planned dose.

Ribavirin can also cause nasal and chest congestion, pruritus, and precipitation of gout. Combination IFN-ribavirin therapy is more difficult to tolerate than IFN monotherapy and more likely to lead to dose reductions and discontinuation of therapy.

Studies of viral kinetics have shown that despite a virion half-life in serum of only 2–3 h, the level of HCV is maintained by a high replication rate of 10^{12} hepatitis C virions per day. IFN-α blocks virion production or release with an efficacy that increases with increasing drug doses; moreover, the calculated death rate for infected cells during IFN therapy is inversely related to the level of HCV RNA. Patients with the most rapid death rate of infected hepatocytes are more likely to achieve undetectable HCV RNA at 3 months; in practice, failure to achieve an early virologic response (EVR), a ≥2-$\log_{10}$ reduction in HCV RNA by week 12, predicts failure to experience a subsequent SVR. Similarly, patients in whom HCV RNA becomes undetectable within 4 weeks (i.e., who achieve a rapid virologic response [RVR]) have a very high likelihood of achieving an SVR (Fig. 334-2). Surprisingly, however, high-dose induction with IFN-based therapy did not yield higher SVR rates.

For the treatment of chronic hepatitis C, standard IFNs were supplanted beginning in 2001 by PEG IFNs. These have elimination times up to sevenfold longer than standard IFNs (i.e., a substantially longer half-life) and achieve prolonged concentrations, permitting administration once (rather than three times) a week. Instead of the frequent drug peaks (linked to side effects) and troughs (when drug is absent) associated with frequent administration of short-acting IFNs, administration of PEG IFNs results in drug concentrations that are more stable and sustained over time. Once-a-week PEG IFN monotherapy is twice as effective as monotherapy with its standard IFN counterpart, approaches the efficacy of combination standard IFN plus ribavirin, and is as well tolerated as standard IFNs, without more difficult-to-manage thrombocytopenia and leukopenia than standard IFNs. For most of the decade prior to 2011, when protease inhibitors were introduced for HCV genotype 1 (see below), the standard of care was a combination of PEG IFN plus ribavirin for all HCV genotypes.

Two PEG IFNs are available: PEG IFN-α2b, a 12-kD, linear PEG molecule bound to IFN-α2b, and PEG IFN-α2a, a larger, 40-kD, branched PEG molecule bound to IFN-α2a; because of its larger size and smaller volume of extravascular distribution, PEG IFN-α2a can be given at a uniform dose independent of weight, whereas the dose of the smaller PEG IFN-α2b, which has a much wider volume distribution, must be weight-based. The standard dose of PEG IFN α2a was 180 μg and of PEG IFN-α2b 1.5 μg/kg. The ribavirin dose adopted for both PEG IFNs was, for genotype 1, 1000 mg (for patients <75 kg) to 1200 mg (for patients ≥75 kg) and, for genotypes 2 and 3, 800 mg; a broader ribavirin dose/weight range was approved subsequently for PEG IFN-α2b in patients with genotype 1: <65 kg, 800 mg; 65–85 kg, 1000 mg; >85–105 kg, 1200 mg; and >105 kg, 1400 mg. For both drugs, recommended treatment durations were 48 weeks for genotype 1 and 24 weeks for genotypes 2 and 3 (somewhat more refractory, justifying a full 48 weeks especially for advanced hepatic fibrosis or cirrhosis and/or high-level HCV RNA). Between the two PEG IFNs, PEG IFN-α2a appeared to be slightly better tolerated and slightly more effective than PEG IFN-α2b in registration trials (SVRs for genotype 1: 41–51% vs 40–42%, respectively) as well as in subsequent head-to-head trials and a systematic review of randomized trials (SVR in genotypes 1–4: 48–55% vs 32–40%, respectively).

Until the 2011 introduction of protease inhibitors, unless ribavirin was contraindicated (see above), combination PEG IFN plus ribavirin was the recommended course of therapy. Even after the introduction of protease inhibitors for genotypes 1 and 4, however, PEG IFN–ribavirin remained the standard of care for patients with genotypes 2 and 3 until late 2013. For patients treated with combination PEG IFN–ribavirin, measurement of quantitative HCV RNA levels at 12 weeks was helpful in guiding therapy; if a 2-$\log_{10}$ drop in HCV RNA had not been achieved by this time, chances for an SVR were negligible, and additional therapy was futile. If the 12-week HCV RNA had fallen by 2 $\log_{10}$ (EVR), the chances for an SVR at the end of therapy were approximately two-thirds; if the 12-week HCV

RNA was undetectable ("complete" EVR), the chances for an SVR exceeded 80% (Fig. 334-2).

The frequency of an SVR to PEG IFN–ribavirin therapy could be increased by tailoring therapy according to baseline variables and on-treatment virologic responsiveness. In patients with baseline variables weighing against a response (e.g., HCV RNA >800,000 IU/mL, weight >85 kg), by raising the dose of PEG IFN (e.g., to as high as 270 μg of PEG IFN-α2a) and/or the dose of ribavirin to as high as 1600 mg daily (if tolerated or supplemented by erythropoietin); or by extending therapy from 48 to 72 weeks for patients with genotype 1 and a slow virologic response (i.e., failure of HCV RNA to fall rapidly to undetectable levels within 4 weeks [absence of a RVR]), SVR rates could be improved somewhat. In contradistinction, in patients with genotype 1 (and 4) who had a 4-week RVR (which occurred in ≤20%), especially in the subset with low baseline HCV RNA, abbreviating the duration of therapy to 24 weeks, resulted in SVR rates of ~90%. Responsiveness to IFN-ribavirin-based therapy was diminished in immunocompromised patients and in patients with HIV-HCV co-infection and contraindicated in patients with decompensated liver disease or end-stage renal disease. The cumbersome nature of IFN-ribavirin-based therapy (injections, complicated laboratory monitoring, side effects and poor tolerability, modest efficacy, variables and patient subsets associated with poor responsiveness, tailored therapy, futility rules, etc.) was supplanted eventually (in 2016) by DAAs for all genotypes (see below). Most of the variables associated with poor responsiveness to IFN-based therapy became irrelevant, and difficult-to-treat patient subpopulations began to experience responses to DAAs that were indistinguishable from responses in standard patients (see below).

Persons with chronic HCV infection have been shown to suffer increased liver-related mortality. On the other hand, successful antiviral therapy of chronic hepatitis C resulting in an SVR has been shown to improve survival (and to reduce the need for liver transplantation); to lower the risk of liver failure, liver-related death, and all-cause mortality; to slow the progression of chronic hepatitis C; and to reverse fibrosis and even cirrhosis. Whereas the 10-year and 20-year survival in the absence of an SVR is reduced in cirrhotic patients with chronic hepatitis C, survival at these intervals after an SVR has been found to be indistinguishable from that of the general population. Although successful treatment reduces mortality and liver failure (3-4-fold 10-year reduction) in cirrhotic patients (and in those with advanced fibrosis) and reduces the need for liver transplantation and the likelihood of HCC (14-fold 10-year reduction), the risk of liver-related death and HCC persists, albeit at a much reduced level, necessitating continued clinical monitoring and cancer surveillance after SVR in cirrhotics. On the other hand, in the absence of an SVR, IFN-based therapy does not reduce the risk of HCC. Similarly, for nonresponders to PEG IFN–ribavirin therapy, three trials of long-term maintenance therapy with PEG IFN showed no benefit in reducing the risk of histologic progression or clinical decompensation, including the development of HCC. Fortunately, PEG IFN-ribavirin nonresponders can now be retreated with DAAs and experience SVR rates comparable to those in treatment-naïve persons (see below).

FIRST-GENERATION PROTEASE INHIBITORS (2011–2013)

The HCV RNA genome encodes a single polyprotein, which is cleaved during and after translation by host and viral-encoded proteases. One protease involved in the cleavage of the viral polyprotein is an NS3/4A viral protein that has serine protease activity. Telaprevir and boceprevir are serine protease inhibitors that target NS3/4A. In 2011, telaprevir and boceprevir used in combination with PEG IFN and ribavirin were approved by the U.S. Food and Drug Administration (FDA) as the first oral DAA agents for the treatment of hepatitis C genotype 1 (not other genotypes) in adults with stable liver disease, both in patients who had not been treated before or who had failed previous treatment. Although now replaced by more effective, all-oral regimens, these first-in-class agents represented a breakthrough in the treatment of chronic hepatitis C and established milestones against which subsequent therapies could be measured.

Because resistance developed rapidly during monotherapy with telaprevir and boceprevir, these drugs had to be used in combination with PEG IFN and ribavirin. Ribavirin in particular appeared to reduce relapse rates significantly in protease inhibitor-based regimens, such that those who could not take or were intolerant to ribavirin were unlikely to benefit from the addition of these agents. Telaprevir and boceprevir regimens consisted of periods of triple therapy (protease inhibitor plus PEG IFN plus ribavirin) and periods of dual therapy (PEG IFN plus ribavirin). Telaprevir regimens began with 12 weeks of triple therapy followed by dual therapy of a duration based on HCV RNA status at weeks 4 and 12 ("response-guided therapy") and prior treatment status. Boceprevir-based regimens consisted of a 4-week lead-in period of dual (PEG IFN–ribavirin) therapy followed by triple therapy and, in some instances, a further extension of dual therapy, with duration of response-guided therapy based on HCV RNA status at weeks 4, 8, and 24 and prior treatment status.

For patients with HCV genotype 1, protease inhibitors improved the frequency of RVRs and SVRs significantly as compared to PEG IFN plus ribavirin alone. In treatment-naïve patients, telaprevir-based SVRs were achieved in up to 79% of patients who received 12 weeks of triple therapy followed by 12–36 weeks of dual therapy, and among those with EVRs (undetectable HCV RNA at weeks 4 and 12) and response-guided therapy stopped at week 24 (12 weeks of triple therapy, then 12 weeks of dual therapy), SVRs occurred in 83–92%. In studies with boceprevir in treatment-naïve patients, SVRs occurred in 59–66% of patients, and among those with undetectable HCV RNA at 8 weeks, the SVR rate increased to 86–88%. Adding to the complexity of treatment with these protease inhibitors were absolute stopping rules for futility, that is, absence of HCV RNA reductions at critical treatment milestones, which were shown to be invariably predictive of nonresponse (telaprevir: HCV RNA >1000 IU/mL at weeks 4 or 12, or detectable at week 24; boceprevir: HCV RNA ≥100 IU/mL at week 12, or detectable at week 24).

In patients previously treated unsuccessfully with PEG IFN plus ribavirin, telaprevir-based treatment achieved SVRs in 83–88% of prior relapsers, 54–59% of partial responders (HCV RNA reduced by ≥2 $\log_{10}$ IU/mL but not to undetectable levels), and 29–33% of null responders (HCV RNA reduced by <2 $\log_{10}$ IU/mL). With boceprevir, a similar degradation in SVR rate occurred as a function of prior responsiveness—in 75% of prior relapsers, in 40–52% of previous partial responders; in ~30–40% of null responders. In a substantial proportion of protease inhibitor nonresponders, resistance-associated substitutions (RASs, previously referred to as resistance-associated variants, RAVs) could be identified, but these variants were not archived, and wild-type HCV reemerged in almost all cases within 1.5 to 2 years. SVRs to these protease inhibitors were highest in prior relapsers and treatment-naïve patients (white > black ethnicity), lower in prior partial responders, lower still in prior null responders, and lowest in cirrhotic prior null responders, for whom no benefit accrued over PEG IFN/ribavirin treatment. Responses to protease inhibitor triple-drug regimens were higher in patients with *IL28B* C than non-C genotypes, HCV genotype 1b than genotype 1a, less advanced than more advanced fibrosis stage, whites than blacks, lower body mass index (BMI) than elevated BMI, and, for boceprevir, achievement of a >1 $\log_{10}$ HCV RNA reduction during 4 weeks of PEG IFN–ribavirin lead-in therapy. Age and HCV RNA level were less influential and insulin resistance was noninfluential on response to these antiviral agents.

Both of these protease inhibitors had substantial toxicities. Telaprevir was associated with a severe, generalized (trunk and extremities), often confluent, maculopapular, pruritic rash in ~6% of treated patients (that required careful dermatologic monitoring in all patients and systemic corticosteroid therapy in the most severely affected). Other common side effects included pruritus, rectal burning, nausea, diarrhea, fatigue, dysgeusia (altered or unpleasant taste), and anemia, which required close monitoring, could be relatively refractory, occasionally requiring transfusion and even hospitalization (especially in cirrhotic prior nonresponders). Anemia occurred in half of boceprevir-treated patients, neutropenia in up to 30% and thrombocytopenia

in 3–4%. Other side effects of boceprevir include fatigue, nausea, headache, dysgeusia, dry mouth, vomiting, and diarrhea.

Both drugs came with an inconveniently high pill burden and had to be administered every 8 hours with food (TVR with a 20-g fat meal). Use of protease inhibitors was further complicated by numerous drug-drug interactions. As telaprevir and boceprevir are both eliminated by and inhibit CYP3A4, these agents could not be administered with other medications that induce CYP3A4 or are dependent on CYP3A4 for elimination. Care had to be taken to examine for any potential interactions between these protease inhibitors and other medications the patient was taking, and a convenient website became available to check for such drug-drug interactions (*www.hep-druginteractions.org*).

Despite the improvement in SVRs with protease-inhibitor-based regimens for genotype 1 compared to PEG IFN-ribavirin (e.g., in treatment-naïve patients 66–79% vs 38–44%), triple-drug protease-inhibitor therapy was hampered by amplified intolerability, the complexity of response-guided regimens and futility stopping rules, the inconvenience of thrice-daily dosing with meals and a high pill burden, the need for PEG IFN injections and ribavirin with all their intolerability, and multiple drug-drug interactions. Moreover, side effects appeared to be more severe and burdensome once these drugs entered practice, especially in cirrhotic nonresponders, in whom studies reported from Europe showed serious adverse events in up to 45% and deaths in up to 3%. All these issues, as well as rapidly accelerating progress on next-generation and all-oral DAA therapy (see below), conspired to temper enthusiasm for these new antivirals; after a brief stint as recommended therapy (2011–2013), these drugs became obsolete and are no longer recommended.

CONTEMPORARY DIRECT-ACTING ANTIVIRAL COMBINATION THERAPY (2013–)

Since late 2013, the number of new antiviral agents for hepatitis C has expanded substantially, and, currently, PEG IFN-based treatments have been supplanted by six therapeutic regimens: all oral, IFN-free, highly efficacious (>95% SVR), well tolerated, with high barriers to resistance, simple dosing and low pill burdens, treatment durations as brief as 8 to 12 weeks, and, in many cases, pangenotypic efficacy (**Table 334-6**). These drugs are distributed among three classes of DAAs: NS3/4 protease inhibitors (which cleave the single HCV polyprotein into constituent structural and nonstructural proteins), NS5B nucleoside and nonnucleoside polymerase inhibitors (which interfere with the RNA-dependent RNA polymerase [a replicase] involved in synthesis of viral RNA), and NS5A inhibitors (which interfere with a membrane-associated phosphoprotein essential to the HVC RNA replication complex).

The first of the new DAA agents (approved in November 2013) was simeprevir, a second-generation protease inhibitor for genotype 1, followed shortly thereafter (December 2013) by sofosbuvir, a pangenotypic nucleoside polymerase inhibitor. For genotype 1, both of these agents had to be combined with PEG IFN and ribavirin; for genotypes 2 and 3, sofosbuvir was administered with ribavirin, without PEG IFN; however, these treatment regimens have been supplanted by combinations of all-oral, IFN-free, DAAs, and ribavirin is rarely needed, retained only for very limited indications.

Simeprevir: When simeprevir was used with PEG IFN, its efficacy (genotype 1b > 1a) was similar to that of first-generation protease inhibitors, but required only once-a-day dosing without the complexity of response-guided therapy. Similar to first-generation protease inhibitors, simeprevir was hampered by many drug-drug interactions and side effects (including photosensitivity, rash, and mild hyperbilirubinemia); moreover, patients, with HCV NS3 polymorphism Q80K had markedly reduced drug efficacy, necessitating pretreatment genetic testing and disqualifying a substantial proportion (approximately a third) of potential treatment candidates. Little about simeprevir supported its adoption in combination with PEG IFN and ribavirin. On the other hand, the combination of simeprevir (150 mg) along with sofosbuvir (400 mg) for 12 weeks was found to be effective in treatment-naïve (97% SVR_{12}) or treatment-experienced (95% SVR_{12}) patients without cirrhosis and in treatment-naïve (88% SVR_{12}) or treatment-refractory (79% SVR_{12}) patients with cirrhosis (it remains one of the recommended regimens for genotype 1).

Sofosbuvir: Sofosbuvir, the first nonprotease inhibitor DAA to be approved, has an excellent profile—high potency, high barrier to resistance, pangenotypic activity, very well tolerated with limited adverse effects (most commonly mild fatigue, insomnia, headache, and nausea), once-daily oral administration, and relative freedom from major drug-drug interactions. Sofosbuvir has efficacy in all genotypes (1 to 6); in treatment-naïve subjects and prior nonresponders to PEG IFN-based and protease-inhibitor-based therapy; with PEG IFN-RBV or in IFN-free regimens; in combination with RBV or with NS5A inhibitors; and for treatment periods as brief as 8 to 12 weeks to as long as 24 weeks. Currently, sofosbuvir is used in combination with either the protease inhibitor simeprevir (as described above) or, more commonly, with one of three NS5A inhibitors. Thus, sofosbuvir is a component of four of the six recommended DAA regimens for genotype 1, two of the four regimens for genotype 4, and both of the regimens for genotypes 2, 3, 5, and 6 (Table 334-6).

Sofosbuvir/ledipasvir: The DAA combination that has had a dominant role in the treatment of hepatitis C is sofosbuvir (400 mg) plus the NS5A inhibitor ledipasvir (90 mg) in a once-a-day, fixed-dose, single pill, approved in October 2014 for genotype 1 and in November 2015 for genotypes 4, 5, and 6. Phase-III trials were conducted in treatment-naïve noncirrhotic patients, in treatment-naïve cirrhotic and noncirrhotic patients, and in treatment-experienced cirrhotic and noncirrhotic patients treated for 8, 12, or 24 weeks, both with and without ribavirin. In treatment-naïve noncirrhotics, an SVR_{12} was achieved in 97–99% of subjects, and no benefit was observed by extending therapy from 12 to 24 weeks or by adding ribavirin. Moreover, for treatment-naïve, noncirrhotic patients with baseline HCV RNA $<6 \times 10^6$ IU/mL, a treatment duration of 8 weeks was as effective as one of 12 weeks (94–95% SVR_{12}), which may be a consideration for a proportion of patients. In cirrhotic patients, SVR_{12} was achieved in 97–100% of treatment-naïve subjects (no advantage of extending therapy from 12 to 24 weeks or of adding ribavirin); however, for cirrhotic prior nonresponders to IFN-based therapy, 12 weeks of therapy was inferior (86% SVR_{12}) to 24 weeks of therapy (100% SVR_{12}). This combination, which is equally effective in patients with HIV-HCV co-infection and in African-American patients, has been shown to be highly effective in patients with decompensated cirrhosis and in patients with hepatitis C after liver transplantation and after kidney transplantation. On the other hand, the safety and efficacy of sofosbuvir/ledipasvir in patients with advanced renal failure have not been established, and all sofosbuvir-containing regimens can be associated with severe bradycardia in patients taking the antiarrhythmic agent amiodarone, especially along with beta blockers; sofosbuvir-containing combinations are contraindicated with amiodarone. Drug-drug interactions are few, but P-gp inducers, like St. John's wort and rifampin, and proton-pump gastric acid inhibitors, like omeprazole, may reduce sofosbuvir/ledipasvir concentrations. Generally, responsiveness to sofosbuvir/ledipasvir is not reduced in patients with baseline RASs to these agents, with the exception of treatment-experienced patients who have baseline NS5A RASs (for whom EASL recommends adding ribavirin or, if ribavirin in contraindicated, extending treatment to 24 weeks).

Paritaprevir/ritonavir, ombitasvir, and dasabuvir: The combination of ritonavir (100 mg)-boosted paritaprevir (150 mg), a protease inhibitor; ombitasvir (25 mg), an NS5A inhibitor; dasabuvir (250 mg), a nonnucleoside polymerase inhibitor; ± weight-based ribavirin (total of five drugs) was approved in December 2014 for genotypes 1 and 4. Paritaprevir/ritonavir and ombitasvir, formulated in a single tablet, are taken once daily, and both dasabuvir (a separate pill) and weight-based ribavirin (when included in the regimen) are taken twice daily. In clinical trials, this combination achieved SVR_{12} rates of 87–100% in treatment-naïve and treatment-experienced patients with genotype 1; without ribavirin, this combination in

TABLE 334-6 Indications and Recommendations for Antiviral Therapy of Chronic Hepatitis C[a]

Standard Indications for Therapy

All patients with chronic HCV infection (detectable HCV RNA, with or without elevated ALT) except for those with short life expectancies owing to comorbid conditions.

Any stage of fibrosis; highest priority for advanced fibrosis [METAVIR stage 3]/cirrhosis [METAVIR stage 4] (pretreatment biopsy is no longer embraced and has been supplanted by noninvasive measures of fibrosis, e.g., imaging to determine liver elasticity)

Responsiveness in groups previously refractory to interferon-based therapy (HIV-HCV co-infection, renal insufficiency, African American and Latino ethnicity, *IL28B* non-C haplotype, obesity, insulin resistance, hepatic decompensation, etc.) is not diminished to contemporary direct-acting oral combination regimens.

Retreatment Recommended

Relapsers, partial responders, or nonresponders after a previous course of interferon-based therapy or prior direct-acting antiviral therapy (see genotype-specific recommendations below).

Antiviral Therapy Not Recommended

Pregnancy: No clinical studies of direct-acting antivirals during pregnancy are available. Ribavirin is contraindicated during pregnancy; therefore, any regimen including ribavirin should not be used. Sofosbuvir; sofosbuvir + ledipasvir; and paritaprevir/ritonavir + ombitasvir + dasabuvir are classified as pregnancy category B, but the other direct-acting antivirals do not have a pregnancy classification. Therefore, these therapies are not indicated routinely in pregnancy and should be used, with caution, only if the benefit of treatment outweighs the potential for fetal risk.

Therapeutic Regimens (based on AASLD-IDSA recommendations, *www.hcvguidelines.org*)[b]

The European Association for the Study of the Liver (EASL) issued recommendations in 2016; divergences from AASLD-IDSA recommendations are summarized as a footnote below.[c]

TREATMENT-NAÏVE OR RELAPSED AFTER PRIOR PEG IFN/RIBAVIRIN THERAPY

Genotype 1a

ledipasvir + sofosbuvir 12 weeks (consider 8 weeks for noncirrhotic patients with HCV RNA <6 × 10^6 IU/mL)

paritaprevir/ritonavir + ombitasvir + dasabuvir + RBV 12 weeks (no cirrhosis) or 24 weeks (cirrhosis)

sofosbuvir + simeprevir 12 weeks (no cirrhosis) or ± RBV 24 weeks (cirrhosis)

daclatasvir + sofosbuvir 12 weeks (no cirrhosis) or ± RBV 24 weeks (cirrhosis)

grazoprevir + elbasvir 12 weeks (no cirrhosis or cirrhosis sans ELB NS5A RASs) or + RBV × 16 weeks (ELB NS5A RASs)

sofosbuvir + velpatasvir 12 weeks

Genotype 1b

ledipasvir + sofosbuvir 12 weeks (consider 8 weeks for noncirrhotic patients with HCV RNA <6 × 10^6 IU/mL)

paritaprevir/ritonavir + ombitasvir + dasabuvir 12 weeks

sofosbuvir + simeprevir 12 weeks (no cirrhosis) or ± RBV 24 weeks (cirrhosis)

daclatasvir + sofosbuvir 12 weeks (no cirrhosis) or ± RBV 24 weeks (cirrhosis)

grazoprevir + elbasvir 12 weeks

sofosbuvir + velpatasvir 12 weeks

Genotype 2

sofosbuvir + velpatasvir 12 weeks

daclatasvir + sofosbuvir (no cirrhosis) 12 weeks or 16–24 weeks (cirrhosis)

Genotype 3

sofosbuvir + velpatasvir 12 weeks

daclatasvir + sofosbuvir 12 weeks (no cirrhosis) or ± RBV 24 weeks (cirrhosis)

Genotype 4

sofosbuvir + velpatasvir 12 weeks

ledipasvir + sofosbuvir 12 weeks

paritaprevir/r + ombitasvir + RBV 12 weeks (no dasabuvir)

grazoprevir + elbasvir 12 weeks

Genotypes 5, 6

sofosbuvir + velpatasvir 12 weeks

ledipasvir + sofosbuvir 12 weeks

FAILED PRIOR PEG IFN/RIBAVIRIN THERAPY, NO CIRRHOSIS

Genotype 1a

ledipasvir + sofosbuvir 12 weeks

paritaprevir/ritonavir + ombitasvir + dasabuvir + RBV 12 weeks

sofosbuvir + simeprevir 12 weeks

daclatasvir + sofosbuvir 12 weeks

grazoprevir + elbasvir 12 weeks (without ELB NS5A RASs) or + RBV x 16 weeks (ELB NS5A RASs)

sofosbuvir + velpatasvir 12 weeks

Genotype 1b

ledipasvir + sofosbuvir 12 weeks

paritaprevir/ritonavir + ombitasvir + dasabuvir 12 weeks

sofosbuvir + simeprevir 12 weeks

daclatasvir + sofosbuvir 12 weeks

grazoprevir + elbasvir 12 weeks

sofosbuvir + velpatasvir 12 weeks

Genotype 2

sofosbuvir + velpatasvir 12 weeks

daclatasvir + sofosbuvir 12 weeks

Genotype 3

sofosbuvir + velpatasvir 12 weeks

daclatasvir + sofosbuvir 12 weeks

Genotype 4

sofosbuvir + velpatasvir 12 weeks

ledipasvir + sofosbuvir 12 weeks

paritaprevir/r + ombitasvir + RBV 12 weeks (no dasabuvir)

grazoprevir + elbasvir 12 weeks (prior relapse) or + RBV 16 weeks (prior nonresponse)

Genotypes 5, 6

sofosbuvir + velpatasvir 12 weeks

ledipasvir + sofosbuvir 12 weeks

FAILED PRIOR PEG IFN/RIBAVIRIN THERAPY, COMPENSATED CIRRHOSIS

Genotype 1a

ledipasvir + sofosbuvir + RBV 12 weeks

ledipasvir + sofosbuvir 24 weeks

sofosbuvir + velpatasvir 12 weeks

grazoprevir + elbasvir 12 weeks (without ELB NS5A RASs) or + RBV × 16 weeks (ELB NS5A RASs)

paritaprevir/ritonavir + ombitasvir + dasabuvir + RBV 24 weeks

sofosbuvir + simeprevir ± RBV 24 weeks (no Q80K variant)

daclatasvir + sofosbuvir ± RBV 24 weeks

Genotype 1b

ledipasvir + sofosbuvir + RBV 12 weeks

ledipasvir + sofosbuvir 24 weeks

sofosbuvir + velpatasvir 12 weeks

grazoprevir + elbasvir 12 weeks

paritaprevir/ritonavir + ombitasvir + dasabuvir 12 weeks

sofosbuvir + simeprevir ± RBV 24 weeks

daclatasvir + sofosbuvir ± RBV 24 weeks

Genotype 2

sofosbuvir + velpatasvir 12 weeks

sofosbuvir + daclatasvir 16 or 24 weeks

Genotype 3

sofosbuvir + velpatasvir 12 weeks

daclatasvir + sofosbuvir + RBV 24 weeks

Genotype 4

sofosbuvir + velpatasvir 12 weeks

ledipasvir + sofosbuvir + RBV 12 weeks

(*Continued*)

TABLE 334-6 Indications and Recommendations for Antiviral Therapy of Chronic Hepatitis C[a] (Continued)

paritaprevir/ritonavir + ombitasvir + RBV 12 weeks (no dasabuvir)	**FEATURES ASSOCIATED WITH REDUCED RESPONSIVENESS TO DIRECT-ACTING ANTIVIRAL COMBINATION THERAPY**
grazoprevir + elbasvir 12 weeks (prior relapse) or + RBV 16 weeks (prior nonresponse)	Genotype and subtype (genotype 1a less responsive than genotype 1b for several drugs)
ledipasvir + sofosbuvir 24 weeks	Treatment experience
Genotypes 5, 6	Advanced fibrosis (bridging fibrosis, cirrhosis)
sofosbuvir + velpatasvir 12 weeks	Reduced adherence
ledipasvir + sofosbuvir 12 weeks	

[a]Rapidly evolving new recommendations continue to be issued; for up-to-date treatment recommendations, please see *www.hcvguidelines.org*. [b]Class-I recommendations in **bold** font, all others are Class-II recommendations [c]The following EASL recommendations differ from those of AASLD-IDSA (Please note that, although mentioned in EASL recommendations, testing for baseline RASs is not recommended routinely, but, if reliable resistance testing available, results can be used to guide therapy.):

Genotype 1

For genotype 1, simeprevir + sofosbuvir is not recommended.

For genotype 1a, treatment-experienced patients (IFN-based regimen failures) treated with sofosbuvir + ledipasvir should have weight-based ribavirin added. If reliable testing for RASs is available, ribavirin is needed only if baseline RASs are present, and, in such patients, if ribavirin is contraindicated, sofosbuvir + ledipasvir should be extended to 24 weeks.

For genotype 1b, in treatment-naïve, noncirrhotic patients receiving paritaprevir/ritonavir + ombitasvir + dasabuvir a treatment duration of 8 weeks can be considered.

For genotype 1a, in patients treatment with grazoprevir + elbasvir, EASL recommends testing for ELB RASs even in noncirrhotics. If resistance testing is not done, the level of baseline HCV RNA should determine whether ribavirin is added and the duration of therapy. If HCV RNA >800,000 IU/mL, add ribavirin and treat for 16 weeks; if HCV RNA ≤800,000 IU/mL, ribavirin is not added, and treatment for 12 weeks suffices. If baseline testing for RASs is available, patients with HCV RNA >800,000 IU/mL and detectable RASs should be treated with ribavirin for 16 weeks. Treatment without ribavirin and for 12 weeks suffices if HCV RNA ≤800,000 IU/mL even with detectable RASs or even if HCV RNA >800,000 IU/mL with undetectable RASs.

For genotype 1a, in treatment-experienced patients (IFN-based regimen failures) treated with daclatasvir + sofosbuvir, follow the same recommendations described above for ledipasvir + sofosbuvir regarding the addition of ribavirin.

Genotype 2

EASL recommendations are the same as those of AASLD-IDSA.

Genotype 3

For treatment-experienced patients (IFN-based regimen failures) treated with sofosbuvir + velpatasvir or sofosbuvir + daclatasvir, if testing for baseline RASs is not available, add weight-based ribavirin. If resistance testing is available, ribavirin is needed only if baseline RASs are present, and, in such patients, if ribavirin is contraindicated, treatment should be extended to 24 weeks.

Genotype 4

Treatment-experienced patients (IFN-based regimen failures) treated with sofosbuvir + ledipasvir should have weight-based ribavirin added, and, in such patients, if ribavirin is contraindicated, treatment should be extended to 24 weeks.

In treatment-experienced patients (IFN-based regimen failures) treated with grazoprevir + elbasvir, if HCV RNA >800,000 IU/mL, weight-based ribavirin should be added, and treatment should be extended to 16 weeks.

EASL recommends two additional treatment options for genotype 4 (noncirrhotic or cirrhotic) that are not included in AASLD-IDSA guidelines: sofosbuvir + daclatasvir and sofosbuvir + simeprevir. For both these options, treatment-naïve patients should be treated for 12 weeks without ribavirin; treatment-experienced (IFN-based regimen failures) patients should be treated with ribavirin for 12 weeks or, if ribavirin is contraindicated, without ribavirin for 24 weeks.

Genotypes 5 and 6

Treatment-experienced patients (IFN-based regimen failures) treated with sofosbuvir + ledipasvir should have weight-based ribavirin added, and, in such patients, if ribavirin is contraindicated, treatment should be extended to 24 weeks.

EASL recommends an additional treatment option for genotype 5 and 6 (noncirrhotic or cirrhotic) that is not included in AASLD-IDSA guidelines: sofosbuvir + daclatasvir. Treatment-naïve patients should be treated for 12 weeks without ribavirin; treatment-experienced (IFN-based regimen failures) patients should be treated with ribavirin for 12 weeks or, if ribavirin is contraindicated, without ribavirin for 24 weeks.

Drug doses: sofosbuvir 400 mg; ledipasvir 90 mg; paritaprevir 150 mg; ritonavir 100 mg; ombitasvir 25 mg; dasabuvir 250 mg; ribavirin, weight-based: 1000 mg (<75 Kg)–1200 mg (≥75 kg); simeprevir 150 mg; daclatasvir 60 mg; elbasvir 50 mg; grazoprevir 100 mg; velpatasvir 100 mg.

Abbreviations: AASLD, American Association for the Study of Liver Diseases; ALT, alanine aminotransferase; ELB NS5A RASs, elbasvir NS5A resistance-associated substitutions; HCV, hepatitis C virus; IFN, interferon; IDSA, Infectious Diseases Society of America; PEG IFN, pegylated interferon; IU, international units (1 IU/mL is equivalent to ~2.5 copies/mL); RASs, resistance-associated substitutions; RBV, ribavirin.

genotype 1a is ~7% less responsive than genotype 1b. Therefore, in treatment-naïve patients with genotype 1a, this combination is administered *with* ribavirin for 12 weeks in the absence of cirrhosis (95–97% SVR_{12}) or for 24 weeks in the presence of compensated cirrhosis (94% SVR_{12}), while in patients with genotype 1b, the combination does not require ribavirin, and the duration of therapy is 12 weeks for both noncirrhotics and cirrhotics (99–100% SVR_{12}). In prior nonresponders without cirrhosis, the combination is administered for 12 weeks, *with* ribavirin in genotype 1a (96% SVR_{12}), *without* ribavirin in genotype 1b (100% SVR_{12}). In prior nonresponders with cirrhosis, the combination is administered for 24 weeks *with* ribavirin in genotype 1a (SVR_{12} 100% in prior relapsers and partial responders, 95% in prior null responders [in whom treatment without ribavirin was associated with an 80% SVR_{12}]), but only for 12 weeks and *without* ribavirin in genotype 1b (100% SVR_{12}). For genotype 4, the regimen is given for 12 weeks with ribavirin, but without dasabuvir in treatment-naïve and treatment-experienced patients (100% SVR_{12}), including those with compensated cirrhosis. In July 2016, the FDA approved a long-acting formulation of dasabuvir, allowing once-a-day instead of twice-a-day treatment; for genotype 1a, twice-daily ribavirin dosing remains.

This combination is well tolerated with generally mild side effects, for example, fatigue, asthenia, insomnia, headache, and pruritus. Hyperbilirubinemia (primarily unconjugated) and elevations in alanine aminotransferase activity may occur but resolve during or shortly after treatment. Because of occasional hyperbilirubinemia and potential hepatotoxicity (FDA warning letter issued October 2015 regarding hepatic failure/decompensation reported in treated cirrhotic patients), this combination is not recommended in patients with decompensated cirrhosis, and treated cirrhotic patients should be monitored closely for decompensation; however, the safety and efficacy of this combination have been demonstrated for patients with advanced renal insufficiency. Similar to other regimens containing protease inhibitors, drug-drug interactions are common with other drugs that induce CYP3A4 or are dependent on CYP3A4 for elimination. Checking for potential drug-drug interactions is important prior to initiating therapy with this drug combination (*www.hep-druginteractions.org*). Responsiveness to this multidrug

regimen is not reduced in patients with baseline RASs to these agents.

Compared to sofosbuvir/ledipasvir, this regimen has the disadvantage of requiring twice-a-day ribavirin therapy for genotype 1a and of being contraindicated in decompensated cirrhosis; however, it has the advantage of offering a 12-week, ribavirin-free regimen for prior null responders with cirrhosis and providing an option for patients with renal failure.

Sofosbuvir and Daclatasvir: Daclatasvir, an NS5A inhibitor, along with the polymerase inhibitor sofosbuvir, was approved by the FDA in July 2015 for genotype 3 and in February 2016 for genotype 1 (AASLD-Infectious Diseases Society of America [IDSA] guidelines [see below] include its recommendation as well for genotype 2; in August 2014, this combination was approved in Europe for genotypes 1, 2, 3, and 4, and EASL recommends it for all these genotypes as well as for genotypes 5 and 6). At the time of its approval for genotype 3, daclatasvir filled a need inadequately met by other available combination DAAs. Although data on genotype 3 are the most robust, clinical trials of this combination in genotypes 1 and 2 support its efficacy and recommendations for first-line (genotype 1) and alternative (genotype 2) treatment, in some cases with ribavirin (Table 334-6). Daclatasvir, a 60-mg tablet, and sofosbuvir, a separate 400 mg tablet are taken once-a-day for 12 to 24 weeks.

In clinical trials among treatment-naïve or treatment-experienced patients, SVR_{12} rates for 12 weeks of daclatasvir plus sofosbuvir were 98% with genotype 1 (comparable results in genotypes 1a and 1b), 92% for genotype 2, and 89% for genotype 3. For noncirrhotic patients, the addition of ribavirin or the extension of therapy to 24 weeks did not improve efficacy. In patients with compensated cirrhosis, limited prospective data and data from observational cohorts suggested that extending therapy to 24 weeks, with or without ribavirin, improved efficacy. In cirrhotics, SVR_{12} was achieved in 93% with Child Class-Pugh A and B but in only 56% with Class-C decompensated cirrhosis. For patients with genotype 3 and cirrhosis, the combination was effective in treatment-naïve patients (94% SVR_{12}), but less so in prior nonresponders (69% SVR_{12}). Outcomes in patients with HIV-HCV co-infection were comparable.

Like other sofosbuvir-NS5A inhibitor combinations, daclatasvir plus sofosbuvir is well tolerated (mild fatigue, headache, nausea, diarrhea in 5–14%), but can cause severe bradycardia when administered with amiodarone (contraindicated), especially along with beta blockers. Because daclatasvir is a substrate for CYP3A, CYP3A inducers can reduce daclatasvir levels, and CYP3A inhibitors reduce daclatasvir levels. Similarly, daclatasvir, an inhibitor of P-gp, OATP1B1 and 1B3, and BCP, can increase the levels of drugs that are substrates of these transporters. As noted above for other DAAs, checking for potential drug-drug interactions is advisable prior to initiating therapy (*www.hep-druginteractions.org*). Responsiveness to daclatasvir-containing drug-combination therapy is reduced in cirrhotic patients with genotype 1a and in both cirrhotic and noncirrhotic patients with genotype 3 who have baseline daclatasvir-associated NS5A RASs.

Although daclatasvir-sofosbuvir is approved for genotypes 1 and 3 and recommended as an alternative for genotype 2, better documented efficacy and simplicity of other regimens have limited the popularity of this drug combination.

Elbasvir/Grazoprevir: Elbasvir (50 mg), an NS5A inhibitor, combined in a single, fixed-dose pill with grazoprevir (100 mg), an NS3/4 protease inhibitor, was approved in January 2016 as a once-a-day (with or without food) treatment for genotypes 1 and 4. In clinical trials, a 12-week course was effective in treatment-naïve and treatment-experienced patients without cirrhosis or with compensated cirrhosis. In treatment-naïve patients, this combination yielded an SVR_{12} in 92% of patients with genotype 1a, 99% with genotype 1b, and 100% with genotype 4 (very small numbers, however); 10 patients with genotype 6 were included, but only 80% achieved SVR_{12}. Cirrhotic and noncirrhotic patients had comparable rates of SVR_{12}, 97% and 94%, respectively. For this drug combination, however, ~11% of patients with genotype 1a harbor NS5A polymorphisms, that is, RASs, at baseline. If present, these NS5A RASs reduce efficacy of elbasvir/grazoprevir (unlike baseline RASs to the most of the other combination DAA regimens described above and below) from 99% to 58% in treatment-naïve patients. Therefore, all patients with genotype 1a require baseline RAS testing; if these RASs are present, treatment extension to 16 weeks and the addition of weight-based ribavirin bring the SVR_{12} up to expected levels of close to 100%. In treatment-experienced patients, both extending treatment to 16 weeks and adding ribavirin were studied; however, generally, in the absence of baseline NS5A RASs, SVR_{12} rates were not increased over those without ribavirin for 12 weeks (94–97%). For genotype 1a, among prior nonresponders to PEG IFN/ribavirin, 12 weeks of elbasvir/grazoprevir suffices without ribavirin except for patients with baseline NS5A RASs, who require 16 weeks of therapy and ribavirin. Among nonresponders to prior protease-inhibitor therapy, even in the absence of baseline NS5A RASs, ribavirin should be added to a 12-week regimen; in the presence of baseline NS5A RASs, treatment should be extended to 16 weeks and ribavirin added. For genotype 1b, NS5A RASs are not an issue, and the only subgroup requiring modification of a 12-week course of therapy are prior nonreponders to protease-inhibitor regimens, for whom ribavirin is added. For genotype 4, the recommended regimen for all prior nonresponders (whether to PEG IFN/ribavirin or protease inhibitor regimens) is 16 weeks of elbasvir/grazoprevir plus ribavirin (Table 334-6).

This combination is just as effective in patients with HIV-HCV co-infection and in patients with advanced renal failure (including those requiring hemodialysis); however, it is contraindicated in decompensated cirrhosis. Like other protease inhibitor regimens, elbasvir/grazoprevir can be associated with aminotransferase elevations and potential hepatotoxicity; because these drugs are excreted by the liver, in decompensated liver disease, plasma drug concentrations may become elevated substantially. Therefore, all treated patients should have alanine aminotransferase screening periodically during therapy, and the drug should be stopped for elevations exceeding 10-fold or for elevations of conjugated bilirubin, alkaline phosphatase, or prothrombin time.

Elbasvir/grazoprevir is well tolerated, with only low levels of mild adverse effects (fatigue, headache, nausea in 5–11%) seen just as frequently in placebo recipients. Both elbasvir and grazoprevir are substrates for CYP3A and are subject to multiple potential drug-drug interactions. Therefore, this combination should not be used with potent CYP3A inducers; conversely, CYP3A and OATP1B1 inhibitors can lead to untoward elevations of plasma elbasvir/grazoprevir concentrations. Checking for potential drug-drug interactions is advisable prior to initiating therapy (*www.hep-druginteractions.org*).

Compared to other available regimens for genotypes 1 and 4, elbasvir/grazoprevir has the disadvantage/inconvenience of requiring baseline NS5A RAS testing but the advantages of a comparable regimen for cirrhotics and noncirrhotics, for treatment-naïve and treatment-experienced patients, and for patients with normal renal function and with renal failure.

Sofosbuvir/velpatasvir: The combination in a single, fixed-dose pill of velpatasvir (100 mg), a highly potent, pangenotypic NS5A inhibitor, along with the polymerase inhibitor sofosbuvir (400 mg) was approved in June 2016 for genotypes 1–6, in treatment-naïve and treatment-experienced noncirrhotics and cirrhotics. Ribavirin is not required, including in patients with genotypes 2 and 3, except in patients with decompensated cirrhosis.

In a series of clinical trials, this combination for 12 weeks in the absence of ribavirin was shown to yield 99% SVR_{12} (range 97–100%) in genotypes 1, 2, 4, 5, and 6 and 95% in genotype 3. Baseline NS5A RASs had no impact on responsiveness.

Prior to the availability of this drug combination, patients with genotype 3, especially those with cirrhosis and prior null response to other therapies, proved to be the most refractory subset of patients. In treatment-naïve patients with genotype 3, 12 weeks of sofosbuvir/velpatasvir (95% SVR_{12}) was superior to 24 weeks of sofosbuvir plus ribavirin (80% SVR_{12}). In patients with genotype 3, the combination of sofosbuvir/velpatasvir for 12 weeks was comparable in noncirrhotics (97% SVR_{12}) and cirrhotics (91% SVR_{12}) and in

treatment-naïve (97% SVR_{12}) and treatment-experienced (90% SVR_{12}) patients, superior in all these categories to 24 weeks of sofosbuvir plus ribavirin (87%, 66%, 86%, and 63%, respectively). In cirrhotic null responders, most available IFN-free regimens for genotype 3 (including daclatasvir plus sofosbuvir, approved specifically for this genotype) achieved SVR_{12} rates in the range of ~60–75%, while the combination of PEG IFN, ribavirin, and sofosbuvir could boost SVR_{12} to the mid-80% range. For treatment-experienced patients with genotype 3, sofosbuvir/velpatasvir in noncirrhotics and cirrhotics had similarly high efficacy (91% and 89% SVR_{12}, respectively); this was the highest recorded SVR_{12} for genotype-3 cirrhotic null responders treated with IFN-free DAA regimens. Finally, in patients with genotypes 1–4 and 6 and with decompensated, Class-B cirrhosis (55% treatment-experienced), sofosbuvir/velpatasvir plus ribavirin for 12 weeks yielded an SVR_{12} in 94%; this result was better than sofosbuvir/velpatasvir without ribavirin for 12 weeks (83% SVR_{12}) or 24 weeks (86% SVR_{12}).

Like other all-oral DAAs, sofosbuvir/velpatasvir was very well tolerated; in noncirrhotic and compensated cirrhotic patients, mild headache and fatigue was seen in >10%—this occurred in a comparable proportion of placebo recipients; in decompensated cirrhosis, mild fatigue, headache, nausea, insomnia, diarrhea, and anemia (ribavirin was part of the regimen) was seen in >10%. Like other sofosbuvir-containing regimens, sofosbuvir/velpatasvir should not be administered along with amiodarone (potential serious bradycardia); in addition, P-gp inducers and moderate-to-potent CYP3A inducers can reduce plasma levels of sofosbuvir and/or velpatasvir. Checking for drug-drug interactions prior to therapy is advisable (*www.hep-druginteractions.org*). Baseline RASs do not influence responsiveness to this combination.

FUTURE DIRECT-ACTING ANTIVIRAL COMBINATION THERAPY (2017–)

Most treatment needs have been met by contemporary DAA regimens described above; however, several additional, highly potent, pangenotypic drug combinations are in development. For example, an investigative protease inhibitor (voxilaprevir) added to the polymerase inhibitor/NS5A inhibitor combination of sofosbuvir/velpatasvir yields a very well tolerated *triple-drug* combination with 97% SVR_{12} across all HCV genotypes and patient subgroups. These include noncirrhotic/cirrhotic, treatment-naïve/treatment-experienced groups, including those who had prior NS5A treatment and results were independent of the number of prior DAA drug classes received; no effects of baseline NS5A RASs were noted. Several experimental combinations may allow even briefer durations of therapy. In a small, exploratory trial, a *6-week* combination of sofosbuvir plus an experimental pangenotypic, very high potency, very low resistance NS5A inhibitor (odalasvir) achieved SVR_{12} in 100% of 12 patients with genotype 1. Similarly, in a *6-week* triple combination of odalasvir with the protease inhibitor simeprevir and an experimental polymerase inhibitor ("AL-335"), SVR_{12} was observed in 100% of 20 treatment-naïve noncirrhotic patients with genotype 1. In phase-II clinical trials, *8 weeks* of an experimental combination of two high-potency, pangenotypic DAAs, a protease inhibitor ("ABT-493") plus an NS5A inhibitor ("ABT-530"), yielded 100% SVR_{12} in treatment-naïve noncirrhotic patients with genotypes 1, 2, and 3. In cirrhotics with genotype 3 and in patients with genotypes 4, 5, and 6, 12 weeks of therapy with this DAA combination yielded 100% SVR_{12}. In patients with prior DAA treatment failure, 12 weeks of this double-combination sufficed to achieve a ≥95% SVR_{12}; neither baseline NS5A nor protease inhibitor RASs influenced SVR_{12} rates. No safety issues have been encountered, and the potential for drug-drug interactions is limited. These promising combinations are undergoing phase-II and phase-III trials.

Less advanced is the development of inhibitors of host proteins, such as oral, nonimmunosuppressive inhibitors of cyclophilin A (which interacts with NS5A during HCV replication) and subcutaneous antisense antagonists of host liver-expressed micro-RNA-122 (which promotes HCV replication). Given the accelerated progress of all-oral, short-treatment-duration, high-efficacy, DAAs, these alternative approaches may not be practical or competitive; moreover, development of both approaches has been retarded by emerging toxicities such as pancreatitis associated with cyclophilin inhibitors and jaundice associated with micro-RNA-122.

Although data on the impact of DAAs on the natural history of chronic hepatitis C are still limited, preliminary findings are that successful therapy is associated with a gradual reduction in fibrosis progression and a regression of advanced fibrosis (cirrhosis), improvement in survival among patients with decompensated cirrhosis, and a decline in the number of patients with hepatitis C being referred for liver transplantation. Based on the known prevalence, natural history, and rate of progression of chronic hepatitis C and on the efficacy of DAA therapies and their impact on the complications of hepatitis C, modeling estimates have suggested that the availability and application of these therapies have the potential to reduce the hepatitis C-associated disease burden including liver-related death, HCC, decompensated cirrhosis, and liver transplantation by 50–70% between 2015 and 2050.

TREATMENT RECOMMENDATIONS

Because the pace of new drug development and approval has been so rapid, the AASLD and the IDSA have been providing a consensus of updated treatment recommendations for patients with hepatitis C; these recommendations, which continue to be revised regularly based on new data, are available online at *www.hcvguidelines.org* and should be consulted before initiating therapy (Table 334-6). The EASL issues similar (but not identical) treatment recommendations annually for hepatitis C (*www.easl.eu*), most recently in September 2016. Divergences between AASLD-IDSA and EASL recommendations are noted in Table 334-6.

Prior to therapy, HCV genotype should be determined, because the genotype dictates which treatment regimens are indicated (Table 334-6). Monitoring of serum HCV RNA levels pretreatment, during treatment, and posttreatment is crucial in assessing response to therapy; moreover, the baseline level may contribute to determining the duration of therapy (e.g., in noncirrhotic patients with genotype 1 and HCV RNA $<6 \times 10^6$ IU/mL, 8 [instead of the usual 12] weeks of sofosbuvir/ledipasvir may be a consideration). The goal of treatment is to eradicate HCV RNA during therapy and to document that the virus remains undetectable for at least 12 weeks after completion of therapy (SVR_{12}). Several reports have appeared describing hepatitis B reactivation, often severe, during and after DAA therapy in patients coinfected with HCV and HBV who were not being treated for their HBV infections. Therefore, screening for HBV infection is recommended prior to initiating DAA therapy for hepatitis C (which should have been done to determine HBV-immunity status as a prelude to recommended hepatitis B vaccination in patients with chronic hepatitis C), and therapy for HBV infection (for those meeting HBV treatment criteria, see above) should be initiated prior to or simultaneously with HCV therapy.

INDICATIONS FOR ANTIVIRAL THERAPY

Patients with chronic hepatitis C who have detectable HCV RNA in serum, whether or not aminotransferase levels are increased, and chronic hepatitis of any grade and stage are candidates for antiviral therapy with DAA agents. The only exception would be patients with short life expectancies, for whom treating hepatitis C would have no influence on longevity. Certainly, for patients with advanced liver disease, early treatment merits a high priority. Although patients with persistently normal aminotransferase activity tend to progress histologically very slowly or not at all, they respond to antiviral therapy just as well as do patients with elevated aminotransferase levels; therefore, such patients are potential candidates for antiviral therapy. As noted above, antiviral therapy has been shown to improve survival and complication-free survival and to slow progression of and to reverse fibrosis.

HCV genotype determines the regimen to be selected (Table 334-6). Similarly, the absence or presence of cirrhosis/advanced fibrosis determines the treatment options from which to select, including

the antiviral agents to be used, the duration of therapy, and the need for ribavirin (Table 334-6). A pretreatment liver biopsy to assess histologic grade and stage provides substantial information about progression of hepatitis C in the past, has prognostic value for future progression, and can identify such histologic factors as steatosis and stage of fibrosis, which can influence responsiveness to therapy. As therapy has improved for patients with a broad range of histologic severity, and as noninvasive measures of the stage of fibrosis (e.g., assessment of liver elasticity by imaging) have gained in accuracy and popularity, noninvasive approaches have supplanted histology in most cases. If cirrhosis/advanced fibrosis is present prior to therapy, the risk of HCC, although reduced substantially by successful therapy, is not eliminated, and twice yearly posttreatment imaging for HCC surveillance (and endoscopic surveillance for esophageal varices at intervals of 1–3 years) is indicated even after an SVR. In patients with low-level fibrosis at baseline, achievement of an SVR allows the cessation of such surveillance.

Patients who have relapsed after, or failed to respond to, a course of IFN-based or DAA agent-based therapy are candidates for retreatment with a DAA therapy regimen (Table 334-6). For patients who have failed to respond to a DAA combination, options include increasing the duration of therapy with the failed regimen, adding ribavirin, or changing the drug class (e.g., after failed protease and polymerase inhibitors, switching to an NS5A-containing combination). In the presence of cirrhosis or a need for urgent retreatment, patients who have failed protease inhibitor plus polymerase inhibitor combination therapy or who have failed an NS5A combination are candidates for RAS testing and tailored therapy based on such resistance testing. If reliable RAS testing is not available, adding ribavirin or extending the duration of therapy are options. For prior nonresponders to IFN-based therapy, NS5A inhibitor-containing regimens are highly effective; however, reduced responsiveness can be encountered, especially in cirrhotic patients. For this relatively refractory group, ideally, the most potent/effective NS5A regimen should be selected to give such patients the best chance of responding and to avoid treatment-emergent NS5A RASs. Additional details for treatment of such patient subgroups can be found at *www.hcvguidelines.org*.

Persons with acute hepatitis C are also candidates for antiviral therapy (**Chap. 332**) with the same DAA agents approved for chronic hepatitis C; delaying the initiation of therapy for an observation period of 12–16 weeks (and even up to 6 months) has been recommended to allow for spontaneous recovery, especially in light of the fact that most cases of acute hepatitis C are not clinically severe or rapidly progressive. The duration of therapy for acute hepatitis C has not been determined definitively; however, in a small study of 20 patients, 6 weeks of sofosbuvir/ledipasvir sufficed for a 100% SVR_{12}. According to 2016 EASL recommendations, patients with acute hepatitis C should be treated for 8 weeks with a genotype-appropriate DAA regimen consisting of sofosbuvir plus one of the three approved NS5A inhibitors without ribavirin (extended to 12 weeks for patients with acute hepatitis C and HIV co-infection or for patients with acute hepatitis C and a baseline HCV RNA level >1 million IU/mL). In patients with biochemically and histologically mild chronic hepatitis C, the rate of progression is slow; however, such patients respond just as well to antiviral therapy as those with elevated aminotransferase levels and more histologically severe hepatitis. Because of the high cost of DAA treatments, initially a higher priority was assigned to patients with advanced fibrosis/cirrhosis; however, this controversial approach was relied upon by some medical insurers and pharmacy benefit management organizations to withhold therapy from patients with low-level fibrosis. Unfortunately, delaying therapy until fibrosis becomes advanced misses the opportunity to prevent all the dire consequences of chronic hepatitis C (liver failure, death/transplantation, HCC), which can be reduced, but not eliminated completely once advanced fibrosis is established. Therefore, therapy for patients with mild disease is justified as well as cost-effective.

Patients with compensated cirrhosis can respond to therapy, and their likelihood of a sustained response with DAAs is comparable to that in noncirrhotics. Patients with decompensated cirrhosis, who were not candidates for IFN-based antiviral therapy, respond well to DAA therapy regimens consisting of combinations of polymerase inhibitors and NS5A inhibitors (e.g., sofosbuvir/ledipasvir, sofosbuvir/velpatasvir); however, protease-inhibitor-containing combinations have been associated with potential hepatotoxicity and hepatic decompensation and are contraindicated in this patient subset. Patients with decompensated cirrhosis should be referred to a liver transplantation center. DAAs are highly effective not only for patients with end-stage liver disease awaiting liver transplantation but also for patients with recurrent hepatitis C after liver transplantation. Ideally, patients should be treated prior to liver transplantation; however, a concern is that eradication of HCV infection will disqualify such patients from accepting donor livers from persons with HCV infection, thus contracting the potential donor pool and limiting accessibility to donor organs and timely transplantation. In addition, responsiveness to DAA therapy appears to be reduced in patients with decompensated cirrhosis and with high model for end-stage liver disease (MELD) scores; in this subgroup, responsiveness after liver transplantation would be substantially better. Therefore, advocacy has been expressed (recommended by EASL) for postponing DAA therapy in patients with high-MELD HCV-associated end-stage liver disease until after liver transplantation; the decision whether to treat pretransplantation or posttransplantation should be individualized thoughtfully for each patient, based on such factors as MELD score, time anticipated prior to availability of a donor organ, relative clinical stability, and co-morbidities **(Chap. 338)**. The cutaneous and renal vasculitis of HCV-associated essential mixed cryoglobulinemia **(Chap. 332)** may respond to antiviral therapy, but sustained responses were rare after discontinuation of therapy in the IFN era, and prolonged, potentially indefinite, therapy was recommended. Now that more effective DAAs are available, a 12-week course of sofosbuvir-based combination therapy has been shown to yield an SVR_{12} rate exceeding 80% in cryoglobulinemic vasculitis. Anecdotal reports suggest that IFN-based antiviral therapy may be effective in porphyria cutanea tarda or lichen planus associated with hepatitis C; whether the more appealing DAAs are effective in these groups remains to be documented.

In patients with HCV/HIV co-infection, hepatitis C is more progressive and severe than in HCV-monoinfected patients. Although patients with HCV/HIV co-infection responded less well to IFN-based antiviral therapy for hepatitis C, they respond as well as patients with HCV infection alone to DAA combination regimens. In HCV/HIV-infected patients, ribavirin can potentiate the toxicity of didanosine (e.g., lactic acidosis) and the lipoatrophy of stavudine, and zidovudine can exacerbate ribavirin-associated hemolytic anemia; therefore, these drug combinations should be avoided.

Patients with a history of injection drug use and alcoholism can be treated successfully for chronic hepatitis C, preferably in conjunction with drug and alcohol treatment programs. Moreover, because injection-drug users, as a source of transmission to others, account disproportionately for perpetuating the spread of HCV infection in the population, the impact of treating active injection-drug users is amplified by reducing such transmission. The approved oral combinations of DAAs are effective in patients with mild-modest renal failure and require no dose adjustments; however, in patients with severe renal impairment (creatinine clearances <30 mL/minute), data are limited on the use of sofosbuvir-containing combinations. For such patients, including those undergoing hemodialysis, recommended combinations are 12 weeks of elbasvir/grazoprevir for genotypes 1a, 1b, and 4 or 12 weeks of paritaprevir/ritonavir, ombitasvir, and dasabuvir for genotype 1b. In genotype 1a, the addition of 200 mg/day of ribavirin to paritaprevir/ritonavir, ombitasvir, and dasabuvir, if the hemoglobin level exceeds 10 g/dL, is an alternative regimen but requires vigilance for the onset of ribavirin-induced hemolytic anemia. For patients with severe renal impairment and HCV genotypes 2, 3, 5, or 6, PEG IFN with low-dose ribavirin (200 mg daily, if the hemoglobin exceeds 10 g/dL) is recommended. After renal transplantation, levels of SVR_{12} in patients treated with the approved oral DAA combinations have approached 100%.

No clinical studies of the use of DAAs during pregnancy are available. Ribavirin is contraindicated during pregnancy; therefore, any regimen including ribavirin should not be used. Sofosbuvir; sofosbuvir + ledipasvir; and paritaprevir/ritonavir, ombitasvir, and dasabuvir are classified as pregnancy category B; the other DAAs do not have a pregnancy classification. Therefore, these therapies are not indicated routinely in pregnancy and should be used, with caution, only if the benefit of treatment is compelling and justified compared to the potential for fetal risk.

Choosing among available treatment options: The large number of recommended all-oral DAA combinations can be daunting to treating clinicians. In some instances, the combination approved is determined by insurance payers; however, cost considerations aside, how is the clinician to choose among the options? The most popular of the regimens has been fixed-dose, single-pill sofosbuvir/ledipasvir, which is effective for all genotypes except 2 and 3, which requires no baseline RAS testing, and which can be used in noncirrhotic patients with genotype 1 and low-level viremia for as brief a period as 8 weeks. For genotypes 2 and 3, fixed-dose, single-pill sofosbuvir/velpatasvir appears to be the combination of choice; because this combination is so effective across all genotypes, in the future, for simplicity, clinicians may resort to a "one-size-fits-all" regimen such as this one in all patients (except for those with advanced renal failure). In addition, this regimen is the only one that can be used in almost all situations (independent of genotype, treatment experience, and cirrhosis) without ribavirin, and the duration of which is almost always 12 weeks; exceptions: (a) ribavirin recommended for decompensated cirrhosis, (b) EASL recommends adding ribavirin in treatment-experienced patients with genotype 3 or, if ribavirin is contraindicated, extending treatment to 24 weeks (Table 334-6, footnote c). As noted above, protease-inhibitor-containing DAA regimens (elbasvir/grazoprevir; paritaprevir/ritonavir, ombitasvir, and dasabuvir; simeprevir and sofosbuvir) are contraindicated in decompensated cirrhosis. For advanced renal failure, safety and efficacy have been documented for elbasvir/grazoprevir and paritaprevir/ritonavir, ombitasvir, and dasabuvir, but not for sofosbuvir-NS5A combinations.

AUTOIMMUNE HEPATITIS

DEFINITION

Autoimmune hepatitis is a chronic disorder characterized by continuing hepatocellular necrosis and inflammation, usually with fibrosis, which can progress to cirrhosis and liver failure. When fulfilling criteria of severity, this type of chronic hepatitis, when untreated, may have a 6-month mortality of as high as 40%. Based on contemporary estimates of the natural history of autoimmune hepatitis, the 10-year survival is 80–98% for treated and 67% for untreated patients. The prominence of extrahepatic features of autoimmunity and seroimmunologic abnormalities in this disorder supports an autoimmune process in its pathogenesis; this concept is reflected in the prior labels *lupoid* and *plasma cell hepatitis*. Autoantibodies and other typical features of autoimmunity, however, do not occur in all cases; among the broader categories of "idiopathic" or cryptogenic chronic hepatitis, many, perhaps the majority, are probably autoimmune in origin. Cases in which hepatotropic viruses, metabolic/genetic derangements (including nonalcoholic fatty liver disease), and hepatotoxic drugs have been excluded represent a spectrum of heterogeneous liver disorders of unknown cause, a proportion of which are most likely autoimmune hepatitis.

IMMUNOPATHOGENESIS

The weight of evidence suggests that the progressive liver injury in patients with autoimmune hepatitis is the result of a cell-mediated immunologic attack directed against liver cells in the setting of a loss of, or failed, immunologic tolerance for self liver antigens. In all likelihood, predisposition to autoimmunity is inherited, whereas the liver specificity of this injury is triggered by environmental (e.g., chemical, drug [e.g., minocycline], or viral) factors. For example, patients have been described in whom apparently self-limited cases of acute hepatitis A, B, or C led to autoimmune hepatitis, presumably because of genetic susceptibility or predisposition. Evidence to support an autoimmune pathogenesis in this type of hepatitis includes the following: (1) in the liver, the histopathologic lesions are composed predominantly of cytotoxic T cells and plasma cells; (2) circulating autoantibodies (nuclear, smooth muscle, thyroid, etc.; see below), rheumatoid factor, and hyperglobulinemia are common; (3) other autoimmune disorders—such as autoimmune thyroiditis, rheumatoid arthritis, autoimmune hemolytic anemia, ulcerative colitis, membranoproliferative glomerulonephritis, juvenile diabetes mellitus, vitiligo, celiac disease, and Sjögren's syndrome—occur with increased frequency in patients and in their relatives who have autoimmune hepatitis; (4) histocompatibility haplotypes associated with autoimmune diseases, such as HLA-B1, B8, DR3, and DR4 as well as extended haplotype *DRB1*0301* and *DRB1*0401* alleles, are common in patients with autoimmune hepatitis; and (5) this type of chronic hepatitis is responsive to glucocorticoid/immunosuppressive therapy, effective in a variety of autoimmune disorders.

Cellular immune mechanisms appear to be important in the pathogenesis of autoimmune hepatitis. In vitro studies have suggested that in patients with this disorder, $CD4^+$ T lymphocytes are capable of becoming sensitized to hepatocyte membrane proteins and of destroying liver cells. Molecular mimicry by cross-reacting antigens that contain epitopes similar to liver antigens is postulated to activate these T cells, which infiltrate, and result in injury to, the liver. Abnormalities of immunoregulatory control over cytotoxic lymphocytes (impaired regulatory CD4+CD25+ T cell influences) may play a role as well. Studies of genetic predisposition to autoimmune hepatitis demonstrate that certain haplotypes are associated with the disorder, as enumerated above, as are polymorphisms in cytotoxic T lymphocyte antigens (*CTLA-4*) and tumor necrosis factor α (*TNFA*2*). The precise triggering factors, genetic influences, and cytotoxic and immunoregulatory mechanisms involved in this type of liver injury remain incompletely defined.

Intriguing clues into the pathogenesis of autoimmune hepatitis come from the observation that circulating autoantibodies are prevalent in patients with this disorder. Among the autoantibodies described in these patients are antibodies to nuclei (so-called antinuclear antibodies [ANAs], primarily in a homogeneous pattern) and smooth muscle (so-called anti-smooth-muscle antibodies, directed at actin, vimentin, and skeletin), antibodies to F-actin, anti-LKM (see below), antibodies to "soluble liver antigen" (directed against a uracil-guanine-adenine transfer RNA suppressor protein), antibodies to α-actinin, and antibodies to the liver-specific asialoglycoprotein receptor (or "hepatic lectin") and other hepatocyte membrane proteins. Although some of these provide helpful diagnostic markers, their involvement in the pathogenesis of autoimmune hepatitis has not been established.

Humoral immune mechanisms have been shown to play a role in the extrahepatic manifestations of autoimmune and idiopathic hepatitis. Arthralgias, arthritis, cutaneous vasculitis, and glomerulonephritis occurring in patients with autoimmune hepatitis appear to be mediated by the deposition of circulating immune complexes in affected tissue vessels, followed by complement activation, inflammation, and tissue injury. While specific viral antigen-antibody complexes can be identified in acute and chronic viral hepatitis, the nature of the immune complexes in autoimmune hepatitis has not been defined.

CLINICAL FEATURES

Many of the *clinical features* of autoimmune hepatitis are similar to those described for chronic viral hepatitis. The onset of disease may be insidious or abrupt; the disease may present initially like, and be confused with, acute viral hepatitis; a history of recurrent bouts of what had been labeled *acute hepatitis* is not uncommon. In approximately a quarter of patients, the diagnosis is made in the absence of symptoms, based on abnormal liver laboratory tests. A subset of patients with autoimmune hepatitis has distinct features. Such patients are predominantly young to middle-aged women with marked hyperglobulinemia and high-titer circulating ANAs. This is the group with positive lupus erythematosus (LE) preparations (initially labeled "*lupoid*" hepatitis) in whom other autoimmune features are common. Fatigue, malaise,

anorexia, amenorrhea, acne, arthralgias, and jaundice are common. Occasionally, arthritis, maculopapular eruptions (including cutaneous vasculitis), erythema nodosum, colitis, pleurisy, pericarditis, anemia, azotemia, and sicca syndrome (keratoconjunctivitis, xerostomia) occur. In some patients, complications of cirrhosis, such as ascites and edema (associated with portal hypertension and hypoalbuminemia), encephalopathy, hypersplenism, coagulopathy, or variceal bleeding may bring the patient to initial medical attention.

The course of autoimmune hepatitis may be variable. In patients with mild disease or limited histologic lesions (e.g., piecemeal necrosis without bridging), progression to cirrhosis is limited, but, even in this subset, clinical monitoring is important to identify progression; up to half left untreated can progress to cirrhosis over the course of 15 years. In North America, cirrhosis at presentation is more common in African Americans than in whites. In those with severe symptomatic autoimmune hepatitis (aminotransferase levels >10 times normal, marked hyperglobulinemia, "aggressive" histologic lesions—bridging necrosis or multilobular collapse, cirrhosis), the 6-month mortality without therapy may be as high as 40%. Such severe disease accounts for only 20% of cases; the natural history of milder disease is variable, often accentuated by spontaneous remissions and exacerbations. Especially poor prognostic signs include the presence histologically of multilobular collapse at the time of initial presentation and failure of serum bilirubin to improve after 2 weeks of therapy. Death may result from hepatic failure, hepatic coma, other complications of cirrhosis (e.g., variceal hemorrhage), and intercurrent infection. In patients with established cirrhosis, HCC may be a late complication **(Chap. 78)** but occurs less frequently than in cirrhosis associated with viral hepatitis.

Laboratory features of autoimmune hepatitis are similar to those seen in chronic viral hepatitis. Liver biochemical tests are invariably abnormal but may not correlate with the clinical severity or histopathologic features in individual cases. Many patients with autoimmune hepatitis have normal serum bilirubin, alkaline phosphatase, and globulin levels with only minimal aminotransferase elevations. Serum AST and ALT levels are increased and fluctuate in the range of 100–1000 units. In severe cases, the serum bilirubin level is moderately elevated (51–171 μmol/L [3–10 mg/dL]). Hypoalbuminemia occurs in patients with very active or advanced disease. Serum alkaline phosphatase levels may be moderately elevated or near normal. In a small proportion of patients, marked elevations of alkaline phosphatase activity occur; in such patients, clinical and laboratory features overlap with those of primary biliary cirrhosis **(Chap. 337)**. The prothrombin time is often prolonged, particularly late in the disease or during active phases.

Polyclonal hypergammaglobulinemia (>2.5 g/dL) is common in autoimmune hepatitis, as is the presence of rheumatoid factor. As noted above, circulating autoantibodies are also prevalent, most characteristically ANAs in a homogeneous staining pattern. Smooth-muscle antibodies are less specific, seen just as frequently in chronic viral hepatitis. Because of the high levels of globulins achieved in the circulation of some patients with autoimmune hepatitis, occasionally the globulins may bind nonspecifically in solid-phase binding immunoassays for viral antibodies. This has been recognized most commonly in tests for antibodies to hepatitis C virus, as noted above. In fact, studies of autoantibodies in autoimmune hepatitis have led to the recognition of new categories of autoimmune hepatitis. *Type I autoimmune hepatitis* is the classic syndrome prevalent in North America and northern Europe occurring in young women, associated with marked hyperglobulinemia, lupoid features, circulating ANAs, and HLA-DR3 or HLA-DR4 (especially *B8-DRB1*03*). Also associated with type I autoimmune hepatitis are autoantibodies against actin and atypical perinuclear antineutrophilic cytoplasmic antibodies (pANCA). Included in the spectrum of type-I autoimmune hepatitis is a subset of patients who lack ANA and anti-LKM1, but who have circulating antibodies to soluble liver antigen. Most of these patients are women and have clinical features similar to, perhaps more severe than, those of other patients with type I autoimmune hepatitis.

Type II autoimmune hepatitis, often seen in children, more common in Mediterranean populations, and linked to HLA- DRB1 and HLA-DQB1 haplotypes, is associated not with ANA but with anti-LKM. Actually, anti-LKM represent a heterogeneous group of antibodies. In type II autoimmune hepatitis, the antibody is anti-LKM1, directed against cytochrome P450 2D6. This is the same anti-LKM seen in some patients with chronic hepatitis C. Anti-LKM2 is seen in drug-induced hepatitis, and anti-LKM3 (directed against uridine diphosphate glucuronyltransferases) is seen in patients with chronic hepatitis D. Another autoantibody observed in type II autoimmune hepatitis is directed against liver cytosol formiminotransferase cyclodeaminase (anti-liver cytosol 1).

Liver biopsy abnormalities are similar to those described for chronic viral hepatitis. Expanding portal tracts and extending beyond the plate of periportal hepatocytes into the parenchyma (designated *interface hepatitis* or *piecemeal necrosis*) is a mononuclear cell infiltrate that, in autoimmune hepatitis, may include the presence of plasma cells. Necroinflammatory activity characterizes the lobular parenchyma, and evidence of hepatocellular regeneration is reflected by "rosette" formation, the occurrence of thickened liver cell plates, and regenerative "pseudolobules." Septal fibrosis, bridging fibrosis, and cirrhosis are frequent. In patients with early autoimmune hepatitis presenting as an acute-hepatitis-like illness, lobular and centrilobular (as opposed to the more common periportal) necrosis has been reported. Bile duct injury and granulomas are uncommon; however, a subgroup of patients with autoimmune hepatitis has histologic, biochemical, and serologic features overlapping those of primary biliary cirrhosis **(Chap. 337)**.

DIAGNOSTIC CRITERIA

An international group has suggested a set of criteria for establishing a diagnosis of autoimmune hepatitis. Exclusion of liver disease caused by genetic disorders, viral hepatitis, drug hepatotoxicity, and alcohol are linked with such inclusive diagnostic criteria as hyperglobulinemia, autoantibodies, and characteristic histologic features. This international group has also suggested a comprehensive diagnostic scoring system that, rarely required for typical cases, may be helpful when typical features are not present. Factors that weigh in favor of the diagnosis include female gender; predominant aminotransferase elevation; presence and level of globulin elevation; presence of nuclear, smooth muscle, LKM1, and other autoantibodies; concurrent other autoimmune diseases; characteristic histologic features (interface hepatitis, plasma cells, rosettes); HLA-DR3 or DR4 markers; and response to treatment (see below). A more simplified, more specific scoring system relies on four variables: autoantibodies, serum IgG level, typical or compatible histologic features, and absence of viral hepatitis markers. Weighing against the diagnosis are predominant alkaline phosphatase elevation, mitochondrial antibodies, markers of viral hepatitis, history of hepatotoxic drugs or excessive alcohol, histologic evidence of bile duct injury, or such atypical histologic features as fatty infiltration, iron overload, and viral inclusions.

DIFFERENTIAL DIAGNOSIS

Early during the course of chronic hepatitis, autoimmune hepatitis may resemble typical *acute viral hepatitis* **(Chap. 332)**. Without histologic assessment, severe chronic hepatitis cannot be readily distinguished based on clinical or biochemical criteria from mild chronic hepatitis. In adolescence, *Wilson's disease* **(Chaps. 337 and 408)** may present with features of chronic hepatitis long before neurologic manifestations become apparent and before the formation of Kayser-Fleischer rings (copper deposition in Descemet's membrane in the periphery of the cornea). In this age group, serum ceruloplasmin and serum and urinary copper determinations plus measurement of liver copper levels establish the correct diagnosis. *Postnecrotic* or *cryptogenic cirrhosis* and *primary biliary cirrhosis* **(Chap. 337)** share clinical features with autoimmune hepatitis, and both alcoholic hepatitis **(Chap. 335)** and nonalcoholic steatohepatitis **(Chap. 336)** may present with many features common to autoimmune hepatitis; historic, biochemical, serologic, and histologic assessments are usually sufficient to allow these entities to be distinguished from autoimmune hepatitis. Of course, the distinction between autoimmune and chronic viral hepatitis is not always straightforward, especially when viral antibodies occur in patients with autoimmune disease or when autoantibodies occur in patients with viral disease. Furthermore, the presence of extrahepatic features such as arthritis, cutaneous vasculitis, or pleuritis—not to

mention the presence of circulating autoantibodies—may cause confusion with *rheumatologic disorders* such as rheumatoid arthritis and systemic LE. The existence of clinical and biochemical features of progressive necroinflammatory liver disease distinguishes chronic hepatitis from these other disorders, which are not associated with severe liver disease. Rarely, hepatic venous outflow obstruction (Budd-Chiari syndrome) may present with features suggestive of autoimmune hepatitis, but painful hepatomegaly, ascites, and vascular imaging provide distinguishing diagnostic clues. Other diagnostic considerations would include celiac disease and ischemic liver disease, which would be readily distinguishable by clinical and laboratory features from autoimmune hepatitis.

Finally, occasionally, features of autoimmune hepatitis overlap with features of autoimmune biliary disorders such as primary biliary cirrhosis, primary sclerosing cholangitis **(Chaps. 337 and 339)**, or, even more rarely, mitochondrial antibody-negative autoimmune cholangitis. Such overlap syndromes are difficult to categorize, and often response to therapy may be the distinguishing factor that establishes the diagnosis.

TREATMENT

Autoimmune Hepatitis

The mainstay of management in autoimmune hepatitis is glucocorticoid therapy. Several controlled clinical trials have documented that such therapy leads to symptomatic, clinical, biochemical, and histologic improvement as well as increased survival. A therapeutic response can be expected in up to 80% of patients. Unfortunately, therapy has not been shown in clinical trials to prevent ultimate progression to cirrhosis; however, instances of reversal of fibrosis and cirrhosis have been reported in patients responding to treatment, and rapid treatment responses within 1 year do translate into a reduction in progression to cirrhosis. Although some advocate the use of prednisolone (the hepatic metabolite of prednisone), prednisone is just as effective and is favored by most authorities. Therapy may be initiated at 20 mg/d, but a popular regimen in the United States relies on an initiation dose of 60 mg/d. This high dose is tapered successively over the course of a month down to a maintenance level of 20 mg/d. An alternative, but equally effective, more appealing approach is to begin with half the prednisone dose (30 mg/d) along with azathioprine (50 mg/d). With azathioprine maintained at 50 mg/d, the prednisone dose is tapered over the course of a month down to a maintenance level of 10 mg/d. The advantage of the combination approach is a reduction, over the span of an 18-month course of therapy, in serious, life-threatening complications of steroid therapy (e.g., cushingoid features, hypertension, diabetes, osteoporosis) from 66% down to under 20%. Genetic analysis for thiopurine S-methyltransferase allelic variants does not correlate with azathioprine-associated cytopenias or efficacy and is not assessed routinely in patients with autoimmune hepatitis. In combination regimens, 6-mercaptopurine may be substituted for its prodrug azathioprine, but this is rarely required. Azathioprine alone, however, is not effective in achieving remission, nor is alternate-day glucocorticoid therapy. Limited experience with budesonide in noncirrhotic patients suggests that this steroid side effect–sparing drug may be effective; however, the few randomized controlled trials of budesonide have not consistently shown efficacy. Although therapy has been shown to be effective for severe autoimmune hepatitis (AST ≥10 × the upper limit of normal or ≥5 × the upper limit of normal in conjunction with serum globulin greater than or equal to twice normal; bridging necrosis or multilobular necrosis on liver biopsy; presence of symptoms), therapy is not indicated for mild forms of chronic hepatitis, and the efficacy of therapy in mild or asymptomatic autoimmune hepatitis has not been established.

Improvement of fatigue, anorexia, malaise, and jaundice tends to occur within days to several weeks; biochemical improvement occurs over the course of several weeks to months, with a fall in serum bilirubin and globulin levels and an increase in serum albumin. Serum aminotransferase levels usually drop promptly, but improvements in AST and ALT alone do not appear to be reliable markers of recovery in individual patients; histologic improvement, characterized by a decrease in mononuclear infiltration and in hepatocellular necrosis, may be delayed for 6–24 months. Still, if interpreted cautiously, aminotransferase levels are valuable indicators of relative disease activity, and, although recommended, many authorities do *not* advocate for serial liver biopsies to assess therapeutic success or to guide decisions to alter or stop therapy. Rapidity of response is more common in older patients (≥69 years) and those with HLA *DBR1*04*; although rapid responders may progress less slowly to cirrhosis and liver transplantation, they are no less likely than slower responders to relapse after therapy. Therapy should continue for at least 12–18 months. After tapering and cessation of therapy, the likelihood of relapse is at least 50%, even if posttreatment histology has improved to show mild chronic hepatitis, and the majority of patients require therapy at maintenance doses indefinitely. Continuing azathioprine alone (2 mg/kg body weight daily) after cessation of prednisone therapy has been shown to reduce the frequency of relapse. Long-term maintenance with low-dose prednisone (≤10 mg daily) has also been shown to keep autoimmune hepatitis in check without the theoretical risk of azathioprine marrow suppression and, in young women of child-bearing age, teratogenicity; however, maintenance azathioprine is more effective in preserving remission.

In medically refractory cases, an attempt should be made to intensify treatment with high-dose glucocorticoid monotherapy (60 mg daily) or combination glucocorticoid (30 mg daily) plus high-dose azathioprine (150 mg daily) therapy. After a month, doses of prednisone can be reduced by 10 mg a month, and doses of azathioprine can be reduced by 50 mg a month toward ultimate, conventional maintenance doses. Patients refractory to this regimen may be treated with cyclosporine, tacrolimus, or mycophenolate mofetil. Similarly, in exploratory studies, infusions of monoclonal antibodies directed at tumor necrosis factor (infliximab) and against the B-lymphocyte antigen CD20 (rituximab) have been reported to be of clinical benefit (improved aminotransferase levels, immunoglobulin G levels, histologic inflammatory activity) as rescue therapy for refractory autoimmune hepatitis. To date, however, only limited, often anecdotal, data in small numbers of patients support these alternative approaches. If medical therapy fails, or when chronic hepatitis progresses to cirrhosis and is associated with life-threatening complications of liver decompensation, liver transplantation is the only recourse **(Chap. 338)**; in patients with severe autoimmune hepatitis, failure of the bilirubin to improve after 2 weeks of therapy should prompt early consideration of the patient for liver transplantation. Recurrence of autoimmune hepatitis in the new liver occurs rarely in most experiences but in as many as 35–40% of cases in others; nonetheless, 5-year patient and graft survival exceed 80%.

Like all patients with chronic liver disease, patients with autoimmune hepatitis should be vaccinated against hepatitis A and B, ideally before immunosuppressive therapy is begun, if practical. Patients with autoimmune hepatitis and cirrhosis should be screened for HCC with ultrasound at 6-month intervals and for gastroesophageal varices with upper gastrointestinal endoscopy at intervals of 1–3 years, based on severity of liver disease.

FURTHER READING

AASLD/IDSA HCV GUIDANCE PANEL: Hepatitis C guidance: AASLD-IDSA recommendations for testing, managing, and treating adults infected with hepatitis C virus. Hepatology 62:932, 2015. Updated regularly and available at *http://www.hcvguidelines.org*. Accessed September 22, 2016.

CARBONE M, NEUBERGER JM: Autoimmune liver disease, autoimmunity and liver transplantation. J Hepatol 60:210, 2014.

EUROPEAN ASSOCIATION FOR THE STUDY OF THE LIVER: EASL clinical practice guidelines: Management of chronic hepatitis B virus infection. J Hepatol 57:167, 2012.

EUROPEAN ASSOCIATION FOR THE STUDY OF THE LIVER: EASL clinical practice guidelines: autoimmune hepatitis. J Hepatol 63:971, 2015.

European Association for the Study of the Liver: EASL Recommendations on treatment of hepatitis C 2016. J Hepatol 2016 in press. Available at *http://dx.doi.org/10.1016/j. jhep.2016.09.001*. Accessed September 22, 2016.

Lok ASG et al: Antiviral therapy for chronic hepatitis B viral infection in adults: A systematic review and meta-analysis. Hepatology 63:284, 2016.

Manns MP et al: AASLD Practice Guidelines: Diagnosis and management of autoimmune hepatitis. Hepatology 51;2193, 2010.

Pawlotsky J-M et al: From non-A, non-B hepatitis to hepatitis C virus cure. J Hepatol 62:S87, 2015.

Terrault N et al: AASLD guidelines for treatment of chronic hepatitis B. Hepatology 63:261, 2016.

Wedemeyer H et al: Peginterferon plus adefovir versus either drug alone for hepatitis delta. N Engl J Med 364:322, 2011.

TABLE 335-1 Risk Factors for Alcoholic Liver Disease

RISK FACTOR	COMMENT
Quantity	In men, 40–80 g/d of ethanol produces fatty liver; 160 g/d for 10–20 years causes hepatitis or cirrhosis. Only 15% of alcoholics develop alcoholic liver disease.
Gender	Women exhibit increased susceptibility to alcoholic liver disease at amounts >20 g/d; two drinks per day is probably safe.
Hepatitis C	HCV infection concurrent with alcoholic liver disease is associated with younger age for severity, more advanced histology, and decreased survival.
Genetics	Patatin-like phospholipase domain-containing protein 3 (PNPLA3) has been associated with alcoholic cirrhosis.
Fatty liver	Alcohol injury does not require malnutrition, but obesity and nonalcoholic fatty liver are risk factors. Patients should receive vigorous attention to nutritional support.

335 Alcoholic Liver Disease

Mark E. Mailliard, Michael F. Sorrell

Chronic and excessive alcohol ingestion is a major cause of liver disease and is responsible for nearly 50% of the mortality from all cirrhosis. The pathology of alcoholic liver disease consists of three major lesions, with the progressive injury rarely existing in a pure form: (1) fatty liver, (2) alcoholic hepatitis, and (3) cirrhosis. Fatty liver is present in >90% of daily as well as binge drinkers. A much smaller percentage of heavy drinkers will progress to alcoholic hepatitis, thought to be a precursor to cirrhosis. The prognosis of severe alcoholic liver disease is dismal; the mortality of patients with alcoholic hepatitis concurrent with cirrhosis is nearly 60% at 4 years. Although alcohol is considered a direct hepatotoxin, only between 10 and 20% of alcoholics will develop alcoholic hepatitis. The explanation for this apparent paradox is unclear but involves the complex interaction of facilitating factors such as drinking patterns, diet, obesity, and gender. There are no diagnostic tools that can predict individual susceptibility to alcoholic liver disease.

GLOBAL CONSIDERATIONS

Alcohol is the world's third largest risk factor for disease burden. The harmful use of alcohol results in about 3.5 million deaths worldwide each year. Most of the mortality attributed to alcohol is secondary to cirrhosis. Mortality from cirrhosis is directly related to alcohol consumption, with the Eastern European countries the most significantly burdened. Cirrhosis and its complications are closely correlated with volume of alcohol consumed per capita population and are regardless of gender.

ETIOLOGY AND PATHOGENESIS

Quantity and duration of alcohol intake are the most important risk factors involved in the development of alcoholic liver disease **(Table 335-1)**. The roles of beverage type(s), i.e., wine, beer, or spirits, and pattern of drinking (daily versus binge drinking) are less clear. Progress beyond the fatty liver stage seems to require additional risk factors that remain incompletely defined. Although there are genetic predispositions for alcoholism **(Chap. 445)**, gender is a strong determinant for alcoholic liver disease. Women are more susceptible to alcoholic liver injury when compared to men. They develop advanced liver disease with substantially less alcohol intake. In general, the time it takes to develop liver disease is directly related to the amount of alcohol consumed. It is useful in estimating alcohol consumption to understand that one beer, four ounces of wine, or one ounce of 80% spirits all contain ~12 g of alcohol. The threshold for developing alcoholic liver disease is higher in men (>14 drinks per week), while women are at increased risk for liver injury by consuming >7 drinks per week. Gender-dependent differences result from poorly understood effects of estrogen, proportion of body fat, and the gastric metabolism of alcohol. Obesity, a high-fat diet, and the protective effect of coffee have been postulated to play a part in the development of the pathogenic process.

Chronic infection with hepatitis C virus (HCV) **(Chap. 334)** is an important comorbidity in the progression of alcoholic liver disease to cirrhosis in chronic drinkers. Even light to moderate alcohol intake of 15–30 g/d increases the risk of cirrhosis and hepatocellular cancer in HCV-infected individuals. Patients with both alcoholic liver injury and HCV infection develop decompensated liver disease at a younger age and have poorer overall survival. Increased liver iron stores and, rarely, porphyria cutanea tarda can occur as a consequence of the overlapping injurious processes secondary to alcohol and HCV infection.

The pathogenesis of alcoholic liver injury is unclear. The present conceptual foundation is that alcohol acts as a direct hepatotoxin and that malnutrition does not have a major role. Ingestion of alcohol initiates an inflammatory cascade by its metabolism, resulting in a variety of metabolic responses. Steatosis from lipogenesis, fatty acid synthesis, and depression of fatty acid oxidation appears secondary to effects on sterol regulatory transcription factor and peroxisome proliferator-activated receptor α (PPAR-α). Intestinal-derived endotoxin initiates a pathogenic process through toll-like receptor 4 and tumor necrosis factor α (TNF-α) that facilitates hepatocyte apoptosis and necrosis. The cell injury and endotoxin release initiated by ethanol and its metabolites also activate innate and adaptive immunity pathways releasing proinflammatory cytokines (e.g., TNF-α), chemokines, and proliferation of T and B cells. The effect of chronic ethanol ingestion on intestinal permeability influences liposaccharide hepatic influx as well as microbiome dysbiosis, further contributing to the pathogenic process. The production of toxic protein-aldehyde adducts, generation of reducing equivalents, and oxidative stress also play a role. Hepatocyte injury and impaired regeneration following chronic alcohol ingestion are ultimately associated with stellate cell activation and collagen production; key events in fibrogenesis. The resulting fibrosis from continuing alcohol use determines the architectural derangement of the liver and associated pathophysiology.

PATHOLOGY

The liver has a limited repertoire in response to injury. Fatty liver is the initial and most common histologic response to hepatotoxic stimuli, including excessive alcohol ingestion. The accumulation of fat within the perivenular hepatocytes coincides with the location of alcohol dehydrogenase, the major enzyme responsible for alcohol metabolism. Continuing alcohol ingestion results in fat accumulation throughout the entire hepatic lobule. Despite extensive fatty change and distortion of the hepatocytes with macrovesicular fat, the cessation of drinking results in normalization of hepatic architecture and fat content. Alcoholic fatty liver has traditionally been regarded as entirely benign, but similar to the spectrum of nonalcoholic fatty liver disease, the appearance of steatohepatitis and certain

pathologic features such as giant mitochondria, perivenular fibrosis, and macrovesicular fat may be associated with progressive liver injury.

The transition between fatty liver and the development of alcoholic hepatitis is blurred. The hallmark of alcoholic hepatitis is hepatocyte injury characterized by ballooning degeneration, spotty necrosis, polymorphonuclear infiltrate, and fibrosis in the perivenular and perisinusoidal space of Disse. Mallory-Denk bodies are often present in florid cases but are neither specific nor necessary to establish the diagnosis. Alcoholic hepatitis is thought to be a precursor to the development of cirrhosis. However, like fatty liver, it is potentially reversible with cessation of drinking. Cirrhosis is present in up to 50% of patients with biopsy-proven alcoholic hepatitis, and its regression is uncertain, even with abstention.

CLINICAL FEATURES

The clinical manifestations of alcoholic fatty liver are subtle and characteristically detected as a consequence of the patient's visit for a seemingly unrelated matter. Previously unsuspected hepatomegaly is often the only clinical finding. Occasionally, patients with fatty liver will present with right upper quadrant discomfort, nausea, and, rarely, jaundice. Differentiation of alcoholic fatty liver from nonalcoholic fatty liver is difficult unless an accurate drinking history is ascertained. In every instance where liver disease is present, a thoughtful and sensitive drinking history should be obtained. Standard, validated questions accurately detect alcohol-related problems **(Chap. 445)**. Alcoholic hepatitis is associated with a wide gamut of clinical features. Fever, spider nevi, jaundice, and abdominal pain simulating an acute abdomen represent the extreme end of the spectrum, while many patients will be entirely asymptomatic. Portal hypertension, ascites, or variceal bleeding can occur in the absence of cirrhosis. Recognition of the clinical features of alcoholic hepatitis is central to the initiation of an effective and appropriate diagnostic and therapeutic strategy. It is important to recognize that patients with alcoholic cirrhosis often exhibit clinical features identical to other causes of cirrhosis.

LABORATORY FEATURES

Patients with alcoholic liver disease are often identified through routine screening tests. The typical laboratory abnormalities seen in fatty liver are nonspecific and include modest elevations of aspartate aminotransferase (AST), alanine aminotransferase (ALT), and γ-glutamyl transpeptidase (GGTP), often accompanied by hypertriglyceridemia and hyperbilirubinemia. In alcoholic hepatitis and in contrast to other causes of fatty liver, AST and ALT are usually elevated two- to sevenfold. They are rarely >400 IU, and the AST/ALT ratio is >1 **(Table 335-2)**. Hyperbilirubinemia is accompanied by modest increases in the alkaline phosphatase level. Derangement in hepatocyte synthetic function indicates more serious disease. Hypoalbuminemia and coagulopathy are common in advanced liver injury. Ultrasonography is useful in detecting fatty infiltration of the liver and determining liver size. The demonstration by ultrasound of portal vein flow reversal, ascites, and intraabdominal venous collaterals indicates serious liver injury with less potential for complete reversal.

PROGNOSIS

Critically ill patients with alcoholic hepatitis have short-term (30-day) mortality rates >50%. Severe alcoholic hepatitis is heralded by coagulopathy (prothrombin time increased >5 s), anemia, serum albumin concentrations <25 g/L (2.5 mg/dL), serum bilirubin levels >137 μmol/L (8 mg/dL), renal failure, and ascites. A discriminant function calculated as 4.6 X (the prolongation of the prothrombin time above control [seconds]) + serum bilirubin (mg/dL) can identify patients with a poor prognosis (discriminant function >32). A Model for End-Stage Liver Disease (MELD) score **(Chap. 338)** ≥21 also is associated with significant mortality in alcoholic hepatitis. The presence of ascites, variceal hemorrhage, deep encephalopathy, or hepatorenal syndrome predicts a dismal prognosis. The pathologic stage of the injury can be helpful in predicting prognosis. Liver biopsy should be performed whenever possible to establish the diagnosis and to guide the therapeutic decisions.

TABLE 335-2 Laboratory Diagnosis of Alcoholic Fatty Liver and Alcoholic Hepatitis

TEST	COMMENT
AST	Increased two- to sevenfold, <400 IU/L, greater than ALT
ALT	Increased two- to sevenfold, <400 IU/L
AST/ALT	Usually >1
GGTP	Not specific to alcohol, easily inducible, elevated in all forms of fatty liver
Bilirubin	May be markedly increased in alcoholic hepatitis despite modest elevation in alkaline phosphatase

Abbreviations: ALT, alanine aminotransferase; AST, aspartate aminotransferase; GGTP, γ-glutamyl transpeptidase.

TREATMENT

Alcoholic Liver Disease

Complete abstinence from alcohol is the cornerstone in the treatment of alcoholic liver disease. Improved survival and the potential for reversal of histologic injury regardless of the initial clinical presentation are associated with total avoidance of alcohol ingestion. Referral of patients to experienced alcohol counselors and/or alcohol treatment programs should be routine in the management of patients with alcoholic liver disease. Attention should be directed to the nutritional and psychosocial states during the evaluation and treatment periods. Because of data suggesting that the pathogenic mechanisms in alcoholic hepatitis involve cytokine release and the perpetuation of injury by immunologic processes, glucocorticoids have been extensively evaluated in the treatment of alcoholic hepatitis. Patients with severe alcoholic hepatitis, defined as a discriminant function >32 or MELD >20, should be given prednisone, 40 mg/d, or prednisolone, 32 mg/d, for 4 weeks, followed by a steroid taper **(Fig. 335-1)**. Exclusion criteria include active gastrointestinal bleeding, renal failure, or pancreatitis. Patients with infection can be concurrently treated with antibiotics and steroids. Women with encephalopathy from severe alcoholic hepatitis may be particularly good candidates for glucocorticoids. A Lille score >0.45, at *http://www.lillemodel.com*, uses pretreatment variables plus the change in total bilirubin at day 7 of glucocorticoids to identify those patients unresponsive to therapy.

The role of TNF-α expression and receptor activity in alcoholic liver injury has led to an examination of pentoxifylline, the nonspecific TNF inhibitor, either by itself, or with glucocorticoids for severe alcoholic hepatitis. In one study, pentoxifylline demonstrated an improved survival in the therapy of severe alcoholic hepatitis, primarily due to a decrease in hepatorenal syndrome. Subsequent clinical trials failed to find an increased benefit from pentoxifylline, either by itself or in combination with prednisolone **(Fig. 335-2)**.

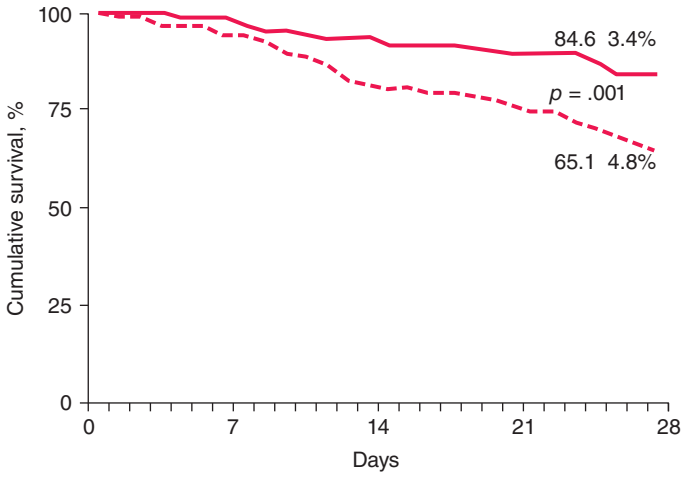


FIGURE 335-1 Effect of glucocorticoid therapy of severe alcoholic hepatitis on short-term survival: the result of a meta-analysis of individual data from three studies. Prednisolone, solid line; placebo, dotted line. *(Adapted from P Mathurin et al: J Hepatol 36:480, 2002, with permission from Elsevier Science.)*

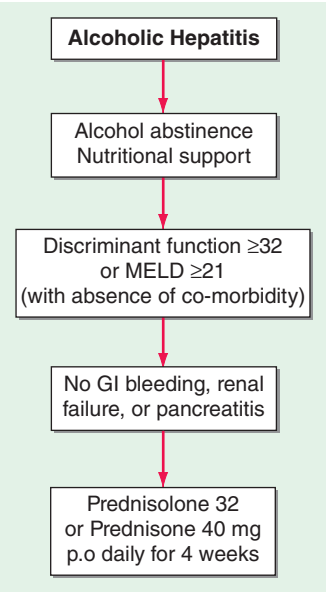


FIGURE 335-2 Treatment algorithm for alcoholic hepatitis. As identified by a calculated discriminant function >32 or MELD >20 (see text), patients with severe alcoholic hepatitis, without the presence of gastrointestinal bleeding, renal failure or pancreatitis would be candidates for glucocorticoids.

Liver transplantation is an accepted indication for treatment in select patients with complications of cirrhosis secondary to alcohol abuse. Outcomes are equal or superior to other indications for transplantation. In general, transplant candidacy should be reevaluated after a defined period of sobriety. Patients presenting with alcoholic hepatitis have been largely excluded from transplant candidacy because of the perceived risk of increased surgical mortality and high rates of recidivism following transplantation. A European multidisciplinary group has reported excellent long-term transplant outcomes in highly selected patients with florid alcoholic hepatitis. General application of transplantation in such patients must await confirmatory outcomes.

FURTHER READING

Mathurin P et al: Corticosteroids improves short-term survival in patients with severe alcoholic hepatitis (AH): Individual data analysis of the last three randomized placebo controlled double blind trials of corticosteroids in severe AH. J Hepatol 36:480, 2002.

Sanyal AJ et al: Alcoholic and nonalcoholic fatty liver disease. Gastroenterology 150:8 (suppl), 2016.

Thurz MR et al: Prednisolone or pentoxifylline for alcoholic hepatitis. N Engl J Med 372:1619, 2015.

336 Nonalcoholic Fatty Liver Diseases and Nonalcoholic Steatohepatitis

Manal F. Abdelmalek, Anna Mae Diehl

INCIDENCE, PREVALENCE, AND NATURAL HISTORY

Nonalcoholic fatty liver disease (NAFLD) is the most common chronic liver disease in many parts of the world, including the United States. Population-based abdominal imaging studies have demonstrated fatty liver in at least 25% of American adults. Because the vast majority of these subjects deny hazardous levels of alcohol consumption (defined as greater than one drink per day in women or two drinks per day in men), they are considered to have NAFLD. NAFLD is strongly associated with overweight/obesity and insulin resistance. However, it can also occur in lean individuals and is particularly common in those with a paucity of adipose depots (i.e., lipodystrophy). Ethnic/racial factors also appear to influence liver fat accumulation; the documented prevalence of NAFLD is lowest in African Americans (~25%), highest in Americans of Hispanic ancestry (~50%), and intermediate in American whites (~33%).

NAFLD encompasses a spectrum of liver pathology with different clinical prognoses. The simple accumulation of triglyceride within hepatocytes (hepatic steatosis) is on the most clinically benign extreme of the spectrum. On the opposite, most clinically ominous extreme, are cirrhosis **(Chap. 337)** and primary liver cancer **(Chap. 78)**. The risk of developing cirrhosis is extremely low in individuals with chronic hepatic steatosis, but increases as steatosis becomes complicated by histologically conspicuous hepatocyte death and inflammation (i.e., nonalcoholic steatohepatitis [NASH]). NASH itself is also a heterogeneous condition; sometimes it improves to steatosis or normal histology, sometimes it remains relatively stable for years, but sometimes it results in progressive accumulation of fibrous scar that eventuates in cirrhosis. Once NAFLD-related cirrhosis develops, the annual incidence of primary liver cancer can be as high as 3%.

Abdominal imaging is not able to determine which individuals with NAFLD have associated liver cell death and inflammation (i.e., NASH), and specific blood tests to diagnose NASH are not yet available. However, population-based studies that have used elevated serum ALT as a marker of liver injury indicate that about 6–8% of American adults have serum ALT elevations that cannot be explained by excessive alcohol consumption, other known causes of fatty liver disease **(Table 336-1)**, viral hepatitis, or drug-induced or congenital liver diseases. Because the prevalence of such "cryptogenic" ALT elevations increases with

TABLE 336-1 Alternative Causes of Hepatic Steatosis

- Alcoholic liver disease
- Hepatitis C (particularly genotype 3)
- Inborn errors of metabolism
  - Abetalipoproteinemia
  - Cholesterol ester storage disease
  - Galactosemia
  - Glycogen storage disease
  - Hereditary fructose intolerance
  - Homocystinuria
  - Systemic carnitine deficiency
  - Tyrosinemia
  - Weber-Christian syndrome
  - Wilson's disease
  - Wolman's disease
- Medications **(see Table 336-2)**
- Miscellaneous
  - Industrial exposure to petrochemical
  - Inflammatory bowel disease
  - Lipodystrophy
  - Bacterial overgrowth
  - Starvation
  - Parenteral nutrition
- Surgical procedures
  - Bilopancreatic diversion
  - Extensive small-bowel resection
  - Gastric bypass
  - Jejunoileal bypass
- Reye's syndrome
- Acute fatty liver of pregnancy
- HELLP syndrome (*h*emolytic anemia, elevated *l*iver enzymes, *l*ow *p*latelet count)

body mass index, it is presumed that they are due to NASH. Hence, at any given point in time, NASH is present in about 25% of individuals who have NAFLD (i.e., about 6% of the general U.S. adult population has NASH). Smaller cross-sectional studies in which liver biopsies have been performed on NASH patients at tertiary referral centers consistently demonstrate advanced fibrosis or cirrhosis in about 25% of those cohorts. By extrapolation, therefore, cirrhosis develops in about 6% of individuals with NAFLD (i.e., in about 1.5–2% of the general U.S. population). The risk for advanced liver fibrosis is highest in individuals with NASH who are aged >45–50 years and overweight/obese or afflicted with type 2 diabetes.

Heritable factors clearly impact susceptibility to hepatic steatosis, NASH, liver fibrosis, and liver cancer. Indeed, recent twin studies suggest that inheritance accounts for about half the risk for developing cirrhosis. Certain variants in PNPLA3 (a gene that encodes an enzyme involved in intracellular trafficking of lipids) consistently correlates with susceptibility to hepatic steatosis, cirrhosis, and liver cancer. Polymorphisms in other genes involved in lipid homeostasis (e.g., TM6SF2 and MBOAT7) are also emerging as potential genetic risk factors for NAFLD. Epigenetic factors (i.e., heritable traits that do not result from direct changes in DNA) may also influence NAFLD pathogenesis and/or progression based on evidence that intra-uterine exposures influence susceptibility to obesity and the metabolic syndrome in adolescence.

Experts have predicted that NAFLD will be the leading indication for liver transplantation in the United States by 2020. Similar to cirrhosis caused by other liver diseases, cirrhosis caused by NAFLD increases the risk for primary liver cancer. Both hepatocellular carcinoma and intrahepatic cholangiocarcinoma (ICC) have also been reported to occur in NAFLD patients without cirrhosis, suggesting that NAFLD per se may be a premalignant condition. NAFLD, NASH, and NAFLD-related cirrhosis are not limited to adults. All have been well documented in children. As in adults, obesity and insulin resistance are the main risk factors for pediatric NAFLD. Thus, the rising incidence and prevalence of childhood obesity suggests that NAFLD is likely to become an even greater contributor to society's burden of liver disease in the future.

PATHOGENESIS

The mechanisms underlying the pathogenesis and progression of NAFLD are not entirely clear. The best-understood mechanisms pertain to hepatic steatosis. This is proven to result when hepatocyte mechanisms for triglyceride synthesis (e.g., lipid uptake and de novo lipogenesis) overwhelm mechanisms for triglyceride disposal (e.g., degradative metabolism and lipoprotein export), leading to accumulation of fat (i.e., triglyceride) within hepatocytes. Obesity stimulates hepatocyte triglyceride accumulation by altering the intestinal microbiota to enhance both energy harvest from dietary sources and intestinal permeability. Reduced intestinal barrier function increases hepatic exposure to gut-derived products, which stimulate liver cells to generate inflammatory mediators that inhibit insulin actions. Obese adipose depots also produce excessive soluble factors (adipokines) that inhibit tissue insulin sensitivity. Insulin resistance promotes hyperglycemia. This drives the pancreas to produce more insulin to maintain glucose homeostasis. However, hyperinsulinemia also promotes lipid uptake, fat synthesis, and fat storage. The net result is hepatic triglyceride accumulation (i.e., steatosis).

Triglyceride per se is not hepatotoxic. However, its precursors (e.g., fatty acids and diacylglycerols) and metabolic by-products (e.g., reactive oxygen species) may damage hepatocytes, leading to hepatocyte lipotoxicity. Lipotoxicity also triggers the generation of other factors (e.g., inflammatory cytokines, hormonal mediators) that deregulate systems that normally maintain hepatocyte viability. The net result is increased hepatocyte death. Dying hepatocytes, in turn, release various factors that trigger wound healing responses that aim to replace (regenerate) lost hepatocytes. Such repair involves transient expansion of other cell types, such as myofibroblasts and progenitor cells, that make and degrade matrix, remodel the vasculature, and generate replacement hepatocytes, as well as the recruitment of immune cells that release factors that modulate liver injury and repair. NASH is the morphologic manifestation of lipotoxicity and resultant wound healing responses. Because the severity and duration of lipotoxic liver injury dictate the intensity and duration of repair, the histologic features and outcomes of NASH are variable. Cirrhosis and liver cancer are potential outcomes of chronic NASH. Cirrhosis results from futile repair, i.e., progressive accumulation of wound healing cells, fibrous matrix, and abnormal vasculature (scarring), rather than efficient reconstruction/regeneration of healthy hepatic parenchyma. Primary liver cancers develop when malignantly transformed liver cells escape mechanisms that normally control regenerative growth. The mechanisms responsible for futile repair (cirrhosis) and liver carcinogenesis are not well understood. Because normal liver regeneration is a very complex process, there are multiple opportunities for deregulation and, thus, pathogenic heterogeneity. To date, this heterogeneity has confounded development of both diagnostic tests and treatments for defective/deregulated liver repair (i.e., cirrhosis and cancer). Hence, current strategies focus on circumventing misrepair by preventing and/or reducing lipotoxic liver injury.

DIAGNOSIS

Diagnosing NAFLD requires demonstration of increased liver fat in the absence of hazardous levels of alcohol consumption. Thresholds for potentially dangerous alcohol ingestion have been set at more than one drink per day in women and two drinks per day in men based on epidemiologic evidence that the prevalence of serum aminotransferase elevations increases when alcohol consumption habitually exceeds these levels. In those studies, one drink was defined as having 10 g of ethanol and, thus, is equivalent to one can of beer, 4 ounces of wine, or 1.5 ounces (one shot) of distilled spirits. Other causes of liver fat accumulation (particularly exposure to certain drugs; **Table 336-2**) and liver injury (e.g., viral hepatitis, autoimmune liver disease, iron or copper overload, α_1 antitrypsin deficiency) must also be excluded. Thus,

TABLE 336-2 Medications Associated with Hepatic Steatosis

- Cytotoxic and cytostatic drugs
  - L-Asparaginase
  - Azacitidine
  - Azaserine
  - Bleomycin
  - Methotrexate
  - Puromycin
  - Tetracycline
  - Doxycycline
- Metals
  - Antimony
  - Barium salts
  - Chromates
  - Phosphorus
  - Rare earths of low atomic number
  - Thallium compounds
  - Uranium compounds
- Other drugs and toxins
  - Amiodarone
  - 4,4′-Diethylaminoethoxyhexesterol
  - Ethionine
  - Ethyl bromide
  - Estrogens
  - Glucocorticoids
  - Highly active antiretroviral therapy
  - Hydralazine
  - Hypoglycin
  - Orotate
  - Perhexiline maleate
  - Safrole
  - Tamoxifen

establishing the diagnosis of NAFLD does not require invasive testing: it can be accomplished by history and physical examination, liver imaging (ultrasound is an acceptable first-line test; computed tomography [CT] or magnetic resonance imaging [MRI] enhances sensitivity for liver fat detection but adds expense), and blood tests to exclude other liver diseases. It is important to emphasize that the liver may not be enlarged, and serum aminotransferases and liver function tests (e.g., bilirubin, albumin, prothrombin time) may be completely normal, in individuals with NAFLD. Because there is yet no one specific blood test for NAFLD, confidence in the diagnosis of NAFLD is increased by identification of NAFLD risk factors. The latter include increased body mass index, insulin resistance/type 2 diabetes mellitus, and other parameters indicative of the metabolic syndrome (e.g., systemic hypertension, dyslipidemia, hyperuricemia/gout, cardiovascular disease; **Chap. 401**) in the patient or family members.

Establishing the severity of NAFLD-related liver injury and related scarring (i.e., staging NAFLD) is more difficult than simply diagnosing NAFLD. Staging is critically important, however, because it is necessary to define prognosis and thereby determine treatment recommendations. The goal of staging is to distinguish patients with NASH from those with simple steatosis and to identify which of the NASH patients have advanced fibrosis. The 10-year probability of developing liver-related morbidity or mortality in steatosis is negligible, and hence, this subgroup of NAFLD patients tends to be managed conservatively (see below). In contrast, more intensive follow-up and therapy are justified in NASH patients, and the subgroup with advanced fibrosis merits the most intensive scrutiny and intervention because their 10-year risk of liver-related morbidity and mortality is clearly increased.

Staging approaches can be separated into noninvasive testing (i.e., blood testing, physical examination, and imaging) and invasive approaches (i.e., liver biopsy). Blood test evidence of hepatic dysfunction (e.g., hyperbilirubinemia, hypoalbuminemia, prothrombin time prolongation) or portal hypertension (e.g., thrombocytopenia) and stigmata of portal hypertension on physical examination (e.g., spider angiomata, palmar erythema, splenomegaly, ascites, clubbing, encephalopathy) suggest a diagnosis of advanced NAFLD. Currently, however, liver biopsy is the gold standard for establishing the severity of liver injury and fibrosis because it is both more sensitive and specific than these other tests for establishing NAFLD severity. Although invasive, liver biopsy is seldom complicated by serious adverse sequelae such as significant bleeding, pain, or inadvertent puncture of other organs and thus is relatively safe. However, biopsy suffers from potential sampling error unless tissue cores of 2 cm or longer are acquired. Also, examination of tissue at a single point in time is not reliable for determining whether the pathologic processes are progressing or regressing. The risk of serial liver biopsies within short time intervals is generally deemed as unacceptable outside of research studies. These limitations of liver biopsy have stimulated efforts to develop noninvasive approaches to stage NAFLD. As is true for many other types of chronic liver disease, in NAFLD the levels of serum aminotransferases (aspartate aminotransferase [AST] and alanine aminotransferase [ALT]) do not reliably reflect the severity of liver cell injury, extent of liver cell death, or related liver inflammation and fibrosis. Thus, they are imperfect for determining which individuals with NAFLD have NASH. This has prompted efforts to identify superior markers of NASH, and particularly liver fibrosis, because fibrosis stage predicts eventual liver outcomes and mortality in NASH. Algorithms that combine various laboratory tests (e.g., ELF score, BARD Score, NAFLD Fibrosis Score, APRI score) are somewhat helpful in separating NASH patients with advanced versus mild liver fibrosis. Combining these tests with new imaging approaches that permit noninvasive quantification of liver fat (e.g., MRI using proton density fat fraction [MRI-PDFF]) and liver stiffness, a surrogate marker of liver fibrosis (e.g., magnetic resonance elastography, MRE, and transient elastography, FibroScan) improves their predictive power **(Chap. 330)**. Increasingly, these new tools are being used serially to monitor fibrosis progression and regression in NAFLD patients. As a result, liver biopsy staging is becoming restricted to patients who cannot be stratified reliably using these noninvasive assessments.

CLINICAL FEATURES OF NAFLD

Most subjects with NAFLD are asymptomatic. The diagnosis is often made when abnormal liver aminotransferases or features of fatty liver are noted during an evaluation performed for other reasons. NAFLD may also be diagnosed during the workup of vague right upper quadrant abdominal pain, hepatomegaly, or an abnormal-appearing liver at time of abdominal surgery. Obesity is present in 50–90% of subjects. Most patients with NAFLD also have other features of the metabolic syndrome **(Chap. 401)**. Some have subtle stigmata of chronic liver disease, such as spider angiomata, palmer erythema, or splenomegaly. In a small minority of patients with advanced NAFLD, complications of end-stage liver disease (e.g., jaundice, features of portal hypertension such as ascites or variceal hemorrhage) may be the initial findings.

The association of NAFLD with obesity, diabetes, hypertriglyceridemia, hypertension, and cardiovascular disease is well known. Other associations include chronic fatigue, mood alterations, obstructive sleep apnea, thyroid dysfunction, and chronic pain syndrome. NAFLD is an independent risk factor for metabolic syndrome **(Chap. 401)**. Longitudinal studies suggest that patients with NASH are at two- to threefold increased risk for the development of metabolic syndrome. Similarly, studies have shown that patients with NASH have a higher risk for the development of hypertension and diabetes mellitus. The presence of NAFLD is also independently associated with endothelial dysfunction, increased carotid intimal thickness, and the number of plaques in carotid and coronary arteries. Such data indicate that NAFLD has many deleterious effects on health in general.

TREATMENT OF NAFLD

Treatment of NAFLD can be divided into three components: (1) specific therapy of NAFLD-related liver disease; (2) treatment of NAFLD-associated comorbidities; and (3) treatment of the complications of advanced NAFLD. The subsequent discussion focuses on specific therapies for NAFLD, with some mention of their impact on major NAFLD comorbidities (insulin resistance/diabetes, obesity, and dyslipidemia). Treatment of the complications of advanced NAFLD involves management of the complications of cirrhosis and portal hypertension, including primary liver cancers. Approaches to accomplish these objectives are similar to those used in other chronic liver diseases and are covered elsewhere in the textbook **(Chaps. 337 and 78)**.

At present, there are no Food and Drug Administration (FDA)-approved therapies for the treatment of NAFLD. Thus, the current approach to NAFLD management focuses on treatment to improve the risk factors for NASH (i.e., obesity, insulin resistance, metabolic syndrome, dyslipidemia). Based on our understanding of the natural history of NAFLD, only patients with NASH or those with features of hepatic fibrosis on liver biopsy are considered currently for targeted pharmacologic therapies. This approach may change as our understanding of disease pathophysiology improves and potential targets of therapy evolve.

Diet and Exercise Lifestyle changes and dietary modification are the foundation for NAFLD treatment. Many studies indicate that lifestyle modification can improve serum aminotransferases and hepatic steatosis, with loss of at least 3–5% of body weight improving steatosis, but greater weight loss (up to 10%) necessary to improve steatohepatitis. The benefits of different dietary macronutrient contents (e.g., low-carbohydrate vs low-fat diets, saturated vs unsaturated fat diets) and different intensities of calorie restriction appear to be comparable. In adults with NAFLD, exercise regimens that improve fitness may be sufficient to reduce hepatic steatosis, but their impact on other aspects of liver histology remains unknown. Unfortunately, most NAFLD patients are unable to achieve sustained weight loss. Although pharmacologic therapies such as orlistat, topiramate, and phentermine to facilitate weight loss are available, their role in the treatment of NAFLD remains experimental.

Pharmacologic Therapies Several drug therapies have been tried in both research and clinical settings. No agent has yet been approved by the FDA for the treatment of NAFLD. Hence, this remains an area of active research. Because NAFLD is strongly associated with

the metabolic syndrome and type 2 diabetes (Chaps. 396 and 397), the efficacy of various insulin-sensitizing agents has been examined. *Metformin*, an agent that mainly improves hepatic insulin sensitivity, has been evaluated in several small, open-label studies in adults and a recent larger, prospectively randomized trial in children (dubbed the TONIC study). Although several of the adult NASH studies suggested improvements in aminotransferases and/or liver histology, metformin did not improve liver histology in the TONIC study of children with NASH. Thus, it is not currently recommended as a treatment for NASH. Uncontrolled open-label studies have also investigated *thiazolidinediones* (*pioglitazone and rosiglitazone*) in adults with NASH. This class of drugs is known to improve systemic insulin resistance. Both pioglitazone and rosiglitazone reduced aminotransferases and improved some of the histologic features of NASH in small, uncontrolled studies. A large, National Institutes of Health–sponsored, randomized placebo-controlled clinical trial, the PIVENS Study (Pioglitazone vs Vitamin E vs Placebo for the Treatment of 247 Nondiabetic Adults with NASH), demonstrated that resolution of histologic NASH occurred more often in subjects treated with pioglitazone (30 mg/d) than with placebo for 18 months (47 vs 21%, $p = .001$). However, many subjects in the pioglitazone group gained weight, and liver fibrosis did not improve. Also, it should be noted that the long-term safety and efficacy of thiazolidinediones in patients with NASH has not been established. Five-year follow-up of subjects treated with rosiglitazone demonstrated no reduction in liver fibrosis, and rosiglitazone has been associated with increased long-term risk for cardiovascular mortality. Hence, it is not recommended as a treatment for NAFLD. Pioglitazone may be safer because in a recent large meta-analysis it was associated with reduced overall morality, myocardial infarction, and stroke. However, caution must be exercised when considering its use in patients with impaired myocardial function.

Antioxidants have also been evaluated for the treatment of NAFLD because oxidant stress is thought to contribute to the pathogenesis of NASH. *Vitamin E*, an inexpensive yet potent antioxidant, has been examined in several small pediatric and adult studies with varying results. In all of those studies, vitamin E was well tolerated, and most showed modest improvements in aminotransferase levels, radiographic features of hepatic steatosis, and/or histologic features of NASH. Vitamin E (800 IU/d) was also compared to placebo in the PIVENS and TONIC studies. In PIVENS, vitamin E was the only agent that achieved the predetermined primary endpoint (i.e., improvement in steatohepatitis, lobular inflammation, and steatosis score, without an increase in the fibrosis score). This endpoint was met in 43% of patients in the vitamin E group ($p = .001$ vs placebo), 34% in the pioglitazone group ($p = .04$ vs placebo), and 19% in the placebo group. Vitamin E also improved NASH histology in pediatric patients with NASH (TONIC trial). However, a recent population-based study suggested that chronic vitamin E therapy may increase the risk for cardiovascular mortality. Thus, vitamin E should only be considered as a first-line pharmacotherapy for nondiabetic NASH patients. Also, given its potentially negative effects on cardiovascular health, caution should be exercised until the risk-to-benefit ratio and long-term therapeutic efficacy of vitamin E are better defined. Ursodeoxycholic acid (a bile acid that improves certain cholestatic liver diseases) and *betaine* (metabolite of choline that raises SAM levels and decreases cellular oxidative damage) offer no histologic benefit over placebo in patients with NASH. Experimental evidence to support the use of *omega-3 fatty acids* in NAFLD exists; however, a recent large, multicenter, placebo-controlled study failed to demonstrate a histologic benefit. Other pharmacotherapies are also being evaluated in NAFLD (e.g., *probiotics, farnesoid X receptor agonists, intestinal bile acid transport inhibitors, fibroblast growth factor agonists, anticytokine agents, glucagon-like peptide agonists, dipeptidyl IV antagonists, dual PPAR-alpha/PPAR-delta agonists, modulators of liver fibrosis*); however, sufficient data do not yet exist to justify their use as NASH treatments in standard clinical practice.

Statins are an important class of agents to treat dyslipidemia and decrease cardiovascular risk. There is no evidence to suggest that statins cause liver failure in patients with any chronic liver disease, including NAFLD. The incidence of liver enzyme elevations in NAFLD patients taking statins is also no different than that of healthy controls or patients with other chronic liver diseases. Moreover, several studies have suggested that statins may improve aminotransferases and histology in patients with NASH. Yet, there is continued reluctance to use statins in patients with NAFLD. The lack of evidence that statins harm the liver in NAFLD patients, combined with the increase risk for cardiovascular morbidity and mortality in NAFLD patients, warrants the use of statins to treat dyslipidemia in patients with NAFLD/NASH.

Bariatric Surgery Although interest in bariatric surgery as a treatment for NAFLD exists, a recently published Cochrane review concluded that lack of randomized clinical trials or adequate clinical studies prevents definitive assessment of benefits and harms of bariatric surgery as a treatment for NASH. Most studies of bariatric surgery have shown that bariatric surgery is generally safe in individuals with well-compensated chronic liver disease and improves hepatic steatosis and necroinflammation (i.e., features of NAFLD/NASH); however, effects on hepatic fibrosis have been variable. Concern lingers because some of the largest prospective studies suggest that hepatic fibrosis might progress after bariatric surgery. Thus, the Cochrane review deemed it premature to recommend bariatric surgery as a primary treatment for NASH. This opinion was challenged by a recently study which demonstrated that fibrosis stage had improved by 5 years after surgery in about half the patients in one large bariatric surgery cohort. However, most of those individuals had relatively mild fibrosis initially and thus, it is unclear if similar outcomes would occur in individuals with more advanced liver disease. Indeed there is general agreement that patients with NAFLD-related cirrhosis and portal hypertension should be excluded as candidates for bariatric surgery. However, given growing evidence for the benefits of bariatric surgery on metabolic syndrome complications in individuals with refractory obesity, it is not contraindicated in otherwise eligible patients with NAFLD or NASH.

Liver Transplantation Patients with NAFLD in whom end-stage liver disease develops should be evaluated for liver transplantation (Chap. 338). The outcomes of liver transplantation in well-selected patients with NAFLD are generally good, but comorbid medical conditions associated with NAFLD, such as diabetes mellitus, obesity, and cardiovascular disease, often limit transplant candidacy. NAFLD may recur after liver transplantation. The risk factors for recurrent or de novo NAFLD after liver transplantation are multifactorial and include hypertriglyceridemia, obesity, diabetes mellitus, and immunosuppressive therapies, particularly glucocorticoids.

GLOBAL HEALTH CONSIDERATIONS

The epidemic of obesity is now a global and accelerating phenomenon. Worldwide, there are >1 billion overweight adults, of whom at least 300 million are obese. The worldwide prevalence of obesity has more than doubled since 1980. In the wake of the obesity epidemic follow numerous comorbidities, including NAFLD. NAFLD is the most common liver disease identified in Western countries and the fastest rising form of chronic liver disease worldwide. The economic burden directly attributable to NAFLD is already enormous (estimated as $103 billion/year in the United States and nearly 35 billion Euros/year for four European Union countries) and predicted to increase tenfold by the year 2025. Present understanding of NAFLD natural history is based mainly on studies in whites who became overweight/obese and developed the metabolic syndrome in adulthood. The impact of the global childhood obesity epidemic on NAFLD pathogenesis/progression is unknown. Emerging evidence demonstrates that advanced NAFLD, including cirrhosis and primary liver cancer, can occur in children, prompting concerns that childhood-onset NAFLD might follow a more aggressive course than typical adult-acquired NAFLD. Some of the most populated parts of the world are in the midst of industrial revolutions, and certain environmental pollutants seem to exacerbate NAFLD. Some studies also suggest that the risk for NASH and NAFLD-related cirrhosis may be higher in certain ethnic groups such as Asians, certain Hispanics, and Native Americans and lower in others such as African Americans, compared with whites. Although all of these variables confound efforts to predict the net impact of this obesity-related liver disease on global health, it seems likely that

NAFLD will remain a major cause of chronic liver disease worldwide for the foreseeable future.

FURTHER READING

ANGULO P et al: Fibrosis in nonalcoholic fatty liver disease: Mechanisms and clinical implications. Semin Liver Dis 35:132, 2015.

CHALASANI N et al: The diagnosis and management of non-alcoholic fatty liver disease: Practice guideline by the American Gastroenterological Association, American Association for the Study of Liver Diseases, and American College of Gastroenterology. Gastroenterology 142:1592, 2012.

EUROPEAN ASSOCIATION FOR THE STUDY OF THE LIVER (EASL); EUROPEAN ASSOCIATION FOR THE STUDY OF DIABETES (EASD); EUROPEAN ASSOCIATION FOR THE STUDY OF OBESITY (EASO): EASL-EASD-EASO Clinical Practice Guidelines for the management of non-alcoholic fatty liver disease. J Hepatol 64:1388, 2016.

SANYAL AJ et al: Pioglitazone, vitamin E, or placebo for nonalcoholic steatohepatitis. N Engl J Med 362:1675, 2010.

337 Cirrhosis and Its Complications

Bruce R. Bacon

Cirrhosis is a condition that is defined histopathologically and has a variety of clinical manifestations and complications, some of which can be life-threatening. In the past, it has been thought that cirrhosis was never reversible; however, it has become apparent that when the underlying insult that has caused the cirrhosis has been removed, there can be reversal of fibrosis. This is most apparent with the successful treatment of chronic hepatitis C; however, reversal of fibrosis is also seen in patients with hemochromatosis who have been successfully treated and in patients with alcoholic liver disease who have discontinued alcohol use.

Regardless of the cause of cirrhosis, the pathologic features consist of the development of fibrosis to the point that there is architectural distortion with the formation of regenerative nodules. This results in a decrease in hepatocellular mass, and thus function, and an alteration of blood flow. The induction of fibrosis occurs with activation of hepatic stellate cells, resulting in the formation of increased amounts of collagen and other components of the extracellular matrix.

Clinical features of cirrhosis are the result of pathologic changes and mirror the severity of the liver disease. Most hepatic pathologists provide an assessment of grading and staging when evaluating liver biopsy samples. These grading and staging schemes vary between disease states and have been developed for most conditions, including chronic viral hepatitis, nonalcoholic fatty liver disease, and primary biliary cholangitis. Advanced fibrosis usually includes bridging fibrosis with nodularity designated as stage 3 and cirrhosis designated as stage 4. Patients who have cirrhosis have varying degrees of compensated liver function, and clinicians need to differentiate between those who have stable, compensated cirrhosis and those who have decompensated cirrhosis. Patients who have developed complications of their liver disease and have become decompensated should be considered for liver transplantation. Many of the complications of cirrhosis will require specific therapy. *Portal hypertension* is a significant complicating feature of decompensated cirrhosis and is responsible for the development of ascites and bleeding from esophagogastric varices, two complications that signify decompensated cirrhosis. Loss of hepatocellular function results in jaundice, coagulation disorders, and hypoalbuminemia and contributes to the causes of portosystemic encephalopathy.

TABLE 337-1 Causes of Cirrhosis

Alcoholism	Cardiac cirrhosis
Chronic viral hepatitis	Inherited metabolic liver disease
Hepatitis B	Hemochromatosis
Hepatitis C	Wilson's disease
Autoimmune hepatitis	α_1 Antitrypsin deficiency
Nonalcoholic steatohepatitis	Cystic fibrosis
Biliary cirrhosis	Cryptogenic cirrhosis
Primary biliary cholangitis	
Primary sclerosing cholangitis	
Autoimmune cholangiopathy	

The complications of cirrhosis are basically the same regardless of the etiology. Nonetheless, it is useful to classify patients by the cause of their liver disease (Table 337-1); patients can be divided into broad groups with alcoholic cirrhosis, cirrhosis due to chronic viral hepatitis, biliary cirrhosis, and other, less common causes such as cardiac cirrhosis, cryptogenic cirrhosis, and other miscellaneous causes.

ALCOHOLIC CIRRHOSIS

Excessive chronic alcohol use can cause several different types of chronic liver disease, including alcoholic fatty liver, alcoholic hepatitis, and alcoholic cirrhosis. Furthermore, use of excessive alcohol can contribute to liver damage in patients with other liver diseases, such as hepatitis C, hemochromatosis, and fatty liver disease related to obesity. Chronic alcohol use can produce fibrosis in the absence of accompanying inflammation and/or necrosis. Fibrosis can be centrilobular, pericellular, or periportal. When fibrosis reaches a certain degree, there is disruption of the normal liver architecture and replacement of liver cells by regenerative nodules. In alcoholic cirrhosis, the nodules are usually <3 mm in diameter; this form of cirrhosis is referred to as *micronodular*. With cessation of alcohol use, larger nodules may form, resulting in a mixed micronodular and macronodular cirrhosis.

Pathogenesis Alcohol is the most commonly used drug in the United States, and more than two-thirds of adults drink alcohol each year. Thirty percent have had a binge within the past month, and over 7% of adults regularly consume more than two drinks per day. Unfortunately, more than 14 million adults in the United States meet the diagnostic criteria for alcohol abuse or dependence. In the United States, chronic liver disease is the tenth most common cause of death in adults, and alcoholic cirrhosis accounts for ~40% of deaths due to cirrhosis.

Ethanol is mainly absorbed by the small intestine and, to a lesser degree, through the stomach. Gastric alcohol dehydrogenase (ADH) initiates alcohol metabolism. Three enzyme systems account for metabolism of alcohol in the liver. These include cytosolic ADH, the microsomal ethanol oxidizing system (MEOS), and peroxisomal catalase. The majority of ethanol oxidation occurs via ADH to form acetaldehyde, which is a highly reactive molecule that may have multiple effects. Ultimately, acetaldehyde is metabolized to acetate by aldehyde dehydrogenase (ALDH). Intake of ethanol increases intracellular accumulation of triglycerides by increasing fatty acid uptake and by reducing fatty acid oxidation and lipoprotein secretion. Protein synthesis, glycosylation, and secretion are impaired. Oxidative damage to hepatocyte membranes occurs due to the formation of reactive oxygen species; acetaldehyde is a highly reactive molecule that combines with proteins to form protein-acetaldehyde adducts. These adducts may interfere with specific enzyme activities, including microtubular formation and hepatic protein trafficking. With acetaldehyde-mediated hepatocyte damage, certain reactive oxygen species can result in Kupffer cell activation. As a result, profibrogenic cytokines are produced that initiate and perpetuate stellate cell activation, with the resultant production of excess collagen and extracellular matrix. Connective tissue appears in both periportal and pericentral zones and eventually connects portal triads with central veins forming regenerative nodules. Hepatocyte loss occurs, and with increased collagen production and deposition,

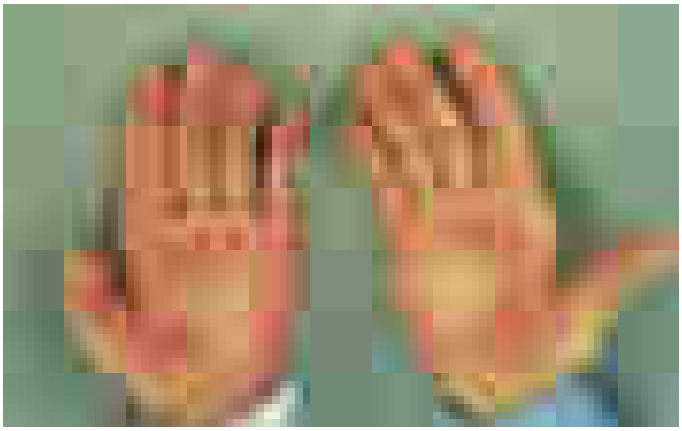


FIGURE 337-1 Palmar erythema. This figure shows palmar erythema in a patient with alcoholic cirrhosis. The erythema is peripheral over the palm with central pallor.

together with continuing hepatocyte destruction, the liver contracts and shrinks in size. This process generally takes from years to decades to occur and requires repeated insults.

Clinical Features The diagnosis of alcoholic liver disease requires an accurate history regarding both amount and duration of alcohol consumption. Patients with alcoholic liver disease can present with nonspecific symptoms such as vague right upper quadrant abdominal pain, fever, nausea and vomiting, diarrhea, anorexia, and malaise. Alternatively, they may present with more specific complications of chronic liver disease, including ascites, edema, or upper gastrointestinal (GI) hemorrhage. Many cases present incidentally at the time of autopsy or elective surgery. Other clinical manifestations include the development of jaundice or encephalopathy. The abrupt onset of any of these complications may be the first event prompting the patient to seek medical attention. Other patients may be identified in the course of an evaluation of routine laboratory studies that are found to be abnormal. On physical examination, the liver and spleen may be enlarged, with the liver edge being firm and nodular. Other frequent findings include scleral icterus, palmar erythema **(Fig. 337-1)**, spider angiomas **(Fig. 337-2)**, parotid gland enlargement, digital clubbing, muscle wasting, or the development of edema and ascites. Men may have decreased body hair and gynecomastia as well as testicular atrophy, which may be a consequence of hormonal abnormalities or a direct toxic effect of alcohol on the testes. In women with advanced alcoholic cirrhosis, menstrual irregularities usually occur, and some women may be amenorrheic. These changes are often reversible following cessation of alcohol ingestion.

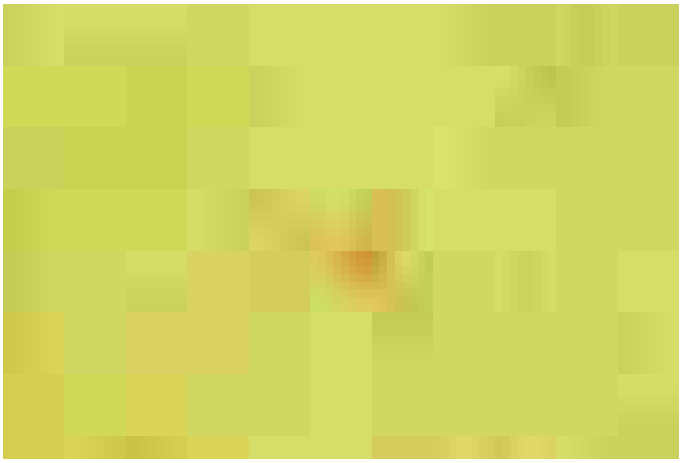


FIGURE 337-2 Spider angioma. This figure shows a spider angioma in a patient with hepatitis C cirrhosis. With release of central compression, the arteriole fills from the center and spreads out peripherally.

Laboratory tests may be completely normal in patients with early compensated alcoholic cirrhosis. Alternatively, in advanced liver disease, many abnormalities usually are present. Patients may be anemic either from chronic GI blood loss, nutritional deficiencies, or hypersplenism related to portal hypertension, or as a direct suppressive effect of alcohol on the bone marrow. A unique form of hemolytic anemia (with spur cells and acanthocytes) called *Zieve's syndrome* can occur in patients with severe alcoholic hepatitis. Platelet counts are often reduced early in the disease, reflective of portal hypertension with hypersplenism. Serum total bilirubin can be normal or elevated with advanced disease. Direct bilirubin is frequently mildly elevated in patients with a normal total bilirubin, but the abnormality typically progresses as the disease worsens. Prothrombin times are often prolonged and usually do not respond to administration of parenteral vitamin K. Serum sodium levels are usually normal unless patients have ascites and then can be depressed, largely due to ingestion of excess free water. Serum alanine and aspartate aminotransferases (ALT, AST) are typically elevated, particularly in patients who continue to drink, with AST levels being higher than ALT levels, usually by a 2:1 ratio.

Diagnosis Patients who have any of the above-mentioned clinical features, physical examination findings, or laboratory studies should be considered to have alcoholic liver disease. The diagnosis, however, requires accurate knowledge that the patient is continuing to use and abuse alcohol. Furthermore, other forms of chronic liver disease (e.g., chronic viral hepatitis or metabolic or autoimmune liver diseases) must be considered or ruled out, or if present, an estimate of relative causality along with the alcohol use should be determined. Liver biopsy can be helpful to confirm a diagnosis, but generally when patients present with alcoholic hepatitis and are still drinking, liver biopsy is withheld until abstinence has been maintained for at least 6 months to determine residual, nonreversible disease.

In patients who have had complications of cirrhosis and who continue to drink, there is a <50% 5-year survival. In contrast, in patients who are able to remain abstinent, the prognosis is significantly improved. In patients with advanced liver disease, the prognosis remains poor; however, in individuals who are able to remain abstinent, liver transplantation is a viable option.

TREATMENT

Alcoholic Cirrhosis

Abstinence is the cornerstone of therapy for patients with alcoholic liver disease. In addition, patients require good nutrition and long-term medical supervision to manage underlying complications that may develop. Complications such as the development of ascites and edema, variceal hemorrhage, or portosystemic encephalopathy all require specific management and treatment. Glucocorticoids are occasionally used in patients with severe alcoholic hepatitis in the absence of infection. Survival has been shown to improve in certain studies. Treatment is restricted to patients with a discriminant function (DF) value of >32. The DF is calculated as the serum total bilirubin plus the difference in the patient's prothrombin time compared to control (in seconds) multiplied by 4.6. In patients for whom this value is >32, there is improved survival at 28 days with the use of glucocorticoids.

Other therapies that have been used include oral pentoxifylline, which decreases the production of tumor necrosis factor α (TNF-α) and other proinflammatory cytokines. In contrast to glucocorticoids, with which complications can occur, pentoxifylline is relatively easy to administer and has few, if any, side effects. A variety of nutritional therapies have been tried with either parenteral or enteral feedings; however, it is unclear whether any of these modalities have significantly improved survival.

Recent studies have used parenterally administered inhibitors of TNF-α such as infliximab or etanercept. Early results have shown no adverse events; however, there was no clear-cut improvement in

survival. Anabolic steroids, propylthiouracil, antioxidants, colchicine, and penicillamine have all been used but do not show clear-cut benefits and are not recommended.

As mentioned above, the cornerstone to treatment is cessation of alcohol use. Recent experience with medications that reduce craving for alcohol, such as acamprosate calcium, has been favorable. Patients may take other necessary medications even in the presence of cirrhosis. Acetaminophen use is often discouraged in patients with liver disease; however, if no more than 2 g of acetaminophen per day are consumed, there generally are no problems.

CIRRHOSIS DUE TO CHRONIC VIRAL HEPATITIS B OR C

Of patients exposed to the hepatitis C virus (HCV), ~80% develop chronic hepatitis C, and of those, about 20–30% will develop cirrhosis over 20–30 years. Many of these patients have had concomitant alcohol use, and the true incidence of cirrhosis due to hepatitis C alone is unknown. Nonetheless, this represents a significant number of patients. It is expected that an even higher percentage will go on to develop cirrhosis over longer periods of time. In the United States, ~5 to 6 million people have been exposed to HCV, with about 4–5 million who are chronically viremic. Worldwide, about 170 million individuals have hepatitis C, with some areas of the world (e.g., Egypt) having up to 15% of the population infected. HCV is a noncytopathic virus, and liver damage is probably immune-mediated. Progression of liver disease due to chronic hepatitis C is characterized by portal-based fibrosis with bridging fibrosis and nodularity developing, ultimately culminating in the development of cirrhosis. In cirrhosis due to chronic hepatitis C, the liver is small and shrunken with characteristic features of a mixed micro- and macronodular cirrhosis seen on liver biopsy. In addition to the increased fibrosis that is seen in cirrhosis due to hepatitis C, an inflammatory infiltrate is found in portal areas with interface hepatitis and occasionally some lobular hepatocellular injury and inflammation. In patients with HCV genotype 3, steatosis is often present.

Similar findings are seen in patients with cirrhosis due to chronic hepatitis B. Of adult patients exposed to hepatitis B, about 5% develop chronic hepatitis B, and about 20% of those patients will go on to develop cirrhosis. Special stains for hepatitis B core (HBc) and hepatitis B surface (HBs) antigen will be positive, and ground-glass hepatocytes signifying HBs antigen (HBsAg) may be present. In the United States, there are about 2 million carriers of hepatitis B, whereas in other parts of the world where hepatitis B virus (HBV) is endemic (i.e., Asia, Southeast Asia, sub-Saharan Africa), up to 15% of the population may be infected, having acquired the infection vertically at the time of birth. Thus, over 300–400 million individuals are thought to have hepatitis B worldwide. Approximately 25% of these individuals may ultimately develop cirrhosis.

Clinical Features and Diagnosis Patients with cirrhosis due to either chronic hepatitis C or B can present with the usual symptoms and signs of chronic liver disease. Fatigue, malaise, vague right upper quadrant pain, and laboratory abnormalities are frequent presenting features. Diagnosis requires a thorough laboratory evaluation, including quantitative HCV RNA testing and analysis for HCV genotype, or hepatitis B serologies to include HBsAg, anti-HBs, HBeAg (hepatitis B e antigen), anti-HBe, and quantitative HBV DNA levels.

TREATMENT

Cirrhosis due to Chronic Viral Hepatitis B or C

Management of complications of cirrhosis revolves around specific therapy for treatment of whatever complications occur (e.g., esophageal variceal hemorrhage, development of ascites and edema, or encephalopathy). In patients with chronic hepatitis B, numerous studies have shown beneficial effects of antiviral therapy, which is effective at viral suppression, as evidenced by reducing aminotransferase levels and HBV DNA levels, and improving histology by reducing inflammation and fibrosis. Several clinical trials and case series have demonstrated that patients with decompensated liver disease can become compensated with the use of antiviral therapy directed against hepatitis B. Currently available therapy includes lamivudine, adefovir, telbivudine, entecavir, and tenofovir. Interferon α can also be used for treating hepatitis B, but it should not be used in cirrhotics. The majority of patients being treated for hepatitis B are receiving either entecavir or tenofovir **(see Chap. 334)**.

Treatment of patients with cirrhosis due to hepatitis C used to be more difficult because the side effects of pegylated interferon and ribavirin therapy were difficult to manage. Over the last several years, interferon-based regimens have been replaced by direct-acting antiviral protocols that are highly successful (>95% cure rate), well tolerated, usually of short duration (8–12 weeks), but costly. These medications have truly revolutionized the treatment of hepatitis C **(see Chap. 334)**.

CIRRHOSIS FROM AUTOIMMUNE HEPATITIS AND NONALCOHOLIC FATTY LIVER DISEASE

Other causes of posthepatitic cirrhosis include autoimmune hepatitis and cirrhosis due to nonalcoholic steatohepatitis. Many patients with autoimmune hepatitis (AIH) present with cirrhosis that is already established. Typically, these patients will not benefit from immunosuppressive therapy with glucocorticoids or azathioprine because the AIH is "burned out." In this situation, liver biopsy does not show a significant inflammatory infiltrate. Diagnosis in this setting requires positive autoimmune markers such as antinuclear antibody (ANA) or anti-smooth-muscle antibody (ASMA). When patients with AIH present with cirrhosis and active inflammation accompanied by elevated liver enzymes, there can be considerable benefit from the use of immunosuppressive therapy.

Patients with nonalcoholic steatohepatitis are increasingly being found to have progressed to cirrhosis. With the epidemic of obesity that continues in Western countries, more and more patients are identified with nonalcoholic fatty liver disease **(Chap. 336)**. Of these, a significant subset has nonalcoholic steatohepatitis and can progress to increased fibrosis and cirrhosis. Over the past several years, it has been increasingly recognized that many patients who were thought to have cryptogenic cirrhosis in fact have nonalcoholic steatohepatitis. As their cirrhosis progresses, they become catabolic and then lose the telltale signs of steatosis seen on biopsy. Management of complications of cirrhosis due to either AIH or nonalcoholic steatohepatitis is similar to that for other forms of cirrhosis.

BILIARY CIRRHOSIS

Biliary cirrhosis has pathologic features that are different from either alcoholic cirrhosis or posthepatitic cirrhosis, yet the manifestations of end-stage liver disease are the same. Cholestatic liver disease may result from necroinflammatory lesions, congenital or metabolic processes, or external bile duct compression. Thus, two broad categories reflect the anatomic sites of abnormal bile retention: *intrahepatic* and *extrahepatic*. The distinction is important for obvious therapeutic reasons. Extrahepatic obstruction may benefit from surgical or endoscopic biliary tract decompression, whereas intrahepatic cholestatic processes will not improve with such interventions and require a different approach.

The major causes of chronic cholestatic syndromes are primary biliary cholangitis (PBC), autoimmune cholangitis (AIC), primary sclerosing cholangitis (PSC), and idiopathic adulthood ductopenia. These syndromes are usually clinically distinguished from each other by antibody testing, cholangiographic findings, and clinical presentation. However, they all share the histopathologic features of chronic cholestasis, such as cholate stasis; copper deposition; xanthomatous transformation of hepatocytes; and irregular, so-called biliary fibrosis. In addition, there may be chronic portal inflammation, interface activity, and chronic lobular inflammation. Ductopenia is a result of this progressive disease as patients develop cirrhosis.

PRIMARY BILIARY CHOLANGITIS

PBC is seen in about 100–200 individuals per million, with a strong female preponderance and a median age of around 50 years at the

time of diagnosis. The cause of PBC is unknown; it is characterized by portal inflammation and necrosis of cholangiocytes in small- and medium-sized bile ducts. Cholestatic features prevail, and biliary cirrhosis is characterized by an elevated bilirubin level and progressive liver failure. Liver transplantation is the treatment of choice for patients with decompensated cirrhosis due to PBC. A variety of therapies have been proposed, but ursodeoxycholic acid (UDCA) has been the only approved treatment that has some degree of efficacy by slowing the rate of progression of the disease. In 2016, obeticholic acid was approved for use in PBC patients with an inadequate response to UDCA. Several other agents are in varying stages of development.

Antimitochondrial antibodies (AMA) are present in about 90% of patients with PBC. These autoantibodies recognize intermitochondrial membrane proteins that are enzymes of the pyruvate dehydrogenase complex (PDC), the branched-chain 2-oxoacid dehydrogenase complex, and the 2-oxogluterate dehydrogenase complex. Most relate to pyruvate dehydrogenase. These autoantibodies are not pathogenic but rather are useful markers for making a diagnosis of PBC.

Pathology Histopathologic analyses of liver biopsies of patients with PBC have resulted in identifying four distinct stages of the disease as it progresses. The earliest lesion is termed *chronic nonsuppurative destructive cholangitis* and is a necrotizing inflammatory process of the portal tracts. Medium and small bile ducts are infiltrated with lymphocytes and undergo duct destruction. Mild fibrosis and sometimes bile stasis can occur. With progression, the inflammatory infiltrate becomes less prominent, but the number of bile ducts is reduced and there is proliferation of smaller bile ductules. Increased fibrosis ensues with the expansion of periportal fibrosis to bridging fibrosis. Finally, cirrhosis, which may be micronodular or macronodular, develops.

Clinical Features Currently, most patients with PBC are diagnosed well before the end-stage manifestations of the disease are present, and, as such, most patients are actually asymptomatic. When symptoms are present, they most prominently include a significant degree of fatigue out of proportion to what would be expected for either the severity of the liver disease or the age of the patient. Pruritus is seen in ~50% of patients at the time of diagnosis, and it can be debilitating. It might be intermittent and usually is most bothersome in the evening. In some patients, pruritus can develop toward the end of pregnancy, and there are examples of patients having been diagnosed with cholestasis of pregnancy rather than PBC. Pruritus that presents prior to the development of jaundice indicates severe disease and a poor prognosis.

Physical examination can show jaundice and other complications of chronic liver disease including hepatomegaly, splenomegaly, ascites, and edema. Other features that are unique to PBC include hyperpigmentation, xanthelasma, and xanthomata, which are related to the altered cholesterol metabolism seen in this disease. Hyperpigmentation is evident on the trunk and the arms and is seen in areas of exfoliation and lichenification associated with progressive scratching related to the pruritus. Bone pain resulting from osteopenia or osteoporosis is occasionally seen at the time of diagnosis.

Laboratory Findings Laboratory findings in PBC show cholestatic liver enzyme abnormalities with an elevation in γ-glutamyl transpeptidase and alkaline phosphatase (ALP) along with mild elevations in aminotransferases (ALT and AST). Immunoglobulins, particularly IgM, are typically increased. Hyperbilirubinemia usually is seen once cirrhosis has developed. Thrombocytopenia, leukopenia, and anemia may be seen in patients with portal hypertension and hypersplenism. Liver biopsy shows characteristic features as described above and should be evident to any experienced hepatopathologist. Up to 10% of patients with characteristic PBC will have features of AIH as well and are defined as having "overlap" syndrome. These patients are usually treated as PBC patients and may progress to cirrhosis with the same frequency as typical PBC patients. Some patients require immunosuppressive medications as well.

Diagnosis PBC should be considered in patients with chronic cholestatic liver enzyme abnormalities. It is most often seen in middle-aged women. AMA testing may be negative, and it should be remembered that as many as 10% of patients with PBC may be AMA-negative. Liver biopsy is most important in this setting of AMA-negative PBC. In patients who are AMA-negative with cholestatic liver enzymes, PSC should be ruled out by way of cholangiography.

TREATMENT

Primary Biliary Cholangitis

Treatment of the typical manifestations of cirrhosis is no different for PBC than for other forms of cirrhosis. UDCA has been shown to improve both biochemical and histologic features of the disease. Improvement is greatest when therapy is initiated early; the likelihood of significant improvement with UDCA is low in patients with PBC who present with manifestations of cirrhosis. UDCA is given in doses of 13–15 mg/kg per day; the medication is usually well-tolerated, although some patients have worsening pruritus with initiation of therapy. A small proportion of patients may have diarrhea or headache as a side effect of the drug. UDCA has been shown to slow the rate of progression of PBC, but it does not reverse or cure the disease. About 30 to 40% of patients with PBC do not have a satisfactory response to UCDA; about half of these patients will have significant improvement with obeticholic acid. Patients with PBC require long-term follow-up by a physician experienced with the disease. Certain patients may need to be considered for liver transplantation should their liver disease decompensate.

The main symptoms of PBC are fatigue and pruritus, and symptom management is important. Several therapies have been tried for treatment of fatigue, but none of them has been successful; frequent naps should be encouraged. Pruritus is treated with antihistamines, narcotic receptor antagonists (naltrexone), and rifampin. Cholestyramine, a bile salt–sequestering agent, has been helpful in some patients but is somewhat tedious and difficult to take. Plasmapheresis has been used rarely in patients with severe intractable pruritus. There is an increased incidence of osteopenia and osteoporosis in patients with cholestatic liver disease, and bone density testing should be performed. Treatment with a bisphosphonate should be instituted when bone disease is identified.

PRIMARY SCLEROSING CHOLANGITIS

As in PBC, the cause of PSC remains unknown. PSC is a chronic cholestatic syndrome that is characterized by diffuse inflammation and fibrosis involving the entire biliary tree, resulting in chronic cholestasis. This pathologic process ultimately results in obliteration of both the intra- and extrahepatic biliary tree, leading to biliary cirrhosis, portal hypertension, and liver failure. The cause of PSC remains unknown despite extensive investigation into various mechanisms related to bacterial and viral infections, toxins, genetic predisposition, and immunologic mechanisms, all of which have been postulated to contribute to the pathogenesis and progression of this syndrome.

Pathologic changes that can occur in PSC show bile duct proliferation as well as ductopenia and fibrous cholangitis (pericholangitis). Often, liver biopsy changes in PSC are not pathognomonic, and establishing the diagnosis of PSC must involve imaging of the biliary tree. Periductal fibrosis is occasionally seen on biopsy specimens and can be quite helpful in making the diagnosis. As the disease progresses, biliary cirrhosis is the final, end-stage manifestation of PSC.

Clinical Features The usual clinical features of PSC are those found in cholestatic liver disease, with fatigue, pruritus, steatorrhea, deficiencies of fat-soluble vitamins, and the associated consequences. As in PBC, the fatigue is profound and nonspecific. Pruritus can often be debilitating and is related to the cholestasis. The severity of pruritus does not correlate with the severity of the disease. Metabolic bone disease, as seen in PBC, can occur with PSC and should be treated (see above).

Laboratory Findings Patients with PSC typically are identified in the course of an evaluation of abnormal liver enzymes. Most

patients have at least a twofold increase in ALP and may have elevated aminotransferases as well. Albumin levels may be decreased, and prothrombin times are prolonged in a substantial proportion of patients at the time of diagnosis. Some degree of correction of a prolonged prothrombin time may occur with parenteral vitamin K. A small subset of patients have aminotransferase elevations greater than five times the upper limit of normal and may have features of AIH on biopsy. These individuals are thought to have an overlap syndrome between PSC and AIH. Autoantibodies are frequently positive in patients with the overlap syndrome, but are typically negative in patients who only have PSC. One autoantibody, the perinuclear antineutrophil cytoplasmic antibody (p-ANCA), is positive in about 65% of patients with PSC. Over 50% of patients with PSC also have ulcerative colitis (UC); accordingly, once a diagnosis of PSC is established, colonoscopy should be performed to look for evidence of UC.

Diagnosis The definitive diagnosis of PSC requires cholangiographic imaging. Over the last several years, magnetic resonance imaging (MRI) with magnetic resonance cholangiopancreatography (MRCP) has been used as the imaging technique of choice for initial evaluation. Once patients are screened in this manner, some investigators feel that endoscopic retrograde cholangiopancreatography (ERCP) should also be performed to be certain whether or not a dominant stricture is present. Typical cholangiographic findings in PSC are multifocal stricturing and beading involving both the intrahepatic and extrahepatic biliary tree. However, although involvement may be of the intrahepatic bile ducts alone or of the extrahepatic bile ducts alone, more commonly, both are involved. These strictures are typically short and with intervening segments of normal or slightly dilated bile ducts that are distributed diffusely, producing the classic beaded appearance. The gallbladder and cystic duct can be involved in up to 15% of cases. Patients with high-grade, diffuse stricturing of the intrahepatic bile ducts have an overall poor prognosis. Gradually, biliary cirrhosis develops, and patients will progress to decompensated liver disease with all the manifestations of ascites, esophageal variceal hemorrhage, and encephalopathy.

TREATMENT

Primary Sclerosing Cholangitis

There is no specific proven treatment for PSC. A recently completed study of high-dose (20 mg/kg per day) UDCA was found to be harmful. Some clinicians use UDCA at "PBC dosages" of 13–15 mg/kg per day with anecdotal improvement. Endoscopic dilatation of dominant strictures can be helpful, but the ultimate treatment is liver transplantation. A dreaded complication of PSC is the development of cholangiocarcinoma, which is a relative contraindication to liver transplantation. Symptoms of pruritus are common, and the approach is as mentioned previously for this problem in patients with PBC (see above).

CARDIAC CIRRHOSIS

Definition Patients with long-standing right-sided congestive heart failure may develop chronic liver injury and cardiac cirrhosis. This is an increasingly uncommon, if not rare, cause of chronic liver disease given the advances made in the care of patients with heart failure.

Etiology and Pathology In the case of long-term right-sided heart failure, there is an elevated venous pressure transmitted via the inferior vena cava and hepatic veins to the sinusoids of the liver, which become dilated and engorged with blood. The liver becomes enlarged and swollen, and with long-term passive congestion and relative ischemia due to poor circulation, centrilobular hepatocytes can become necrotic, leading to pericentral fibrosis. This fibrotic pattern can extend to the periphery of the lobule outward until a unique pattern of fibrosis causing cirrhosis can occur.

Clinical Features Patients typically have signs of congestive heart failure and will manifest an enlarged firm liver on physical examination. ALP levels are characteristically elevated, and aminotransferases may be normal or slightly increased with AST usually higher than ALT. It is unlikely that patients will develop variceal hemorrhage or encephalopathy.

Diagnosis The diagnosis is usually made in someone with clear-cut cardiac disease who has an elevated ALP and an enlarged liver. Liver biopsy shows a pattern of fibrosis that can be recognized by an experienced hepatopathologist. Differentiation from Budd-Chiari syndrome (BCS) can be made by seeing extravasation of red blood cells in BCS, but not in cardiac hepatopathy. Venoocclusive disease can also affect hepatic outflow and has characteristic features on liver biopsy. Venoocclusive disease can be seen under the circumstances of conditioning for bone marrow transplant with radiation and chemotherapy; it can also be seen with the ingestion of certain herbal teas as well as pyrrolizidine alkaloids. This is typically seen in Caribbean countries and rarely in the United States. Treatment is based on management of the underlying cardiac disease.

OTHER TYPES OF CIRRHOSIS

There are several other less common causes of chronic liver disease that can progress to cirrhosis. These include inherited metabolic liver diseases such as hemochromatosis, Wilson's disease, α_1 antitrypsin (α_1AT) deficiency, and cystic fibrosis. For all of these disorders, the manifestations of cirrhosis are similar, with some minor variations, to those seen in other patients with other causes of cirrhosis.

Hemochromatosis is an inherited disorder of iron metabolism that results in a progressive increase in hepatic iron deposition, which, over time, can lead to a portal-based fibrosis progressing to cirrhosis, liver failure, and hepatocellular cancer. While the frequency of hemochromatosis is relatively common, with genetic susceptibility occurring in 1 in 250 individuals, the frequency of end-stage manifestations due to the disease is relatively low, and fewer than 5% of those patients who are genotypically susceptible will go on to develop severe liver disease from hemochromatosis. Diagnosis is made with serum iron studies showing an elevated transferrin saturation and an elevated ferritin level, along with abnormalities identified by *HFE* mutation analysis. Treatment is straightforward, with regular therapeutic phlebotomy.

Wilson's disease is an inherited disorder of copper homeostasis with failure to excrete excess amounts of copper, leading to an accumulation in the liver. This disorder is relatively uncommon, affecting 1 in 30,000 individuals. Wilson's disease typically affects adolescents and young adults. Prompt diagnosis before end-stage manifestations become irreversible can lead to significant clinical improvement. Diagnosis requires determination of ceruloplasmin levels, which are low; 24-h urine copper levels, which are elevated; typical physical examination findings, including Kayser-Fleischer corneal rings; and characteristic liver biopsy findings. Treatment consists of copper-chelating medications.

α_1AT deficiency results from an inherited disorder that causes abnormal folding of the α_1AT protein, resulting in failure of secretion of that protein from the liver. It is unknown how the retained protein leads to liver disease. Patients with α_1AT deficiency at greatest risk for developing chronic liver disease have the ZZ phenotype, but only about 10–20% of such individuals will develop chronic liver disease. Diagnosis is made by determining α_1AT levels and phenotype. Characteristic periodic acid–Schiff (PAS)-positive, diastase-resistant globules are seen on liver biopsy. The only effective treatment is liver transplantation, which is curative.

Cystic fibrosis is an uncommon inherited disorder affecting whites of northern European descent. A biliary-type cirrhosis can occur, and some patients derive benefit from the chronic use of UDCA.

MAJOR COMPLICATIONS OF CIRRHOSIS

The clinical course of patients with advanced cirrhosis is often complicated by a number of important sequelae that can occur regardless of the underlying cause of the liver disease. These include portal hypertension and its consequences of gastroesophageal variceal hemorrhage, splenomegaly, ascites, hepatic encephalopathy, spontaneous bacterial peritonitis (SBP), hepatorenal syndrome (HRS), and hepatocellular carcinoma **(Table 337-2)**.

TABLE 337-2 Complications of Cirrhosis

Portal hypertension	Coagulopathy
Gastroesophageal varices	Factor deficiency
Portal hypertensive gastropathy	Fibrinolysis
Splenomegaly, hypersplenism	Thrombocytopenia
Ascites	Bone disease
Spontaneous bacterial peritonitis	Osteopenia
Hepatorenal syndrome	Osteoporosis
Type 1	Osteomalacia
Type 2	Hematologic abnormalities
Hepatic encephalopathy	Anemia
Hepatopulmonary syndrome	Hemolysis
Portopulmonary hypertension	Thrombocytopenia
Malnutrition	Neutropenia

PORTAL HYPERTENSION

Portal hypertension is defined as the elevation of the hepatic venous pressure gradient (HVPG) to >5 mmHg. Portal hypertension is caused by a combination of two simultaneously occurring hemodynamic processes: (1) increased intrahepatic resistance to the passage of blood flow through the liver due to cirrhosis and regenerative nodules, and (2) increased splanchnic blood flow secondary to vasodilation within the splanchnic vascular bed. Portal hypertension is directly responsible for the two major complications of cirrhosis: variceal hemorrhage and ascites. *Variceal hemorrhage* is an immediate life-threatening problem with a 20–30% mortality rate associated with each episode of bleeding. The portal venous system normally drains blood from the stomach, intestines, spleen, pancreas, and gallbladder, and the portal vein is formed by the confluence of the superior mesenteric and splenic veins. Deoxygenated blood from the small bowel drains into the superior mesenteric vein along with blood from the head of the pancreas, the ascending colon, and part of the transverse colon. Conversely, the splenic vein drains the spleen and the pancreas and is joined by the inferior mesenteric vein, which brings blood from the transverse and descending colon as well as from the superior two-thirds of the rectum. Thus, the portal vein normally receives blood from almost the entire GI tract.

The causes of portal hypertension are usually subcategorized as prehepatic, intrahepatic, and posthepatic **(Table 337-3)**. Prehepatic causes of portal hypertension are those affecting the portal venous system before it enters the liver; they include portal vein thrombosis and splenic vein thrombosis. Posthepatic causes encompass those affecting the hepatic veins and venous drainage to the heart; they include BCS, venoocclusive disease, and chronic right-sided cardiac congestion. Intrahepatic causes account for over 95% of cases of portal hypertension and are represented by the major forms of cirrhosis. Intrahepatic causes of portal hypertension can be further subdivided into presinusoidal, sinusoidal, and postsinusoidal causes. Postsinusoidal causes include venoocclusive disease, whereas presinusoidal causes include congenital hepatic fibrosis and schistosomiasis. Sinusoidal causes are related to cirrhosis from various causes.

TABLE 337-3 Classification of Portal Hypertension

Prehepatic
Portal vein thrombosis
Splenic vein thrombosis
Massive splenomegaly (Banti's syndrome)
Hepatic
Presinusoidal
Schistosomiasis
Congenital hepatic fibrosis
Sinusoidal
Cirrhosis—many causes
Alcoholic hepatitis
Postsinusoidal
Hepatic sinusoidal obstruction (venoocclusive syndrome)
Posthepatic
Budd-Chiari syndrome
Inferior vena caval webs
Cardiac causes
Restrictive cardiomyopathy
Constrictive pericarditis
Severe congestive heart failure

Cirrhosis is the most common cause of portal hypertension in the United States, and clinically significant portal hypertension is present in >60% of patients with cirrhosis. Portal vein obstruction may be idiopathic or can occur in association with cirrhosis or with infection, pancreatitis, or abdominal trauma.

Coagulation disorders that can lead to the development of portal vein thrombosis include polycythemia vera; essential thrombocytosis; deficiencies in protein C, protein S, antithrombin 3, and factor V Leiden; and abnormalities in the gene-regulating prothrombin production. Some patients may have a subclinical myeloproliferative disorder.

Clinical Features The three primary complications of portal hypertension are gastroesophageal varices with hemorrhage, ascites, and hypersplenism. Thus, patients may present with upper GI bleeding, which, on endoscopy, is found to be due to esophageal or gastric varices; with the development of ascites along with peripheral edema; or with an enlarged spleen with associated reduction in platelets and white blood cells on routine laboratory testing.

ESOPHAGEAL VARICES Over the last decade, it has become common practice to screen known cirrhotics with endoscopy to look for esophageal varices. Such screening studies have shown that approximately one-third of patients with histologically confirmed cirrhosis have varices. Approximately 5–15% of cirrhotics per year develop varices, and it is estimated that the majority of patients with cirrhosis will develop varices over their lifetimes. Furthermore, it is anticipated that roughly one-third of patients with varices will develop bleeding. Several factors predict the risk of bleeding, including the severity of cirrhosis (Child's class, MELD score); the height of wedged-hepatic vein pressure; the size of the varix; the location of the varix; and certain endoscopic stigmata, including red wale signs, hematocystic spots, diffuse erythema, bluish color, cherry red spots, or white-nipple spots. Patients with tense ascites are also at increased risk for bleeding from varices.

Diagnosis In patients with cirrhosis who are being followed chronically, the development of portal hypertension is usually revealed by the presence of thrombocytopenia; the appearance of an enlarged spleen; or the development of ascites, encephalopathy, and/or esophageal varices with or without bleeding. In previously undiagnosed patients, any of these features should prompt further evaluation to determine the presence of portal hypertension and liver disease. Varices should be identified by endoscopy. Abdominal imaging, either by computed tomography (CT) or MRI, can be helpful in demonstrating a nodular liver and in finding changes of portal hypertension with intraabdominal collateral circulation. If necessary, interventional radiologic procedures can be performed to determine wedged and free hepatic vein pressures that will allow for the calculation of a wedged-to-free gradient, which is equivalent to the portal pressure. The average normal wedged-to-free gradient is 5 mmHg, and patients with a gradient >12 mmHg are at risk for variceal hemorrhage.

TREATMENT

Variceal Hemorrhage

Treatment for variceal hemorrhage as a complication of portal hypertension is divided into two main categories: (1) primary prophylaxis and (2) prevention of rebleeding once there has been an initial variceal hemorrhage. Primary prophylaxis requires routine screening by endoscopy of all patients with cirrhosis. Once varices that are at

increased risk for bleeding are identified, primary prophylaxis can be achieved either through nonselective beta blockade or by variceal band ligation. Numerous placebo-controlled clinical trials of either propranolol or nadolol have been reported in the literature. The most rigorous studies were those that only included patients with significantly enlarged varices or with hepatic vein pressure gradients >12 mmHg. Patients treated with beta blockers have a lower risk of variceal hemorrhage than those treated with placebo over 1 and 2 years of follow-up. There is also a decrease in mortality related to variceal hemorrhage. Unfortunately, overall survival was improved in only one study. Further studies have demonstrated that the degree of reduction of portal pressure is a significant feature to determine success of therapy. Therefore, it has been suggested that repeat measurements of hepatic vein pressure gradients may be used to guide pharmacologic therapy; however, this may be cost-prohibitive. Several studies have evaluated variceal band ligation and variceal sclerotherapy as methods for providing primary prophylaxis.

Endoscopic variceal ligation (EVL) has achieved a level of success and comfort with most gastroenterologists who see patients with these complications of portal hypertension. Thus, in patients with cirrhosis who are screened for portal hypertension and are found to have large varices, it is recommended that they receive either beta blockade or primary prophylaxis with EVL.

The approach to patients once they have had a variceal bleed is first to treat the acute bleed, which can be life-threatening, and then to prevent further bleeding. Prevention of further bleeding is usually accomplished with repeated variceal band ligation until varices are obliterated. Treatment of acute bleeding requires both fluid and blood-product replacement as well as prevention of subsequent bleeding with EVL.

The medical management of acute variceal hemorrhage includes the use of vasoconstricting agents, usually somatostatin or octreotide. Vasopressin was used in the past but is no longer commonly used. Balloon tamponade (Sengstaken-Blakemore tube or Minnesota tube) can be used in patients who cannot get endoscopic therapy immediately or who need stabilization prior to endoscopic therapy. Control of bleeding can be achieved in the vast majority of cases; however, bleeding recurs in the majority of patients if definitive endoscopic therapy has not been instituted. Octreotide, a direct splanchnic vasoconstrictor, is given at dosages of 50–100 μg/h by continuous infusion. Endoscopic intervention is used as first-line treatment to control bleeding acutely. Some endoscopists will use variceal injection therapy (sclerotherapy) as initial therapy, particularly when bleeding is vigorous. Variceal band ligation is used to control acute bleeding in over 90% of cases and should be repeated until obliteration of all varices is accomplished. When esophageal varices extend into the proximal stomach, band ligation is less successful. In these situations, when bleeding continues from gastric varices, consideration for a transjugular intrahepatic portosystemic shunt (TIPS) should be made. This technique creates a portosystemic shunt by a percutaneous approach using an expandable metal stent, which is advanced under angiographic guidance to the hepatic veins and then through the substance of the liver to create a direct portocaval shunt. This offers an alternative to surgery for acute decompression of portal hypertension. Encephalopathy can occur in as many as 20% of patients after TIPS and is particularly problematic in elderly patients and in patients with preexisting encephalopathy. TIPS should be reserved for individuals who fail endoscopic or medical management or who are poor surgical risks. TIPS can sometimes be used as a bridge to transplantation. Surgical esophageal transsection is a procedure that is rarely used and generally is associated with a poor outcome.

MANAGEMENT OF RECURRENT VARICEAL HEMORRHAGE

Once patients have had an acute bleed and have been managed successfully, attention should be paid to preventing recurrent bleeding. This usually requires repeated variceal band ligation until varices are obliterated. Beta blockade may be of adjunctive benefit in patients who are having recurrent variceal band ligation; however, once varices have been obliterated, the need for beta blockade is lessened. Despite successful variceal obliteration, many patients will still have portal hypertensive gastropathy from which bleeding can occur. Nonselective beta blockade may be helpful to prevent further bleeding from portal hypertensive gastropathy once varices have been obliterated **(Fig. 337-3)**.

Portosystemic shunt surgery is less commonly performed with the advent of TIPS; nonetheless, this procedure should be considered for patients with good hepatic synthetic function who could benefit by having portal decompressive surgery.

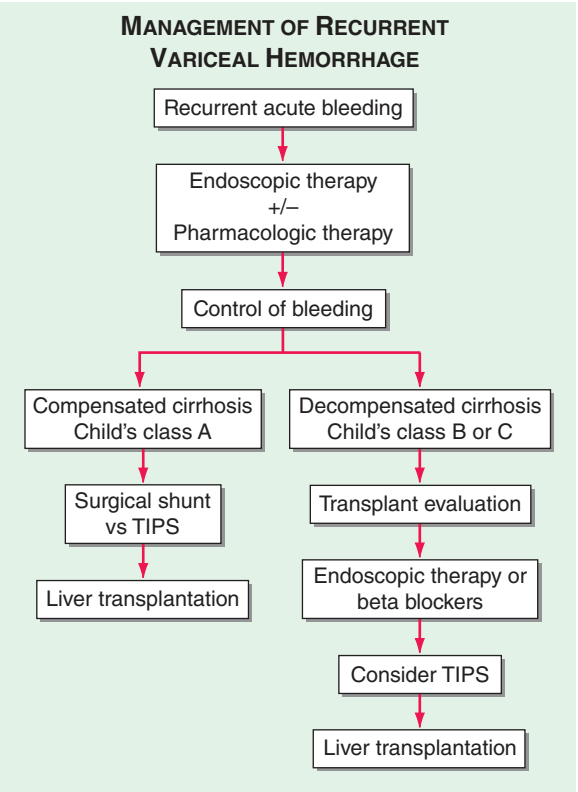


FIGURE 337-3 Management of recurrent variceal hemorrhage. This algorithm describes an approach to management of patients who have recurrent bleeding from esophageal varices. Initial therapy is generally with endoscopic therapy often supplemented by pharmacologic therapy. With control of bleeding, a decision needs to be made as to whether patients should go on to a surgical shunt or TIPS (if they are Child's class A) and be considered for transplant, or if they should have TIPS and be considered for transplant (if they are Child's class B or C). TIPS, transjugular intrahepatic portosystemic shunt.

SPLENOMEGALY AND HYPERSPLENISM

Congestive splenomegaly is common in patients with portal hypertension. Clinical features include the presence of an enlarged spleen on physical examination and the development of thrombocytopenia and leukopenia in patients who have cirrhosis. Some patients will have fairly significant left-sided and left upper quadrant abdominal pain related to an enlarged and engorged spleen. Splenomegaly itself usually requires no specific treatment, although splenectomy can be successfully performed under very special circumstances.

Hypersplenism with the development of thrombocytopenia is a common feature of patients with cirrhosis and is usually the first indication of portal hypertension.

ASCITES

Definition Ascites is the accumulation of fluid within the peritoneal cavity. Overwhelmingly, the most common cause of ascites is portal hypertension related to cirrhosis; however, clinicians should remember that malignant or infectious causes of ascites can be present

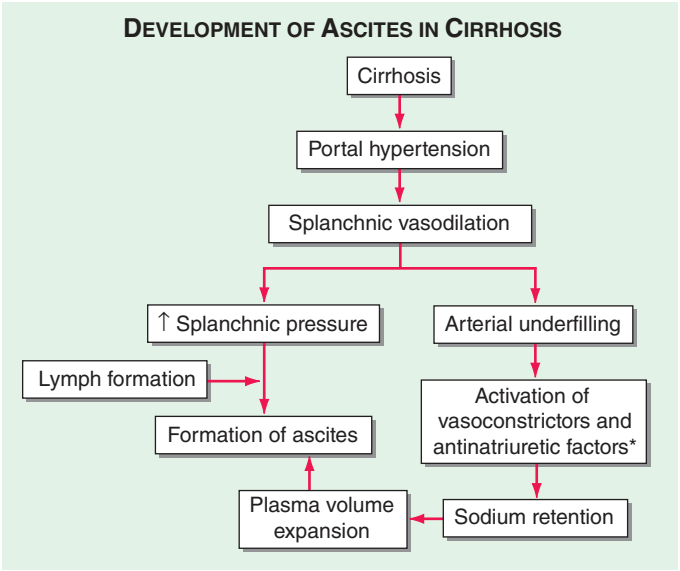


FIGURE 337-4 Development of ascites in cirrhosis. This flow diagram illustrates the importance of portal hypertension with splanchnic vasodilation in the development of ascites. *Antinatriuretic factors include the renin-angiotensin-aldosterone system and the sympathetic nervous system.

as well, and careful differentiation of these other causes are obviously important for patient care.

Pathogenesis The presence of portal hypertension contributes to the development of ascites in patients who have cirrhosis (Fig. 337-4). There is an increase in intrahepatic resistance, causing increased portal pressure, but there is also vasodilation of the splanchnic arterial system, which, in turn, results in an increase in portal venous inflow. Both of these abnormalities result in increased production of splanchnic lymph. Vasodilating factors such as nitric oxide are responsible for the vasodilatory effect. These hemodynamic changes result in sodium retention by causing activation of the renin-angiotensin-aldosterone system with the development of hyperaldosteronism. The renal effects of increased aldosterone leading to sodium retention also contribute to the development of ascites. Sodium retention causes fluid accumulation and expansion of the extracellular fluid volume, which results in the formation of peripheral edema and ascites. Sodium retention is the consequence of a homeostatic response caused by underfilling of the arterial circulation secondary to arterial vasodilation in the splanchnic vascular bed. Because the retained fluid is constantly leaking out of the intravascular compartment into the peritoneal cavity, the sensation of vascular filling is not achieved, and the process continues. Hypoalbuminemia and reduced plasma oncotic pressure also contribute to the loss of fluid from the vascular compartment into the peritoneal cavity. Hypoalbuminemia is due to decreased synthetic function in a cirrhotic liver.

Clinical Features Patients typically note an increase in abdominal girth that is often accompanied by the development of peripheral edema. The development of ascites is often insidious, and it is surprising that some patients wait so long and become so distended before seeking medical attention. Patients usually have at least 1–2 L of fluid in the abdomen before they are aware that there is an increase. If ascitic fluid is massive, respiratory function can be compromised, and patients will complain of shortness of breath. Hepatic hydrothorax may also occur in this setting, contributing to respiratory symptoms. Patients with massive ascites are often malnourished and have muscle wasting and excessive fatigue and weakness.

Diagnosis Diagnosis of ascites is by physical examination and is often aided by abdominal imaging. Patients will have bulging flanks, may have a fluid wave, or may have the presence of shifting dullness. This is determined by taking patients from a supine position to lying on either their left or right side and noting the movement of the dullness to percussion. Subtle amounts of ascites can be detected by ultrasound or CT scanning. Hepatic hydrothorax is more common on the right side and implicates a rent in the diaphragm with free flow of ascitic fluid into the thoracic cavity.

When patients present with ascites for the first time, it is recommended that a diagnostic paracentesis be performed to characterize the fluid. This should include the determination of total protein and albumin content, blood cell counts with differential, and cultures. In the appropriate setting, amylase may be measured and cytology performed. In patients with cirrhosis, the protein concentration of the ascitic fluid is quite low, with the majority of patients having an ascitic fluid protein concentration <1 g/dL. The development of the serum ascites-to-albumin gradient (SAAG) has replaced the description of exudative or transudative fluid. When the gradient between the serum albumin level and the ascitic fluid albumin level is >1.1 g/dL, the cause of the ascites is most likely due to portal hypertension; this is usually in the setting of cirrhosis. When the gradient is <1.1 g/dL, infectious or malignant causes of ascites should be considered. When levels of ascitic fluid proteins are very low, patients are at increased risk for developing SBP. A high level of red blood cells in the ascitic fluid signifies a traumatic tap or perhaps a hepatocellular cancer or a ruptured omental varix. When the absolute level of polymorphonuclear leukocytes is >250/μL, the question of ascitic fluid infection should be strongly considered. Ascitic fluid cultures should be obtained using bedside inoculation of culture media.

TREATMENT

Ascites

Patients with small amounts of ascites can usually be managed with dietary sodium restriction alone. Most average diets in the United States contain 6–8 g of sodium per day, and if patients eat at restaurants or fast-food outlets, the amount of sodium in their diet can exceed this amount. Thus, it is often extremely difficult to get patients to change their dietary habits to ingest <2 g of sodium per day, which is the recommended amount. Patients are frequently surprised to realize how much sodium is in the standard U.S. diet; thus, it is important to make educational pamphlets available to the patient. Often, a simple recommendation is to eat fresh or frozen foods, avoiding canned or processed foods, which are usually preserved with sodium. When a moderate amount of ascites is present, diuretic therapy is usually necessary. Traditionally, spironolactone at 100–200 mg/d as a single dose is started, and furosemide may be added at 40–80 mg/d, particularly in patients who have peripheral edema. In patients who have never received diuretics before, the failure of the above-mentioned dosages suggests that they are not being compliant with a low-sodium diet. If compliance is confirmed and ascitic fluid is not being mobilized, spironolactone can be increased to 400–600 mg/d and furosemide increased to 120–160 mg/d. If ascites is still present with these dosages of diuretics in patients who are compliant with a low-sodium diet, then they are defined as having *refractory ascites*, and alternative treatment modalities including repeated large-volume paracentesis or a TIPS procedure should be considered (Fig. 337-5). Recent studies have shown that TIPS, while managing the ascites, does not improve survival in these patients. Unfortunately, TIPS is often associated with an increased frequency of hepatic encephalopathy and must be considered carefully on a case-by-case basis. The prognosis for patients with cirrhosis with ascites is poor, and some studies have shown that <50% of patients survive 2 years after the onset of ascites. Thus, there should be consideration for liver transplantation in patients with the onset of ascites.

SPONTANEOUS BACTERIAL PERITONITIS

SBP is a common and severe complication of ascites characterized by spontaneous infection of the ascitic fluid without an intraabdominal source. In patients with cirrhosis and ascites severe enough for hospitalization, SBP can occur in up to 30% of individuals and can have a

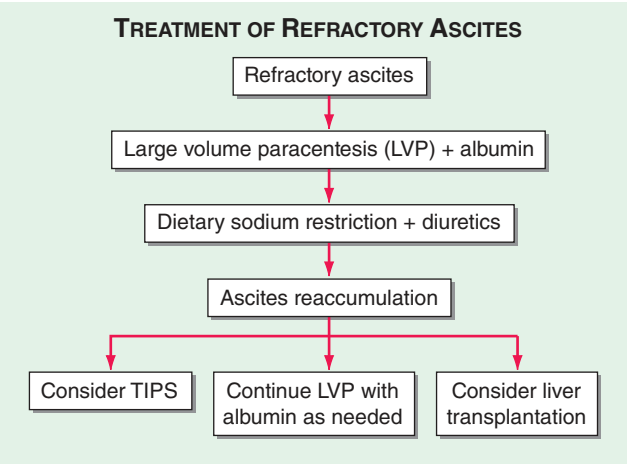


FIGURE 337-5 Treatment of refractory ascites. In patients who develop azotemia in the course of receiving diuretics in the management of their ascites, some will require repeated large-volume paracentesis (LVP), some may be considered for transjugular intrahepatic portosystemic shunt (TIPS), and some would be good candidates for liver transplantation. These decisions are all individualized.

25% in-hospital mortality rate. Bacterial translocation is the presumed mechanism for development of SBP, with gut flora traversing the intestine into mesenteric lymph nodes, leading to bacteremia and seeding of the ascitic fluid. The most common organisms are *Escherichia coli* and other gut bacteria; however, gram-positive bacteria, including *Streptococcus viridans*, *Staphylococcus aureus*, and *Enterococcus* sp., can also be found. If more than two organisms are identified, secondary bacterial peritonitis due to a perforated viscus should be considered. The diagnosis of SBP is made when the fluid sample has an absolute neutrophil count >250/μL. Bedside cultures should be obtained when ascitic fluid is tapped. Patients with ascites may present with fever, altered mental status, elevated white blood cell count, and abdominal pain or discomfort, or they may present without any of these features. Therefore, it is necessary to have a high degree of clinical suspicion, and peritoneal taps are important for making the diagnosis. Treatment is commonly with a third-generation cephalosporin. In patients with variceal hemorrhage, the frequency of SBP is significantly increased, and prophylaxis against SBP is recommended when a patient presents with upper GI bleeding. Furthermore, in patients who have had an episode(s) of SBP and recovered, once-weekly administration of antibiotics is used as prophylaxis for recurrent SBP.

HEPATORENAL SYNDROME

HRS is a form of functional renal failure without renal pathology that occurs in about 10% of patients with advanced cirrhosis or acute liver failure. There are marked disturbances in the arterial renal circulation in patients with HRS; these include an increase in vascular resistance accompanied by a reduction in systemic vascular resistance. The reason for renal vasoconstriction is most likely multifactorial and is poorly understood. The diagnosis is made usually in the presence of a large amount of ascites in patients who have a stepwise progressive increase in creatinine. Type 1 HRS is characterized by a progressive impairment in renal function and a significant reduction in creatinine clearance within 1–2 weeks of presentation. Type 2 HRS is characterized by a reduction in glomerular filtration rate with an elevation of serum creatinine level, but it is fairly stable and is associated with a better outcome than that of type 1 HRS.

HRS is often seen in patients with refractory ascites and requires exclusion of other causes of acute renal failure. Treatment has, unfortunately, been difficult, and in the past, dopamine or prostaglandin analogues were used as renal vasodilating medications. Carefully performed studies have failed to show clear-cut benefit from these therapeutic approaches. Currently, patients are treated with midodrine, an α-agonist, along with octreotide and intravenous albumin. The best therapy for HRS is liver transplantation; recovery of renal function is typical in this setting. In patients with either type 1 or type 2 HRS, the prognosis is poor unless transplant can be achieved within a short period of time.

HEPATIC ENCEPHALOPATHY

Portosystemic encephalopathy is a serious complication of chronic liver disease and is broadly defined as an alteration in mental status and cognitive function occurring in the presence of liver failure. In acute liver injury with fulminant hepatic failure, the development of encephalopathy is a requirement for a diagnosis of fulminant failure. Encephalopathy is much more commonly seen in patients with chronic liver disease. Gut-derived neurotoxins that are not removed by the liver because of vascular shunting and decreased hepatic mass get to the brain and cause the symptoms that we know of as hepatic encephalopathy. Ammonia levels are typically elevated in patients with hepatic encephalopathy, but the correlation between severity of liver disease and height of ammonia levels is often poor, and most hepatologists do not rely on ammonia levels to make a diagnosis. Other compounds and metabolites that may contribute to the development of encephalopathy include certain false neurotransmitters and mercaptans.

Clinical Features In acute liver failure, changes in mental status can occur within weeks to months. Brain edema can be seen in these patients, with severe encephalopathy associated with swelling of the gray matter. Cerebral herniation is a feared complication of brain edema in acute liver failure, and treatment is meant to decrease edema with mannitol and judicious use of intravenous fluids.

In patients with cirrhosis, encephalopathy is often found as a result of certain precipitating events such as hypokalemia, infection, an increased dietary protein load, or electrolyte disturbances. Patients may be confused or exhibit a change in personality. They may actually be quite violent and difficult to manage; alternatively, patients may be very sleepy and difficult to rouse. Because precipitating events are so commonly found, they should be sought carefully. If patients have ascites, this should be tapped to rule out infection. Evidence of GI bleeding should be sought, and patients should be appropriately hydrated. Electrolytes should be measured and abnormalities corrected. In patients presenting with encephalopathy, asterixis is often present. Asterixis can be elicited by having patients extend their arms and bend their wrists back. In this maneuver, patients who are encephalopathic have a "liver flap"—that is, a sudden forward movement of the wrist. This requires patients to be able to cooperate with the examiner and obviously cannot be elicited in patients who are severely encephalopathic or in hepatic coma.

The diagnosis of hepatic encephalopathy is clinical and requires an experienced clinician to recognize and put together all of the various features. Often when patients have encephalopathy for the first time, they (and/or their caregivers) are unaware of what is transpiring, but once they have been through the experience for the first time, they can identify when this is developing in subsequent situations and can often self-medicate to impair the development or worsening of encephalopathy.

TREATMENT

Hepatic Encephalopathy

Treatment is multifactorial and includes management of the above-mentioned precipitating factors. Sometimes hydration and correction of electrolyte imbalance are all that is necessary. In the past, restriction of dietary protein was considered for patients with encephalopathy; however, the negative impact of that maneuver on overall nutrition is thought to outweigh the benefit when treating encephalopathy, and it is thus discouraged. There may be some benefit to replacing animal-based protein with vegetable-based protein in some patients with encephalopathy that is difficult to manage. The mainstay of treatment for encephalopathy, in addition to correcting precipitating factors, is to use lactulose, a nonabsorbable disaccharide, which results in colonic acidification. Catharsis ensues, contributing to the elimination of nitrogenous products in the gut that are responsible for the development of encephalopathy.

The goal of lactulose therapy is to promote 2–3 soft stools per day. Patients are asked to titrate their amount of ingested lactulose to achieve the desired effect. Poorly absorbed antibiotics are often used as adjunctive therapies for patients who have a difficult time with lactulose. The alternating administration of neomycin and metronidazole has been used in the past to reduce the individual side effects of each: neomycin for renal insufficiency and ototoxicity and metronidazole for peripheral neuropathy. More recently, rifaximin at 550 mg twice daily has been very effective in treating encephalopathy without the known side effects of neomycin or metronidazole. Zinc supplementation is sometimes helpful in patients with encephalopathy and is relatively harmless. The development of encephalopathy in patients with chronic liver disease is a poor prognostic sign, but this complication can be managed in the vast majority of patients.

MALNUTRITION IN CIRRHOSIS

Because the liver is principally involved in the regulation of protein and energy metabolism in the body, it is not surprising that patients with advanced liver disease are commonly malnourished. Once patients become cirrhotic, they are more catabolic, and muscle protein is metabolized. There are multiple factors that contribute to the malnutrition of cirrhosis, including poor dietary intake, alterations in gut nutrient absorption, and alterations in protein metabolism. Dietary supplementation for patients with cirrhosis is helpful in preventing patients from becoming catabolic.

ABNORMALITIES IN COAGULATION

Coagulopathy is almost universal in patients with cirrhosis. There is decreased synthesis of clotting factors and impaired clearance of anticoagulants. In addition, patients may have thrombocytopenia from hypersplenism due to portal hypertension. Vitamin K–dependent clotting factors are factors II, VII, IX, and X. Vitamin K requires biliary excretion for its subsequent absorption; thus, in patients with chronic cholestatic syndromes, vitamin K absorption is frequently diminished. Intravenous or intramuscular vitamin K can quickly correct this abnormality. More commonly, the synthesis of vitamin K–dependent clotting factors is diminished because of a decrease in hepatic mass, and, under these circumstances, administration of parenteral vitamin K does not improve the clotting factors or the prothrombin time. Platelet function is often abnormal in patients with chronic liver disease, in addition to decreases in platelet levels due to hypersplenism.

BONE DISEASE IN CIRRHOSIS

Osteoporosis is common in patients with chronic cholestatic liver disease because of malabsorption of vitamin D and decreased calcium ingestion. The rate of bone resorption exceeds that of new bone formation in patients with cirrhosis, resulting in bone loss. Dual x-ray absorptiometry (DEXA) is a useful method for determining osteoporosis or osteopenia in patients with chronic liver disease. When a DEXA scan shows decreased bone mass, treatment should be administered with bisphosphonates that are effective at inhibiting resorption of bone and efficacious in the treatment of osteoporosis.

HEMATOLOGIC ABNORMALITIES IN CIRRHOSIS

Numerous hematologic manifestations of cirrhosis are present, including anemia from a variety of causes including hypersplenism, hemolysis, iron deficiency, and perhaps folate deficiency from malnutrition. Macrocytosis is a common abnormality in red blood cell morphology seen in patients with chronic liver disease, and neutropenia may be seen as a result of hypersplenism.

FURTHER READING

AASLD/IDSA HCV Guidance Panel: Hepatitis C guidance: AASLD-IDSA recommendations for testing, managing, and treating adults infected with hepatitis C virus. Hepatology 62:932, 2015.

Garcia-Tsao G et al: Prevention and management of gastroesophageal varices and variceal hemorrhage in cirrhosis. Am J Gastroenterol 102:2086, 2007.

Runyon BA: Introduction to the revised American Association for the Study of Liver Diseases practice guideline management of adult patients with ascites due to cirrhosis: Update 2012. Hepatology 57:1651, 2013.

Terrault NA et al: AASLD guidelines for treatment of chronic hepatitis B. Hepatology 63:261, 2016.

Vilstrup H et al: Hepatic encephalopathy in chronic liver disease: 2014 Practice Guideline by the American Association for the Study of Liver Diseases and the European Association for the Study of the Liver. Hepatology 60:715, 2014.

338 Liver Transplantation

Raymond T. Chung, Jules L. Dienstag

Liver transplantation—the replacement of the native, diseased liver by a normal organ (allograft)—has matured from an experimental procedure reserved for desperately ill patients to an accepted, lifesaving operation applied more optimally in the natural history of end-stage liver disease. The preferred and technically most advanced approach is *orthotopic transplantation*, in which the native organ is removed and the donor organ is inserted in the same anatomic location. Pioneered in the 1960s by Thomas Starzl at the University of Colorado and, later, at the University of Pittsburgh and by Roy Calne in Cambridge, England, liver transplantation is now performed routinely worldwide. Success measured as 1-year survival has improved from ~30% in the 1970s to >90% today. These improved prospects for prolonged survival resulted from refinements in operative technique, improvements in organ procurement and preservation, advances in immunosuppressive therapy, and, perhaps most influentially, more enlightened patient selection and timing. Despite the perioperative morbidity and mortality, the technical and management challenges of the procedure, and its costs, liver transplantation has become the approach of choice for selected patients whose chronic or acute liver disease is progressive, life-threatening, and unresponsive to medical therapy. Based on the current level of success, the number of liver transplants has continued to grow each year; in 2017, 8082 patients received liver allografts in the United States. Still, the demand for new livers continues to outpace availability; as of 2018, 13,925 patients in the United States were on a waiting list for a donor liver. In response to this drastic shortage of donor organs, many transplantation centers supplement cadaver-organ liver transplantation with living-donor transplantation.

INDICATIONS

Potential candidates for liver transplantation are children and adults who, in the absence of contraindications (see below), suffer from severe, irreversible liver disease for which alternative medical or surgical treatments have been exhausted or are unavailable. *Timing of the operation is of critical importance.* Indeed, improved timing and better patient selection are felt to have contributed more to the increased success of liver transplantation in the 1980s and beyond than all the impressive technical and immunologic advances combined. Although the disease should be advanced, and although opportunities for spontaneous or medically induced stabilization or recovery should be allowed, the procedure should be done sufficiently early to give the surgical procedure a fair chance for success. Ideally, transplantation should be considered in patients with end-stage liver disease who are experiencing or have experienced a life-threatening complication of hepatic decompensation or whose quality of life has deteriorated to unacceptable levels. Although patients with well-compensated cirrhosis can survive for many years, many patients with quasi-stable chronic liver disease have much more advanced disease than may be apparent. As discussed below, the better the status of the patient prior to transplantation, the higher will be its anticipated success rate. The decision about *when* to

TABLE 338-1 Indications for Liver Transplantation

CHILDREN	ADULTS
Biliary atresia	Primary biliary cirrhosis
Neonatal hepatitis	Secondary biliary cirrhosis
Congenital hepatic fibrosis	Primary sclerosing cholangitis
Alagille's syndrome[a]	Autoimmune hepatitis
Byler's disease[b]	Caroli's disease[c]
α_1-Antitrypsin deficiency	Cryptogenic cirrhosis
Inherited disorders of metabolism	Chronic hepatitis with cirrhosis
Wilson's disease	Hepatic vein thrombosis
Tyrosinemia	Fulminant hepatitis
Glycogen storage diseases	Alcoholic cirrhosis
Lysosomal storage diseases	Chronic viral hepatitis
Protoporphyria	Primary hepatocellular malignancies
Crigler-Najjar disease type I	Hepatic adenomas
Familial hypercholesterolemia	Nonalcoholic steatohepatitis
Primary hyperoxaluria type I	Familial amyloid polyneuropathy
Hemophilia	

[a]Arteriohepatic dysplasia, with paucity of bile ducts, and congenital malformations, including pulmonary stenosis. [b]Intrahepatic cholestasis, progressive liver failure, and mental and growth retardation. [c]Multiple cystic dilatations of the intrahepatic biliary tree.

transplant is complex and requires the combined judgment of an experienced team of hepatologists, transplant surgeons, anesthesiologists, and specialists in support services, not to mention the well-informed consent of the patient and the patient's family.

TRANSPLANTATION IN CHILDREN

Indications for transplantation in children are listed in Table 338-1. The most common is *biliary atresia*. *Inherited or genetic disorders of metabolism* associated with liver failure constitute another major indication for transplantation in children and adolescents. In Crigler-Najjar disease type I and in certain hereditary disorders of the urea cycle and of amino acid or lactate-pyruvate metabolism, transplantation may be the only way to prevent impending deterioration of central nervous system function, despite the fact that the native liver is structurally normal. Combined heart and liver transplantation has yielded dramatic improvement in cardiac function and in cholesterol levels in children with homozygous familial hypercholesterolemia; combined liver and kidney transplantation has been successful in patients with primary hyperoxaluria type I. In hemophiliacs with transfusion-associated hepatitis and liver failure, liver transplantation has been associated with recovery of normal factor VIII synthesis.

TRANSPLANTATION IN ADULTS

Liver transplantation is indicated for end-stage *cirrhosis* of all causes (Table 338-1). In *sclerosing cholangitis* and *Caroli's disease* (multiple cystic dilatations of the intrahepatic biliary tree), recurrent infections and sepsis associated with inflammatory and fibrotic obstruction of the biliary tree may be an indication for transplantation. Because prior biliary surgery complicates and is a relative contraindication for liver transplantation, surgical diversion of the biliary tree has been all but abandoned for patients with sclerosing cholangitis. In patients who undergo transplantation for *hepatic vein thrombosis* (*Budd-Chiari syndrome*), postoperative anticoagulation is essential; underlying myeloproliferative disorders may have to be treated but are not a contraindication to liver transplantation. If a donor organ can be located quickly, before life-threatening complications—including cerebral edema—set in, patients with acute liver failure are candidates for liver transplantation. Routine candidates for liver transplantation are patients with *alcoholic cirrhosis*, *chronic viral hepatitis*, and *primary hepatocellular malignancies*. Although all three of these categories are considered to be high risk, liver transplantation can be offered to carefully selected patients. Currently, chronic hepatitis C and alcoholic liver disease are the most common indications for liver transplantation, accounting for over 40% of all adult candidates who undergo the procedure. Patients with alcoholic cirrhosis can be considered as candidates for transplantation if they meet strict criteria for abstinence and reform; however, these criteria still do not prevent recidivism in up to a quarter of cases. In highly selected cases in a limited number of centers, transplantation for severe *acute* alcoholic hepatitis has been performed with success; however, because patients with acute alcoholic hepatitis are still actively using alcohol, and because continued alcohol abuse remains a concern, acute alcoholic hepatitis is not a routine indication for liver transplantation. Patients with chronic hepatitis C have early allograft and patient survival comparable to those of other subsets of patients after transplantation; however, reinfection in the donor organ is universal, recurrent hepatitis C is insidiously progressive, allograft cirrhosis develops in 20–30% at 5 years, and cirrhosis and late organ failure occur at a higher frequency beyond 5 years. With the introduction of highly effective direct acting antiviral agents targeting hepatitis C virus (HCV), allograft outcomes have already improved substantially. In patients with chronic hepatitis B, in the absence of measures to prevent recurrent hepatitis B, survival after transplantation is reduced by ~10–20%; however, prophylactic use of hepatitis B immune globulin (HBIg) during and after transplantation increases the success of transplantation to a level comparable to that seen in patients with nonviral causes of liver decompensation. Specific oral antiviral drugs (e.g., entecavir, tenofovir disoproxil fumarate) (Chap. 334) can be used both for prophylaxis against and for treatment of recurrent hepatitis B, facilitating further the management of patients undergoing liver transplantation for end-stage hepatitis B; most transplantation centers rely on antiviral drugs with or without HBIg to manage patients with hepatitis B. Issues of disease recurrence are discussed in more detail below. Patients with nonmetastatic primary hepatobiliary tumors—primary hepatocellular carcinoma (HCC), cholangiocarcinoma, hepatoblastoma, angiosarcoma, epithelioid hemangioendothelioma, and multiple or massive hepatic adenomata—have undergone liver transplantation; however, for some hepatobiliary malignancies, overall survival is significantly lower than that for other categories of liver disease. Most transplantation centers have reported 5-year recurrence-free survival rates in patients with unresectable HCC for single tumors <5 cm in diameter or for three or fewer lesions all <3 cm comparable to those seen in patients undergoing transplantation for nonmalignant indications. Consequently, liver transplantation is currently restricted to patients whose hepatic malignancies meet these criteria. Expanded criteria for patients with HCC continue to be evaluated. Because the likelihood of recurrent cholangiocarcinoma is very high, only highly selected patients with limited disease are being evaluated for transplantation after intensive chemotherapy and radiation.

CONTRAINDICATIONS

Absolute contraindications for transplantation include life-threatening systemic diseases, uncontrolled extrahepatic bacterial or fungal infections, preexisting advanced cardiovascular or pulmonary disease, multiple uncorrectable life-threatening congenital anomalies, metastatic malignancy, and active drug or alcohol abuse (Table 338–2). Because carefully selected patients in their sixties and even seventies have undergone transplantation successfully, advanced age per se is no longer considered an absolute contraindication; however, in older patients a more thorough preoperative evaluation should be undertaken to exclude ischemic cardiac disease and other comorbid conditions. Advanced age (>70 years), however, should be considered a *relative contraindication*—that is, a factor to be taken into account with other relative contraindications. Other relative contraindications include portal vein thrombosis, preexisting renal disease not associated with liver disease (which may prompt consideration of combined liver and kidney transplantation), intrahepatic or biliary sepsis, severe hypoxemia (Po_2 <50 mmHg) resulting from right-to-left intrapulmonary shunts, portopulmonary hypertension with high mean pulmonary artery pressures (>35 mmHg), previous extensive hepatobiliary surgery, any uncontrolled serious psychiatric disorder, and lack of sufficient social supports. Any one of these relative contraindications is insufficient in and of itself to preclude transplantation. For example, the problem of portal vein thrombosis can be overcome by constructing a graft

TABLE 338-2 Contraindications to Liver Transplantation

ABSOLUTE	RELATIVE
Uncontrolled extrahepatobiliary infection	Age >70
Active, untreated sepsis	Prior extensive hepatobiliary surgery
Uncorrectable, life-limiting congenital anomalies	Portal vein thrombosis
Active substance or alcohol abuse	Renal failure not attributable to liver disease
Advanced cardiopulmonary disease	Previous extrahepatic malignancy (not including nonmelanoma skin cancer)
Extrahepatobiliary malignancy (not including nonmelanoma malignancy skin cancer)	Severe obesity
Metastatic malignancy to the liver	Severe malnutrition/wasting
Cholangiocarcinoma	Medical noncompliance
AIDS	HIV seropositivity with failure to control HIV viremia or CD4 <100/μL
Life-threatening systemic diseases	Intrahepatic sepsis Severe hypoxemia secondary to right-to-left intrapulmonary shunts (Po_2 <50 mmHg) Severe pulmonary hypertension (mean pulmonary artery pressure >35 mmHg) Uncontrolled psychiatric disorder

from the donor liver portal vein to the recipient's superior mesenteric vein. Now that combination antiretroviral therapy has dramatically improved the survival of persons with HIV infection **(Chap. 197)**, and because end-stage liver disease caused by chronic hepatitis C and B has emerged as a serious source of morbidity and mortality in the HIV-infected population, liver transplantation has now been performed successfully in selected HIV-positive persons who have excellent control of HIV infection. Selected patients with CD4± T cell counts >100/μL and with pharmacologic suppression of HIV viremia have undergone transplantation for end-stage liver disease. HIV-infected persons who have received liver allografts for end-stage liver disease resulting from chronic hepatitis B have experienced survival rates compared to those of HIV-negative persons undergoing transplantation for the same indication. In contrast, recurrent HCV in the allograft has until recently limited long-term success in persons with HCV-related end-stage liver disease. Again, it is expected that the availability of direct-acting antiviral (DAA) agents targeting HCV (see below and **Chap 334**) will significantly improve allograft outcomes.

TECHNICAL CONSIDERATIONS

CADAVER DONOR SELECTION

Cadaver donor livers for transplantation are procured primarily from victims of head trauma. Organs from brain-dead donors up to age 60 are acceptable if the following criteria are met: hemodynamic stability, adequate oxygenation, absence of bacterial or fungal infection, absence of abdominal trauma, absence of hepatic dysfunction, and serologic exclusion of hepatitis B (HBV) and C viruses and HIV. Occasionally, organs from donors with hepatitis B and C are used (e.g., for recipients with prior hepatitis B and C, respectively). Organs from donors with antibodies to hepatitis B core antigen (anti-HBc) can also be used when the need is especially urgent, and recipients of these organs are treated prophylactically with antiviral drugs. Cardiovascular and respiratory functions are maintained artificially until the liver can be removed. Transplantation of organs procured from deceased donors who have succumbed to cardiac death can be performed successfully under selected circumstances, when ischemic time is minimized and liver histology preserved. Compatibility in ABO blood group and organ size between donor and recipient are important considerations in donor selection; however, ABO-incompatible, split liver, or reduced-donor-organ transplants can be performed in emergencies or marked donor scarcity. Tissue typing for human leukocyte antigen (HLA) matching is not required, and preformed cytotoxic HLA antibodies do not preclude liver transplantation. Following perfusion with cold electrolyte solution, the donor liver is removed and packed in ice. The use of University of Wisconsin (UW) solution, rich in lactobionate and raffinose, has permitted the extension of cold ischemic time up to 20 h; however, 12 h may be a more reasonable limit. Improved techniques for harvesting multiple organs from the same donor have increased the availability of donor livers, but the availability of donor livers is far outstripped by the demand. Currently in the United States, all donor livers are distributed through a nationwide organ-sharing network (United Network for Organ Sharing [UNOS]) designed to allocate available organs based on regional considerations and recipient acuity. Recipients who have the highest disease severity generally have the highest priority, but allocation strategies that balance highest urgency against best outcomes continue to evolve to distribute cadaver organs most effectively. Allocation based on the Child-Turcotte-Pugh (CTP) score, which uses five clinical variables (encephalopathy stage, ascites, bilirubin, albumin, and prothrombin time) and waiting time, has been replaced by allocation based on urgency alone, calculated by the Model for End-Stage Liver Disease (MELD) score. The MELD score is based on a mathematical model that includes bilirubin, creatinine, and prothrombin time expressed as international normalized ratio (INR) **(Table 338-3)**. Neither waiting time (except as a tie breaker between two potential recipients with the same MELD scores) nor posttransplantation outcome is taken into account, but use of the MELD score has been shown to reduce waiting list mortality, to reduce waiting time prior to transplantation, to be the best predictor of pretransplantation mortality, to satisfy the prevailing view that medical need should be the decisive determinant, and to eliminate both the subjectivity inherent in the CTP scoring system (presence and degree of ascites and hepatic encephalopathy) and the differences in waiting times among different regions of the country. Recent data indicate that liver recipients with MELD scores <15 experienced higher posttransplantation mortality rates than similarly classified patients who remained on the wait list. This observation led to the modification of UNOS policy to allocate donor organs to candidates with MELD scores exceeding 15 within the local or regional procurement organization before offering the organ to local patients whose scores are <15. In 2016, the MELD score was modified to incorporate serum sodium, another important predictor of survival in liver transplantation candidates (the MELD-Na score).

The highest priority (status 1) continues to be reserved for patients with fulminant hepatic failure or primary graft nonfunction. Because candidates for liver transplantation who have HCC may not be

TABLE 338-3 United Network for Organ Sharing (UNOS) Liver Transplantation Waiting List Criteria

Status 1	Fulminant hepatic failure (including primary graft nonfunction and hepatic artery thrombosis within 7 days after transplantation as well as acute decompensated Wilson's disease)[a]
The Model for End-Stage Liver Disease (MELD)-Na score, on a continuous scale,[b] determines allocation of the remainder of donor organs. This model is based on the following calculation:	
MELD = 3.78 × $\log_e$ bilirubin (mg/100 mL) ± 11.2 × $\log_e$ international normalized ratio (INR) ± 9.57 × $\log_e$ creatinine (mg/100 mL) + 6.43.[c,d,e]	
MELD-Na = MELD + 1.59 * (135 – Na [mEq/L])	
Online calculators to determine MELD scores are available, such as the following: *https://optn.transplant.hrsa.gov/resources/allocation-calculators/meld-calculator/*	

[a]For children <18 years of age, status 1 includes acute or chronic liver failure plus hospitalization in an intensive care unit or inborn errors of metabolism. Status 1 is retained for those persons with fulminant hepatic failure and supersedes the MELD score. [b]The MELD scale is continuous, with 34 levels ranging between 6 and 40 (scores above 40 are categorized as 40). Donor organs usually do not become available unless the MELD score exceeds 20. [c]Patients with stage T2 hepatocellular carcinoma receive 22 disease-specific points. [d]Creatinine is included because renal function is a validated predictor of survival in patients with liver disease. For adults undergoing dialysis twice a week, the creatinine in the equation is set to 4 mg/100 mL. [e]For children <18 years of age, the Pediatric End-Stage Liver Disease (PELD) scale is used. This scale is based on albumin, bilirubin, INR, growth failure, and age. Status 1 is retained.

sufficiently decompensated to compete for donor organs based on urgency criteria alone, and because protracted waiting for cadaver donor organs often results in tumor growth beyond acceptable limits for transplantation, such patients are assigned disease-specific MELD points (Table 338–3). Other disease-specific MELD exceptions include portopulmonary hypertension, hepatopulmonary syndrome, familial amyloid polyneuropathy, primary hyperoxaluria (necessitating liver-kidney transplantation), cystic fibrosis liver disease, and highly selected cases of hilar cholangiocarcinoma.

LIVING DONOR TRANSPLANTATION

Occasionally, especially for liver transplantation in children, one cadaver organ can be split between two recipients (one adult and one child). A more viable alternative, transplantation of the right lobe of the liver from a healthy adult donor into an adult recipient, has gained increased popularity. Living donor transplantation of the left lobe (left lateral segment), introduced in the early 1990s to alleviate the extreme shortage of donor organs for small children, accounts currently for approximately one-third of all liver transplantation procedures in children. Driven by the shortage of cadaver organs, living donor transplantation involving the more sizable right lobe is being considered with increasing frequency in adults; however, living donor liver transplantation cannot be expected to solve the donor organ shortage; 367 such procedures were done in 2017, representing only about 5% of all liver transplant operations done in the United States.

Living donor transplantation can reduce waiting time and cold-ischemia time; it is done under elective, rather than emergency, circumstances; and it may be lifesaving in recipients who cannot afford to wait for a cadaver donor. The downside, of course, is the risk to the healthy donor (a mean of 10 weeks of medical disability; biliary complications in ~5%; postoperative complications such as wound infection, small-bowel obstruction, and incisional hernias in 9–19%; and even, in 0.2–0.4%, death) as well as the increased frequency of biliary (15–32%) and vascular (10%) complications in the recipient. Potential donors must participate voluntarily without coercion, and transplantation teams should go to great lengths to exclude subtle coercive or inappropriate psychological factors as well as outline carefully to both donor and recipient the potential benefits and risks of the procedure. Donors for the procedure should be 18–60 years old; have a compatible blood type with the recipient; have no chronic medical problems or history of major abdominal surgery; should be related genetically or emotionally to the recipient; and pass an exhaustive series of clinical, biochemical, and serologic evaluations to unearth disqualifying medical disorders. The recipient should meet the same UNOS criteria for liver transplantation as recipients of a cadaver donor allograft. Comprehensive outcome data on adult-to-adult living donor liver transplantation are being collected (*www.nih-a2all.org*).

SURGICAL TECHNIQUE

Removal of the recipient's native liver is technically difficult, particularly in the presence of portal hypertension with its associated collateral circulation and extensive varices and especially in the presence of scarring from previous abdominal operations. The combination of portal hypertension and coagulopathy (elevated prothrombin time and thrombocytopenia) may translate into large blood product transfusion requirements. After the portal vein and infrahepatic and suprahepatic inferior vena cavae are dissected, the hepatic artery and common bile duct are dissected. Then the native liver is removed and the donor organ inserted. During the anhepatic phase, coagulopathy, hypoglycemia, hypocalcemia, and hypothermia are encountered and must be managed by the anesthesiology team. Caval, portal vein, hepatic artery, and bile duct anastomoses are performed in succession, the last by end-to-end suturing of the donor and recipient common bile ducts **(Fig. 338-1)** or by choledochojejunostomy to a Roux-en-Y loop if the recipient common bile duct cannot be used for reconstruction (e.g., in sclerosing cholangitis). A typical transplant operation lasts 8 h, with a range of 6–18 h. Because of excessive bleeding, large volumes of blood, blood products, and volume expanders may be required during surgery; however, blood requirements have fallen sharply with improvements in surgical technique, blood-salvage interventions, and experience.

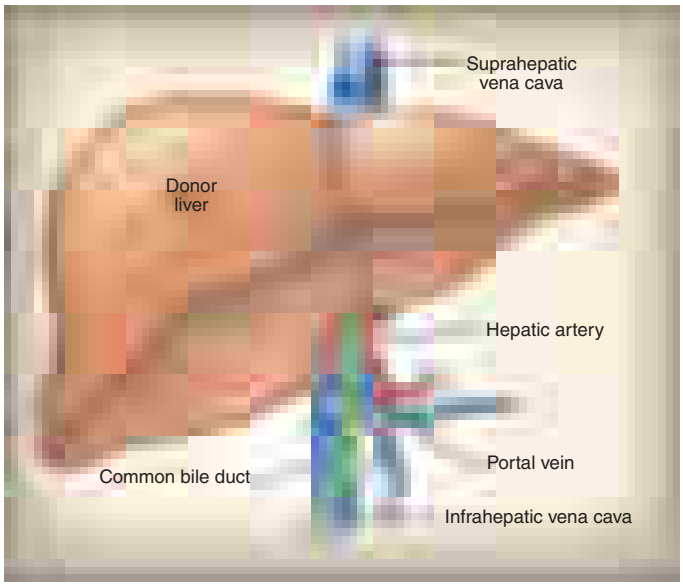


FIGURE 338-1 The anastomoses in orthotopic liver transplantation. The anastomoses are performed in the following sequence: (1) suprahepatic and infrahepatic vena cava, (2) portal vein, (3) hepatic artery, and (4) common bile duct-to-duct anastomosis. *(Adapted from JL Dienstag, AB Cosimi: N Engl J Med 367:1483, 2012.)*

As noted above, emerging alternatives to orthotopic liver transplantation include split-liver grafts, in which one donor organ is divided and inserted into two recipients; and living donor procedures, in which part of the left (for children), the left (for children or small adults), or the right (for adults) lobe of the liver is harvested from a living donor for transplantation into the recipient. In the adult procedure, once the right lobe is removed from the donor, the donor right hepatic vein is anastomosed to the recipient right hepatic vein remnant, followed by donor-to-recipient anastomoses of the portal vein and then the hepatic artery. Finally, the biliary anastomosis is performed, duct-to-duct if practical or via Roux-en-Y anastomosis. Heterotopic liver transplantation, in which the donor liver is inserted without removal of the native liver, has met with very limited success and acceptance, except in a very small number of centers. In attempts to support desperately ill patients until a suitable donor organ can be identified, several transplantation centers are studying extracorporeal perfusion with bioartificial liver cartridges constructed from hepatocytes bound to hollow fiber systems and used as temporary hepatic-assist devices, but their efficacy remains to be established. Areas of research with the potential to overcome the shortage of donor organs include hepatocyte transplantation and xenotransplantation with genetically modified organs of nonhuman origin (e.g., swine).

POSTOPERATIVE COURSE AND MANAGEMENT

IMMUNOSUPPRESSIVE THERAPY

The introduction in 1980 of cyclosporine as an immunosuppressive agent contributed substantially to the improvement in survival after liver transplantation. Cyclosporine, a calcineurin inhibitor, blocks early activation of T cells and is specific for T cell functions that result from the interaction of the T cell with its receptor and that involve the calcium-dependent signal transduction pathway. As a result, the activity of cyclosporine leads to inhibition of lymphokine gene activation, blocking interleukins 2, 3, and 4, tumor necrosis factor a, and other lymphokines. Cyclosporine also inhibits B cell functions. This process occurs without affecting rapidly dividing cells in the bone marrow, which may account for the reduced frequency of posttransplantation systemic infections. The most common and important side effect of cyclosporine therapy is nephrotoxicity. Cyclosporine causes dose-dependent renal tubular injury and direct renal artery vasospasm. Following renal function is therefore important in monitoring cyclosporine therapy, perhaps

even a more reliable indicator than blood levels of the drug. Nephrotoxicity is reversible and can be managed by dose reduction. Other adverse effects of cyclosporine therapy include hypertension, hyperkalemia, tremor, hirsutism, glucose intolerance, and gingival hyperplasia.

Tacrolimus, a macrolide lactone antibiotic isolated from a Japanese soil fungus, *Streptomyces tsukubaensis*, has the same mechanism of action as cyclosporine but is 10–100 times more potent. Initially applied as "rescue" therapy for patients in whom rejection occurred despite the use of cyclosporine, tacrolimus was shown to be associated with a reduced frequency of acute, refractory, and chronic rejection. Although patient and graft survival are the same with these two drugs, the advantage of tacrolimus in minimizing episodes of rejection, reducing the need for additional glucocorticoid doses, and reducing the likelihood of bacterial and cytomegalovirus (CMV) infection has simplified the management of patients undergoing liver transplantation. In addition, the oral absorption of tacrolimus is more predictable than that of cyclosporine, especially during the early postoperative period when T-tube drainage interferes with the enterohepatic circulation of cyclosporine. As a result, in most transplantation centers, tacrolimus has now supplanted cyclosporine for primary immunosuppression, and many centers rely on oral rather than IV administration from the outset. For transplantation centers that prefer cyclosporine, a better-absorbed microemulsion preparation is available.

Although more potent than cyclosporine, tacrolimus is also more toxic and more likely to be discontinued for adverse events. The toxicity of tacrolimus is similar to that of cyclosporine; nephrotoxicity and neurotoxicity are the most commonly encountered adverse effects, and neurotoxicity (tremor, seizures, hallucinations, psychoses, coma) is more likely and more severe in tacrolimus-treated patients. Both drugs can cause diabetes mellitus, but tacrolimus does not cause hirsutism or gingival hyperplasia. Because of overlapping toxicity between cyclosporine and tacrolimus, especially nephrotoxicity, and because tacrolimus reduces cyclosporine clearance, these two drugs should not be used together. Because 99% of tacrolimus is metabolized by the liver, hepatic dysfunction reduces its clearance; in primary graft nonfunction (when, for technical reasons or because of ischemic damage prior to its insertion, the allograft is defective and does not function normally from the outset), tacrolimus doses have to be reduced substantially, especially in children. Both cyclosporine and tacrolimus are metabolized by the cytochrome P450 IIIA system, and, therefore, drugs that induce cytochrome P450 (e.g., phenytoin, phenobarbital, carbamazepine, rifampin) reduce available levels of cyclosporine and tacrolimus; and drugs that inhibit cytochrome P450 (e.g., erythromycin, fluconazole, ketoconazole, clotrimazole, itraconazole, verapamil, diltiazem, danazol, metoclopramide, the HIV protease inhibitor ritonavir, and the HCV protease inhibitor paritaprevir) increase cyclosporine and tacrolimus blood levels. Indeed, itraconazole is used occasionally to help boost tacrolimus levels. Like azathioprine, cyclosporine and tacrolimus appear to be associated with a risk of lymphoproliferative malignancies (see below), which may occur earlier after cyclosporine or tacrolimus than after azathioprine therapy. Because of these side effects, combinations of cyclosporine or tacrolimus with prednisone and an antimetabolite (azathioprine or mycophenolic acid, see below)—all at reduced doses—are preferable regimens for immunosuppressive therapy.

Mycophenolic acid, a nonnucleoside purine metabolism inhibitor derived as a fermentation product from several *Penicillium* species, is another immunosuppressive drug being used for patients undergoing liver transplantation. Mycophenolate has been shown to be better than azathioprine, when used with other standard immunosuppressive drugs, in preventing rejection after renal transplantation and has been adopted widely as well for use in liver transplantation. The most common adverse effects of mycophenolate are bone marrow suppression and gastrointestinal complaints.

In patients with pretransplantation renal dysfunction or renal deterioration that occurs intraoperatively or immediately postoperatively, tacrolimus or cyclosporine therapy may not be practical; under these circumstances, induction or maintenance of immunosuppression with antithymocyte globulin (ATG, thymoglobulin) or monoclonal antibodies to T cells, OKT3, may be appropriate. Therapy with these agents has been especially effective in reversing acute rejection in the posttransplantation period and is the standard treatment for acute rejection that fails to respond to methylprednisolone boluses. Available data support the use of thymoglobulin induction to delay calcineurin inhibitor use and its attendant nephrotoxicity. IV infusions of thymoglobulin may be complicated by fever and chills, which can be ameliorated by premedication with antipyretics and a low dose of glucocorticoids. Infusions of OKT3 may be complicated by fever, chills, and diarrhea, or by pulmonary edema, which can be fatal. Because OKT3 is such a potent immunosuppressive agent, its use is also more likely to be complicated by opportunistic infection or lymphoproliferative disorders; therefore, because of the availability of alternative immunosuppressive drugs, OKT3 is now used sparingly.

Sirolimus, an inhibitor of the mammalian target of rapamycin (mTOR), blocks later events in T cell activation, is approved for use in kidney transplantation, but is not formally approved for use in liver transplant recipients because of the reported association with an increased frequency of hepatic artery thrombosis in the first month posttransplantation. In patients with calcineurin inhibitor–related nephrotoxicity, conversion to sirolimus has been demonstrated to be effective in preventing rejection with accompanying improvements in renal function. Because of its profound antiproliferative effects, sirolimus has also been suggested to be a useful immunosuppressive agent in patients with a prior or current history of malignancy, such as HCC. Side effects include hyperlipidemia, peripheral edema, oral ulcers, and interstitial pneumonitis. Everolimus is a hydroxyethyl derivative of sirolimus that, when used in conjunction with low-dose tacrolimus, also provides successful protection against acute rejection, with decreased renal impairment compared to that associated with standard tacrolimus dosing. Everolimus and sirolimus share a similar adverse events profile. The most important principle of immunosuppression is that the ideal approach strikes a balance between immunosuppression and immunologic competence. In general, given sufficient immunosuppression, acute liver allograft rejection is nearly always reversible. On one hand, incompletely treated acute rejection predisposes to the development of chronic rejection, which can threaten graft survival. On the other hand, if the cumulative dose of immunosuppressive therapy is too large, the patient may succumb to opportunistic infection. In hepatitis C, pulse glucocorticoids or OKT3 use accelerate recurrent allograft hepatitis, although the routine use of DAA therapy to clear the allograft of HCV should remove or greatly diminish this concern. Further complicating matters, acute rejection can be difficult to distinguish histologically from recurrent hepatitis C. Therefore, immunosuppressive drugs must be used judiciously, with strict attention to the infectious consequences of such therapy and careful confirmation of the diagnosis of acute rejection. In this vein, efforts have been made to minimize the use of glucocorticoids, a mainstay of immunosuppressive regimens, and steroid-free immunosuppression can be achieved in some instances. Patients who undergo liver transplantation for autoimmune diseases such as primary biliary cirrhosis, autoimmune hepatitis, and primary sclerosing cholangitis are less likely to achieve freedom from glucocorticoids.

POSTOPERATIVE COMPLICATIONS

Complications of liver transplantation can be divided into nonhepatic and hepatic categories **(Tables 338-4 and 338-5)**. In addition, both immediate postoperative and late complications are encountered. As a rule, patients who undergo liver transplantation have been chronically ill for protracted periods and may be malnourished and wasted. The impact of such chronic illness and the multisystem failure that accompanies liver failure continue to require attention in the postoperative period. Because of the massive fluid losses and fluid shifts that occur during the operation, patients may remain fluid-overloaded during the immediate postoperative period, straining cardiovascular reserve; this effect can be amplified in the face of transient renal dysfunction and pulmonary capillary vascular permeability. Continuous monitoring of cardiovascular and pulmonary function, measures to maintain the

TABLE 338-4 Nonhepatic Complications of Liver Transplantation

CATEGORY	COMPLICATION
Cardiovascular instability	Arrhythmias
	Congestive heart failure
	Cardiomyopathy
Pulmonary compromise	Pneumonia
	Pulmonary capillary vascular permeability
	Fluid overload
Renal dysfunction	Prerenal azotemia
	Hypoperfusion injury (acute tubular necrosis)
	Drug nephrotoxicity
	↓ Renal blood flow secondary to ↑ intraabdominal pressure
Hematologic	Anemia secondary to gastrointestinal and/or intraabdominal bleeding
	Hemolytic anemia, aplastic anemia
	Thrombocytopenia
Infection	Bacterial: early, common postoperative infections
	Fungal/parasitic: late, opportunistic infections
	Viral: late, opportunistic infections, recurrent hepatitis
Neuropsychiatric	Seizures
	Metabolic encephalopathy
	Depression
	Difficult psychosocial adjustment
Diseases of donor	Infectious
	Malignant
Malignancy	B cell lymphoma (posttransplantation lymphoproliferative disorders)
	De novo neoplasms (particularly squamous cell skin carcinoma)

integrity of the intravascular compartment and to treat extravascular volume overload, and scrupulous attention to potential sources and sites of infection are of paramount importance. Cardiovascular instability may also result from the electrolyte imbalance that may accompany reperfusion of the donor liver as well as from restoration of systemic vascular resistance following implantation. Pulmonary function may be compromised further by paralysis of the right hemidiaphragm associated with phrenic nerve injury. The hyperdynamic state with increased cardiac output that is characteristic of patients with liver failure reverses rapidly after successful liver transplantation.

TABLE 338-5 Hepatic Complications of Liver Transplantation

Hepatic Dysfunction Common after Major Surgery	
Prehepatic	Pigment load
	Hemolysis
	Blood collections (hematomas, abdominal collections)
Intrahepatic	
Early	Hepatotoxic drugs and anesthesia
	Hypoperfusion (hypotension, shock, sepsis)
	Benign postoperative cholestasis
Late	Transfusion-associated hepatitis
	Exacerbation of primary hepatic disease
Posthepatic	Biliary obstruction
	↓ Renal clearance of conjugated bilirubin (renal dysfunction)
Hepatic Dysfunction Unique to Liver Transplantation	
Primary graft nonfunction	
Vascular compromise	Portal vein obstruction
	Hepatic artery thrombosis
	Anastomotic leak with intraabdominal bleeding
Bile duct disorder	Stenosis, obstruction, leak
Rejection	
Recurrent primary hepatic disease	

Other immediate management issues include renal dysfunction. Prerenal azotemia, acute kidney injury associated with hypoperfusion (acute tubular necrosis), and renal toxicity caused by antibiotics, tacrolimus, or cyclosporine are encountered frequently in the postoperative period, sometimes necessitating dialysis. Hemolytic-uremic syndrome can be associated with cyclosporine, tacrolimus, or OKT3. Occasionally, postoperative intraperitoneal bleeding may be sufficient to increase intraabdominal pressure, which, in turn, may reduce renal blood flow; this effect is rapidly reversible when abdominal distention is relieved by exploratory laparotomy to identify and ligate the bleeding site and to remove intraperitoneal clot.

Anemia may also result from acute upper gastrointestinal bleeding or from transient hemolytic anemia, which may be autoimmune, especially when blood group O livers are transplanted into blood group A or B recipients. This autoimmune hemolytic anemia is mediated by donor intrahepatic lymphocytes that recognize red blood cell A or B antigens on recipient erythrocytes. Transient in nature, this process resolves once the donor liver is repopulated by recipient bone marrow–derived lymphocytes; the hemolysis can be treated by transfusing blood group O red blood cells and/or by administering higher doses of glucocorticoids. Transient thrombocytopenia is also commonly encountered. Aplastic anemia, a late occurrence, is rare but has been reported in almost 30% of patients who underwent liver transplantation for acute, severe hepatitis of unknown cause.

Bacterial, fungal, or viral infections are common and may be life-threatening postoperatively. Early after transplant surgery, common postoperative infections predominate—pneumonia, wound infections, infected intraabdominal collections, urinary tract infections, and IV line infections—rather than opportunistic infections; these infections may involve the biliary tree and liver as well. Beyond the first postoperative month, the toll of immunosuppression becomes evident, and opportunistic infections—CMV, herpes viruses, fungal infections (*Aspergillus*, *Candida*, cryptococcal disease), mycobacterial infections, parasitic infections (*Pneumocystis*, *Toxoplasma*), bacterial infections (*Nocardia*, *Legionella*, *Listeria*)—predominate. Rarely, early infections represent those transmitted with the donor liver, either infections present in the donor or infections acquired during procurement processing. De novo viral hepatitis infections acquired from the donor organ or, almost unheard of now, from transfused blood products occur after typical incubation periods for these agents (well beyond the first month). Obviously, infections in an immunosuppressed host demand early recognition and prompt management; prophylactic antibiotic therapy is administered routinely in the immediate postoperative period. Use of sulfamethoxazole with trimethoprim reduces the incidence of postoperative *Pneumocystis carinii* pneumonia. Antiviral prophylaxis for CMV with ganciclovir should be administered in patients at high risk (e.g., when a CMV-seropositive donor organ is implanted into a CMV-seronegative recipient).

Neuropsychiatric complications include seizures (commonly associated with cyclosporine and tacrolimus toxicity), metabolic encephalopathy, depression, and difficult psychosocial adjustment. Rarely, diseases are transmitted by the allograft from the donor to the recipient. In addition to viral and bacterial infections, malignancies of donor origin have occurred. Posttransplantation lymphoproliferative disorders, especially B cell lymphoma, are a recognized complication associated with immunosuppressive drugs such as azathioprine, tacrolimus, and cyclosporine (see above). Epstein-Barr virus has been shown to play a contributory role in some of these tumors, which may regress when immunosuppressive therapy is reduced. De novo neoplasms appear at increased frequency after liver transplantation, particularly squamous cell carcinomas of the skin. Routine screening should be performed.

Long-term complications after liver transplantation attributable primarily to immunosuppressive medications include diabetes mellitus and osteoporosis (associated with glucocorticoids and calcineurin inhibitors) as well as hypertension, hyperlipidemia, and chronic renal insufficiency (associated with cyclosporine and tacrolimus). Monitoring and treating these disorders are routine components of posttransplantation care; in some cases, they respond to changes in immunosuppressive regimen, while in others, specific treatment of the disorder is introduced. Data from a large U.S. database showed that the prevalence of renal failure was 18% at year 5 and 25% at year 10 after liver transplantation. Similarly, the high frequency of diabetes, hypertension, hyperlipidemia, obesity, and the metabolic syndrome renders patients susceptible to cardiovascular disease after liver transplantation; although hepatic complications account for most of the mortality after liver transplantation, renal failure and cardiovascular disease are the other leading causes of late mortality after liver transplantation.

HEPATIC COMPLICATIONS

Hepatic dysfunction after liver transplantation is similar to the hepatic complications encountered after major abdominal and cardiothoracic surgery; however, in addition, hepatic complications include primary graft failure, vascular compromise, failure or stricture of the biliary anastomoses, and rejection. As in nontransplantation surgery, postoperative jaundice may result from prehepatic, intrahepatic, and posthepatic sources. *Prehepatic* sources represent the massive hemoglobin pigment load from transfusions, hemolysis, hematomas, ecchymoses, and other collections of blood. *Early intrahepatic* liver injury includes effects of hepatotoxic drugs and anesthesia; hypoperfusion injury associated with hypotension, sepsis, and shock; and benign postoperative cholestasis. *Late intrahepatic* sources of liver injury include exacerbation of primary disease. *Posthepatic* sources of hepatic dysfunction include biliary obstruction and reduced renal clearance of conjugated bilirubin. Hepatic complications unique to liver transplantation include primary graft failure associated with ischemic injury to the organ during harvesting; vascular compromise associated with thrombosis or stenosis of the portal vein or hepatic artery anastomoses; vascular anastomotic leak; stenosis, obstruction, or leakage of the anastomosed common bile duct; recurrence of primary hepatic disorder (see below); and rejection.

TRANSPLANT REJECTION

Despite the use of immunosuppressive drugs, rejection of the transplanted liver still occurs in a proportion of patients, beginning 1–2 weeks after surgery. Clinical signs suggesting rejection are fever, right upper quadrant pain, and reduced bile pigment and volume. Leukocytosis may occur, but the most reliable indicators are increases in serum bilirubin and aminotransferase levels. Because these tests lack specificity, distinguishing among rejection, biliary obstruction, primary graft nonfunction, vascular compromise, viral hepatitis, CMV infection, drug hepatotoxicity, and recurrent primary disease may be difficult. Radiographic visualization of the biliary tree and/or percutaneous liver biopsy often help to establish the correct diagnosis. Morphologic features of acute rejection include a mixed portal cellular infiltrate, bile duct injury, and/or endothelial inflammation ("endothelialitis"); some of these findings are reminiscent of graft-versus-host disease, primary biliary cirrhosis, or recurrent allograft hepatitis C. As soon as transplant rejection is suspected, treatment consists of IV methylprednisolone in repeated boluses; if this fails to abort rejection, many centers use thymoglobulin or OKT3. Caution should be exercised when managing acute rejection with pulse glucocorticoids or OKT3 in patients with HCV infection, because of the high risk of triggering recurrent allograft hepatitis C. The availability of DAAs for HCV should greatly alleviate this concern.

Chronic rejection is a relatively rare outcome that can follow repeated bouts of acute rejection or that occurs unrelated to preceding rejection episodes. Morphologically, chronic rejection is characterized by progressive cholestasis, focal parenchymal necrosis, mononuclear infiltration, vascular lesions (intimal fibrosis, subintimal foam cells, fibrinoid necrosis), and fibrosis. This process may be reflected as ductopenia—the vanishing bile duct syndrome, which is more common in patients undergoing liver transplantation for autoimmune liver disease. Reversibility of chronic rejection is limited; in patients with therapy-resistant chronic rejection, retransplantation has yielded encouraging results.

OUTCOME

SURVIVAL

The survival rate for patients undergoing liver transplantation has improved steadily since 1983. One-year survival rates have increased from ~70% in the early 1980s to 85–90% from 2003 to the present time. Currently, the 5-year survival rate exceeds 60%. An important observation is the relationship between clinical status before transplantation and outcome. For patients who undergo liver transplantation when their level of compensation is high (e.g., still working or only partially disabled), a 1-year survival rate of >85% is common. For those whose level of decompensation mandates continuous in-hospital care prior to transplantation, the 1-year survival rate is ~70%, whereas for those who are so decompensated that they require life support in an intensive care unit, the 1-year survival rate is ~50%. Since the adoption by UNOS in 2002 of the MELD system for organ allocation, posttransplantation survival has been found to be affected adversely for candidates with MELD scores >25, considered high disease severity. Thus, irrespective of allocation scheme, high disease severity before transplantation corresponds to diminished posttransplantation survival. Another important distinction in survival has been drawn between high- and low-risk patient categories. For patients who do not fit any "high-risk" designations, 1-year and 5-year survival rates of 85 and 80%, respectively, have been recorded. In contrast, among patients in high-risk categories—cancer, fulminant hepatitis, age >65, concurrent renal failure, respirator dependence, portal vein thrombosis, and history of a portacaval shunt or multiple right upper quadrant operations—survival statistics fall into the range of 60% at 1 year and 35% at 5 years. Survival after retransplantation for primary graft nonfunction is ~50%. Causes of failure of liver transplantation vary with time. Failures within the first 3 months result primarily from technical complications, postoperative infections, and hemorrhage. Transplant failures after the first 3 months are more likely to result from infection, rejection, or recurrent disease (such as malignancy or viral hepatitis).

RECURRENCE OF PRIMARY DISEASE

Features of autoimmune hepatitis, primary sclerosing cholangitis, and primary biliary cirrhosis overlap with those of rejection or posttransplantation bile duct injury. Whether autoimmune hepatitis and sclerosing cholangitis recur after liver transplantation is controversial; data supporting recurrent autoimmune hepatitis (in up to one-third of patients in some series) are more convincing than those supporting recurrent sclerosing cholangitis. Similarly, reports of recurrent primary biliary cirrhosis after liver transplantation have appeared; however, the histologic features of primary biliary cirrhosis and chronic rejection are virtually indistinguishable and occur as frequently in patients with primary biliary cirrhosis as in patients undergoing transplantation for other reasons. The presence of a florid inflammatory bile duct lesion is highly suggestive of the recurrence of primary biliary cirrhosis, but even this lesion can be observed in acute rejection. Hereditary disorders such as Wilson's disease and α_1-antitrypsin deficiency have not recurred after liver transplantation; however, recurrence of disordered iron metabolism has been observed in some patients with hemochromatosis. Hepatic vein thrombosis (Budd-Chiari syndrome) may recur; this can be minimized by treating underlying myeloproliferative disorders and by anticoagulation. Because cholangiocarcinoma recurs almost invariably, few centers now offer transplantation to such patients; however, a few highly selected patients with operatively confirmed stage I or II cholangiocarcinoma who undergo liver

transplantation combined with neoadjuvant chemoradiation may experience excellent outcomes. In patients with intrahepatic HCC who meet criteria for transplantation, 1- and 5-year survivals are similar to those observed in patients undergoing liver transplantation for nonmalignant disease. Finally, metabolic disorders such as nonalcoholic steatohepatitis recur frequently, especially if the underlying metabolic predisposition is not altered. The metabolic syndrome occurs commonly after liver transplantation as a result of recurrent nonalcoholic fatty liver, immunosuppressive medications, and/or, in patients with hepatitis C related to the impact of HCV infection on insulin resistance, diabetes and fatty liver.

Hepatitis A can recur after transplantation for fulminant hepatitis A, but such acute reinfection has no serious clinical sequelae. In fulminant hepatitis B, recurrence is not the rule; however, in the absence of any prophylactic measures, hepatitis B usually recurs after transplantation for end-stage chronic hepatitis B. Before the introduction of prophylactic antiviral therapy, immunosuppressive therapy sufficient to prevent allograft rejection led inevitably to marked increases in hepatitis B viremia, regardless of pretransplantation levels. Overall graft and patient survival were poor, and some patients experienced a rapid recapitulation of severe injury—severe chronic hepatitis or even fulminant hepatitis—after transplantation. Also recognized in the era before availability of antiviral regimens was *fibrosing cholestatic hepatitis*, rapidly progressive liver injury associated with marked hyperbilirubinemia, substantial prolongation of the prothrombin time (both out of proportion to relatively modest elevations of aminotransferase activity), and rapidly progressive liver failure. This lesion has been suggested to represent a "choking off" of the hepatocyte by an overwhelming density of HBV proteins. Complications such as sepsis and pancreatitis were also observed more frequently in patients undergoing liver transplantation for hepatitis B prior to the introduction of antiviral therapy. The introduction of long-term prophylaxis with HBIg revolutionized liver transplantation for chronic hepatitis B. Preoperative hepatitis B vaccination, preoperative or postoperative interferon (IFN) therapy, or short-term (≤2 months) HBIg prophylaxis has not been shown to be effective, but a retrospective analysis of data from several hundred European patients followed for 3 years after transplantation has shown that long-term (≥6 months) prophylaxis with HBIg is associated with a lowering of the risk of HBV reinfection from ~75 to 35% and a reduction in mortality from ~50 to 20%.

As a result of long-term HBIg use following liver transplantation for chronic hepatitis B, similar improvements in outcome have been observed in the United States, with 1-year survival rates between 75 and 90%. Currently, with HBIg prophylaxis, the outcome of liver transplantation for chronic hepatitis B is indistinguishable from that for chronic liver disease unassociated with chronic hepatitis B; essentially, medical concerns regarding liver transplantation for chronic hepatitis B have been eliminated. Passive immunoprophylaxis with HBIg is begun during the anhepatic stage of surgery, repeated daily for the first 6 postoperative days, and then continued with infusions that are given either at regular intervals of 4–6 weeks or, alternatively, when anti-hepatitis B surface (HBs) levels fall below a threshold of 100 mIU/mL. The current approach in most centers is to continue HBIg indefinitely, which can add ~$20,000 per year to the cost of care; some centers are evaluating regimens that shift to less frequent administration or to IM administration in the late posttransplantation period or, in low-risk patients, maintenance with antiviral therapy (see below) alone. Still, "breakthrough" HBV infection occasionally occurs.

Further improving the outcome of liver transplantation for chronic hepatitis B is the current availability of such antiviral drugs as entecavir and tenofovir disoproxil fumarate **(Chap. 334)**. When these drugs are administered to patients with decompensated liver disease, a proportion improves sufficiently to postpone imminent liver transplantation. In addition, antiviral therapy can be used to prevent recurrence of HBV infection when administered *prior to* transplantation; to treat hepatitis B that recurs *after* transplantation, including in patients who break through HBIg prophylaxis; and to reverse the course of otherwise fatal fibrosing cholestatic hepatitis. Clinical trials have shown that entecavir or tenofovir antiviral therapy reduces the level of HBV replication substantially, sometimes even resulting in clearance of hepatitis B surface antigen (HBsAg); reduces alanine aminotransferase (ALT) levels; and improves histologic features of necrosis and inflammation. Currently, most liver transplantation centers combine HBIg plus one of the low-resistance oral nucleoside (entecavir) or nucleotide analogues (tenofovir). In low-risk patients with no detectable hepatitis B viremia at the time of transplantation, a number of clinical trials have suggested that antiviral prophylaxis can suffice, without HBIg or with a finite duration of HBIg, to prevent recurrent HBV infection of the allograft. In patients documented at the time of liver transplantation to have undetectable HBV DNA in serum and cccDNA in the liver (i.e., with low risk for recurrence of HBV infection), a preliminary clinical trial suggested that, after receipt of 5 years of combined therapy, both HBIg and oral-agent therapy can be withdrawn sequentially (over two 6-month periods) with a success rate, as monitored over a median of 6 years postwithdrawal, of 90% and an anti-HBs seroconversion rate of 60% (despite transient reappearance of HBV DNA and/or HBsAg in some of these patients).

Antiviral prophylactic approaches applied to patients undergoing liver transplantation for chronic hepatitis B are being used as well for patients without hepatitis B who receive organs from donors with antibody to hepatitis B core antigen (anti-HBc) but do not have HBsAg. Patients who undergo liver transplantation for chronic hepatitis B plus D are less likely to experience recurrent liver injury than patients undergoing liver transplantation for hepatitis B alone; still, such coinfected patients would also be offered standard posttransplantation prophylactic therapy for hepatitis B.

Until recently, the most common indication for liver transplantation was end-stage liver disease resulting from chronic hepatitis C. For patients undergoing liver transplantation for hepatitis C, because of an aggressive natural history of recurrent allograft hepatitis C, graft and patient survival were diminished substantially compared to other indications for transplantation.

The recent approval of several new DAA agents and of IFN-free DAA regimens against HCV has already had a major impact on the management and outcome of both pretransplantation and posttransplantation HCV infection. Such therapeutic approaches (1) permit the clearance of viremia in a substantial proportion of decompensated cirrhotics, thereby preventing recurrent allograft infection and even improving the clinical status of most of these patients, delaying or obviating the need for liver replacement; and (2) achieve sustained virologic responses in a much higher proportion of persons with allograft HCV infection, because of improvements in antiviral treatment efficacy and tolerability. Ideally, patients should be treated prior to liver transplantation. A concern, however, is that eradication of HCV infection will reduce the MELD score and lower the priority for a donor organ in some patients who still require transplantation because of continued hepatic decompensation and profound reduction in quality of life. In addition, elimination of HCV infection prior to transplantation would disqualify such patients from accepting donor livers from persons with HCV infection, contracting the potential donor pool and limiting accessibility to donor organs and timely transplantation. Therefore, consideration should be given to postponing DAA therapy in patients with high-MELD HCV-associated end-stage liver disease until after liver transplantation; however, a distinct threshold at which to treat pretransplantation or posttransplantation has not yet been established. Regardless, the approach to treatment should be individualized thoughtfully for each patient, based on such factors as MELD score, time anticipated prior to availability of a donor organ, relative clinical stability, and co-morbidities.

Recent DAA combinations that have been used successfully against allograft HCV include ledipasvir + sofosbuvir + ribavirin; velpatasvir + sofosbuvir + ribavirin, and grazoprevir/pibrentasvir. (For updated guidelines, see *www.hcvguidelines.org*). In patients with recurrent HCV infection after liver transplantation, each of these regimens has yielded response rates approaching those seen in compensated non-transplant patient populations.

A small number of allograft recipients have historically succumbed to early HCV-associated liver injury, and a syndrome reminiscent of

fibrosing cholestatic hepatitis (see above) has been observed rarely. Currently, however, the routine use of DAA regimens early after transplantation, before the onset of these variant presentations, should have a profound impact on the frequency of severe recurrent allograft hepatitis C.

Patients who undergo liver transplantation for end-stage alcoholic cirrhosis are at risk of resorting to drinking again after transplantation, a potential source of recurrent alcoholic liver injury. Currently, alcoholic liver disease is one of the more common indications for liver transplantation, accounting for 20–25% of all liver transplantation procedures, and most transplantation centers screen candidates carefully for predictors of continued abstinence. Recidivism is more likely in patients whose sobriety prior to transplantation was <6 months. For abstinent patients with alcoholic cirrhosis, liver transplantation can be undertaken successfully, with outcomes comparable to those for other categories of patients with chronic liver disease, when coordinated by a team approach that includes substance abuse counseling.

POSTTRANSPLANTATION QUALITY OF LIFE

Full rehabilitation is achieved in the majority of patients who survive the early postoperative months and escape chronic rejection or unmanageable infection. Psychosocial maladjustment interferes with medical compliance in a small number of patients, but most manage to adhere to immunosuppressive regimens, which must be continued indefinitely. In one study, 85% of patients who survived their transplant operations returned to gainful activities. In fact, some women have conceived and carried pregnancies to term after transplantation without demonstrable injury to their infants.

FURTHER READING

AASLD/IDSA HCV Guidance Panel: Hepatitis C guidance: AASLD-IDSA recommendations for testing, managing, and treating adults infected with hepatitis C virus. Hepatology 62:932, 2015. (Updated regularly, available at *http://www.hcvguidelines.org*). Accessed Mar 28, 2018.

European Association for the Study of the Liver: EASL clinical practice guidelines: Liver transplantation. J Hepatol 64:433, 2016.

Fung J et al: Outcomes including liver histology after liver transplantation for chronic hepatitis B using oral antiviral therapy alone. Liver Transpl 21:1504, 2015.

Gane EJ et al: Combination of lamivudine and adefovir without hepatitis B immune globulin is safe and effective prophylaxis against hepatitis B virus recurrence in hepatitis B surface antigen-positive liver transplant candidates. Liver Transpl 19:268, 2013.

Kwo PY et al: An interferon-free antiviral regimen for HCV after liver transplantation. N Engl J Med 371:2375, 2014.

Lenci I et al: Complete hepatitis B virus prophylaxis withdrawal in hepatitis B surface antigen-positive liver transplant recipients after longterm minimal immunosuppression. Liver Transpl 22:1205, 2016.

Lucey MR et al: Long-term management of the successful adult liver transplant: 2012 practice guideline by the American Association for the Study of Liver Diseases and the American Society of Transplantation. Liver Transpl 19:3, 2013.

Manns M et al: Ledipasvir and sofosbuvir plus ribavirin in patients with genotype 1 or 4 hepatitis C virus infection and advanced liver disease: A multicentre, open-label, randomised, phase 2 trial. Lancet Infectious Diseases 16:685, 2016.

Martin P et al: Evaluation for liver transplantation in adults: 2013 practice guideline by the American Association for the Study of Liver Diseases and the American Society of Transplantation. Hepatology 59:1145, 2014.

Poordad F et al: Daclatasvir with sofosbuvir and ribavirin for hepatitis C virus infection with advanced cirrhosis or post-liver transplantation recurrence. Hepatology 63:1493, 2016.

339 Diseases of the Gallbladder and Bile Ducts

Norton J. Greenberger, Gustav Paumgartner

PHYSIOLOGY OF BILE PRODUCTION AND FLOW

BILE SECRETION AND COMPOSITION

Bile formed in the hepatic lobules is secreted into a complex network of canaliculi, small bile ductules, and larger bile ducts that run with lymphatics and branches of the portal vein and hepatic artery in portal tracts situated between hepatic lobules. These interlobular bile ducts coalesce to form larger septal bile ducts that join to form the right and left hepatic ducts, which in turn, unite to form the common hepatic duct. The common hepatic duct is joined by the cystic duct of the gallbladder to form the common bile duct (CBD), which enters the duodenum (often after joining the main pancreatic duct) through the ampulla of Vater.

Hepatic bile is an isotonic fluid with an electrolyte composition resembling blood plasma. The electrolyte composition of gallbladder bile differs from that of hepatic bile because most of the inorganic anions, chloride, and bicarbonate have been removed by reabsorption across the gallbladder epithelium. As a result of water reabsorption, total solute concentration of bile increases from 3–4 g/dL in hepatic bile to 10–15 g/dL in gallbladder bile.

Major solute components of bile by moles percent include bile acids (80%), lecithin and traces of other phospholipids (16%), and unesterified cholesterol (4.0%). In the lithogenic state, the cholesterol value can be as high as 8–10%. Other constituents include conjugated bilirubin; proteins (all immunoglobulins, albumin, metabolites of hormones, and other proteins metabolized in the liver); electrolytes; mucus; and, often, drugs and their metabolites.

The total daily basal secretion of hepatic bile is ~500–600 mL. Many substances taken up or synthesized by the hepatocyte are secreted into the bile canaliculi. The canalicular membrane forms microvilli and is associated with microfilaments of actin, microtubules, and other contractile elements. Prior to their secretion into the bile, many substances are taken up into the hepatocyte, while others, such as phospholipids, a portion of primary bile acids, and some cholesterol, are synthesized de novo in the hepatocyte. Three mechanisms are important in regulating bile flow: (1) active transport of bile acids from hepatocytes into the bile canaliculi, (2) active transport of other organic anions, and (3) cholangiocellular secretion. The last is a secretin-mediated and cyclic AMP–dependent mechanism that results in the secretion of a sodium- and bicarbonate-rich fluid into the bile ducts.

Active vectorial secretion of biliary constituents from the portal blood into the bile canaliculi is driven by a set of polarized transport systems at the basolateral (sinusoidal) and the canalicular apical plasma membrane domains of the hepatocyte. Two sinusoidal bile salt uptake systems have been cloned in humans, the Na^+/taurocholate cotransporter (NTCP, SLC10A1) and the organic anion–transporting proteins (OATPs), which also transport a large variety of non-bile salt organic anions. Several ATP-dependent canalicular transport systems, "export pumps," (ATP-binding cassette transport proteins, also known as ABC transporters) have been identified, the most important of which are: the bile salt export pump (BSEP, ABCB11); the anionic conjugate export pump (MRP2, ABCC2), which mediates the canalicular excretion of various amphiphilic conjugates formed by phase II conjugation (e.g., bilirubin mono- and diglucuronides and drugs); the multidrug export pump (MDR1, ABCB1) for hydrophobic cationic compounds; and the phospholipid export pump (MDR3, ABCB4). Two hemitransporters ABCG5/G8, functioning as a couple, constitute the canalicular cholesterol and phytosterol transporter. F1C1 (ATP8B1) is an aminophospholipid transferase ("flippase") essential for maintaining the lipid asymmetry of the canalicular membrane. The canalicular membrane

also contains ATP-independent transport systems such as the Cl/HCO_3 anion exchanger isoform 2 (AE2, SLC4A2) for canalicular bicarbonate secretion. For most of these transporters, genetic defects have been identified that are associated with various forms of cholestasis or defects of biliary excretion. F1C1 is defective in progressive familial intrahepatic cholestasis type 1 (PFIC1) and benign recurrent intrahepatic cholestasis type 1 (BRIC1) and results in ablation of all other ATP-dependent transporter functions. BSEP is defective in PFIC2 and BRIC2. Mutations of MRP2 (ABCC2) cause the Dubin-Johnson syndrome, an inherited form of conjugated hyperbilirubinemia **(Chap. 331)**. A defective MDR3 (ABCB4) results in PFIC3. ABCG5/G8, the canalicular half transporters for cholesterol and other neutral sterols, are defective in sitosterolemia. The cystic fibrosis transmembrane regulator (CFTR, ABCC7) located on bile duct epithelial cells but not on canalicular membranes is defective in cystic fibrosis, which is associated with impaired cholangiocellular pH regulation during ductular bile formation and chronic cholestatic liver disease, occasionally resulting in biliary cirrhosis.

THE BILE ACIDS

The primary bile acids, cholic acid and chenodeoxycholic acid (CDCA), are synthesized from cholesterol in the liver, conjugated with glycine or taurine, and secreted into the bile. Secondary bile acids, including deoxycholate and lithocholate, are formed in the colon as bacterial metabolites of the primary bile acids. However, lithocholic acid is much less efficiently absorbed from the colon than deoxycholic acid. Another secondary bile acid, found in low concentration, is ursodeoxycholic acid (UDCA), a stereoisomer of CDCA. In healthy subjects, the ratio of glycine to taurine conjugates in bile is ~3:1.

Bile acids are detergent-like molecules that in aqueous solutions and above a critical concentration of about 2 m*M* form molecular aggregates called *micelles*. Cholesterol alone is sparingly soluble in aqueous environments, and its solubility in bile depends on both the total lipid concentration and the relative molar percentages of bile acids and lecithin. Normal ratios of these constituents favor the formation of solubilizing *mixed micelles*, while abnormal ratios promote the precipitation of cholesterol crystals in bile via an intermediate liquid crystal phase.

In addition to facilitating the biliary excretion of cholesterol, bile acids facilitate the normal intestinal absorption of dietary fats, mainly cholesterol and fat-soluble vitamins, via a micellar transport mechanism **(Chap. 318)**. Bile acids also serve as a major physiologic driving force for hepatic bile flow and aid in water and electrolyte transport in the small bowel and colon.

ENTEROHEPATIC CIRCULATION

Bile acids are efficiently conserved under normal conditions. Unconjugated, and to a lesser degree also conjugated, bile acids are absorbed by *passive diffusion* along the entire gut. Quantitatively much more important for bile salt recirculation, however, is the *active transport* mechanism for conjugated bile acids in the distal ileum **(Chap. 318)**. The reabsorbed bile acids enter the portal bloodstream and are taken up rapidly by hepatocytes, reconjugated, and resecreted into bile (enterohepatic circulation).

The normal bile acid pool size is ~2–4 g. During digestion of a meal, the bile acid pool undergoes at least one or more enterohepatic cycles, depending on the size and composition of the meal. Normally, the bile acid pool circulates ~5–10 times daily. Intestinal reabsorption of the pool is about 95% efficient; therefore, fecal loss of bile acids is in the range of 0.2–0.4 g/d. In the steady state, this fecal loss is compensated by an equal daily synthesis of bile acids by the liver, and, thus, the size of the bile acid pool is maintained. Bile acids in the intestine release fibroblast growth factor 19 (FGF19) into the circulation, which is transported to the liver where it suppresses synthesis of bile acids from cholesterol by inhibiting the rate-limiting enzyme cytochrome P450 7A1 (CYP7A1) and also promotes gallbladder relaxation. While the loss of bile salts in stool is usually matched by increased hepatic synthesis, the maximum rate of synthesis is ~5 g/d, which may be insufficient to replete the bile acid pool size when there is pronounced impairment of intestinal bile salt reabsorption.

The expression of ABC transporters in the enterohepatic circulation and of the rate-limiting enzymes of bile acid and cholesterol synthesis are regulated in a coordinated fashion by nuclear receptors, which are ligand-activated transcription factors. The hepatic BSEP (ABCB11) is upregulated by the farnesoid X receptor (FXR), a bile acid sensor that also represses bile acid synthesis. The expression of the cholesterol transporter, ABCG5/G8, is upregulated by the liver X receptor (LXR), which is an oxysterol sensor.

GALLBLADDER AND SPHINCTERIC FUNCTIONS

In the fasting state, the sphincter of Oddi (SOD) offers a high-pressure zone of resistance to bile flow from the CBD into the duodenum. Its tonic contraction serves to (1) prevent reflux of duodenal contents into the pancreatic and bile ducts and (2) promote filling of the gallbladder. The major factor controlling the evacuation of the gallbladder is the peptide hormone cholecystokinin (CCK), which is released from the duodenal mucosa in response to the ingestion of fats and amino acids. CCK produces (1) powerful contraction of the gallbladder, (2) decreased resistance of the SOD, and (3) enhanced flow of biliary contents into the duodenum.

Hepatic bile is "concentrated" within the gallbladder by energy-dependent transmucosal absorption of water and electrolytes. Almost the entire bile acid pool may be sequestered in the gallbladder following an overnight fast for delivery into the duodenum with the first meal of the day. The normal capacity of the gallbladder is ~30 mL of bile.

DISEASES OF THE GALLBLADDER

CONGENITAL ANOMALIES

Anomalies of the biliary tract are not uncommon and include abnormalities in number, size, and shape (e.g., agenesis of the gallbladder, duplications, rudimentary or oversized "giant" gallbladders, and diverticula). *Phrygian cap* is a clinically innocuous entity in which a partial or complete septum (or fold) separates the fundus from the body. Anomalies of position or suspension are not uncommon and include left-sided gallbladder, intrahepatic gallbladder, retrodisplacement of the gallbladder, and "floating" gallbladder. The latter condition predisposes to acute torsion, volvulus, or herniation of the gallbladder.

GALLSTONES

Epidemiology and Pathogenesis Gallstones are quite prevalent in most Western countries. Gallstone formation increases after age 50. In the United States, the third National Health and Nutrition Examination Survey (NHANES III) has revealed an overall prevalence of gallstones of 7.9% in men and 16.6% in women. The prevalence was high in Mexican Americans (8.9% in men, 26.7% in women), intermediate for non-Hispanic whites (8.6% in men, 16.6% in women), and low for African Americans (5.3% in men, 13.9% in women).

Gallstones are formed because of abnormal bile composition. They are divided into two major types: cholesterol stones and pigment stones. Cholesterol stones account for >90% of all gallstones in Western industrialized countries. Cholesterol gallstones usually contain >50% cholesterol monohydrate plus an admixture of calcium salts, bile pigments, proteins, and fatty acids. Pigment stones are composed primarily of calcium bilirubinate; they contain <20% cholesterol and are classified into "black" and "brown" types, the latter forming secondary to chronic biliary infection.

CHOLESTEROL STONES AND BILIARY SLUDGE Cholesterol is essentially water-insoluble and requires aqueous dispersion into either micelles or vesicles, both of which require the presence of a second lipid to solubilize the cholesterol. Cholesterol and phospholipids are secreted into bile as unilamellar bilayered vesicles, which are converted into mixed micelles consisting of bile acids, phospholipids, and cholesterol by the action of bile acids. If there is an excess of cholesterol in relation to phospholipids and bile acids, unstable, cholesterol-rich vesicles remain, which aggregate into large multilamellar vesicles from which cholesterol crystals precipitate **(Fig. 339-1)**.

There are several important mechanisms in the formation of lithogenic (stone-forming) bile. The most important is increased biliary secretion of cholesterol. This may occur in association with

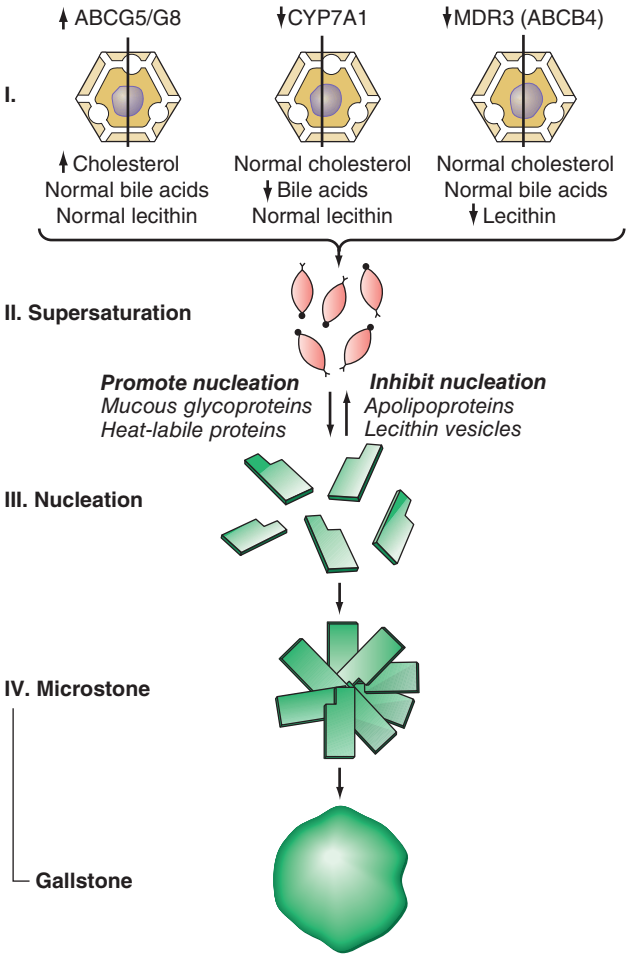


FIGURE 339-1 Scheme showing pathogenesis of cholesterol gallstone formation. Conditions or factors that increase the ratio of cholesterol to bile acids and phospholipids (lecithin) favor gallstone formation. ABCB4, ATP-binding cassette transporter; ABCG5/8, ATP-binding cassette (ABC) transporter G5/G8; CYP7A1, cytochrome P450 7A1; MDR3, multidrug resistance protein 3, also called phospholipid export pump.

obesity, the metabolic syndrome, high-caloric and cholesterol-rich diets, or drugs (e.g., clofibrate) and may result from increased activity of hydroxymethylglutaryl-coenzyme A (HMG-CoA) reductase, the rate-limiting enzyme of hepatic cholesterol synthesis, and increased hepatic uptake of cholesterol from blood. In patients with gallstones, dietary cholesterol *increases* biliary cholesterol secretion. This does not occur in non-gallstone patients on high-cholesterol diets. In addition to environmental factors such as high-caloric and cholesterol-rich diets, genetic factors play an important role in gallstone disease. A large study of symptomatic gallstones in Swedish twins provided strong evidence for a role of genetic factors in gallstone pathogenesis. Genetic factors accounted for 25%, shared environmental factors for 13%, and individual environmental factors for 62% of the phenotypic variation among monozygotic twins. A single nucleotide polymorphism of the gene encoding the hepatic cholesterol transporter ABCG5/G8 has been found in 21% of patients with gallstones, but only in 9% of the general population. It is thought to cause a gain of function of the cholesterol transporter and to contribute to cholesterol hypersecretion. A high prevalence of gallstones is found among first-degree relatives of gallstone carriers and in certain ethnic populations such as American Indians, Chilean Indians, and Chilean Hispanics. A common genetic trait has been identified for some of these populations by mitochondrial DNA analysis. In some patients, impaired hepatic conversion of cholesterol to bile acids may also occur, resulting in an increase of the lithogenic cholesterol/bile acid ratio. Although most cholesterol stones have a polygenic basis, there are rare monogenic (Mendelian) causes. Recently, a mutation in the *CYP7A1* gene has been described that results in a deficiency of the enzyme cholesterol 7-hydroxylase, which catalyzes the initial step in cholesterol catabolism and bile acid synthesis. The homozygous state is associated with hypercholesterolemia and gallstones. Because the phenotype is expressed in the heterozygote state, mutations in the *CYP7A1* gene may contribute to the susceptibility to cholesterol gallstone disease in the population. Mutations in the *MDR3* (ABCB4) gene, which encodes the phospholipid export pump in the canalicular membrane of the hepatocyte, may cause defective phospholipid secretion into bile, resulting in cholesterol supersaturation of bile and formation of cholesterol gallstones in the gallbladder and in the bile ducts. Thus, an excess of biliary cholesterol in relation to bile acids and phospholipids is primarily due to hypersecretion of cholesterol, but hyposecretion of bile acids or phospholipids may contribute. An additional disturbance of bile acid metabolism that is likely to contribute to supersaturation of bile with cholesterol is enhanced conversion of cholic acid to deoxycholic acid, with replacement of the cholic acid pool by an expanded deoxycholic acid pool. It may result from enhanced dehydroxylation of cholic acid and increased absorption of newly formed deoxycholic acid. An increased deoxycholate secretion is associated with hypersecretion of cholesterol into bile.

While supersaturation of bile with cholesterol is an important prerequisite for gallstone formation, it is generally not sufficient by itself to produce cholesterol precipitation in vivo. Most individuals with supersaturated bile do not develop stones because the time required for cholesterol crystals to nucleate and grow is longer than the time bile remains in the gallbladder.

An important mechanism is *nucleation* of cholesterol monohydrate crystals, which is greatly accelerated in human lithogenic bile. Accelerated nucleation of cholesterol monohydrate in bile may be due to either an *excess of pronucleating factors* or a *deficiency of antinucleating factors*. Mucin and certain nonmucin glycoproteins, principally immunoglobulins, appear to be pronucleating factors, while apolipoproteins A-I and A-II and other glycoproteins appear to be antinucleating factors. Pigment particles may possibly play a role as nucleating factors. In a genome-wide analysis of serum bilirubin levels, the uridine diphosphate-glucuronyltransferase 1A1 (*UGT1A1*) Gilbert's syndrome gene variant was associated with the presence of gallstone disease. Because most gallstones associated with the *UGT1A1* variant were cholesterol stones, this finding points to the role of pigment particles in the pathogenesis of gallbladder stones. Cholesterol monohydrate crystal nucleation and crystal growth probably occur within the mucin gel layer. Vesicle fusion leads to liquid crystals, which, in turn, nucleate into solid cholesterol monohydrate crystals. Continued growth of the crystals occurs by direct nucleation of cholesterol molecules from supersaturated unilamellar or multilamellar biliary vesicles.

A third important mechanism in cholesterol gallstone formation is *gallbladder hypomotility*. If the gallbladder emptied all supersaturated or crystal-containing bile completely, stones would not be able to grow. A high percentage of patients with gallstones exhibits abnormalities of gallbladder emptying. Ultrasonographic studies show that gallstone patients display an increased gallbladder volume during fasting and also after a test meal (residual volume) and that fractional emptying after gallbladder stimulation is decreased. The incidence of gallstones is increased in conditions associated with infrequent or impaired gallbladder emptying such as fasting, parenteral nutrition, or pregnancy and in patients using drugs that inhibit gallbladder motility.

Biliary sludge is a thick, mucous material that, upon microscopic examination, reveals lecithin-cholesterol liquid crystals, cholesterol monohydrate crystals, calcium bilirubinate, and mucin gels. Biliary sludge typically forms a crescent-like layer in the most dependent portion of the gallbladder and is recognized by characteristic echoes on ultrasonography (see below). The presence of biliary sludge implies two abnormalities: (1) the normal balance between gallbladder mucin secretion and elimination has become deranged, and (2) nucleation of biliary solutes has occurred. That biliary sludge may be a precursor form of gallstone disease is evident from several observations. In one study, 96 patients with gallbladder sludge were followed prospectively by serial ultrasound studies. In 18%, biliary sludge disappeared and did not recur for at least 2 years. In 60%, biliary sludge disappeared and reappeared; in 14%, gallstones (8% asymptomatic, 6% symptomatic)

developed; and in 6%, severe biliary pain with or without acute pancreatitis occurred. In 12 patients, cholecystectomies were performed, 6 for gallstone-associated biliary pain and 3 in symptomatic patients with sludge but without gallstones who had prior attacks of pancreatitis; the latter did not recur after cholecystectomy. It should be emphasized that biliary sludge can develop with disorders that cause gallbladder hypomotility; that is, surgery, burns, total parenteral nutrition, pregnancy, and oral contraceptives—all of which are associated with gallstone formation. However, the presence of biliary sludge implies supersaturation of bile with either cholesterol or calcium bilirubinate.

Two other conditions are associated with cholesterol-stone or biliary-sludge formation: pregnancy and rapid weight reduction through a very-low-calorie diet. There appear to be two key changes during pregnancy that contribute to a "cholelithogenic state": (1) a marked increase in cholesterol saturation of bile during the third trimester and (2) sluggish gallbladder contraction in response to a standard meal, resulting in impaired gallbladder emptying. That these changes are related to pregnancy per se is supported by several studies that show reversal of these abnormalities quite rapidly after delivery. During pregnancy, gallbladder sludge develops in 20–30% of women and gallstones in 5–12%. Although biliary sludge is a common finding during pregnancy, it is usually asymptomatic and often resolves spontaneously after delivery. Gallstones, which are less common than sludge and frequently associated with biliary colic, may also disappear after delivery because of spontaneous dissolution related to bile becoming unsaturated with cholesterol postpartum.

Approximately 10–20% of persons with rapid weight reduction achieved through very-low-calorie dieting develop gallstones. In a study involving 600 patients who completed a 3-month, 520-kcal/d diet, UDCA in a dosage of 600 mg/d proved highly effective in preventing gallstone formation; gallstones developed in only 3% of UDCA recipients, compared to 28% of placebo-treated patients. In obese patients treated by gastric banding, 500 mg/d of UDCA reduced the risk of gallstone formation from 30 to 8% within a follow-up of 6 months.

To summarize, cholesterol gallstone disease occurs because of several defects, which include (1) bile supersaturation with cholesterol, (2) nucleation of cholesterol monohydrate with subsequent crystal retention and stone growth, and (3) abnormal gallbladder motor function with delayed emptying and stasis. Other important factors known to predispose to cholesterol-stone formation are summarized in Table 339-1.

TABLE 339-1 Predisposing Factors for Cholesterol and Pigment Gallstone Formation

Cholesterol Stones

1. Demographic/genetic factors: Prevalence highest in North American Indians, Chilean Indians, and Chilean Hispanics, greater in Northern Europe and North America than in Asia, lowest in Japan; familial disposition; hereditary aspects
2. Obesity, metabolic syndrome: Normal bile acid pool and secretion but increased biliary secretion of cholesterol
3. Weight loss: Mobilization of tissue cholesterol leads to increased biliary cholesterol secretion while enterohepatic circulation of bile acids is decreased
4. Female sex hormones
   a. Estrogens stimulate hepatic lipoprotein receptors, increase uptake of dietary cholesterol, and increase biliary cholesterol secretion
   b. Natural estrogens, other estrogens, and oral contraceptives lead to decreased bile salt secretion and decreased conversion of cholesterol to cholesteryl esters
5. Pregnancy: Impaired gallbladder emptying caused by progesterone combined with the influence of estrogens, which increase biliary cholesterol secretion
6. Increasing age: Increased biliary secretion of cholesterol, decreased size of bile acid pool, decreased secretion of bile salts
7. Gallbladder hypomotility leading to stasis and formation of sludge
   a. Prolonged parenteral nutrition
   b. Fasting
   c. Pregnancy
   d. Drugs such as octreotide
8. Clofibrate therapy: Increased biliary secretion of cholesterol
9. Decreased bile acid secretion
   a. Primary biliary cirrhosis
   b. Genetic defect of the *CYP7A1* gene
10. Decreased phospholipid secretion: Genetic defect of the *MDR3* gene
11. Miscellaneous
    a. High-calorie, high-fat diet
    b. Spinal cord injury

Pigment Stones

1. Demographic/genetic factors: Asia, rural setting
2. Chronic hemolysis
3. Alcoholic liver cirrhosis
4. Pernicious anemia
5. Cystic fibrosis
6. Chronic biliary tract infection, parasite infections
7. Increasing age
8. Ileal disease, ileal resection or bypass

PIGMENT STONES Black pigment stones are composed of either pure calcium bilirubinate or polymer-like complexes with calcium and mucin glycoproteins. They are more common in patients who have chronic hemolytic states (with increased conjugated bilirubin in bile), liver cirrhosis, Gilbert's syndrome, or cystic fibrosis. Gallbladder stones in patients with ileal diseases, ileal resection, or ileal bypass generally are also black pigment stones. Enterohepatic recycling of bilirubin in ileal disease states contributes to their pathogenesis. Brown pigment stones are composed of calcium salts of unconjugated bilirubin with varying amounts of cholesterol and protein. They are caused by the presence of increased amounts of unconjugated, insoluble bilirubin in bile that precipitates to form stones. Deconjugation of an excess of soluble bilirubin mono- and diglucuronides may be mediated by endogenous β-glucuronidase but may also occur by spontaneous hydrolysis. Sometimes, the enzyme is also produced when bile is chronically infected by bacteria, and such stones are brown. Pigment stone formation is frequent in Asia and is often associated with infections in the gallbladder and biliary tree (Table 339-1).

Diagnosis

Procedures of potential use in the diagnosis of cholelithiasis and other diseases of the gallbladder are detailed in Table 339-2. Ultrasonography of the gallbladder is very accurate in the identification of cholelithiasis and has replaced oral cholecystography (OCG) (Fig. 339-2*A*). Stones as small as 1.5 mm in diameter may be confidently identified provided that firm criteria are used (e.g., acoustic "shadowing" of opacities that are within the gallbladder lumen and that change with the patient's position [by gravity]). In major medical centers, the false-negative and false-positive rates for ultrasound in gallstone patients are ~2–4%. Biliary sludge is material of low echogenic activity that typically forms a layer in the most dependent position of the gallbladder. This layer shifts with postural changes but fails to produce acoustic shadowing; these two characteristics distinguish sludges from gallstones. Ultrasound can also be used to assess the emptying function of the gallbladder.

The plain abdominal film may detect gallstones containing sufficient calcium to be radiopaque (10–15% of cholesterol and ~50% of pigment stones). Plain radiography may also be of use in the diagnosis of emphysematous cholecystitis, porcelain gallbladder, limey bile, and gallstone ileus.

OCG has historically been a useful procedure for the diagnosis of gallstones but has been replaced by ultrasound and is regarded as obsolete. It may be used to assess the patency of the cystic duct and gallbladder emptying function. Further, OCG can also delineate the size and number of gallstones and determine whether they are calcified.

Radiopharmaceuticals such as ^{99m}Tc-labeled *N*-substituted iminodiacetic acids (HIDA, DIDA, DISIDA, etc.) are rapidly extracted from the blood and are excreted into the biliary tree in high concentration even in the presence of mild to moderate serum bilirubin elevations. Failure

TABLE 339-2 Diagnostic Evaluation of the Gallbladder

DIAGNOSTIC ADVANTAGES	DIAGNOSTIC LIMITATIONS	COMMENT
Gallbladder Ultrasound		
Rapid Accurate identification of gallstones (>95%) Simultaneous scanning of GB, liver, bile ducts, pancreas "Real-time" scanning allows assessment of GB volume, contractility Not limited by jaundice, pregnancy May detect very small stones	Bowel gas Massive obesity Ascites	Procedure of choice for detection of stones
Plain Abdominal X-Ray		
Low cost Readily available	Relatively low yield Contraindicated in pregnancy	Pathognomonic findings in: calcified gallstones Limey bile, porcelain GB Emphysematous cholecystitis Gallstone ileus
Radioisotope Scans (HIDA, DIDA, etc.)		
Accurate identification of cystic duct obstruction Simultaneous assessment of bile ducts	Contraindicated in pregnancy Serum bilirubin >103–205 μmol/L (6–12 mg/dL) Cholecystogram of low resolution	Indicated for confirmation of suspected acute cholecystitis; less sensitive and less specific in chronic cholecystitis; useful indiagnosis of acalculous cholecystopathy, especially if given with CCK to assess gallbladder emptying

Abbreviations: CCK, cholecystokinin; GB, gallbladder.

to image the gallbladder in the presence of biliary ductal visualization may indicate cystic duct obstruction, acute or chronic cholecystitis, or surgical absence of the organ. Such scans have some application in the diagnosis of acute cholecystitis.

Symptoms of Gallstone Disease Gallstones usually produce symptoms by causing inflammation or obstruction following their migration into the cystic duct or CBD. The most specific and characteristic symptom of gallstone disease is biliary colic that is a constant and often long-lasting pain (see below). Obstruction of the cystic duct or CBD by a stone produces increased intraluminal pressure and distention of the viscus that cannot be relieved by repetitive biliary contractions. The resultant visceral pain is characteristically a severe, steady ache or fullness in the epigastrium or right upper quadrant (RUQ) of the abdomen with frequent radiation to the interscapular area, right scapula, or shoulder.

Biliary colic begins quite suddenly and may persist with severe intensity for 30 min to 5 h, subsiding gradually or rapidly. It is steady rather than intermittent, as would be suggested by the word *colic*, which must be regarded as a misnomer, although it is in widespread use. An episode of biliary pain persisting beyond 5 h should raise the suspicion of acute cholecystitis (see below). Nausea and vomiting frequently accompany episodes of biliary pain. An elevated level of serum bilirubin and/or alkaline phosphatase suggests a common duct stone. Fever or chills (rigors) with biliary pain usually imply a complication, that is, cholecystitis, pancreatitis, or cholangitis. Complaints of short-lasting, vague epigastric fullness, dyspepsia, eructation, or flatulence, especially following a fatty meal, should not be confused with biliary pain. Such symptoms are frequently elicited from patients with or without gallstone disease but are not specific for biliary calculi. Biliary colic may be precipitated by eating a fatty meal, by consumption of a large meal following a period of prolonged fasting, or by eating a normal meal; it is frequently nocturnal, occurring within a few hours of retiring.

Natural History Gallstone disease discovered in an asymptomatic patient or in a patient whose symptoms are not referable to cholelithiasis is a common clinical problem. Sixty to 80% of persons with asymptomatic gallstones remain asymptomatic over follow-up periods of up to 25 years. The probability of developing symptoms within 5 years after diagnosis is 2–4% per year and decreases in the years thereafter to 1–2%. The yearly incidence of complications is about 0.1–0.3%. Patients remaining asymptomatic for 15 years were found to be unlikely to develop symptoms during further follow-up, and most patients who did develop complications from their gallstones experienced *prior* warning symptoms. Similar conclusions apply to diabetic patients with silent gallstones. Decision analysis has suggested that (1) the cumulative risk of death due to gallstone disease while on expectant management is small, and (2) prophylactic cholecystectomy is not warranted.

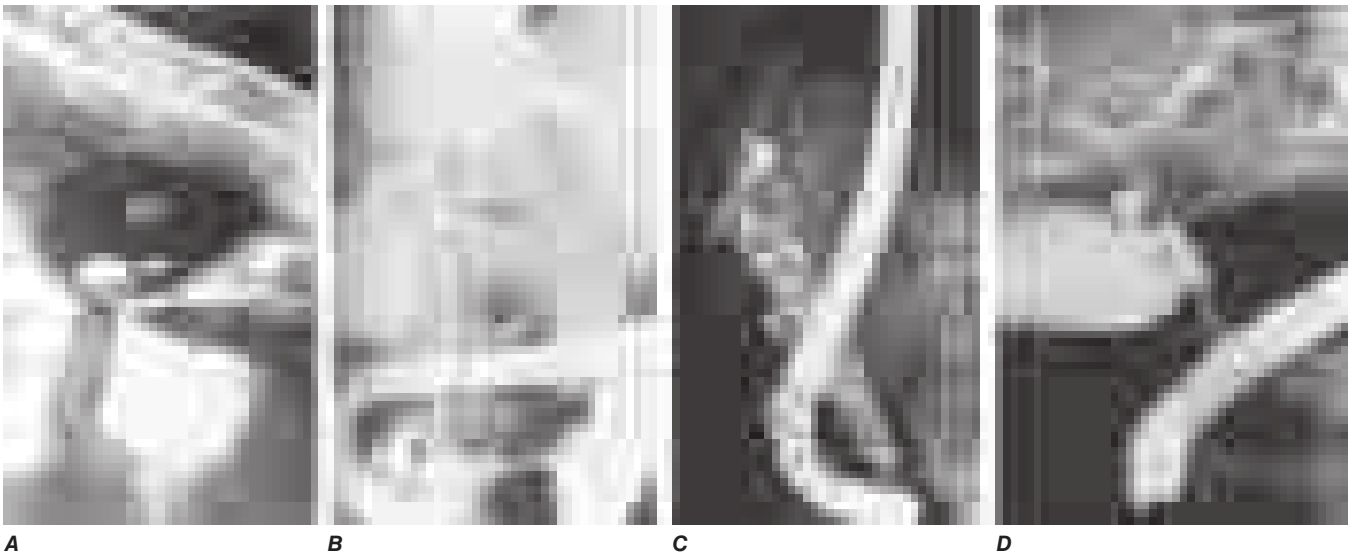


FIGURE 339-2 Examples of ultrasound and radiologic studies of the biliary tract. *A.* An ultrasound study showing a distended gallbladder (GB) containing a single large stone (*arrow*), which casts an acoustic shadow. ***B.*** Endoscopic retrograde cholangiopancreatogram (ERCP) showing normal biliary tract anatomy. In addition to the endoscope and large vertical gallbladder filled with contrast dye, the common hepatic duct (CHD), common bile duct (CBD), and pancreatic duct (PD) are shown. The *arrow* points to the ampulla of Vater. ***C.*** Endoscopic retrograde cholangiogram (ERC) showing choledocholithiasis. The biliary tract is dilated and contains multiple radiolucent calculi. ***D.*** ERCP showing sclerosing cholangitis. The CBD shows areas that are strictured and narrowed.

Complications requiring cholecystectomy are much more common in gallstone patients who have developed symptoms of biliary pain. Patients found to have gallstones at a young age are more likely to develop symptoms from cholelithiasis than are patients >60 years at the time of initial diagnosis. Patients with diabetes mellitus and gallstones may be somewhat more susceptible to septic complications, but the magnitude of risk of septic biliary complications in diabetic patients is incompletely defined.

TREATMENT

Gallstones

SURGICAL THERAPY

In asymptomatic gallstone patients, the risk of developing symptoms or complications requiring surgery is quite small (see above). Thus, a recommendation for cholecystectomy in a patient with gallstones should probably be based on assessment of three factors: (1) the presence of symptoms that are frequent enough or severe enough to interfere with the patient's general routine; (2) the presence of a prior complication of gallstone disease, that is, history of acute cholecystitis, pancreatitis, gallstone fistula, etc.; or (3) the presence of an underlying condition predisposing the patient to increased risk of gallstone complications (e.g., calcified or porcelain gallbladder and/or a previous attack of acute cholecystitis regardless of current symptomatic status). Patients with very large gallstones (>3 cm in diameter) and patients harboring gallstones in a congenitally anomalous gallbladder might also be considered for prophylactic cholecystectomy. Although young age is a worrisome factor in asymptomatic gallstone patients, few authorities would now recommend routine cholecystectomy in all young patients with silent stones. Laparoscopic cholecystectomy is a minimal-access approach for the removal of the gallbladder together with its stones. Its advantages include a markedly shortened hospital stay, minimal disability, and decreased cost, and it is the procedure of choice for most patients referred for elective cholecystectomy.

From several studies involving >4000 patients undergoing laparoscopic cholecystectomy, the following key points emerge: (1) complications develop in ~4% of patients, (2) conversion to laparotomy occurs in 5%, (3) the death rate is remarkably low (i.e., <0.1%), and (4) the rate of bile duct injuries is low (i.e., 0.2–0.6%) and comparable with open cholecystectomy. These data indicate why laparoscopic cholecystectomy has become the "gold standard" for treating symptomatic cholelithiasis.

MEDICAL THERAPY—GALLSTONE DISSOLUTION

In carefully selected patients with a functioning gallbladder and with radiolucent stones <10 mm in diameter, complete dissolution can be achieved in ~50% of patients within 6 months to 2 years. For good results within a reasonable time period, this therapy should be limited to radiolucent stones <5 mm in diameter. The dose of UDCA should be 10–15 mg/kg per day. Stones >10 mm in size rarely dissolve. Pigment stones are not responsive to UDCA therapy. Probably ≤10% of patients with *symptomatic* cholelithiasis are candidates for such treatment. However, in addition to the vexing problem of recurrent stones (30–50% over 3–5 years of follow-up), there is also the factor of taking an expensive drug for up to 2 years. The advantages and success of laparoscopic cholecystectomy have largely reduced the role of gallstone dissolution to patients who wish to avoid or are not candidates for elective cholecystectomy. However, patients with cholesterol gallstone disease who develop recurrent choledocholithiasis after cholecystectomy should be on long-term treatment with UDCA.

ACUTE AND CHRONIC CHOLECYSTITIS

Acute Cholecystitis Acute inflammation of the gallbladder wall usually follows obstruction of the cystic duct by a stone. Inflammatory response can be evoked by three factors: (1) *mechanical inflammation* produced by increased intraluminal pressure and distention with resulting ischemia of the gallbladder mucosa and wall, (2) *chemical inflammation* caused by the release of lysolecithin (due to the action of phospholipase on lecithin in bile) and other local tissue factors, and (3) *bacterial inflammation*, which may play a role in 50–85% of patients with acute cholecystitis. The organisms most frequently isolated by culture of gallbladder bile in these patients include *Escherichia coli*, *Klebsiella* spp., *Streptococcus* spp., and *Clostridium* spp.

Acute cholecystitis often begins as an attack of biliary pain that progressively worsens. Approximately 60–70% of patients report having experienced prior attacks that resolved spontaneously. As the episode progresses, however, the pain of acute cholecystitis becomes more generalized in the right upper abdomen. As with biliary colic, the pain of cholecystitis may radiate to the interscapular area, right scapula, or shoulder. Peritoneal signs of inflammation such as increased pain with jarring or on deep respiration may be apparent. The patient is anorectic and often nauseated. Vomiting is relatively common and may produce symptoms and signs of vascular and extracellular volume depletion. Jaundice is unusual early in the course of acute cholecystitis but may occur when edematous inflammatory changes involve the bile ducts and surrounding lymph nodes.

A low-grade fever is characteristically present, but shaking chills or rigors are not uncommon. The RUQ of the abdomen is almost invariably tender to palpation. An enlarged, tense gallbladder is palpable in 25–50% of patients. Deep inspiration or cough during subcostal palpation of the RUQ usually produces increased pain and inspiratory arrest (Murphy's sign). Localized rebound tenderness in the RUQ is common, as are abdominal distention and hypoactive bowel sounds from paralytic ileus, but generalized peritoneal signs and abdominal rigidity are usually lacking, in the absence of perforation.

The diagnosis of acute cholecystitis is usually made on the basis of a characteristic history and physical examination. The triad of sudden onset of RUQ tenderness, fever, and leukocytosis is highly suggestive. Typically, leukocytosis in the range of 10,000–15,000 cells per microliter with a left shift on differential count is found. The serum bilirubin is mildly elevated (<85.5 μmol/L [5 mg/dL]) in fewer than half of patients, whereas about one-fourth have modest elevations in serum aminotransferases (usually less than a fivefold elevation). Ultrasound will demonstrate calculi in 90–95% of cases and is useful for detection of signs of gallbladder inflammation including thickening of the wall, pericholecystic fluid, and dilatation of the bile duct. The radionuclide (e.g., HIDA) biliary scan may be confirmatory if bile duct imaging is seen without visualization of the gallbladder.

Approximately 75% of patients treated medically have remission of acute symptoms within 2–7 days following hospitalization. In 25%, however, a complication of acute cholecystitis will occur despite conservative treatment (see below). In this setting, prompt surgical intervention is required. Of the 75% of patients with acute cholecystitis who undergo remission of symptoms, ~25% will experience a recurrence of cholecystitis within 1 year, and 60% will have at least one recurrent bout within 6 years. In view of the natural history of the disease, acute cholecystitis is best treated by early surgery whenever possible. Mirizzi's syndrome is a rare complication in which a gallstone becomes impacted in the cystic duct or neck of the gallbladder causing compression of the CBD, resulting in CBD obstruction and jaundice. Ultrasound shows gallstone(s) lying outside the hepatic duct. Endoscopic retrograde cholangiopancreatography (ERCP) **(Fig. 339-2*B*)**, percutaneous transhepatic cholangiography (PTC), or magnetic resonance cholangiopancreatography (MRCP) will usually demonstrate the characteristic extrinsic compression of the CBD. Surgery consists of removing the cystic duct, diseased gallbladder, and the impacted stone. The preoperative diagnosis of Mirizzi's syndrome is important to avoid CBD injury.

ACALCULOUS CHOLECYSTITIS In 5–10% of patients with acute cholecystitis, calculi obstructing the cystic duct are not found at surgery. In >50% of such cases, an underlying explanation for acalculous inflammation is not found. An increased risk for the development of acalculous cholecystitis is especially associated with prolonged fasting, serious trauma or burns, with the postpartum period following prolonged labor, and with orthopedic and other nonbiliary major surgical

operations in the postoperative period. It may possibly complicate periods of prolonged parenteral hyperalimentation. For some of these cases, biliary sludge in the cystic duct may be responsible. Other precipitating factors include vasculitis, obstructing adenocarcinoma of the gallbladder, diabetes mellitus, torsion of the gallbladder, "unusual" bacterial infections of the gallbladder (e.g., *Leptospira*, *Streptococcus*, *Salmonella*, or *Vibrio cholerae*), and parasitic infestation of the gallbladder. Acalculous cholecystitis may also be seen with a variety of other systemic disease processes (e.g., sarcoidosis, cardiovascular disease, tuberculosis, syphilis, actinomycosis).

Although the clinical manifestations of acalculous cholecystitis are indistinguishable from those of calculous cholecystitis, the setting of acute gallbladder inflammation complicating severe underlying illness is characteristic of acalculous disease. Ultrasound or computed tomography (CT) examinations demonstrating a large, tense, static gallbladder without stones and with evidence of poor emptying over a prolonged period may be diagnostically useful in some cases. The complication rate for acalculous cholecystitis exceeds that for calculous cholecystitis. Successful management of acute acalculous cholecystitis appears to depend primarily on early diagnosis and surgical intervention, with meticulous attention to postoperative care.

ACALCULOUS CHOLECYSTOPATHY Disordered motility of the gallbladder can produce recurrent biliary pain in patients without gallstones. Infusion of an octapeptide of CCK can be used to measure the gallbladder ejection fraction during cholescintigraphy. The surgical findings have included abnormalities such as chronic cholecystitis, gallbladder muscle hypertrophy, and/or a markedly narrowed cystic duct. Some of these patients may well have had antecedent gallbladder disease. The following criteria can be used to identify patients with acalculous cholecystopathy: (1) recurrent episodes of typical RUQ pain characteristic of biliary tract pain, (2) abnormal CCK cholescintigraphy demonstrating a gallbladder ejection fraction of <40%, and (3) infusion of CCK reproducing the patient's pain. An additional clue would be the identification of a large gallbladder on ultrasound examination. Importantly, it should be noted that SOD dysfunction can also give rise to recurrent RUQ pain and CCK-scintigraphic abnormalities.

EMPHYSEMATOUS CHOLECYSTITIS So-called emphysematous cholecystitis is thought to begin with acute cholecystitis (calculous or acalculous) followed by ischemia or gangrene of the gallbladder wall and infection by gas-producing organisms. Bacteria most frequently cultured in this setting include anaerobes, such as *Clostridium welchii* or *C. perfringens*, and aerobes, such as *E. coli.* This condition occurs most frequently in elderly men and in patients with diabetes mellitus. The clinical manifestations are essentially indistinguishable from those of nongaseous cholecystitis. The diagnosis is usually made on plain abdominal film by finding gas within the gallbladder lumen, dissecting within the gallbladder wall to form a gaseous ring, or in the pericholecystic tissues. The morbidity and mortality rates with emphysematous cholecystitis are considerable. Prompt surgical intervention coupled with appropriate antibiotics is mandatory.

Chronic Cholecystitis Chronic inflammation of the gallbladder wall is almost always associated with the presence of gallstones and is thought to result from repeated bouts of subacute or acute cholecystitis or from persistent mechanical irritation of the gallbladder wall by gallstones. The presence of bacteria in the bile occurs in >25% of patients with chronic cholecystitis. The presence of infected bile in a patient with *chronic* cholecystitis undergoing elective cholecystectomy probably adds little to the operative risk. Chronic cholecystitis may be asymptomatic for years, which may progress to symptomatic gallbladder disease or to acute cholecystitis, or may present with complications (see below).

Complications of Cholecystitis • **EMPYEMA AND HYDROPS** Empyema of the gallbladder usually results from progression of acute cholecystitis with persistent cystic duct obstruction to superinfection of the stagnant bile with a pus-forming bacterial organism. The clinical picture resembles that of cholangitis with high fever; severe RUQ pain; marked leukocytosis; and often, prostration. Empyema of the gallbladder carries a high risk of gram-negative sepsis and/or perforation. Emergency surgical intervention with proper antibiotic coverage is required as soon as the diagnosis is suspected.

Hydrops or mucocele of the gallbladder may also result from prolonged obstruction of the cystic duct, usually by a large solitary calculus. In this instance, the obstructed gallbladder lumen is progressively distended, over a period of time, by mucus (mucocele) or by a clear transudate (hydrops) produced by mucosal epithelial cells. A visible, easily palpable, nontender mass sometimes extending from the RUQ into the right iliac fossa may be found on physical examination. The patient with hydrops of the gallbladder frequently remains asymptomatic, although chronic RUQ pain may also occur. Cholecystectomy is indicated, because empyema, perforation, or gangrene may complicate the condition.

GANGRENE AND PERFORATION Gangrene of the gallbladder results from ischemia of the wall and patchy or complete tissue necrosis. Underlying conditions often include marked distention of the gallbladder, vasculitis, diabetes mellitus, empyema, or torsion resulting in arterial occlusion. Gangrene usually predisposes to perforation of the gallbladder, but perforation may also occur in chronic cholecystitis without premonitory warning symptoms. *Localized perforations* are usually contained by the omentum or by adhesions produced by recurrent inflammation of the gallbladder. Bacterial superinfection of the walled-off gallbladder contents results in abscess formation. Most patients are best treated with cholecystectomy, but some seriously ill patients may be managed with cholecystostomy and drainage of the abscess. *Free perforation* is less common but is associated with a mortality rate of ~30%. Such patients may experience a sudden transient relief of RUQ pain as the distended gallbladder decompresses; this is followed by signs of generalized peritonitis.

FISTULA FORMATION AND GALLSTONE ILEUS Fistula formation into an adjacent organ adherent to the gallbladder wall may result from inflammation and adhesion formation. Fistulas into the duodenum are most common, followed in frequency by those involving the hepatic flexure of the colon, stomach or jejunum, abdominal wall, and renal pelvis. Clinically "silent" biliary-enteric fistulas occurring as a complication of acute cholecystitis have been found in up to 5% of patients undergoing cholecystectomy. Asymptomatic cholecystoenteric fistulas may sometimes be diagnosed by finding gas in the biliary tree on plain abdominal films. Barium contrast studies or endoscopy of the upper gastrointestinal tract or colon may demonstrate the fistula. Treatment in the symptomatic patient usually consists of cholecystectomy, CBD exploration, and closure of the fistulous tract.

Gallstone ileus refers to mechanical intestinal obstruction resulting from the passage of a large gallstone into the bowel lumen. The stone customarily enters the duodenum through a cholecystoenteric fistula at that level. The site of obstruction by the impacted gallstone is usually at the ileocecal valve, provided that the more proximal small bowel is of normal caliber. The majority of patients do not give a history of either prior biliary tract symptoms or complaints suggestive of acute cholecystitis or fistula formation. Large stones, >2.5 cm in diameter, are thought to predispose to fistula formation by gradual erosion through the gallbladder fundus. Diagnostic confirmation may occasionally be found on the plain abdominal film (e.g., small-intestinal obstruction with gas in the biliary tree [pneumobilia] and a calcified, ectopic gallstone) or following an upper gastrointestinal series (cholecystoduodenal fistula with small-bowel obstruction at the ileocecal valve). Laparotomy with stone extraction (or propulsion into the colon) remains the procedure of choice to relieve obstruction. Evacuation of large stones within the gallbladder should also be performed. In general, the gallbladder and its attachment to the intestines should be left alone.

LIMEY (MILK OF CALCIUM) BILE AND PORCELAIN GALLBLADDER Calcium salts in the lumen of the gallbladder in sufficient concentration may produce calcium precipitation and diffuse, hazy opacification of bile or a layering effect on plain abdominal roentgenography. This so-called limey bile, or milk of calcium bile, is usually clinically innocuous, but cholecystectomy is recommended, especially when it occurs in a hydropic gallbladder. In the entity called *porcelain gallbladder*, calcium salt deposition within the wall of a chronically inflamed gallbladder may be detected on the plain abdominal film. Cholecystectomy is

advised in all patients with porcelain gallbladder because in a high percentage of cases this finding appears to be associated with the development of carcinoma of the gallbladder.

TREATMENT

Acute Cholecystitis

MEDICAL THERAPY

Although surgical intervention remains the mainstay of therapy for acute cholecystitis and its complications, a period of in-hospital stabilization may be required before cholecystectomy. Oral intake is eliminated, nasogastric suction may be indicated, and extracellular volume depletion and electrolyte abnormalities are repaired. Meperidine or nonsteroidal anti-inflammatory drugs (NSAIDs) such as ketoralac or opioids, that is, morphine and hydromorphone, are usually employed for analgesia. Intravenous antibiotic therapy is usually indicated in patients with severe acute cholecystitis, even though bacterial superinfection of bile may not have occurred in the early stages of the inflammatory process. Antibiotic therapy is guided by the most common gram-negative organisms and anaerobes likely to be present, which are *E. coli*, *Klebsiella* spp., and *Streptococcus* spp. Effective antibiotics include piperacillin plus tazobactam, ceftriaxone plus metronidazole, levofloxacin plus metronidazole. Anaerobic coverage by a drug such as metronidazole should be added if gangrenous or emphysematous cholecystitis is suspected. Imipenem and meropenem should be reserved for the most severe, life-threatening infections when other regimens have failed **(Chap. 156)**. Postoperative complications of wound infection, abscess formation, and sepsis are reduced in antibiotic-treated patients.

SURGICAL THERAPY

The optimal timing of surgical intervention in patients with acute cholecystitis depends on stabilization of the patient. The clear trend is toward earlier surgery, and this is due in part to requirements for shorter hospital stays. Urgent (emergency) cholecystectomy or cholecystostomy is probably appropriate in most patients in whom a complication of acute cholecystitis such as empyema, emphysematous cholecystitis, or perforation is suspected or confirmed. Patients with uncomplicated acute cholecystitis should undergo early elective laparoscopic cholecystectomy, ideally within 48–72 h after diagnosis. The complication rate is not increased in patients undergoing early as opposed to delayed (>6 weeks after diagnosis) cholecystectomy. Delayed surgical intervention is probably best reserved for (1) patients in whom the overall medical condition imposes an unacceptable risk for early surgery and (2) patients in whom the diagnosis of acute cholecystitis is in doubt. Thus, early cholecystectomy (within 72 h) is the treatment of choice for most patients with acute cholecystitis. Mortality figures for emergency cholecystectomy in most centers range from 1 to 3%, whereas the mortality risk for early elective cholecystectomy is ~0.5% in patients under age 60. Of course, the operative risks increase with age-related diseases of other organ systems and with the presence of long- or short-term complications of gallbladder disease. Seriously ill or debilitated patients with cholecystitis may be managed with cholecystostomy and tube drainage of the gallbladder. Elective cholecystectomy may then be done at a later date.

Postcholecystectomy Complications Early complications following cholecystectomy include atelectasis and other pulmonary disorders, abscess formation (often subphrenic), external or internal hemorrhage, biliary-enteric fistula, and bile leaks. Jaundice may indicate absorption of bile from an intraabdominal collection following a biliary leak or mechanical obstruction of the CBD by retained calculi, intraductal blood clots, or extrinsic compression.

Overall, cholecystectomy is a very successful operation that provides total or near-total relief of preoperative symptoms in 75–90% of patients. The most common cause of persistent postcholecystectomy symptoms is an overlooked symptomatic nonbiliary disorder (e.g., reflux esophagitis, peptic ulceration, pancreatitis, or—most often—irritable bowel syndrome). In a small percentage of patients, however, a disorder of the extrahepatic bile ducts may result in persistent symptomatology. These so-called postcholecystectomy syndromes may be due to (1) biliary strictures, (2) retained biliary calculi, (3) cystic duct stump syndrome, (4) stenosis or dyskinesia of the SOD, or (5) bile salt–induced diarrhea or gastritis.

CYSTIC DUCT STUMP SYNDROME In the absence of cholangiographically demonstrable retained stones, symptoms resembling biliary pain or cholecystitis in the postcholecystectomy patient have frequently been attributed to disease in a long (>1 cm) cystic duct remnant (cystic duct stump syndrome). Careful analysis, however, reveals that postcholecystectomy complaints are attributable to other causes in almost all patients in whom the symptom complex was originally thought to result from the existence of a long cystic duct stump. Accordingly, considerable care should be taken to investigate the possible role of other factors in the production of postcholecystectomy symptoms before attributing them to cystic duct stump syndrome.

SOD STENOSIS AND SOD DYSKINESIA, AND BILIARY DYSKINESIA Symptoms of biliary colic accompanied by signs of recurrent, intermittent biliary obstruction may be produced by acalculous cholecystopathy, SOD stenosis, or SOD dyskinesia. SOD stenosis is thought to result from acute or chronic inflammation of the papilla of Vater or from glandular hyperplasia of the papillary segment. Five criteria have been used to define SOD stenosis: (1) upper abdominal pain, usually RUQ or epigastric; (2) abnormal liver tests; (3) dilatation of the CBD upon MRCP or ERCP examination; (4) delayed (>45 min) drainage of contrast material from the duct; and (5) increased basal pressure of the SOD. After exclusion of acalculous cholecystopathy, treatment consists of endoscopic or surgical sphincteroplasty to ensure wide patency of the distal portions of both the bile and pancreatic ducts. The greater the number of the preceding criteria present, the greater is the likelihood that a patient does have a degree of SOD sufficient to justify correction. The factors usually considered as indications for sphincterotomy include (1) prolonged duration of symptoms, (2) lack of response to symptomatic treatment, (3) presence of severe disability, and (4) the patient's choice of sphincterotomy over surgery (given a clear understanding on his or her part of the risks involved in both procedures).

Biliary SOD disorders are characterized by three criteria: (1) biliary pain, (2) absence of bile duct stones or other abnormalities, and (3) elevated liver enzymes or a dilated CBD, but not both. In this setting, either hepatobiliary scintigraphy or SOD manometry can support the diagnosis. Importantly, the presence of both elevated liver enzymes and a dilated CBD should raise the question of obstruction. Proposed mechanisms to account for SOD dysfunction include spasm of the sphincter, denervation sensitivity resulting in hypertonicity, and abnormalities in the sequencing or frequency rates of the sphincteric-contraction waves. When thorough evaluation has failed to demonstrate another cause for the pain and when cholangiographic and manometric criteria suggest a diagnosis of SOD dyskinesia, medical treatment with nitrites or anticholinergics to attempt pharmacologic relaxation of SOD has been proposed but not evaluated in detailed studies. Endoscopic biliary sphincterotomy (EBS) or surgical sphincterotomy may be indicated in patients who fail to respond to a 2–3 month trial of medical therapy, especially if SOD pressures are elevated. Approximately 45% of such patients have long-term pain relief after EBS. Endoscopic biliary sphincterotomy has become the procedure of choice for removing bile duct stones and for other biliary and pancreatic problems.

BILE SALT–INDUCED DIARRHEA AND GASTRITIS Postcholecystectomy patients may develop symptoms of dyspepsia, which have been attributed to duodenogastric reflux of bile. However, firm data linking these symptoms to bile gastritis after surgical removal of the gallbladder are lacking. Cholecystectomy induces persistent changes in gut transit, and these changes effect a noticeable modification of bowel habits. Cholecystectomy shortens gut transit time by accelerating passage of the fecal bolus through the colon with marked acceleration in the right colon, thus causing an increase in colonic bile acid output and a shift in bile acid composition toward the more diarrheagenic secondary bile acids,

that is, deoxycholic acid. Diarrhea that is severe enough, that is, three or more watery movements per day, can be classified as postcholecystectomy diarrhea, and this occurs in 5–10% of patients undergoing elective cholecystectomy. Treatment with bile acid–sequestering agents such as cholestyramine or colestipol is often effective in ameliorating troublesome diarrhea.

THE HYPERPLASTIC CHOLECYSTOSES

The term *hyperplastic cholecystoses* is used to denote a group of disorders of the gallbladder characterized by excessive proliferation of normal tissue components.

Adenomyomatosis is characterized by a benign proliferation of gallbladder surface epithelium with glandlike formations, extramural sinuses, transverse strictures, and/or fundal nodule ("adenoma" or "adenomyoma") formation.

Cholesterolosis is characterized by abnormal deposition of lipid, especially cholesteryl esters, within macrophages in the lamina propria of the gallbladder wall. In its diffuse form ("strawberry gallbladder"), the gallbladder mucosa is brick red and speckled with bright yellow flecks of lipid. The localized form shows solitary or multiple "cholesterol polyps" studding the gallbladder wall. Cholesterol stones of the gallbladder are found in nearly half the cases. Cholecystectomy is indicated in both adenomyomatosis and cholesterolosis when symptomatic or when cholelithiasis is present.

The prevalence of gallbladder polyps in the adult population is 1–4% with a marked male predominance. Types of gallbladder polyps include cholesterol polyps, adenomyomas, inflammatory polyps, and adenomas (rare). No significant changes have been found over a 5-year period in asymptomatic patients with gallbladder polyps <6 mm and few changes in polyps 7–9 mm. Cholecystectomy is recommended in symptomatic patients as well as in asymptomatic patients >50 years whose polyps are >10 mm or associated with gallstones or polyp growth on serial ultrasonography.

DISEASES OF THE BILE DUCTS

CONGENITAL ANOMALIES

Biliary Atresia and Hypoplasia Atretic and hypoplastic lesions of the extrahepatic and large intrahepatic bile ducts are the most common biliary anomalies of clinical relevance encountered in infancy. The clinical picture is one of severe obstructive jaundice during the first month of life, with pale stools. When biliary atresia is suspected on the basis of clinical, laboratory, and imaging findings, the diagnosis is confirmed by surgical exploration and operative cholangiography. Approximately 10% of cases of biliary atresia are treatable with Roux-en-Y choledochojejunostomy, with the Kasai procedure (hepatic portoenterostomy) being attempted in the remainder in an effort to restore some bile flow. Most patients, even those having successful biliary-enteric anastomoses, eventually develop chronic cholangitis, extensive hepatic fibrosis, and portal hypertension.

Choledochal Cysts Cystic dilatation may involve the free portion of the CBD, that is, choledochal cyst, or may present as diverticulum formation in the intraduodenal segment. In the latter situation, chronic reflux of pancreatic juice into the biliary tree can produce inflammation and stenosis of the extrahepatic bile ducts leading to cholangitis or biliary obstruction. Because the process may be gradual, ~50% of patients present with onset of symptoms after age 10. The diagnosis may be made by ultrasound, abdominal CT, MRC, or cholangiography. Only one-third of patients show the classic triad of abdominal pain, jaundice, and an abdominal mass. Ultrasonographic detection of a cyst separate from the gallbladder should suggest the diagnosis of choledochal cyst, which can be confirmed by demonstrating the entrance of extrahepatic bile ducts into the cyst. Surgical treatment involves excision of the "cyst" and biliary-enteric anastomosis. Patients with choledochal cysts are at increased risk for the subsequent development of cholangiocarcinoma.

Congenital Biliary Ectasia Dilatation of intrahepatic bile ducts may involve either the major intrahepatic radicles (Caroli's disease), the inter- and intralobular ducts (congenital hepatic fibrosis), or both. In Caroli's disease, clinical manifestations include recurrent cholangitis, abscess formation in and around the affected ducts, and, often, brown pigment gallstone formation within portions of ectatic intrahepatic biliary radicles. Ultrasound, MRC, and CT are of great diagnostic value in demonstrating cystic dilatation of the intrahepatic bile ducts. Treatment with ongoing antibiotic therapy is usually undertaken in an effort to limit the frequency and severity of recurrent bouts of cholangitis. Progression to secondary biliary cirrhosis with portal hypertension, extrahepatic biliary obstruction, cholangiocarcinoma, or recurrent episodes of sepsis with hepatic abscess formation is common.

CHOLEDOCHOLITHIASIS

Pathophysiology and Clinical Manifestations Passage of gallstones into the CBD occurs in ~10–15% of patients with cholelithiasis. The incidence of common duct stones increases with increasing age of the patient, so that up to 25% of elderly patients may have calculi in the common duct at the time of cholecystectomy. Undetected duct stones are left behind in ~1–5% of cholecystectomy patients. The overwhelming majority of bile duct stones are cholesterol stones formed in the gallbladder, which then migrate into the extrahepatic biliary tree through the cystic duct. Primary calculi arising de novo in the ducts are usually brown pigment stones developing in patients with (1) hepatobiliary parasitism or chronic, recurrent cholangitis; (2) congenital anomalies of the bile ducts (especially Caroli's disease); (3) dilated, sclerosed, or strictured ducts; or (4) an *MDR3* (ABCB4) gene defect leading to impaired biliary phospholipids secretion (low phospholipid–associated cholesterol cholelithiasis). Common duct stones may remain asymptomatic for years, may pass spontaneously into the duodenum, or (most often) may present with biliary colic or a complication.

Complications • **CHOLANGITIS** Cholangitis may be acute or chronic, and symptoms result from inflammation, which usually is caused by at least partial obstruction to the flow of bile. Bacteria are present on bile culture in ~75% of patients with acute cholangitis early in the symptomatic course. The characteristic presentation of acute cholangitis involves biliary pain, jaundice, and spiking fevers with chills (Charcot's triad). Blood cultures are frequently positive, and leukocytosis is typical. *Nonsuppurative acute cholangitis* is most common and may respond relatively rapidly to supportive measures and to treatment with antibiotics. In *suppurative acute cholangitis*, however, the presence of pus under pressure in a completely obstructed ductal system leads to symptoms of severe toxicity—mental confusion, bacteremia, and septic shock. Response to antibiotics alone in this setting is relatively poor, multiple hepatic abscesses are often present, and the mortality rate approaches 100% unless prompt endoscopic or surgical relief of the obstruction and drainage of infected bile are carried out. Endoscopic management of bacterial cholangitis is as effective as surgical intervention. ERCP with endoscopic sphincterotomy is safe and the preferred initial procedure for both establishing a definitive diagnosis and providing effective therapy.

OBSTRUCTIVE JAUNDICE Gradual obstruction of the CBD over a period of weeks or months usually leads to initial manifestations of jaundice or pruritus without associated symptoms of biliary colic or cholangitis. Painless jaundice may occur in patients with choledocholithiasis, but is much more characteristic of biliary obstruction secondary to malignancy of the head of the pancreas, bile ducts, or ampulla of Vater.

In patients whose obstruction is secondary to choledocholithiasis, associated chronic calculous cholecystitis is very common, and the gallbladder in this setting may be unable to distend. The absence of a palpable gallbladder in most patients with biliary obstruction from duct stones is the basis for Courvoisier's law, that is, that the presence of a palpably enlarged gallbladder suggests that the biliary obstruction is secondary to an underlying malignancy rather than to calculous disease. Biliary obstruction causes progressive dilatation of the intrahepatic bile ducts as intrabiliary pressures rise. Hepatic bile flow is suppressed, and reabsorption and regurgitation of conjugated bilirubin into the bloodstream lead to jaundice accompanied by dark urine (bilirubinuria) and light-colored (acholic) stools.

CBD stones should be suspected in any patient with cholecystitis whose serum bilirubin level is >85.5 μmol/L (5 mg/dL). The maximum bilirubin level is seldom >256.5 μmol/L (15.0 mg/dL) in patients with choledocholithiasis unless concomitant hepatic or renal disease or another factor leading to marked hyperbilirubinemia exists. Serum bilirubin levels ≥342.0 μmol/L (20 mg/dL) should suggest the possibility of neoplastic obstruction. The serum alkaline phosphatase level is almost always elevated in biliary obstruction. A rise in alkaline phosphatase often precedes clinical jaundice and may be the only abnormality in routine liver function tests. There may be a two- to tenfold elevation of serum aminotransferases, especially in association with acute obstruction. Following relief of the obstructing process, serum aminotransferase elevations usually return rapidly to normal, while the serum bilirubin level may take 1–2 weeks to return to normal. The alkaline phosphatase level usually falls slowly, lagging behind the decrease in serum bilirubin.

PANCREATITIS The most common associated entity discovered in patients with nonalcoholic acute pancreatitis is biliary tract disease. Biochemical evidence of pancreatic inflammation complicates acute cholecystitis in 15% of cases and choledocholithiasis in >30%, and the common factor appears to be the passage of gallstones through the common duct. Coexisting pancreatitis should be suspected in patients with symptoms of cholecystitis who develop (1) back pain or pain to the left of the abdominal midline, (2) prolonged vomiting with paralytic ileus, or (3) a pleural effusion, especially on the left side. Surgical treatment of gallstone disease is usually associated with resolution of the pancreatitis.

SECONDARY BILIARY CIRRHOSIS Secondary biliary cirrhosis may complicate prolonged or intermittent duct obstruction with or without recurrent cholangitis. Although this complication may be seen in patients with choledocholithiasis, it is more common in cases of prolonged obstruction from stricture or neoplasm. Once established, secondary biliary cirrhosis may be progressive even after correction of the obstructing process, and increasingly severe hepatic cirrhosis may lead to portal hypertension or to hepatic failure and death. Prolonged biliary obstruction may also be associated with clinically relevant deficiencies of the fat-soluble vitamins A, D, E, and K.

Diagnosis and Treatment The diagnosis of choledocholithiasis is usually made by cholangiography **(Table 339-3)**, either preoperatively by endoscopic retrograde cholangiogram (ERC) **(Fig. 339-2*C*)** or MRCP or intraoperatively at the time of cholecystectomy. As many as 15% of patients undergoing cholecystectomy will prove to have CBD stones. When CBD stones are suspected prior to laparoscopic cholecystectomy, preoperative ERCP with endoscopic papillotomy and stone extraction is the preferred approach. It not only provides stone clearance but also defines the anatomy of the biliary tree in relationship to the cystic duct. CBD stones should be suspected in gallstone patients who have any of the following risk factors: (1) a history of jaundice or pancreatitis, (2) abnormal tests of liver function, and (3) ultrasonographic or MRCP evidence of a dilated CBD or stones in the duct. Alternatively, if intraoperative cholangiography reveals retained stones, postoperative ERCP can be carried out. The need for preoperative ERCP is expected to decrease further as laparoscopic techniques for bile duct exploration improve.

The widespread use of laparoscopic cholecystectomy and ERCP has decreased the incidence of complicated biliary tract disease and the need for choledocholithotomy and T-tube drainage of the bile ducts. EBS followed by spontaneous passage or stone extraction is the treatment of choice in the management of patients with common duct stones, especially in elderly or poor-risk patients.

TRAUMA, STRICTURES, AND HEMOBILIA

Most benign strictures of the extrahepatic bile ducts result from surgical trauma and occur in about 1 in 500 cholecystectomies. Strictures may present with bile leak or abscess formation in the immediate postoperative period or with biliary obstruction or cholangitis as long as 2 years or more following the inciting trauma. The diagnosis is established by percutaneous or endoscopic cholangiography. Endoscopic brushing of biliary strictures may be helpful in establishing the nature of the lesion and is more accurate than bile cytology alone. When positive exfoliative cytology is obtained, the diagnosis of a neoplastic stricture is established. This procedure is especially important in patients with primary sclerosing cholangitis (PSC) who are predisposed to the development of cholangiocarcinomas. Successful operative correction of non-PSC bile duct strictures by a skillful surgeon with duct-to-bowel anastomosis is usually possible, although mortality rates from surgical complications, recurrent cholangitis, or secondary biliary cirrhosis are high.

Hemobilia may follow traumatic or operative injury to the liver or bile ducts, intraductal rupture of a hepatic abscess or aneurysm of the hepatic artery, biliary or hepatic tumor hemorrhage, or mechanical complications of choledocholithiasis or hepatobiliary parasitism. Diagnostic procedures such as liver biopsy, PTC, and transhepatic biliary drainage catheter placement may also be complicated by hemobilia. Patients often present with a classic triad of biliary pain, obstructive jaundice, and melena or occult blood in the stools. The diagnosis is sometimes made by cholangiographic evidence of blood clot in the biliary tree, but selective angiographic verification may be required. Although minor episodes of hemobilia may resolve without operative intervention, surgical ligation of the bleeding vessel is frequently required.

EXTRINSIC COMPRESSION OF THE BILE DUCTS

Partial or complete biliary obstruction may be produced by extrinsic compression of the ducts. The most common cause of this form of obstructive jaundice is carcinoma of the head of the pancreas. Biliary obstruction may also occur as a complication of either acute or chronic pancreatitis or involvement of lymph nodes in the porta hepatis by lymphoma or metastatic carcinoma. The latter should be distinguished from cholestasis resulting from massive replacement of the liver by tumor.

HEPATOBILIARY PARASITISM

Infestation of the biliary tract by adult helminths or their ova may produce a chronic, recurrent pyogenic cholangitis with or without multiple hepatic abscesses, ductal stones, or biliary obstruction. This condition is relatively rare but does occur in inhabitants of southern China and elsewhere in Southeast Asia. The organisms most commonly involved are trematodes or flukes, including *Clonorchis sinensis*, *Opisthorchis viverrini* or *Opisthorchis felineus*, and *Fasciola hepatica*. The biliary tract also may be involved by intraductal migration of adult *Ascaris lumbricoides* from the duodenum or by intrabiliary rupture of hydatid cysts of the liver produced by *Echinococcus* spp. The diagnosis is made by cholangiography and the presence of characteristic ova on stool examination. When obstruction is present, the treatment of choice is laparotomy under antibiotic coverage, with common duct exploration and a biliary drainage procedure.

SCLEROSING CHOLANGITIS

Primary or idiopathic sclerosing cholangitis (PSC) is characterized by a progressive, inflammatory, sclerosing, and obliterative process affecting the extrahepatic and/or the intrahepatic bile ducts. The disorder occurs up to 90% in association with inflammatory bowel disease, especially ulcerative colitis. It may also be associated with autoimmune pancreatitis; multifocal fibrosclerosis syndromes such as retroperitoneal, mediastinal, and/or periureteral fibrosis; Riedel's struma; or pseudotumor of the orbit.

Immunoglobulin G4 (IgG4)–associated cholangitis is a recently described biliary disease of unknown etiology that presents with biochemical and cholangiographic features indistinguishable from PSC, is often associated with autoimmune pancreatitis and other fibrosing conditions, and is characterized by elevated serum IgG4 and infiltration of IgG4-positive plasma cells in bile ducts and liver tissue. All newly diagnosed PSC patients should have a serum IgG4 level checked. In contrast to PSC, IgG4-associated cholangitis is not associated with inflammatory bowel disease and should be suspected if associated with increased serum IgG4 and unexplained pancreatic disease. Glucocorticoids are regarded as the initial treatment of choice. Relapse is common after steroid withdrawal, especially with proximal strictures. Long-term treatment with glucocorticoids and/or azathioprine may be needed after relapse or for inadequate response **(Chap. 341)**.

Patients with PSC often present with signs and symptoms of chronic or intermittent biliary obstruction: RUQ abdominal pain, pruritus,

TABLE 339-3 Diagnostic Evaluation of the Bile Ducts

DIAGNOSTIC ADVANTAGES	DIAGNOSTIC LIMITATIONS	CONTRAINDICATIONS	COMPLICATIONS	COMMENT
Hepatobiliary Ultrasound				
Rapid Simultaneous scanning of GB, liver, bile ducts, pancreas Accurate identification of dilated bile ducts Not limited by jaundice, pregnancy Guidance for fine-needle biopsy	Bowel gas Massive obesity Ascites Barium Partial bile duct obstruction Poor visualization of distal CBD	None	None	Initial procedure of choice in investigating possible biliary tract obstruction
Computed Tomography				
Simultaneous scanning of GB, liver, bile ducts, pancreas Accurate identification of dilated bile ducts, masses Not limited by jaundice, gas, obesity, ascites High-resolution image Guidance for fine-needle biopsy	Extreme cachexia Movement artifact Ileus Partial bile duct obstruction	Pregnancy	Reaction to iodinated contrast, if used	Indicated for evaluation of hepatic or pancreatic masses Procedure of choice in investigating possible biliary obstruction if diagnostic limitations prevent HBUS
Magnetic Resonance Cholangiopancreatography				
Useful modality for visualizing pancreatic and biliary ducts Has excellent sensitivity for bile duct dilatation, biliary stricture, and intraductal abnormalities Can identify pancreatic duct dilatation or stricture, pancreatic duct stenosis, and pancreas divisum	Cannot offer therapeutic intervention High cost	Claustrophobia Certain metals (iron)	None	
Endoscopic Retrograde Cholangiopancreatography				
Simultaneous pancreatography Best visualization of distal biliary tract Bile or pancreatic cytology Endoscopic sphincterotomy and stone removal Biliary manometry	Gastroduodenal obstruction Roux-en-Ybiliary-entericanastomosis	Pregnancy Acute pancreatitis Severe cardiopulmonary disease	Pancreatitis Cholangitis, sepsis Infected pancreatic pseudocyst Perforation (rare) Hypoxemia, aspiration	Cholangiogram of choice in: Absence of dilated ducts Pancreatic, ampullary or gastroduodenal disease Prior biliary surgery Endoscopic sphincterotomy treatment possibility
Percutaneous Transhepatic Cholangiogram				
Extremely successful when bile ducts dilated Best visualization of proximal biliary tract Bile cytology/culture Percutaneous transhepatic drainage	Nondilated or sclerosed ducts	Pregnancy Uncorrectable coagulopathy Massive ascites Hepatic abscess	Bleeding Hemobilia Bile peritonitis Bacteremia, sepsis	Indicated when ERCP is contraindicated or failed
Endoscopic Ultrasound				
Most sensitive method to detect ampullary stones				Excellent for detecting choledocholithiasis

Abbreviations: CBD, common bile duct; ERCP, endoscopic retrograde cholangiopancreatography; GB, gallbladder; HBUS, hepatobiliary ultrasound.

jaundice, or acute cholangitis. Late in the course, complete biliary obstruction, secondary biliary cirrhosis, hepatic failure, or portal hypertension with bleeding varices may occur. The diagnosis is usually established by finding multifocal, diffusely distributed strictures with intervening segments of normal or dilated ducts, producing a beaded appearance on cholangiography **(Fig. 339-2*D*)**. The cholangiographic techniques of choice in suspected cases are MRCP and ERCP. When a diagnosis of sclerosing cholangitis has been established, a search for associated diseases, especially for chronic inflammatory bowel disease, should be carried out.

A recent study describes the natural history and outcome for 305 patients of Swedish descent with PSC; 134 (44%) of the patients were asymptomatic at the time of diagnosis and, not surprisingly, had a significantly higher survival rate. The independent predictors of a bad prognosis were advanced age, serum bilirubin concentration, and liver histologic changes. Cholangiocarcinoma was found in 24 patients (8%). Inflammatory bowel disease was closely associated with PSC and had a prevalence of 81% in this study population. The PSC Autoimmune Hepatitis (AIH) Overlap syndrome is characterized by clinical, biochemical, and histological features of AIH and cholangiographic features of PSC.

Small duct PSC is defined by the presence of chronic cholestasis and hepatic histology consistent with PSC but with normal findings on cholangiography. Small duct PSC is found in ~5% of patients with PSC and may represent an earlier stage of PSC associated with a

significantly better long-term prognosis. However, such patients may progress to classic PSC and/or end-stage liver disease with consequent necessity of liver transplantation.

In patients with AIDS, cholangiopancreatography may demonstrate a broad range of biliary tract changes as well as pancreatic duct obstruction and occasionally pancreatitis **(Chap. 197)**. Further, biliary tract lesions in AIDS include infection and cholangiopancreatographic changes similar to those of PSC. Changes noted include: (1) diffuse involvement of intrahepatic bile ducts alone, (2) involvement of both intra- and extrahepatic bile ducts, (3) ampullary stenosis, (4) stricture of the intrapancreatic portion of the CBD, and (5) pancreatic duct involvement. Associated infectious organisms include *Cryptosporidium*, *Mycobacterium avium-intracellulare*, cytomegalovirus, *Microsporidia*, and *Isospora*. ERCP sphincterotomy can provide significant pain reduction in patients with AIDS-associated papillary stenosis.

TREATMENT

Sclerosing Cholangitis

Therapy with cholestyramine may help control symptoms of pruritus, and antibiotics are useful when cholangitis complicates the clinical picture. Vitamin D and calcium supplementation may help prevent the loss of bone mass frequently seen in patients with chronic cholestasis. Glucocorticoids, methotrexate, and cyclosporine have not been shown to be efficacious in PSC. UDCA in high dosage (28–30 mg/kg) was not effective in prolonging survival. In cases where high-grade biliary obstruction (dominant strictures) has occurred, balloon dilatation or stenting may be appropriate. Only rarely is surgical intervention indicated. Efforts at biliary-enteric anastomosis or stent placement may, however, be complicated by recurrent cholangitis and further progression of the stenosing process. The prognosis is unfavorable, with a median survival of 9–12 years following the diagnosis, regardless of therapy. Four variables (age, serum bilirubin level, histologic stage, and splenomegaly) predict survival in patients with PSC and serve as the basis for a risk score. PSC is one of the most common indications for liver transplantation.

FURTHER READING

Baron TH et al: Interventional approaches to gallbladder disease. N Engl J Med 373:357, 2015.

Lindor K et al: American College of Gastroenterology (ACG) Guidelines: Primary Sclerosing Cholangitis. Hepatology 51:660, 2010.

Ryl JK et al: Clinical features of acute acalculous cholecystitis. J Clin Gastroenterol 36:166, 2003.

Strasberg S: Clinical practice. Acute calculous cholecystitis. N Engl J Med 358:2804, 2008.

Section 4 Disorders of the Pancreas

340 Approach to the Patient with Pancreatic Disease

Darwin L. Conwell, Norton J. Greenberger, Peter A. Banks

GENERAL CONSIDERATIONS

As emphasized in **Chap. 341**, the etiologies as well as clinical manifestations of pancreatitis are quite varied. Although it is well-appreciated that pancreatitis is frequently secondary to biliary tract disease and alcohol abuse, it can also be caused by drugs, genetic mutations, trauma, and viral infections and is associated with metabolic and connective tissue disorders. In ~30% of patients with acute pancreatitis and 25–40% of patients with chronic pancreatitis, the etiology initially can be obscure.

The incidence of acute pancreatitis is about 5–35 per 100,000 new cases per year worldwide, with a mortality rate of about 3%. The incidence of chronic pancreatitis is about 4–8 per 100,000 persons with a prevalence of 26–42 cases per 100,000. The number of patients admitted to the hospital who suffer with both acute and chronic pancreatitis in the United States is largely increasing and is now estimated to be 274,119 for acute pancreatitis and 19,724 for chronic pancreatitis. Acute pancreatitis is the most common gastrointestinal diagnosis requiring hospitalization in the United States. Acute and chronic pancreatic disease costs an estimated $3 billion annually in health care expenditures. These numbers may underestimate the true incidence and prevalence, because non–alcohol-induced pancreatitis has been largely ignored. At autopsy, the prevalence of chronic pancreatitis ranges from 0.04 to 5%.

The diagnosis of acute pancreatitis is generally clearly defined based on a combination of laboratory, imaging, and clinical symptoms. The diagnosis of chronic pancreatitis, especially in mild disease, is hampered by the relative inaccessibility of the pancreas to direct examination and the nonspecificity of the abdominal pain associated with chronic pancreatitis. Many patients with chronic pancreatitis do not have elevated blood amylase or lipase levels. Some patients with chronic pancreatitis develop signs and symptoms of pancreatic exocrine insufficiency (PEI), and thus, objective evidence for pancreatic disease can be demonstrated. However, there is a very large reservoir of pancreatic exocrine function. More than 90% of the pancreas must be damaged before maldigestion of fat and protein is manifested. Noninvasive, indirect tests of pancreatic exocrine function (fecal elastase) are much more likely to give abnormal results in patients with obvious advanced pancreatic disease (i.e., pancreatic calcification, steatorrhea, or diabetes mellitus) than in patients with occult disease. Invasive, direct tests of pancreatic secretory function (secretin tests) are the most sensitive and specific tests to detect early chronic pancreatic disease when imaging is equivocal or normal.

TESTS USEFUL IN THE DIAGNOSIS OF PANCREATIC DISEASE

Several tests have proved of value in the evaluation of pancreatic disease. Examples of specific tests and their usefulness in the diagnosis of acute and chronic pancreatitis are summarized in **Table 340-1** and **Fig. 340-1**. At some institutions, pancreatic function tests are available and performed if the diagnosis of chronic pancreatic disease remains a possibility after noninvasive tests (ultrasound, computed tomography [CT], magnetic resonance cholangiopancreatography [MRCP]) or invasive tests (endoscopic retrograde cholangiopancreatography [ERCP], endoscopic ultrasonography [EUS]) have given normal or inconclusive results. In this regard, tests using *direct* stimulation of the pancreas with secretin are the most sensitive.

Pancreatic Enzymes in Body Fluids The serum amylase and lipase levels are widely used as screening tests for acute pancreatitis in the patient with acute abdominal pain or back pain. Values greater than three times the upper limit of normal (3× ULN) in combination with epigastric pain strongly suggest the diagnosis of acute pancreatitis. In acute pancreatitis, the serum amylase and lipase are usually elevated within 24 h of onset and remain so for 3–7 days. Levels usually return to normal within 7 days unless there is pancreatic ductal disruption, ductal obstruction, or pseudocyst formation. Approximately 85% of patients with acute pancreatitis have a threefold or greater elevated serum lipase and amylase levels. The values may be normal if (1) there is a delay (2–5 days) before blood samples are obtained, (2) the underlying disorder is chronic pancreatitis rather than acute pancreatitis, or (3) hypertriglyceridemia is present. Patients with hypertriglyceridemia and pancreatitis have been found to have spuriously low levels of amylase and perhaps lipase activity. In the absence of objective evidence of pancreatitis by abdominal ultrasound, CT scan, MRCP, or EUS, mild to moderate elevations of amylase and/or lipase are not helpful in making a diagnosis of chronic pancreatitis.

The serum amylase can be elevated in other conditions **(Table 340-2)**, in part because the enzyme is found in many organs. In addition to the pancreas and salivary glands, small quantities of amylase are found

TABLE 340-1 Tests Useful in the Diagnosis of Acute and Chronic Pancreatitis and Pancreatic Tumors

TEST	PRINCIPLE	COMMENT
Pancreatic Enzymes in Body Fluids		
Serum lipase	Pancreatic inflammation leads to increased serum enzyme levels	Enzyme measurement of choice for diagnosis of acute pancreatitis
Amylase		
1. Serum	Pancreatic inflammation leads to increased serum enzyme levels	Simple; reliable if test results are three times the upper limit of normal (3× ULN)
2. Urine	Renal clearance of amylase is increased in acute pancreatitis	Infrequently used
3. Ascitic fluid	Disruption of gland or main pancreatic duct leads to increased amylase concentration	Can help establish source of ascites; false positives occur with intestinal obstruction and perforated ulcer; can also measure lipase
4. Pleural fluid	Exudative pleural effusion with pancreatitis	False positives occur with carcinoma of the lung and esophageal perforation
Studies Pertaining to Pancreatic Structure		
Radiologic and radionuclide tests		
1. Plain film of the abdomen	Can be abnormal in acute and chronic pancreatitis	Infrequently used
2. Upper gastrointestinal x-rays		Infrequently used
3. Ultrasonography (US)	Can provide information on edema, inflammation, calcification, pseudocysts, and mass lesions	Simple, noninvasive; sequential studies quite feasible; useful in diagnosis of gallstones; pancreas visualization limited by interference from overlying bowel gas
4. Computed tomography (CT) scan	Permits detailed visualization of pancreas and surrounding structures, pancreatic fluid collection, pseudocyst; assessment of necrosis or interstitial disease	Useful in the diagnosis of pancreatic calcification, dilated pancreatic ducts, and pancreatic tumors; may not be able to distinguish between inflammatory and neoplastic mass lesions
5. Magnetic resonance cholangiopancreatography (MRCP)	Three-dimensional imaging has been used to produce very good images of the pancreatic-biliary ductal system by a noninvasive technique	Has replaced ERCP as a diagnostic test; noninvasive
6. Endoscopic ultrasonography (EUS)	High-frequency transducer used with EUS can produce very high-resolution images and depict changes in the pancreatic duct and parenchyma with great detail	Can be used to assess gallstones, chronic pancreatitis, and pancreatic carcinoma
7. Endoscopic retrograde cholangiopancreatography (ERCP)	Cannulation of pancreatic and common bile duct permits visualization of pancreatic-biliary ductal system	Primarily a therapeutic procedure; invasive
Pancreatic biopsy with US or CT guidance	Percutaneous aspiration biopsy of mass-forming lesions of the pancreas	High diagnostic yield; laparotomy avoided; can be done with EUS for the evaluation of chronic pancreatitis, autoimmune pancreatitis, and pancreatic carcinoma
Tests of Exocrine Pancreatic Function		
Direct stimulation of the pancreas with analysis of duodenal contents		
1. Secretin test	Secretin leads to increased output of pancreatic juice and HCO_3^-; pancreatic secretory response is related to the functional mass of pancreatic tissue	Sensitive enough to detect occult disease; involves duodenal intubation and fluoroscopic placement of gastroduodenal tube; poorly defined normal enzyme response; overlap in chronic pancreatitis; large secretory reserve capacity of the pancreas; currently done at only a few medical centers
2. Endoscopic secretin test	Replaces need for tube placement duodenum	Sensitive enough to detect occult disease; high negative predictive value; avoids intubation and fluoroscopy; requires sedation
Measurement of intraluminal digestion products		
1. Quantitative stool fat determination	Lack of lipolytic enzymes brings about impaired fat digestion	Reliable reference standard for defining severity of malabsorption; does not distinguish between maldigestion and malabsorption
Measurement of pancreatic enzymes in feces		
1. Fecal elastase	Pancreatic secretion of proteolytic enzymes; not degraded in intestine	Diagnostic accuracy best if value is <100 μg/g performed on a solid stool

in the tissues of the fallopian tubes, lung, thyroid, and tonsils and can be produced by various tumors (carcinomas of the lung, esophagus, breast, and ovary). Isoamylase determinations do not accurately distinguish elevated blood amylase levels from pancreatic or nonpancreatic sources. In patients with unexplained hyperamylasemia, measurement of macroamylase can avoid numerous tests in patients with this rare disorder.

Elevation of ascitic fluid amylase occurs in acute pancreatitis as well as in (1) ascites due to disruption of the main pancreatic duct or a leaking pseudocyst and (2) other abdominal disorders that simulate pancreatitis (e.g., intestinal obstruction, intestinal infarction, or perforated peptic ulcer). Elevation of pleural fluid amylase can occur in acute pancreatitis, chronic pancreatitis, carcinoma of the lung, and esophageal perforation. Lipase is the single best enzyme to measure for the diagnosis of acute pancreatitis. No single blood test is reliable for the diagnosis of acute pancreatitis in patients with renal failure. Pancreatic enzyme elevations are usually less than three times the upper limit of normal. Determining whether a patient with renal failure and abdominal pain has pancreatitis remains a difficult clinical problem. One study found that serum amylase levels were elevated in patients with renal dysfunction only when creatinine clearance was <0.8 mL/s (<50 mL/min). In such patients, the serum amylase level was invariably <500 IU/L in the absence of objective evidence of acute pancreatitis. In that study, serum lipase and trypsin levels paralleled serum amylase values. With these limitations in mind, the recommended screening test for acute pancreatitis in renal disease is serum lipase.

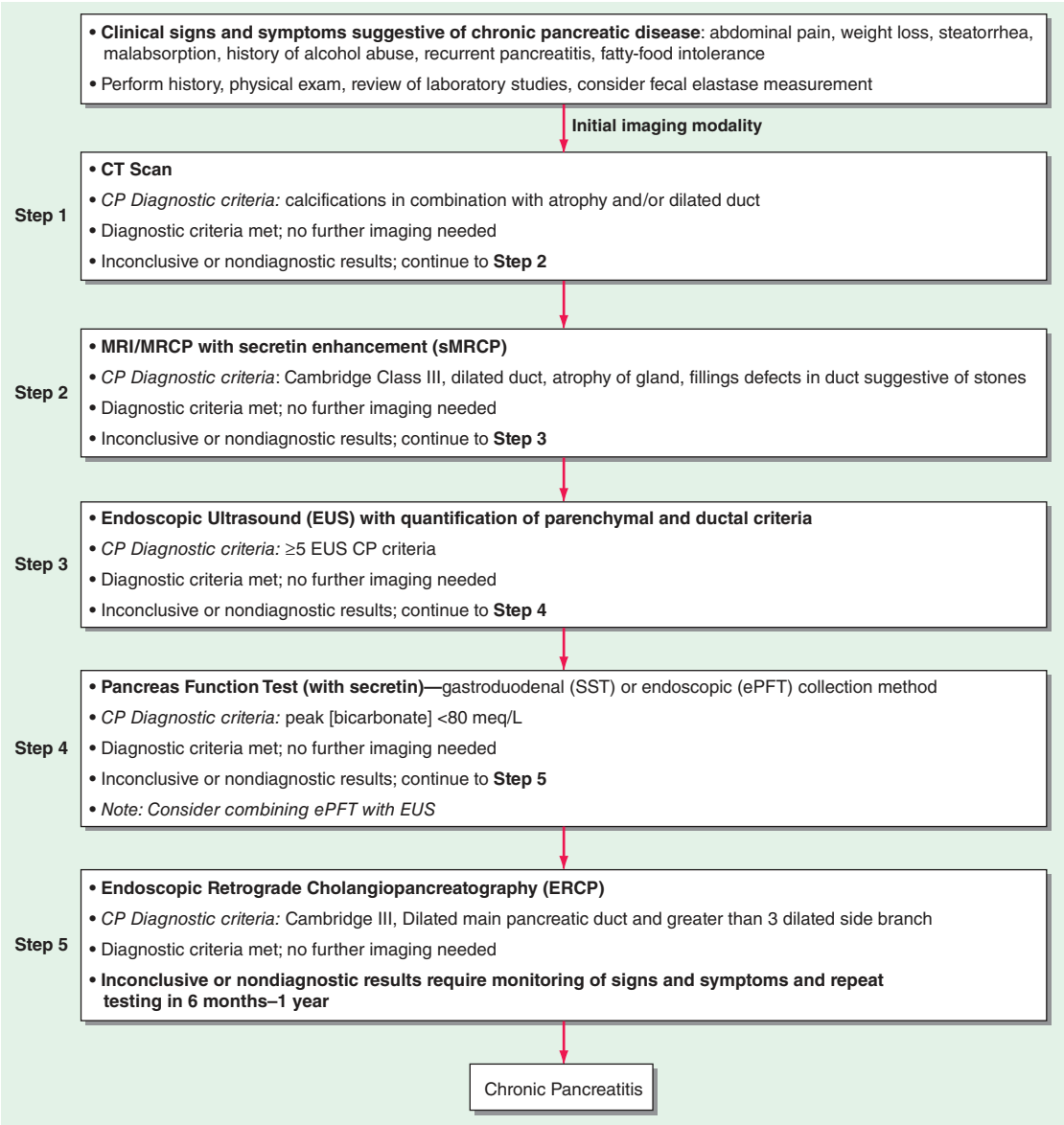


FIGURE 340-1 A stepwise diagnostic approach to the patient with suspected chronic pancreatitis (CP). Endoscopic ultrasonography (EUS) and magnetic resonance cholangiopancreatography (sMRCP/MRCP) are appropriate diagnostic alternatives to endoscopic retrograde cholangiopancreatography (ERCP). CT, computed tomography.

Studies Pertaining to Pancreatic Structure

• **RADIOLOGIC TESTS** Plain films of the abdomen, which once provided useful information in patients with acute and chronic pancreatitis, have been superseded by other more detailed imaging procedures (ultrasound, EUS, CT, MRCP).

Ultrasonography can provide important information in patients with acute pancreatitis, chronic pancreatitis, pseudocysts, and pancreatic carcinoma. Sonographic changes can indicate the presence of edema, inflammation, and calcification (not obvious on plain films of the abdomen), as well as pseudocysts, mass lesions, and gallstones. In acute pancreatitis, the pancreas is characteristically enlarged. In pancreatic pseudocyst, the usual appearance is primarily that of a smooth, round fluid collection. Pancreatic carcinoma distorts the usual landmarks, and mass lesions >3.0 cm are usually detected as localized, solid lesions. US is often the initial investigation for most patients with suspected pancreatic disease. However, obesity and excess intestinal bowel gas can interfere with pancreatic imaging by US studies.

Computed tomography is the best imaging study for initial evaluation of a suspected pancreatic disorder and for the assessment of complications of acute and chronic pancreatitis. It is especially useful in the detection of pancreatic and peripancreatic acute fluid collections, fluid-containing lesions such as pseudocysts, walled-off necrosis, calcium deposits (see Chap. 341, Figs. 341-1, 341-2, and 341-4), and pancreatic neoplasms. Acute pancreatitis is characterized by (1) enlargement of the pancreatic outline, (2) distortion of the pancreatic contour, and/or (3) a pancreatic fluid that has a different attenuation coefficient than normal pancreas. Oral, water-soluble contrast agents are used to opacify the stomach and duodenum during CT scans; this strategy permits more precise delineation of various organs as well as mass lesions. Dynamic CT (using rapid IV administration of contrast) is useful in estimating the extent of pancreatic necrosis and in predicting morbidity and mortality. CT provides clear images much more rapidly and essentially negates artifact caused by patient movement. If acute pancreatitis is confirmed with serology and physical examination findings, CT scan in the first 3 days is not recommended to avoid overuse and minimize costs. The major benefit of CT in acute pancreatitis is the diagnosis of pancreatic necrosis in patients not responding to conservative management within 72 h.

Endoscopic ultrasonography produces high-resolution images of the pancreatic parenchyma and pancreatic duct with a transducer fixed to an endoscope that can be directed onto the surface of the pancreas through the stomach or duodenum. EUS and MRCP have replaced ERCP for diagnostic purposes. EUS allows one to obtain information about the pancreatic duct as well as the parenchyma and has few

TABLE 340-2 Causes of Hyperamylasemia and Hyperamylasuria

Pancreatic Disease
I. Pancreatitis
A. Acute
B. Chronic: ductal obstruction
C. Complications of pancreatitis
1. Pancreatic pseudocyst
2. Ascites caused by pancreatic duct disruption
3. Pancreatic necrosis
II. Pancreatic trauma
III. Pancreatic carcinoma
Non-Pancreatic Disorders
I. Renal insufficiency
II. Salivary gland lesions
A. Mumps
B. Calculus
C. Irradiation sialadenitis
D. Maxillofacial surgery
III. "Tumor" hyperamylasemia
A. Carcinoma of the lung, esophagus, breast, or ovary
IV. Macroamylasemia
V. Burns
VI. Diabetic ketoacidosis
VII. Pregnancy
VIII. Renal transplantation
IX. Cerebral trauma
X. Drugs: opiates
Other Abdominal Disorders
I. Biliary tract disease: cholecystitis, choledocholithiasis
II. Intraabdominal disease
A. Perforated or penetrating peptic ulcer
B. Intestinal obstruction or inflammation
C. Ruptured ectopic pregnancy
D. Peritonitis
E. Aortic aneurysm
F. Postoperative hyperamylasemia

procedure-related complications associated with it, in contrast to the 5–10% of post-ERCP pancreatitis observed. EUS is also helpful in detecting common bile duct stones in acute pancreatitis. Pancreatic masses can also be biopsied via EUS in cases with suspected pancreas cancer, and one can deliver nerve-blocking agents through EUS fine-needle injection in patients suffering from pancreatic pain from chronic pancreatitis or cancer. EUS has been studied as a diagnostic modality for chronic pancreatitis. Criteria for abnormalities on EUS in severe chronic pancreatic disease have been developed. There is general agreement that the presence of five or more of the nine criteria listed in Table 340-3 is highly predictive of chronic pancreatitis. Recent studies comparing EUS and ERCP to the secretin test in patients with unexplained abdominal pain suspected of having chronic pancreatitis show similar diagnostic accuracy in detecting early changes of chronic pancreatitis. The exact role of EUS versus CT, ERCP, or function testing in the early diagnosis of chronic pancreatitis has yet to be clearly defined.

TABLE 340-3 Endoscopic Ultrasonographic Criteria for Chronic Pancreatitis (Total Criteria = 9)

DUCTAL	PARENCHYMAL
Stones	Echogenic strands
Hyperechoic main duct margins	Echogenic foci
Main duct irregularity	Lobular contour
Main duct dilatation	Cyst
Visible side branches	

MRI and MRCP are now being used to view the bile ducts, pancreatic duct, and the pancreas parenchyma in both acute pancreatitis and chronic pancreatitis. For diagnostic imaging in chronic pancreatitis, non-breath-holding and three-dimensional turbo spin-echo techniques are being used to produce superb MRCP images. The main pancreatic duct and common bile duct can be seen well, but there is still a question as to whether changes can be detected consistently in the secondary ducts. The secondary ducts are not visualized in a normal pancreas. Secretin-enhanced MRCP is emerging as a method to better evaluate ductal changes. In anteroposterior imaging, T2 imaging of fluid collections can differentiate necrotic debris from fluid in suspected walled-off necrosis, and T1 imaging can diagnose hemorrhage in suspected pseudoaneurysm rupture.

As mentioned, EUS and MRCP have largely replaced ERCP in the diagnostic evaluation of pancreatic disease. As these techniques become more refined, especially with the administration of secretin, they may well be the diagnostic tests of choice to evaluate the pancreatic duct. ERCP is primarily of therapeutic value after CT, EUS, or MRCP has detected abnormalities requiring invasive endoscopic treatment. ERCP can also be helpful at clarification of equivocal findings discovered with other imaging techniques (see Chap. 341, Fig. 341-1). Pancreatic carcinoma is characterized by stenosis or obstruction of either the pancreatic duct or the common bile duct; both ductal systems are often abnormal (double-duct sign). In chronic pancreatitis, ERCP abnormalities in the main pancreatic duct and side branches have been outlined by the Cambridge classification. The presence of ductal stenosis and irregularity can make it difficult to distinguish chronic pancreatitis from carcinoma. It is important to be aware that ERCP changes interpreted as indicating chronic pancreatitis actually may be due to the effects of aging on the pancreatic duct or sequelae of a recent attack of acute pancreatitis. Although aging may cause impressive ductal alterations, it does not affect the results of pancreatic function tests (i.e., the secretin test). Elevated serum amylase levels after ERCP have been reported in the majority of patients, and clinical pancreatitis in 5–10% of patients. Recent data suggest that pancreatic duct stenting and rectal indomethacin can decrease the incidence of ERCP-induced pancreatitis. ERCP should not be done for diagnostic purposes and should especially be avoided in high-risk patients.

PANCREATIC BIOPSY WITH RADIOLOGIC GUIDANCE Percutaneous aspiration biopsy or a trucut biopsy of a pancreatic mass often distinguishes a pancreatic inflammatory mass from a pancreatic neoplasm.

TESTS OF EXOCRINE PANCREATIC FUNCTION

Pancreatic function tests (Table 340-1) can be divided into the following:

1. *Direct stimulation of the pancreas* by IV infusion of secretin followed by collection and measurement of duodenal contents

   The secretin test, used to detect diffuse pancreatic disease, is based on the physiologic principle that the pancreatic secretory response is directly related to the functional mass of pancreatic tissue. In the standard assay, secretin is given IV in a dose of 0.2 μg/kg of synthetic human secretin as a bolus. Normal values for the standard secretin test are (1) volume output >2 mL/kg per hour, (2) bicarbonate (HCO_3^-) concentration >80 mmol/L, and (3) HCO_3^- output >10 mmol/L in 1 h. The most reproducible measurement, giving the highest level of discrimination between normal subjects and patients with chronic PEI, appears to be the maximal bicarbonate concentration. A cutoff point below 80 mmol/L is considered abnormal and suggestive of abnormal secretory function that is most commonly observed in early chronic pancreatitis.

   There may be a dissociation between the results of the secretin test and other tests of absorptive function. For example, patients with chronic pancreatitis often have abnormally low outputs of HCO_3^- after secretin but have normal fecal fat excretion. Thus the secretin test measures the secretory capacity of ductular epithelium, whereas fecal fat excretion indirectly reflects intraluminal lipolytic activity.

Steatorrhea does not occur until intraluminal levels of lipase are markedly reduced, underscoring the fact that only small amounts of enzymes are necessary for intraluminal digestive activities. It must be emphasized that an abnormal secretin test result suggests that pancreatic ductal secretory function is abnormal. This is commonly observed in chronic pancreatitis.

2. *Measurement of fecal pancreatic enzymes* such as elastase

Measurement of *intraluminal digestion products* (i.e., undigested muscle fibers, stool fat, and fecal nitrogen) is discussed in **Chap. 318**. The amount of human elastase in stool reflects the pancreatic output of this proteolytic enzyme. Decreased elastase-1 activity (FE-1) in stool is an excellent test to detect severe PEI in patients with chronic pancreatitis and cystic fibrosis. FE-1 levels >200 µg/g are normal, levels of 100–200 µg/g are considered mild, and levels <100 µg/g are severe for PEI. Although the test is simple and noninvasive, it can give false-positive results and has a low sensitivity. Fecal levels <50 µg/g are definitive for PEI provided that the stool specimen is solid.

Tests useful in the diagnosis of exocrine pancreatic insufficiency and the differential diagnosis of malabsorption are also discussed in Chaps. 318 and 341.

FURTHER READING

CONWELL DL, WU BU: Chronic pancreatitis: Making the diagnosis. Clin Gastroenterol Hepatol 10:1088, 2012.

CONWELL DL et al: American Pancreatic Association practice guidelines in chronic pancreatitis: Evidence-based report on diagnostic guidelines. Pancreas 43:1143, 2014.

HART PA et al: Endoscopic pancreas fluid collection: Methods and relevance for clinical care and translational science. Am J Gastroenterol 111:1258, 2016.

YADAV D, LOWENFELS AB: The epidemiology of pancreatitis and pancreatic cancer. Gastroenterology 144:1252, 2013.

341 Acute and Chronic Pancreatitis

Darwin L. Conwell, Peter A. Banks, Norton J. Greenberger

BIOCHEMISTRY AND PHYSIOLOGY OF PANCREATIC EXOCRINE SECRETION

GENERAL CONSIDERATIONS

The pancreas secretes 1500–3000 mL of isosmotic alkaline (pH >8) fluid per day containing about 20 enzymes. The pancreatic secretions provide the enzymes and bicarbonate needed to affect the major digestive activity of the gastrointestinal tract and provide an optimal pH for the function of these enzymes.

REGULATION OF PANCREATIC SECRETION

The exocrine pancreas is influenced by intimately interacting hormonal and neural systems. *Gastric acid* is the stimulus for the release of secretin from the duodenal mucosa (S cells), which stimulates the secretion of water and electrolytes from pancreatic ductal cells. Release of cholecystokinin (CCK) from the duodenal and proximal jejunal mucosa (Ito cells) is largely triggered by long-chain fatty acids, essential amino acids (tryptophan, phenylalanine, valine, methionine), and gastric acid itself. CCK evokes an enzyme-rich secretion from acinar cells in the pancreas. The *parasympathetic nervous system* (via the vagus nerve) exerts significant control over pancreatic secretion. Secretion evoked by secretin and CCK depends on permissive roles of vagal afferent and efferent pathways. This is particularly true for enzyme secretion, whereas water and bicarbonate secretions are heavily dependent on the hormonal effects of secretin and to a lesser extent CCK. Also, vagal stimulation affects the release of vasoactive intestinal peptide (VIP), a secretin agonist. Pancreatic exocrine secretion is also influenced by inhibitory neuropeptides such as somatostatin, pancreatic polypeptide, peptide YY, neuropeptide Y, enkephalin, pancreastatin, calcitonin gene–related peptides, glucagon, and galanin. Although pancreatic polypeptide and peptide YY may act primarily on nerves outside the pancreas, somatostatin acts at multiple sites. Nitric oxide (NO) is also an important neurotransmitter.

WATER AND ELECTROLYTE SECRETION

Bicarbonate is the ion of primary physiologic importance within pancreatic secretion. The ductal cells secrete bicarbonate predominantly derived from plasma (93%) more than from intracellular metabolism (7%). Bicarbonate enters the duct lumen through the sodium bicarbonate cotransporter with depolarization caused by chloride efflux through the cystic fibrosis transmembrane conductance regulator (CFTR). Secretin and VIP bind at the basolateral surface and cause an increase in secondary messenger intracellular cyclic AMP, and act on the apical surface of the ductal cells opening the CFTR in promoting secretion. CCK, acting as a neuromodulator, markedly potentiates the stimulatory effects of secretin. Acetylcholine also plays an important role in ductal cell secretion. Intraluminal bicarbonate secreted from the ductal cells helps neutralize gastric acid and creates the appropriate pH for the activity of pancreatic enzymes and bile salts on ingested food.

ENZYME SECRETION

The acinar cell is highly compartmentalized and is concerned with the secretion of pancreatic enzymes. Proteins synthesized by the rough endoplasmic reticulum are processed in the Golgi and then targeted to the appropriate site, whether that be zymogen granules, lysosomes, or other cell compartments. The zymogen granules migrate to the apical region of the acinar cell awaiting the appropriate neural or hormonal stimulatory response. The pancreas secretes amylolytic, lipolytic, and proteolytic enzymes into the duct lumen. *Amylolytic enzymes*, such as amylase, hydrolyze starch to oligosaccharides and to the disaccharide maltose. The *lipolytic enzymes* include lipase, phospholipase A_2, and cholesterol esterase. Bile salts inhibit lipase in isolation, but colipase, another constituent of pancreatic secretion, binds to lipase and prevents this inhibition. Bile salts activate phospholipase A and cholesterol esterase. *Proteolytic enzymes* include endopeptidases (trypsin, chymotrypsin), which act on internal peptide bonds of proteins and polypeptides; exopeptidases (carboxypeptidases, aminopeptidases), which act on the free carboxyl- and amino-terminal ends of peptides, respectively; and elastase. The proteolytic enzymes are secreted as inactive zymogen precursors. Ribonucleases (deoxyribonucleases, ribonuclease) are also secreted. *Enterokinase*, an enzyme found in the duodenal mucosa, cleaves the lysine-isoleucine bond of trypsinogen to form trypsin. Trypsin then activates the other proteolytic zymogens and phospholipase A_2 in a cascade phenomenon. All pancreatic enzymes have pH optima in the alkaline range. The nervous system initiates pancreatic enzyme secretion. The neurologic stimulation is cholinergic, involving extrinsic innervation by the vagus nerve and subsequent innervation by intrapancreatic cholinergic nerves. The stimulatory neurotransmitters are acetylcholine and gastrin-releasing peptides. These neurotransmitters activate calcium-dependent secondary messenger systems, resulting in the release of zymogens into the pancreas duct. VIP is present in intrapancreatic nerves and potentiates the effect of acetylcholine. In contrast to other species, there are no CCK receptors on acinar cells in humans. CCK in physiologic concentrations stimulates pancreatic secretion by stimulating afferent vagal and intrapancreatic nerves.

AUTOPROTECTION OF THE PANCREAS

Autodigestion of the pancreas is prevented by (1) the packaging of pancreatic proteases in precursor (proenzyme) form, (2) intracellular

calcium homeostasis (low intracellular calcium in the cytosol of the acinar cell promotes the destruction of spontaneously activated trypsin), (3) acid-base balance, and (4) the synthesis of protective protease inhibitors (pancreatic secretory trypsin inhibitor [PSTI] or SPINK1), which can bind and inactivate about 20% of intracellular trypsin activity. Chymotrypsin C can also lyse and inactivate trypsin. These protease inhibitors are found in the acinar cell, the pancreatic secretions, and the α_1- and α_2-globulin fractions of plasma. Loss of any of these four protective mechanisms leads to premature enzyme activation, autodigestion, and acute pancreatitis.

ENTEROPANCREATIC AXIS AND FEEDBACK INHIBITION

Pancreatic enzyme secretion is controlled, at least in part, by a negative feedback mechanism induced by the presence of active serine proteases in the duodenum. To illustrate, perfusion of the duodenal lumen with phenylalanine (stimulates early digestion) causes a prompt increase in plasma CCK levels as well as increased secretion of chymotrypsin and other pancreatic enzymes. However, simultaneous perfusion with trypsin (stimulates late digestion) blunts both responses. Conversely, perfusion of the duodenal lumen with protease inhibitors actually leads to enzyme hypersecretion. The available evidence supports the concept that the duodenum contains a peptide called *CCK-releasing factor* (CCK-RF) that is involved in stimulating CCK release. It appears that serine proteases inhibit pancreatic secretion by inactivating a CCK-releasing peptide in the lumen of the small intestine. Thus, the integrative result of both bicarbonate and enzyme secretion depends on a feedback process for both bicarbonate and pancreatic enzymes. Acidification of the duodenum releases secretin, which stimulates vagal and other neural pathways to activate pancreatic duct cells, which secrete bicarbonate. This bicarbonate then neutralizes the duodenal acid, and the feedback loop is completed. Dietary proteins bind proteases, thereby leading to an increase in free CCK-RF. CCK is then released into the blood in physiologic concentrations, acting primarily through the neural pathways (vagal-vagal). This leads to acetylcholine-mediated pancreatic enzyme secretion. Proteases continue to be secreted from the pancreas until the protein within the duodenum is digested. At this point, pancreatic protease secretion is reduced to basic levels, thus completing this step in the feedback process.

TABLE 341-1 Causes of Acute Pancreatitis

Common Causes
Gallstones (including microlithiasis)
Alcohol (acute and chronic alcoholism)
Hypertriglyceridemia
Endoscopic retrograde cholangiopancreatography (ERCP), especially after biliary manometry
Drugs (azathioprine, 6-mercaptopurine, sulfonamides, estrogens, tetracycline, valproic acid, anti-HIV medications, 5-aminosalicylic acid [5-ASA])
Trauma (especially blunt abdominal trauma)
Postoperative (abdominal and nonabdominal operations)
Uncommon Causes
Vascular causes and vasculitis (ischemic-hypoperfusion states after cardiac surgery)
Connective tissue disorders and thrombotic thrombocytopenic purpura (TTP)
Cancer of the pancreas
Hypercalcemia
Periampullary diverticulum
Pancreas divisum
Hereditary pancreatitis
Cystic fibrosis
Renal failure
Infections (mumps, coxsackievirus, cytomegalovirus, echovirus, parasites)
Autoimmune (e.g., type 1 and type 2)
Causes to Consider in Patients with Recurrent Bouts of Acute Pancreatitis without an Obvious Etiology
Occult disease of the biliary tree or pancreatic ducts, especially microlithiasis, biliary sludge
Drugs
Alcohol abuse
Metabolic: Hypertriglyceridemia, hypercalcemia
Anatomic: Pancreas divisum
Pancreatic cancer
Intraductal papillary mucinous neoplasm (IPMN)
Hereditary pancreatitis
Cystic fibrosis
Autoimmune
Idiopathic

ACUTE PANCREATITIS

GENERAL CONSIDERATIONS

Recent U.S. estimates from the Nationwide Inpatient Sample report that acute pancreatitis is the most common inpatient principal gastrointestinal diagnosis. The incidence of acute pancreatitis also varies in different countries and depends on cause (e.g., alcohol, gallstones, metabolic factors, drugs [Table 341-1]). The annual incidence ranges from 13 to 45 cases per 100,000 persons. Acute pancreatitis results in >250,000 hospitalizations per year. The median length of hospital stay is 4 days, with a median hospital cost of $6096 and a mortality of 1%. The estimated cost annually approaches $2.6 billion. Hospitalization rates increase with age, which are 88% higher among blacks, and are higher among males than females. The age-adjusted rate of hospital discharges with an acute pancreatitis diagnosis increased 62% between 1988 and 2004. From 2000 to 2009, the rate increased 30%. Thus, acute pancreatitis is increasing and is a significant burden on health care costs and resource utilization.

ETIOLOGY AND PATHOGENESIS

There are many causes of acute pancreatitis (Table 341-1), but the mechanisms by which these conditions trigger pancreatic inflammation have not been fully elucidated. Gallstones and alcohol account for 80–90% of the acute pancreatitis cases in the United States. Gallstones continue to be the leading cause of acute pancreatitis in most series (30–60%). The risk of acute pancreatitis in patients with at least one gallstone <5 mm in diameter is fourfold greater than that in patients with larger stones. Alcohol is the second most common cause, responsible for 15–30% of cases in the United States. The incidence of pancreatitis in alcoholics is surprisingly low (5/100,000), indicating that in addition to the amount of alcohol ingested, other factors affect a person's susceptibility to pancreatic injury such as cigarette smoking. Acute pancreatitis occurs in 5–10% of patients following endoscopic retrograde cholangiopancreatography (ERCP). Use of a prophylactic pancreatic duct stent and rectal nonsteroidal anti-inflammatory drugs (NSAIDs, indomethicin) has been shown to reduce pancreatitis after ERCP. Risk factors for post-ERCP pancreatitis include minor papilla sphincterotomy, sphincter of Oddi dysfunction, prior history of post-ERCP pancreatitis, age <60 years, >2 contrast injections into the pancreatic duct, and endoscopic trainee involvement.

Hypertriglyceridemia is the cause of acute pancreatitis in 1.3–3.8% of cases; serum triglyceride levels are usually >1000 mg/dL. Most patients with hypertriglyceridemia, when subsequently examined, show evidence of an underlying derangement in lipid metabolism, probably unrelated to pancreatitis. Such patients are prone to recurrent episodes of pancreatitis. Any factor (e.g., drugs or alcohol) that causes an abrupt increase in serum triglycerides can precipitate a bout of acute pancreatitis. Patients with a deficiency of apolipoprotein CII have an increased incidence of pancreatitis; apolipoprotein CII activates lipoprotein lipase, which is important in clearing chylomicrons from the bloodstream. Patients with diabetes mellitus who have developed ketoacidosis and patients who are on certain medications

such as oral contraceptives may also develop high triglyceride levels. Approximately 0.1–2% of cases of acute pancreatitis are drug related. Drugs cause pancreatitis either by a hypersensitivity reaction or by the generation of a toxic metabolite, although in some cases, it is not clear which of these mechanisms is operative (Table 341-1).

Pathologically, acute pancreatitis varies from *interstitial pancreatitis* (pancreas blood supply maintained), which is generally self-limited to *necrotizing pancreatitis* (pancreas blood supply interrupted), in which the extent of necrosis may correlate with the severity of the attack and its systemic complications. Autodigestion is a currently accepted pathogenic theory; according to this theory, pancreatitis results when proteolytic enzymes (e.g., trypsinogen, chymotrypsinogen, proelastase, and lipolytic enzymes such as phospholipase A_2) are activated in the pancreas acinar cell rather than in the intestinal lumen. A number of factors (e.g., endotoxins, exotoxins, viral infections, ischemia, oxidative stress, lysosomal calcium, and direct trauma) are believed to facilitate premature activation of trypsin. Activated proteolytic enzymes, especially trypsin, not only digest pancreatic and peripancreatic tissues but also can activate other enzymes, such as elastase and phospholipase A_2. Spontaneous activation of trypsin also can occur.

ACTIVATION OF PANCREATIC ENZYMES IN THE PATHOGENESIS OF ACUTE PANCREATITIS

Several recent studies have suggested that pancreatitis is a disease that evolves in three phases. The *initial phase* is characterized by intrapancreatic digestive enzyme activation and acinar cell injury. Trypsin activation appears to be mediated by lysosomal hydrolases such as cathepsin B that become colocalized with digestive enzymes in intracellular organelles; it is currently believed that acinar cell injury is the consequence of trypsin activation. The *second phase* of pancreatitis involves the activation, chemoattraction, and sequestration of leukocytes and macrophages in the pancreas, resulting in an enhanced intrapancreatic inflammatory reaction. Neutrophil depletion induced by prior administration of an antineutrophil serum has been shown to reduce the severity of experimentally induced pancreatitis. There is also evidence to support the concept that neutrophils can activate trypsinogen. Thus, intrapancreatic acinar cell activation of trypsinogen could be a two-step process (i.e., an early neutrophil-independent and a later neutrophil-dependent phase). The *third phase* of pancreatitis is due to the effects of activated proteolytic enzymes and cytokines, released by the inflamed pancreas, on distant organs. Activated proteolytic enzymes, especially trypsin, not only digest pancreatic and peripancreatic tissues but also activate other enzymes such as elastase and phospholipase A_2. The active enzymes and cytokines then digest cellular membranes and cause proteolysis, edema, interstitial hemorrhage, vascular damage, coagulation necrosis, fat necrosis, and parenchymal cell necrosis. Cellular injury and death result in the liberation of bradykinin peptides, vasoactive substances, and histamine that can produce vasodilation, increased vascular permeability, and edema with profound effects on many organs. The systemic inflammatory response syndrome (SIRS) and acute respiratory distress syndrome (ARDS), as well as multiorgan failure, may occur as a result of this cascade of local and distant effects.

A number of genetic factors can increase the susceptibility and/or modify the severity of pancreatic injury in acute pancreatitis, recurrent pancreatitis, and chronic pancreatitis. All of the major genetic susceptibility factors center on the control of trypsin activity within the pancreatic acinar cell, in part because they were identified as candidate genes linked to intrapancreatic trypsin control. Five genetic variants have been identified as being associated with susceptibility to pancreatitis. The genes that have been identified include (1) cationic trypsinogen gene (*PRSS1*), (2) pancreatic secretory trypsin inhibitor (*SPINK1*), (3) the cystic fibrosis transmembrane conductance regulator gene (*CFTR*), (4) the chymotrypsin C gene (*CTRC*), and (5) the calcium-sensing receptor (*CASR*). Investigations of other genetic variants are currently under way, and new genes will be added to this list in the future. Multiple medical, ethical, and psychological issues arise when these genes are discovered, and referral to genetic counselors is recommended.

APPROACH TO THE PATIENT

Abdominal Pain

Abdominal pain is the major symptom of acute pancreatitis. Pain may vary from a mild discomfort to severe, constant, and incapacitating distress. Characteristically, the pain, which is steady and boring in character, is located in the epigastrium and periumbilical region, and may radiate to the back, chest, flanks, and lower abdomen. Nausea, vomiting, and abdominal distention due to gastric and intestinal hypomotility and chemical peritonitis are also frequent complaints.

Physical examination frequently reveals a distressed and anxious patient. Low-grade fever, tachycardia, and hypotension are fairly common. Shock is not unusual and may result from (1) hypovolemia secondary to exudation of blood and plasma proteins into the retroperitoneal space; (2) increased formation and release of kinin peptides, which cause vasodilation and increased vascular permeability; and (3) systemic effects of proteolytic and lipolytic enzymes released into the circulation. Jaundice occurs infrequently; when present, it usually is due to edema of the head of the pancreas with compression of the intrapancreatic portion of the common bile duct or passage of a biliary stone or sludge. Erythematous skin nodules due to subcutaneous fat necrosis may rarely occur. In 10–20% of patients, there are pulmonary findings, including basilar rales, atelectasis, and pleural effusion, the latter most frequently left sided. Abdominal tenderness and muscle rigidity are present to a variable degree, but compared with the intense pain, these signs may be less impressive. Bowel sounds are usually diminished or absent. An enlarged pancreas from acute fluid collection, walled off necrosis, or a pseudocyst may be palpable in the upper abdomen later in the course of the disease (i.e., 4–6 weeks). A faint blue discoloration around the umbilicus (Cullen's sign) may occur as the result of hemoperitoneum, and a blue-red-purple or green-brown discoloration of the flanks (Turner's sign) reflects tissue catabolism of hemoglobin from severe necrotizing pancreatitis with hemorrhage.

LABORATORY DATA

Serum amylase and lipase values threefold or more above normal virtually clinch the diagnosis if gut perforation, ischemia, and infarction are excluded. Serum lipase is the preferred test. However, it should be noted that there is no correlation between the severity of pancreatitis and the degree of serum lipase and amylase elevations. After 3–7 days, even with continuing evidence of pancreatitis, total serum amylase values tend to return toward normal. However, pancreatic lipase levels may remain elevated for 7–14 days. It should be recognized that amylase elevations in serum and urine occur in many conditions other than pancreatitis **(see Chap. 340, Table 340-2).** Importantly, patients with *acidemia* (arterial pH ≤7.32) may have spurious elevations in serum amylase. This finding explains why patients with diabetic ketoacidosis may have marked elevations in serum amylase without any other evidence of acute pancreatitis. Serum lipase activity increases in parallel with amylase activity and is more specific than amylase. A serum lipase measurement can be instrumental in differentiating a pancreatic or nonpancreatic cause for hyperamylasemia. *Leukocytosis* (15,000–20,000 leukocytes/μL) occurs frequently. Patients with more severe disease may show hemoconcentration with hematocrit values >44% and/or prerenal azotemia with a blood urea nitrogen (BUN) level >22 mg/dL resulting from loss of plasma into the retroperitoneal space and peritoneal cavity.

Hemoconcentration may be the harbinger of more severe disease (i.e., pancreatic necrosis), whereas azotemia is a significant risk factor for mortality. *Hyperglycemia* is common and is due to multiple factors, including decreased insulin release, increased glucagon release, and an increased output of adrenal glucocorticoids and catecholamines. *Hypocalcemia* occurs in ~25% of patients, and its pathogenesis is incompletely understood. Although earlier studies suggested that the response of the parathyroid gland to a decrease in serum calcium is impaired, subsequent observations have failed to confirm this phenomenon. Intraperitoneal saponification of calcium by fatty acids in areas of fat necrosis

TABLE 341-2 Revised Atlanta Definitions of Morphologic Features of Acute Pancreatitis

MORPHOLOGIC FEATURE	DEFINITION	COMPUTED TOMOGRAPHY CRITERIA
Interstitial pancreatitis	Acute inflammation of the pancreatic parenchyma and peripancreatic tissues, but without recognizable tissue necrosis	Pancreatic parenchyma enhancement by IV contrast agent No findings of peripancreatic necrosis
Necrotizing pancreatitis	Inflammation associated with pancreatic parenchymal necrosis and/or peripancreatic necrosis	Lack of pancreatic parenchymal enhancement by IV contrast agent and/or presence of findings of peripancreatic necrosis (see below—ANC and WON)
Acute pancreatic fluid collection	Peripancreatic fluid associated with interstitial edematous pancreatitis with no associated peripancreatic necrosis. This term applies only to areas of peripancreatic fluid seen within the first 4 weeks after onset of interstitial edematous pancreatitis and without the features of a pseudocyst.	Occurs in the setting of interstitial edematous pancreatitis Homogeneous collection with fluid density Confined by normal peripancreatic fascial planes No definable wall encapsulating the collection Adjacent to pancreas (no intrapancreatic extension)
Pancreatic pseudocyst	An encapsulated collection of fluid with a well-defined inflammatory wall usually outside the pancreas with minimal or no necrosis. This entity usually occurs >4 weeks after onset of interstitial edematous pancreatitis.	Well circumscribed, usually round or oval Homogeneous fluid density No nonliquid component Well-defined wall; that is, completely encapsulated Maturation usually requires >4 weeks after onset of acute pancreatitis; occurs after interstitial edematous pancreatitis
Acute necrotic collection (ANC)	A collection containing variable amounts of both fluid and necrosis associated with necrotizing pancreatitis; the necrosis can involve the pancreatic parenchyma and/or the peripancreatic tissues.	Occurs only in the setting of acute necrotizing pancreatitis Heterogeneous and nonliquid density of varying degrees in different locations (some appear homogeneous early in their course) No definable wall encapsulating the collection Location—intrapancreatic and/or extrapancreatic
Walled-off necrosis (WON)	A mature, encapsulated collection of pancreatic and/or peripancreatic necrosis that has developed a well-defined inflammatory wall. WON usually occurs >4 weeks after onset of necrotizing pancreatitis.	Heterogeneous with liquid and nonliquid density with varying degrees of loculations (some may appear homogeneous) Well-defined wall; that is, completely encapsulated Location—intrapancreatic and/or extrapancreatic Maturation usually requires 4 weeks after onset of acute necrotizing pancreatitis

Source: Data from P Banks et al: Gut 62:102, 2013.

occurs occasionally, with large amounts (up to 6.0 g) dissolved or suspended in ascitic fluid. Such "soap formation" may also be significant in patients with pancreatitis, mild hypocalcemia, and little or no obvious ascites. *Hyperbilirubinemia* (serum bilirubin >4.0 mg/dL) occurs in ~10% of patients. However, jaundice is transient, and serum bilirubin levels return to normal in 4–7 days. Serum alkaline phosphatase and aspartate aminotransferase levels are also transiently elevated, and they parallel serum bilirubin values and may point to gallbladder-related disease or inflammation in the pancreatic head. *Hypertriglyceridemia* occurs in 5–10% of patients, and serum amylase levels in these individuals are often spuriously normal **(Chap. 340)**. Approximately 5–10% of patients have *hypoxemia* (arterial PO_2 ≤60 mm Hg), which may herald the onset of ARDS. Finally, the electrocardiogram is occasionally abnormal in acute pancreatitis with ST-segment and T-wave abnormalities simulating myocardial ischemia.

An abdominal ultrasound is recommended in the emergency ward as the initial diagnostic imaging modality and is most useful to evaluate for gallstone disease and the pancreatic head.

The Revised Atlanta criteria have clearly outlined the morphologic features of acute pancreatitis on computed tomography (CT) scan as follows: (1) interstitial pancreatitis, (2) necrotizing pancreatitis, (3) acute pancreatic fluid collection, (4) pancreatic pseudocyst, (5) acute necrotic collection (ANC), and (6) walled-off necrosis (WON) **(Table 341-2 and Fig. 341-1). Radiologic studies useful in the diagnosis of acute pancreatitis are discussed in Chap. 340 and listed in Table 340-1.**

DIAGNOSIS

Any severe acute pain in the abdomen or back should suggest the possibility of acute pancreatitis. The diagnosis is established by two of the following three criteria: (1) typical abdominal pain in the epigastrium that may radiate to the back, (2) threefold or greater elevation in serum lipase and/or amylase, and (3) confirmatory findings of acute pancreatitis on cross-sectional abdominal imaging. Patients also have associated nausea, emesis, fever, tachycardia, and abnormal findings on abdominal examination. Laboratory studies may reveal leukocytosis, hypocalcemia, and hyperglycemia. Although not required for diagnosis, markers of severity may include hemoconcentration (hematocrit >44%), admission azotemia (BUN >22 mg/dL), SIRS, and signs of organ failure **(Table 341-3)**.

The *differential diagnosis* should include the following disorders: (1) perforated viscus, especially peptic ulcer; (2) acute cholecystitis and biliary colic; (3) acute intestinal obstruction; (4) mesenteric vascular occlusion; (5) renal colic; (6) inferior myocardial infarction; (7) dissecting aortic aneurysm; (8) connective tissue disorders with vasculitis; (9) pneumonia; and (10) diabetic ketoacidosis. It may be difficult to differentiate acute cholecystitis from acute pancreatitis, because an elevated serum amylase may be found in both disorders. Pain of biliary tract origin is more right sided or epigastric than periumbilical or left upper quadrant and can be more severe; ileus is usually absent. Ultrasound is helpful in establishing the diagnosis of cholelithiasis and cholecystitis. Intestinal obstruction due to mechanical factors can be differentiated from pancreatitis by the history of crescendo-decrescendo pain, findings on abdominal examination, and CT of the abdomen showing changes characteristic of mechanical obstruction. Acute mesenteric vascular occlusion is usually suspected in elderly debilitated patients with brisk leukocytosis, abdominal distention, and bloody diarrhea, confirmed by CT or magnetic resonance angiography. Vasculitides secondary to systemic lupus erythematosus and polyarteritis nodosa may be confused with pancreatitis, especially because pancreatitis may develop as a complication of these diseases. Diabetic ketoacidosis is often accompanied by abdominal pain and elevated total serum amylase levels, thus closely mimicking acute pancreatitis. However, the serum lipase level is not elevated in diabetic ketoacidosis.

CLINICAL COURSE, DEFINITIONS, AND CLASSIFICATIONS

The Revised Atlanta Criteria (1) defines phases of acute pancreatitis, (2) outlines severity of acute pancreatitis, and (3) clarifies imaging definitions as outlined below.

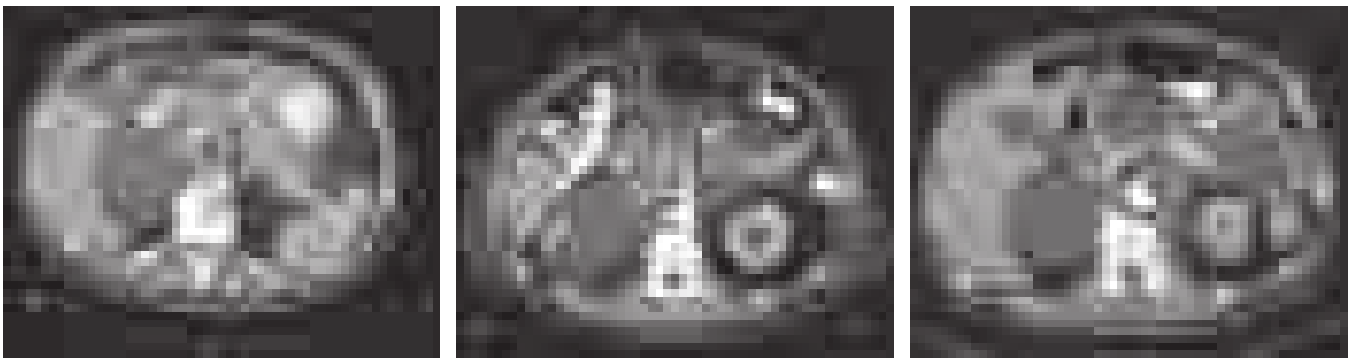


FIGURE 341-1 Acute pancreatitis: computed tomography (CT) evolution. ***A.*** Contrast-enhanced CT scan of the abdomen performed on admission for a patient with clinical and biochemical parameters suggestive of acute pancreatitis. Note the abnormal enhancement of the pancreatic parenchyma (*arrow*) suggestive of interstitial pancreatitis. ***B.*** Contrast-enhanced CT scan of the abdomen performed on the same patient 6 days later for persistent fever and systemic inflammatory response syndrome. The pancreas now demonstrates significant areas of nonenhancement consistent with development of necrosis, particularly in the body and neck region (*arrow*). Note that an early CT scan obtained within the first 48 h of hospitalization may underestimate or miss necrosis. ***C.*** Contrast-enhanced CT scan of the abdomen performed on the same patient 2 months after the initial episode of acute pancreatitis. CT now demonstrates evidence of a fluid collection consistent with walled-off pancreatic necrosis (*arrow*). *(Courtesy of Dr. KJ Mortele, Brigham and Women's Hospital; with permission.)*

Phases of Acute Pancreatitis Two phases of acute pancreatitis have been defined, early (<2 weeks) and late (>2 weeks), which primarily describes the hospital course of the disease. In the *early phase* of acute pancreatitis, which lasts 1–2 weeks, severity is defined by clinical parameters rather than morphologic findings. Most patients exhibit SIRS, and if this persists, patients are predisposed to organ failure. Three organ systems should be assessed to define organ failure: respiratory, cardiovascular, and renal. Organ failure is defined as a score of 2 or more for one of these three organ systems using the modified Marshall scoring system. Persistent organ failure (>48 h) is the most important clinical finding in regard to severity of the acute pancreatitis episode. Organ failure that affects more than one organ is considered multisystem organ failure. CT imaging is usually not needed or recommended during the first 48 h of admission in acute pancreatitis.

The *late phase* is characterized by a protracted course of illness and may require imaging to evaluate for local complications. The important clinical parameter of severity, as in the early phase, is persistent organ failure. These patients may require supportive measures such as renal dialysis, ventilator support, or need for supplemental nutrition via the nasojejunal or parenteral route. The radiographic feature of greatest importance to recognize in this phase is the development of necrotizing pancreatitis on CT imaging. Necrosis generally prolongs hospitalization and, if infected, may require operative, endoscopic, or percutaneous intervention.

TABLE 341-3 Severe Acute Pancreatitis

Risk Factors for Severity
• Age >60 years
• Obesity, BMI >30
• Comorbid disease (Charlson Comorbidity Index)
Markers of Severity at Admission or within 24 h
• SIRS—defined by presence of 2 or more criteria:
• Core temperature <36° or >38°C
• Heart rate >90 beats/min
• Respirations >20/min or Pco_2 <32 mmHg
• White blood cell count >12,000/μL, <4000/μL, or 10% bands
• APACHE II
• Hemoconcentration (hematocrit >44%)
• Admission BUN (>22 mg/dL)
• BISAP Score
• (B) BUN >25 mg/dL
• (I) Impaired mental status
• (S) SIRS: ≥2 of 4 present
• (A) Age >60 years
• (P) Pleural effusion
• Organ failure (Modified Marshall Score)
• Cardiovascular: systolic BP <90 mm Hg, heart rate >130 beats/min
• Pulmonary: Pao_2 <60 mm Hg
• Renal: serum creatinine >2.0 mg %
Markers of Severity during Hospitalization
• Persistent organ failure
• Pancreatic necrosis

Abbreviations: APACHE II, Acute Physiology and Chronic Health Evaluation II; BMI, body mass index; BISAP, Bedside Index of Severity in Acute Pancreatitis; BP, blood pressure; BUN, blood urea nitrogen; SIRS, systemic inflammatory response syndrome.

Severity of Acute Pancreatitis Three severity classifications have also been defined: mild, moderately severe, and severe. *Mild acute pancreatitis* is without local complications or organ failure. Most patients with interstitial acute pancreatitis have mild pancreatitis. In mild acute pancreatitis, the disease is self-limited and subsides spontaneously, usually within 3–7 days after treatment is instituted. Oral intake can be resumed if the patient is hungry, has normal bowel function, and is without nausea and vomiting. Typically, a clear or full liquid diet has been recommended for the initial meal; however, a low-fat solid diet is a reasonable choice following recovery from mild acute pancreatitis.

Moderately severe acute pancreatitis is characterized by transient organ failure (resolves in <48 h) or local or systemic complications in the absence of persistent organ failure. These patients may or may not have necrosis, but may develop a local complication such as a fluid collection that requires a prolonged hospitalization >1 week.

Severe acute pancreatitis is characterized by persistent organ failure (>48 h). Organ failure can be single or multiple. A CT scan or magnetic resonance imaging (MRI) should be obtained to assess for necrosis and/or complications. If a local complication is encountered, management is dictated by clinical symptoms, evidence of infection, maturity of fluid collection, and clinical stability of the patient. Prophylactic antibiotics are not recommended.

Imaging in Acute Pancreatitis Two types of pancreatitis are recognized on imaging as *interstitial* or *necrotizing* based on pancreatic perfusion. CT imaging is best evaluated 3–5 days into hospitalization when patients are not responding to supportive care to look for local complications such as necrosis. Recent studies report the overutilization of CT imaging in acute pancreatitis and its inability to be better than clinical judgment in the early days of acute pancreatitis management. The Revised Atlanta criteria also outline the terminology for local complications and fluid collections along with a CT imaging template to guide reporting of findings. Local morphologic features are summarized in Table 341-1. *Interstitial pancreatitis* occurs in 90–95% of admissions for acute pancreatitis and is characterized by diffuse gland

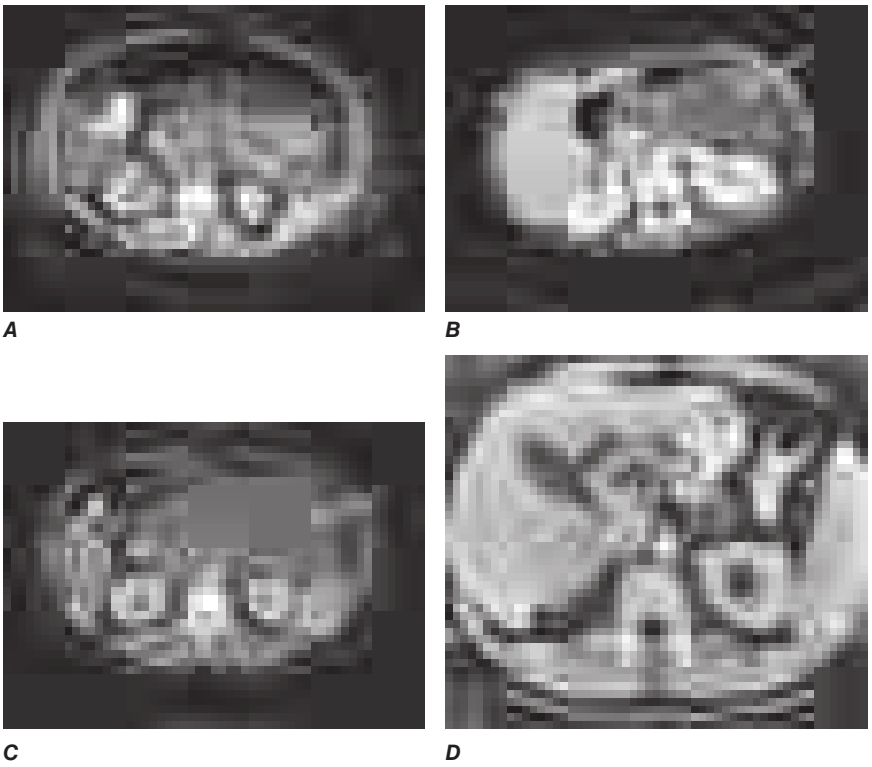


FIGURE 341-2 ***A.*** Acute necrotizing pancreatitis: computed tomography (CT) scan. Contrast-enhanced CT scan showing acute pancreatitis with necrosis. *Arrow* shows partially enhancing body/tail of pancreas surrounded by fluid with decreased enhancement in the neck/body of the pancreas. ***B.*** Acute fluid collection: CT scan. Contrast-enhanced CT scan showing fluid collection in the retroperitoneum (*arrow*) compressing the air-filled stomach arising from the pancreas in a patient with asparaginase-induced acute necrotizing pancreatitis. ***C.*** Walled-off pancreatic necrosis: CT scan. CT scan showing marked walled-off necrosis of the pancreas and peripancreatic area (*arrow*) in a patient with necrotizing pancreatitis. Addendum: In past years, both of these CT findings (Figs. 341-2*B* and 341-2*C*) would have been misinterpreted as pseudocysts. ***D.*** Spiral CT showing a pseudocyst (*small arrow*) with a pseudoaneurysm (*light area in pseudocyst*). Note the demonstration of the main pancreatic duct (*big arrow*), even though this duct is minimally dilated by endoscopic retrograde cholangiopancreatography. *(A, B, C, courtesy of Dr. KJ Mortele, Brigham and Women's Hospital; D, courtesy of Dr. PR Ros, Brigham and Women's Hospital; with permission.)*

enlargement, homogenous contrast enhancement, and mild inflammatory changes or peripancreatic stranding. Symptoms generally resolve with a week of hospitalization. *Necrotizing pancreatitis* occurs in 5–10% of acute pancreatitis admissions and does not evolve until several days of hospitalization. It is characterized by lack of pancreatic parenchymal enhancement by intravenous contrast agent and/or presence of findings of peripancreatic necrosis. According to the Revised Atlanta criteria, the natural history of pancreatic and peripancreatic necrosis is variable because it may remain solid or liquefy, remain sterile or become infected, and persist or disappear over time. CT identification of local complications, particularly necrosis, is critical in patients who are not responding to therapy because patients with infected and sterile necrosis are at greatest risk of mortality (**Figs. 341-1*B*, 341-2, and 341-3**). The median prevalence of organ failure is 54% in necrotizing pancreatitis. The prevalence of organ failure is perhaps slightly higher in infected versus sterile necrosis. With single-organ system failure, the mortality is 3–10% but increases to 47% with multisystem organ failure.

ACUTE PANCREATITIS MANAGEMENT

We will briefly outline the management of patients with acute pancreatitis from the time of diagnosis in the emergency ward to ongoing hospital admission and, finally, to time of discharge, highlighting salient features based on severity and complications. It is important to note that 85–90% of cases of acute pancreatitis are self-limited and subside spontaneously, usually within 3–7 days after initiation of treatment, and do not exhibit organ failure or local complications.

The management of acute pancreatitis begins in the emergency ward. After a diagnosis has been confirmed, aggressive fluid resuscitation is initiated, intravenous analgesics are administered, severity is assessed, and a search for etiologies that may impact acute care is begun. Patients who do not respond to aggressive fluid resuscitation in the emergency ward should be considered for admission to a step-down or intensive care unit for aggressive fluid resuscitation, hemodynamic monitoring, and management of necrosis or organ failure.

Fluid Resuscitation and Monitoring Response to Therapy The most important treatment intervention for acute pancreatitis is safe, aggressive intravenous fluid resuscitation. The patient is made NPO to rest the pancreas and is given intravenous narcotic analgesics to control abdominal pain and supplemental oxygen (2 L) via nasal cannula.

Intravenous fluids of lactated Ringer's or normal saline are initially bolused at 15–20 mL/kg (1050–1400 mL), followed by 2–3 mL/kg per hour (200–250 mL/h), to maintain urine output >0.5 mL/kg per hour. Serial bedside evaluations are required every 6–8 h to assess vital signs, oxygen saturation, and change in physical examination to optimize

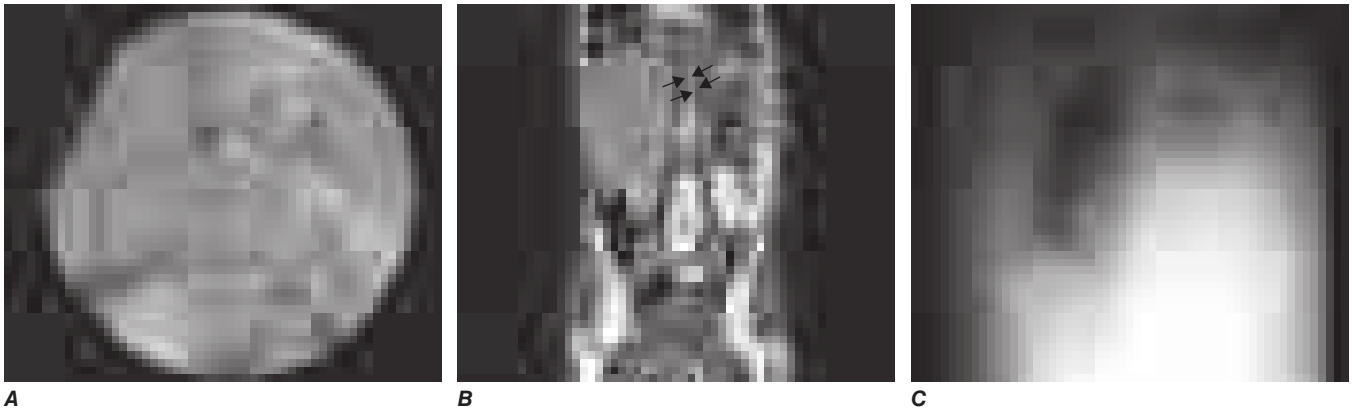


FIGURE 341-3 ***A.*** Pancreaticopleural fistula: pancreatic duct leak on endoscopic retrograde cholangiopancreatography. Pancreatic duct leak (*arrow*) demonstrated at the time of retrograde pancreatogram in a patient with acute exacerbation of alcohol-induced acute or chronic pancreatitis. ***B.*** Pancreaticopleural fistula: computed tomography (CT) scan. Contrast-enhanced CT scan (coronal view) with *arrows* showing fistula tract from pancreatic duct disruption in the pancreatic pleural fistula. ***C.*** Pancreaticopleural fistula: chest x-ray. Large pleural effusion in the left hemithorax from a disrupted pancreatic duct. Analysis of pleural fluid revealed elevated amylase concentration. *(Courtesy of Dr. KJ Mortele, Brigham and Women's Hospital; with permission.)*

fluid resuscitation. Lactated Ringer's solution has been shown to decrease systemic inflammation (lower CRP levels from admission) and may be a better crystalloid than normal saline. A *targeted resuscitation strategy* with measurement of hematocrit and BUN every 8–12 h is recommended to ensure adequacy of fluid resuscitation and monitor response to therapy, noting less aggressive resuscitation strategy may be needed in milder forms of pancreatitis. A rising BUN during hospitalization is not only associated with inadequate hydration but also higher in-hospital mortality.

A decrease in hematocrit and BUN during the first 12–24 h is strong evidence that sufficient fluids are being administered. Serial measurements and bedside assessment for fluid overload are continued, and fluid rates are maintained at the current rate. Adjustments in fluid resuscitation may be required in patients with cardiac, pulmonary, or renal disease. A rise in hematocrit or BUN during serial measurement should be treated with a repeat volume challenge with a 2-L crystalloid bolus followed by increasing the fluid rate by 1.5 mg/kg per hour. If the BUN or hematocrit fails to respond (i.e., remains elevated or does not decrease) to this bolus challenge and increase in fluid rate, consideration of transfer to an intensive care unit is strongly recommended for hemodynamic monitoring.

Assessment of Severity and Hospital Triage Severity of acute pancreatitis should be determined in the emergency ward to assist in patient triage to a regular hospital ward or step-down unit or direct admission to an intensive care unit. The Bedside Index of Severity in Acute Pancreatitis (BISAP) incorporates five clinical and laboratory parameters obtained within the first 24 h of hospitalization (Table 341-3)—BUN >25 mg/dL, impaired mental status (Glasgow coma score <15), SIRS, age >60 years, and pleural effusion on radiography—that can be useful in assessing severity. Presence of three or more of these factors was associated with substantially increased risk for in-hospital mortality among patients with acute pancreatitis. In addition, an elevated hematocrit >44% and admission BUN >22 mg/dL are also associated with more severe acute pancreatitis. Incorporating these indices with the overall patient response to initial fluid resuscitation in the emergency ward can be useful at triaging patients to the appropriate hospital acute care setting.

In general, patients with lower BISAP scores, hematocrits, and admission BUNs tend to respond to initial management and are triaged to a regular hospital ward for ongoing care. If SIRS is not present at 24 h, the patient is unlikely to develop organ failure or necrosis. Therefore, patients with persistent SIRS at 24 h or underlying comorbid illnesses (e.g., chronic obstructive pulmonary disease, congestive heart failure) should be considered for a step-down unit setting if available. Patients with higher BISAP scores and elevations in hematocrit and admission BUN that do not respond to initial fluid resuscitation and exhibit evidence of respiratory failure, hypotension, or organ failure should be considered for direct admission to an intensive care unit.

Special Considerations Based on Etiology A careful history, review of medications, selected laboratory studies (liver profile, serum triglycerides, serum calcium), and an abdominal ultrasound are recommended in the emergency ward to assess for etiologies that may impact acute management. An abdominal ultrasound is the initial imaging modality of choice and will evaluate the gallbladder and common duct and assess the pancreatic head.

GALLSTONE PANCREATITIS Patients with evidence of ascending cholangitis (rising white blood cell count, increasing liver enzymes) should undergo ERCP within 24–48 h of admission. Patients with gallstone pancreatitis are at increased risk of recurrence, and consideration should be given to performing a cholecystectomy during the same admission or within 4–6 weeks of discharge. An alternative for patients who are not surgical candidates would be to perform an endoscopic biliary sphincterotomy before discharge.

HYPERTRIGLYCERIDEMIA Serum triglycerides >1000 mg/dL are associated with acute pancreatitis. Initial therapy may include insulin, heparin, or plasmapheresis. Outpatient therapies include control of diabetes if present, administration of lipid-lowering agents, weight loss, and avoidance of drugs that elevate lipid levels.

Other potential etiologies that may impact acute hospital care include *hypercalcemia*, *autoimmune pancreatitis* (AIP), *post-ERCP pancreatitis*, and *drug-induced pancreatitis*. Treatment of hyperparathyroidism or malignancy is effective at reducing serum calcium. AIP is responsive to glucocorticoid administration. Pancreatic duct stenting and rectal indomethacin administration are effective at decreasing pancreatitis after ERCP. Drugs that cause pancreatitis should be discontinued. Multiple drugs have been implicated, but only about 30 have been challenged (Class 1A) and found to be causative.

Nutritional Therapy A low-fat solid diet can be administered to subjects with mild acute pancreatitis after the abdominal pain has resolved. Enteral nutrition should be considered 2–3 days after admission in subjects with more severe pancreatitis instead of total parenteral nutrition (TPN). Enteral feeding maintains gut barrier integrity, limits bacterial translocation, is less expensive, and has fewer complications than TPN. The choice of gastric versus nasojejunal enteral feeding is currently under investigation.

Management of Local Complications (Table 341-4) Patients exhibiting signs of clinical deterioration despite aggressive fluid resuscitation and hemodynamic monitoring should be assessed for local complications, which may include necrosis, pseudocyst formation, pancreas duct disruption, peripancreatic vascular complications, and extrapancreatic infections. A multidisciplinary team approach is recommended including gastroenterology, surgery, interventional radiology, and intensive care specialists, and consideration should also be made for transfer to a pancreas center.

NECROSIS The management of necrosis requires a multidisciplinary team approach. The benefits of percutaneous aspiration of necrosis with Gram stain and culture should be considered or discussed if there are ongoing signs of possible pancreatic infection such as sustained leukocytosis, fever, or organ failure. There is currently no role for *prophylactic antibiotics* in necrotizing pancreatitis. It is reasonable to start broad-spectrum antibiotics in a patient who appears septic while awaiting the results of Gram stain and cultures. If cultures are negative, the antibiotics should be discontinued to minimize the risk of developing opportunistic or fungal superinfection. Repeated fine-needle aspiration and Gram stain with culture of pancreatic necrosis may be done every 5–7 days in the presence of persistent fever. Repeated CT or MRI imaging should also be considered with any change in clinical course to monitor for complications (e.g., thromboses, hemorrhage, abdominal compartment syndrome).

In general, *sterile necrosis* is most often managed conservatively unless complications arise. Once a diagnosis of *infected necrosis* is established and an organism identified, targeted antibiotics should be instituted. Pancreatic debridement (necrosectomy) should be considered for definitive management of *infected necrosis*, but clinical decisions are generally influenced by response to antibiotic treatment and overall clinical condition. Symptomatic local complications as outlined in the Revised Atlanta criteria may require definitive therapy.

A step-up approach (percutaneous or endoscopic transgastric drainage followed, if necessary, by open necrosectomy) has been successfully reported by some pancreatic centers. One-third of the patients successfully treated with the step-up approach did not require major abdominal surgery. A randomized trial reported advantages to an initial endoscopic approach compared to an initial surgical necrosectomy approach in select patients requiring intervention for symptomatic WON. Taken together, a more conservative approach to the management of infected pancreatic necrosis has evolved under the close supervision of a multidisciplinary team. If conservative therapy can be safely implemented for 4–6 weeks, to allow the pancreatic collections to resolve or "wall-off," surgical or endoscopic intervention is generally much safer and better tolerated by the patient.

PSEUDOCYST The incidence of pseudocyst is low, and most acute collections resolve over time. Less than 10% of patients have persistent fluid collections after 6 weeks that would meet the definition of a pseudocyst. Only symptomatic collections should be drained with surgery or endoscopy or by percutaneous route.

TABLE 341-4 Complications of Acute Pancreatitis

Local
Necrosis
Sterile
Infected
Walled-off necrosis
Pancreatic fluid collections
Pancreatic pseudocyst
Disruption of main pancreatic duct or secondary branches
Pancreatic ascites
Involvement of contiguous organs by necrotizing pancreatitis
Thrombosis of blood vessels (splenic vein, portal vein)
Pancreatic enteric fistula
Bowel infarction
Obstructive jaundice
Systemic
Pulmonary
Pleural effusion
Atelectasis
Mediastinal fluid
Pneumonitis
Acute respiratory distress syndrome
Hypoxemia (unrecognized)
Cardiovascular
Hypotension
Hypovolemia
Nonspecific ST-T changes in electrocardiogram simulating myocardial infarction
Pericardial effusion
Hematologic
Disseminated intravascular coagulation
Gastrointestinal hemorrhage
Peptic ulcer disease
Erosive gastritis
Hemorrhagic pancreatic necrosis with erosion into major blood vessels
Portal vein thrombosis, splenic vein thrombosis, variceal hemorrhage
Renal
Oliguria (<300 mL/d)
Azotemia
Renal artery and/or renal vein thrombosis
Acute tubular necrosis
Metabolic
Hyperglycemia
Hypertriglyceridemia
Hypocalcemia
Encephalopathy
Sudden blindness (Purtscher's retinopathy)
Central nervous system
Psychosis
Fat emboli
Fat necrosis
Subcutaneous tissues (erythematous nodules)
Bone
Miscellaneous (mediastinum, pleura, nervous system)

PANCREATIC DUCT DISRUPTION Pancreatic duct disruption may present with symptoms of increasing abdominal pain or shortness of breath in the setting of an enlarging fluid collection. Diagnosis can be confirmed on magnetic resonance cholangiopancreatography (MRCP) or ERCP. Placement of a bridging pancreatic stent for at least 6 weeks is >90% effective at resolving the leak. Nonbridging stents are less effective (25–50%).

PERIVASCULAR COMPLICATIONS Perivascular complications may include *splenic vein thrombosis* with gastric varices and pseudoaneurysms. *Gastric varices* bleed <5% of the time. Life-threatening bleeding from a ruptured *pseudoaneurysm* can be diagnosed and treated with mesenteric angiography and embolization.

EXTRAPANCREATIC INFECTIONS Hospital-acquired infections occur in up to 20% of patients with acute pancreatitis. Patients should be continually monitored for the development pneumonia, urinary tract infection, and line infection. Continued culturing of urine, monitoring of chest x-rays, and routine changing of intravenous lines are important during hospitalization.

Follow-Up Care Hospitalizations for moderately severe and severe acute pancreatitis can be prolonged and last weeks to months and often involve a period of intensive care unit admission and outpatient rehabilitation or subacute nursing care. Follow-up evaluation should assess for development of diabetes, exocrine insufficiency, recurrent cholangitis, or development of infected fluid collections. As mentioned previously, cholecystectomy should be performed during hospitalization or within 4–6 weeks of discharge if possible for patients with uncomplicated gallstone pancreatitis.

RECURRENT PANCREATITIS

Approximately 25% of patients who have had an attack of acute pancreatitis have a recurrence. The two most common etiologic factors are alcohol and cholelithiasis. In patients with recurrent pancreatitis without an obvious cause, the differential diagnosis should encompass occult biliary tract disease including microlithiasis, hypertriglyceridemia, drugs, pancreatic cancer, pancreas divisum, and cystic fibrosis (Table 341-1). In one series of 31 patients diagnosed initially as having idiopathic or recurrent acute pancreatitis, 23 were found to have occult gallstone disease. Thus, approximately two-thirds of patients with recurrent acute pancreatitis without an obvious cause actually have occult gallstone disease due to microlithiasis. Genetic defects as in hereditary pancreatitis and cystic fibrosis mutations can result in recurrent pancreatitis. Other diseases of the biliary tree and pancreatic ducts that can cause acute pancreatitis include choledochocele; ampullary tumors; pancreas divisum; and pancreatic duct stones, stricture, and tumor. Approximately 2–4% of patients with pancreatic carcinoma present with acute pancreatitis.

PANCREATITIS IN PATIENTS WITH AIDS

The incidence of acute pancreatitis is theoretically increased in patients with AIDS for two reasons: (1) the high incidence of infections involving the pancreas such as infections with cytomegalovirus, *Cryptosporidium*, and the *Mycobacterium avium* complex; and (2) the frequent use by patients with AIDS of medications such as didanosine, pentamidine, trimethoprim-sulfamethoxazole, and protease inhibitors. It should be noted that the incidence has been markedly reduced due to advances in therapy (Chap. 197).

CHRONIC PANCREATITIS AND PANCREATIC EXOCRINE INSUFFICIENCY

PATHOPHYSIOLOGY

Chronic pancreatitis is a disease process characterized by irreversible damage to the pancreas as distinct from the reversible changes noted in acute pancreatitis (Table 341-4). The events that initiate and then perpetuate the inflammatory process in the pancreas are becoming more clearly understood. Irrespective of the mechanism of injury, it is becoming apparent that stellate cell activation that results in cytokine expression and production of extracellular matrix proteins cause acute and chronic inflammation and collagen deposition in the pancreas. Thus, the condition is defined by the presence of histologic abnormalities, including chronic inflammation, fibrosis, and progressive destruction of both exocrine and eventually endocrine tissue (atrophy). A number of etiologies have been associated with chronic pancreatitis resulting in the cardinal manifestations of the disease such as abdominal pain, steatorrhea, weight loss, and diabetes mellitus (Table 341-5).

TABLE 341-5 Chronic Pancreatitis and Pancreatic Exocrine Insufficiency: Tigar-O Classification System

Toxic-metabolic
Alcoholic
Tobacco smoking
Hypercalcemia
Hyperlipidemia
Chronic renal failure
Medications—phenacetin abuse
Toxins—organotin compounds (e.g., dibutylin dichloride, DBTC)
Idiopathic
Early onset
Late onset
Tropical
Genetic
Cationic trypsinogen (*PRSS1*)
Cystic fibrosis transmembrane conductance regulator gene (*CFTR*)
Calcium-sensing receptor (*CASR*)
Chymotrypsin C gene (*CTRC*)
Pancreatic secretory trypsin inhibitor gene (*SPINK1*)
Autoimmune
Type 1 autoimmune chronic pancreatitis
IgG4 systemic
Type 2 autoimmune chronic pancreatitis
Recurrent and severe acute pancreatitis
Postnecrotic (severe acute pancreatitis)
Recurrent acute pancreatitis
Vascular diseases/ischemia
Radiation induced
Obstructive
Pancreas divisum
Duct obstruction (e.g., tumor)
Preampullary duodenal wall cysts
Posttraumatic pancreatic duct scars

Abbreviations: DBTC, dibutylin dichloride; TIGAR-O, toxic-metabolic, idiopathic, genetic, autoimmune, recurrent and severe acute pancreatitis, obstructive.

Although alcohol has been believed to be the primary cause of chronic pancreatitis, other factors contribute to the disease because not all heavy consumers of alcohol develop pancreatic disease. There is also a strong association between smoking and chronic pancreatitis. Cigarette smoke leads to an increased susceptibility to pancreatic autodigestion and predisposes to dysregulation of duct cell CFTR function. Smoking is an independent, dose-dependent risk factor for chronic pancreatitis and recurrent acute pancreatitis. Both continued alcohol and smoking exposure are associated with pancreatic fibrosis, calcifications, and progression of disease.

Recent characterization of pancreatic stellate cells (PSCs) has added insight into the underlying cellular responses behind development of chronic pancreatitis. Specifically, PSCs are believed to play a role in maintaining normal pancreatic architecture that can shift toward fibrogenesis in the case of chronic pancreatitis. The sentinel acute pancreatitis event (SAPE) hypothesis uniformly describes the events in the pathogenesis of chronic pancreatitis. It is believed that alcohol or additional stimuli lead to matrix metalloproteinase–mediated destruction of normal collagen in pancreatic parenchyma, which later allows for pancreatic remodeling. Proinflammatory cytokines, tumor necrosis factor α (TNF-α), interleukin 1 (IL-1), and interleukin 6 (IL-6), as well as oxidant complexes, are able to induce PSC activity with subsequent new collagen synthesis. In addition to being stimulated by cytokines, oxidants, or growth factors, PSCs also possess transforming growth factor β (TGF-β)–mediated self-activating autocrine pathways that may explain disease progression in chronic pancreatitis even after removal of noxious stimuli.

ETIOLOGIC CONSIDERATIONS

Among adults in the United States, alcoholism is the most common cause of clinically apparent chronic pancreatitis, whereas cystic fibrosis is the most frequent cause in children. As many as 25% of adults in the United States with chronic pancreatitis have the *idiopathic* form. Recent investigations have indicated that up to 15% of patients with idiopathic pancreatitis may have pancreatitis due to genetic defects (Table 341-5).

Whitcomb and associates studied several large families with hereditary chronic pancreatitis and were able to identify a genetic defect that affects the gene encoding for trypsinogen. Several additional defects of this gene have also been described. The defect prevents the destruction of prematurely activated trypsin and allows it to be resistant to the intracellular protective effect of trypsin inhibitor. It is hypothesized that this continual activation of digestive enzymes within the gland leads to acute injury and, finally, chronic pancreatitis. Since the initial discovery of the *PRSS1* mutation defect, other genetic diseases have been detected (Table 341-5).

Several other groups of investigators have documented mutations of *CFTR*. This gene functions as a cyclic AMP–regulated chloride channel. In patients with cystic fibrosis, the high concentration of macromolecules can block the pancreatic ducts. It must be appreciated, however, that there is a great deal of heterogeneity in relationship to the *CFTR* gene defect. More than 1000 putative mutations of the *CFTR* gene have been identified. Attempts to elucidate the relationship between the genotype and pancreatic manifestations have been hampered by the number of mutations. The ability to detect *CFTR* mutations has led to the recognition that the clinical spectrum of the disease is broader than previously thought. Two studies have clarified the association between mutations of the *CFTR* gene and another monosymptomatic form of cystic fibrosis (i.e., chronic pancreatitis). It is estimated that in patients with idiopathic pancreatitis, the frequency of a single *CFTR* mutation is 11 times the expected frequency and the frequency of two mutant alleles is 80 times the expected frequency. In these studies, the patients were adults when the diagnosis of pancreatitis was made; none had any clinical evidence of pulmonary disease, and sweat test results were not diagnostic of cystic fibrosis. The prevalence of such mutations is unclear, and further studies are certainly needed. In addition, the therapeutic and prognostic implication of these findings with respect to managing pancreatitis remains to be determined. Long-term follow-up of affected patients is needed. *CFTR* mutations are common in the general population. It is unclear whether the *CFTR* mutation alone can lead to pancreatitis as an autosomal recessive disease. A study evaluated 39 patients with idiopathic chronic pancreatitis to assess the risk associated with these mutations. Patients with two *CFTR* mutations (compound heterozygotes) demonstrated *CFTR* function at a level between that seen in typical cystic fibrosis and cystic fibrosis carriers and had a fortyfold increased risk of pancreatitis. The presence of an *N34S SPINK1* mutation increased the risk twentyfold. A combination of two *CFTR* mutations and an *N34S SPINK1* mutation increased the risk of pancreatitis 900-fold. Knowledge of the genetic defects and downstream alterations in protein expression has led to the development of novel genetic therapy in cystic fibrosis children that potentiates the CFTR channel resulting in improvement in lung function, quality of life, and weight gain. Table 341-5 lists recognized causes of chronic pancreatitis and pancreatic exocrine insufficiency.

AUTOIMMUNE PANCREATITIS (TABLE 341-6)

AIP is an uncommon disorder of presumed autoimmune causation with characteristic laboratory, histologic, and morphologic findings. The nomenclature has recently been simplified to include AIP and idiopathic duct centric pancreatitis (IDCP). In AIP, the pancreas is involved as part of an IgG4 systemic disease (Chap. 361) and meets HISORt criteria as defined below. The characteristic pancreatic histopathologic findings include lymphoplasmacytic infiltrate, storiform fibrosis, and abundant IgG4 cells. IDCP is histologically confirmed IDCP with granulocytic infiltration of the duct wall (termed GEL), but without IgG4 positive cells and systemic involvement. It is a disorder limited to the pancreas only.

TABLE 341-6 Clinical Features of Autoimmune Pancreatitis (AIP)

- Mild symptoms, usually abdominal pain, but without frequent attacks of acute pancreatitis
- Diffuse swelling and enlargement of the pancreas
- Two-thirds of patients present with either obstructive jaundice or a "mass" in the head of the pancreas mimicking carcinoma
- Diffuse irregular narrowing of the pancreatic duct (MRCP or ERCP)
- Increased levels of serum gamma globulins, especially IgG4
- Presence of other autoantibodies (ANA), rheumatoid factor (RF)
- Can occur with other autoimmune diseases: Sjögren's syndrome, primary sclerosing cholangitis, ulcerative colitis, rheumatoid arthritis
- Extrapancreatic bile duct changes such as stricture of the common bile duct and intrahepatic ducts
- Pancreatic calcifications (rare)
- Pancreatic biopsies reveal extensive fibrosis and lymphoplasmacytic infiltration
- Glucocorticoids are effective in alleviating symptoms, decreasing size of the pancreas, and reversing histopathologic changes

Abbreviations: ERCP, endoscopic retrograde cholangiopancreatography; MRCP, magnetic resonance cholangiopancreatography.

Although AIP was initially described as a primary pancreatic disorder, it is now recognized that it is associated with other disorders of presumed autoimmune etiology, and this has been termed IgG4-related disease **(Chap. 361)**. The clinical features include IgG4-associated cholangitis, rheumatoid arthritis, Sjögren's syndrome, ulcerative colitis, mediastinal fibrosis and adenopathy, autoimmune thyroiditis, tubulointerstitial nephritis, retroperitoneal fibrosis, chronic periaortitis, chronic sclerosing sialadenitis, and Mikulicz's disease. Mild symptoms, usually abdominal pain, and recurrent acute pancreatitis are unusual. Furthermore, AIP is not a common cause of idiopathic recurrent pancreatitis.

Weight loss from pancreatic atrophy and new onset of diabetes may also occur in patients that smoke and consume alcohol. An obstructive pattern on liver tests is common (i.e., disproportionately elevated serum alkaline phosphatase and minimally elevated serum aminotransferases). Elevated serum levels of IgG4 provide a marker for the disease, particularly in Western populations. Serum IgG4 is elevated in 2/3 of those with AIP. CT scans reveal abnormalities in the majority of patients and include diffuse enlargement, focal enlargement, and a distinct enlargement at the head of the pancreas. ERCP or MRCP reveals strictures in the bile duct in more than one-third of patients with AIP; these may include common bile duct strictures, intrahepatic bile duct strictures, or proximal bile duct strictures, with accompanying narrowing of the pancreatic portion of the bile duct. This has been termed autoimmune IgG4 cholangitis. Characteristic histologic findings include extensive lymphoplasmacytic infiltrates with dense fibrosis around pancreatic ducts, as well as a lymphoplasmacytic infiltration, resulting in an obliterative phlebitis.

The Mayo Clinic HISORt criteria indicate that AIP can be diagnosed by the presence of at least one or more of the following: (1) *h*istology; (2) *i*maging; (3) *s*erology (elevated serum IgG4 levels); (4) *o*ther organ involvement; and (5) *r*esponse to glucocorticoid *t*herapy, with improvement in pancreatic and extrapancreatic manifestations.

Glucocorticoids have shown efficacy in alleviating symptoms, decreasing the size of the pancreas, and reversing histopathologic features in patients with AIP. Patients may respond dramatically to glucocorticoid therapy within a 2- to 4-week period. Prednisone is usually administered at an initial dose of 40 mg/d for 4 weeks followed by a taper of the daily dosage by 5 mg/wk based on monitoring of clinical parameters. Relief of symptoms, serial changes in abdominal imaging and improvements in liver tests are parameters to follow. A poor response to glucocorticoids over a 2- to 4-week period should raise suspicion of pancreatic cancer or other forms of chronic pancreatitis. A recent multicenter international report reviewed 1064 patients with AIP. Clinical remission was achieved in 99% of AIP and 92% of IDCP patients with steroids. However, disease relapse occurred in 31% of AIP and 9% of IDCP patients. For treatment of disease relapse in AIP, glucocorticoids were successful in 201 of 295 (68%) patients, and azathioprine was successful in 52 of 58 patients (85%). A small number of patients responded favorably to 6-mercaptapurine, rituximab, cyclosporine, and cyclophosphamide. AIP and IDCP are highly responsive to initial glucocorticoid treatment. Relapse is common in AIP patients, especially those with biliary tract strictures. Most relapses occur after glucocorticoids are discontinued. Patients with refractory symptoms and strictures generally require immunomodulator therapy as noted above. Rituximab, a monoclonal antibody directed against B cells has been shown to be very effective at inducing and maintaining remission. Appearance of interval cancers following a diagnosis of AIP is uncommon.

Clinical Features of Chronic Pancreatitis Patients with chronic pancreatitis seek medical attention predominantly because of two symptoms: abdominal pain or maldigestion and weight loss. The abdominal pain may be quite variable in location, severity, and frequency. The pain can be constant or intermittent with frequent pain-free intervals. Eating may exacerbate the pain, leading to a fear of eating with consequent weight loss. The spectrum of abdominal pain ranges from mild to quite severe, with narcotic dependence as a frequent consequence. Maldigestion is manifested as chronic diarrhea, steatorrhea, weight loss, and fatigue. Patients with chronic abdominal pain may or may not progress to maldigestion, and ~20% of patients will present with symptoms of maldigestion without a history of abdominal pain. Patients with chronic pancreatitis have significant morbidity and mortality and use appreciable amounts of societal resources. Despite steatorrhea, clinically apparent deficiencies of fat-soluble vitamins are surprisingly uncommon. Physical findings in these patients are usually unimpressive, so that there is a disparity between the severity of abdominal pain and the physical signs that usually consist of some mild tenderness.

The diagnosis of early or mild chronic pancreatitis can be challenging because there is no biomarker for the disease. In contrast to acute pancreatitis, the serum amylase and lipase levels are usually not strikingly elevated in chronic pancreatitis. Elevation of serum bilirubin and alkaline phosphatase may indicate cholestasis secondary to common bile duct stricture caused by chronic inflammation. Many patients have impaired glucose tolerance with elevated fasting blood glucose levels. The fecal elastase-1 and small-bowel biopsy are useful in the evaluation of patients with suspected pancreatic steatorrhea. The fecal elastase level will be abnormal and small-bowel histology will be normal in such patients. A decrease of fecal elastase level to <100 μg per gram of stool strongly suggests severe pancreatic exocrine insufficiency.

The radiographic evaluation of a patient with suspected chronic pancreatitis usually proceeds from a noninvasive to more invasive approach. Abdominal CT imaging **(Fig. 341-4*A, B*)** is the initial modality of choice, followed by MRI **(Fig. 341-4*C*)**, endoscopic ultrasound, and pancreas function testing. In addition to excluding a pseudocyst and pancreatic cancer, CT may show calcification, dilated ducts, or an atrophic pancreas. Although abdominal CT scanning and MRCP greatly aid in the diagnosis of pancreatic disease, the diagnostic test with the best sensitivity and specificity is the hormone stimulation test using secretin. The secretin test becomes abnormal when ≥60% of the pancreatic exocrine function has been lost. This usually correlates well with the onset of chronic abdominal pain. The role of endoscopic ultrasonography (EUS) in diagnosing early chronic pancreatitis is still being defined. A total of nine endosonographic features have been described in chronic pancreatitis. The presence of five or more features is considered diagnostic of chronic pancreatitis. EUS is not a sensitive enough test for detecting early chronic pancreatitis alone **(Chap. 340)** and may show positive features in patients who have dyspepsia or even normal aging individuals. Recent data suggest that EUS can be combined with endoscopic pancreatic function testing (EUS-ePFT) during a single endoscopy to screen for chronic pancreatitis in patients with chronic abdominal pain. Diffuse calcifications noted on plain film of the abdomen usually indicate significant damage to the pancreas

A

B

C

FIGURE 341-4 ***A.*** Chronic pancreatitis and pancreatic calculi: computed tomography (CT) scan. In this contrast-enhanced CT scan of the abdomen, there is evidence of an atrophic pancreas with multiple calcifications and stones in the parenchyma and dilated pancreatic duct (*arrow*). ***B.*** In this contrast-enhanced CT scan of the abdomen, there is evidence of an atrophic pancreas with multiple calcifications (*arrows*). Note the markedly dilated pancreatic duct seen in this section through the body and tail (*open arrows*). ***C.*** Chronic pancreatitis on magnetic resonance cholangiopancreatography (MRCP): dilated duct with filling defects. Gadolinium-enhanced magnetic resonance imaging/MRCP reveals a dilated pancreatic duct (*arrow*) in chronic pancreatitis with multiple filling defects suggestive of pancreatic duct calculi. *(A, C, courtesy of Dr. KJ Mortele, Brigham and Women's Hospital; with permission.)*

TABLE 341-7 Complications of Chronic Pancreatitis

Chronic abdominal pain	Jaundice
Narcotic addiction	Biliary stricture and/or biliary cirrhosis
Diabetes mellitus/impaired glucose tolerance	Pseudocyst
Gastroparesis	Metabolic bone disease
Malabsorption/maldigestion	Pancreatic cancer

and are pathognomic for chronic pancreatitis (Fig. 341-4*A*). Although alcohol is by far the most common cause of pancreatic calcification, such calcification may also be noted in hereditary pancreatitis, posttraumatic pancreatitis, hypercalcemic pancreatitis, idiopathic chronic pancreatitis, and tropical pancreatitis.

Complications of Chronic Pancreatitis The complications of chronic pancreatitis are protean and are listed in Table 341-7. Although most patients have impaired glucose tolerance, diabetic ketoacidosis and diabetic coma are uncommon. Likewise, end-organ damage (retinopathy, neuropathy, nephropathy) is also uncommon. A nondiabetic retinopathy may be due to either vitamin A and/or zinc deficiency. Gastrointestinal bleeding may occur from peptic ulceration, gastritis, a pseudocyst eroding into the duodenum, arterial bleeding into the pancreatic duct (hemosuccus pancreaticus), or ruptured varices secondary to splenic vein thrombosis due to chronic inflammation of the tail of the pancreas. Jaundice, cholestasis, and biliary cirrhosis may occur from the chronic inflammatory reaction around the intrapancreatic portion of the common bile duct. Twenty years after the diagnosis of calcific chronic pancreatitis, the cumulative risk of pancreatic carcinoma is 4%. Patients with hereditary pancreatitis are at a tenfold higher risk for pancreatic cancer.

TREATMENT

Chronic Pancreatitis

STEATORRHEA

The treatment of steatorrhea with pancreatic enzymes is straightforward even though complete correction of steatorrhea is unusual. Enzyme therapy usually brings diarrhea under control and restores absorption of fat to an acceptable level and affects weight gain. Thus, pancreatic enzyme replacement has been the cornerstone of therapy. In treating steatorrhea, it is important to use a potent pancreatic formulation that will deliver sufficient lipase into the duodenum to correct maldigestion and decrease steatorrhea. In an attempt to standardize the enzyme activity, potency, and bioavailability, the U.S. Food and Drug Administration (FDA) required that all pancreas enzyme drugs in the United States obtain a New Drug Application (NDA) by April 2008. Table 341-8 lists frequently used formulations, but availability will be based on compliance with the FDA mandate. Recent data suggest that dosages up to 80,000–100,000 units of lipase taken during the meal may be necessary to normalize nutritional parameters in malnourished chronic pancreatitis patients, and some may require acid suppression with proton pump inhibitors.

ABDOMINAL PAIN

The management of pain in patients with chronic pancreatitis is problematic. Recent meta-analyses have shown no consistent benefit of enzyme therapy at reducing pain in chronic pancreatitis. In some patients with idiopathic chronic pancreatitis, conventional non-enteric-coated enzyme preparations containing high concentrations of serine proteases may relieve mild abdominal pain or discomfort. The pain relief experienced by these patients actually may be due to improvements in the dyspepsia from maldigestion.

Gastroparesis is also quite common in patients with chronic pancreatitis. It is important to recognize and treat with prokinetic drugs because treatment with enzymes may fail simply because gastric dysmotility is

TABLE 341-8 FDA-Approved Pancreatic Enzyme (Pancrelipase) Preparations

PRODUCT	ENZYME CONTENT/UNIT DOSE, U.S. PHARMACOPEIA UNITS		
	LIPASE[a]	AMYLASE[a]	PROTEASE[a]
Immediate-Release Capsule			
Non-enteric-coated			
Viokace 10,440	10,440	391,550	39,150
Viokace 20,880	20,880	78,300	78,300
Delayed-Release Capsules			
Enteric-coated mini-microspheres			
Creon 3000	3000	15,000	9500
Creon 6000	6000	30,000	19,000
Creon 12,000	12,000	60,000	38,000
Creon 24,000	24,000	120,000	76,000
Enteric-Coated Mini-Tablets			
Ultresa 13,800	13,800	27,600	27,600
Ultresa 20,700	20,700	41,400	41,400
Ultresa 23,000	23,000	46,000	46,000
Enteric-Coated Beads			
Zenpep 3000	3000	16,000	10,000
Zenpep 5000	5000	27,000	17,000
Zenpep 10,000	10,000	55,000	34,000
Zenpep 15,000	15,000	82,000	51,000
Zenpep 20,000	20,000	109,000	68,000
Zenpep 25,000	25,000	136,000	85,000
Enteric-Coated Micro-Tablets			
Pancreaze 4200	4200	17,500	10,000
Pancreaze 10,500	10,500	43,750	25,000
Pancreaze 16,800	16,800	70,000	40,000
Pancreaze 21,000	21,000	61,000	37,000
Bicarbonate-Buffered Enteric-Coated Microspheres			
Pertzye 8000	8000	30,250	28,750
Pertzye 16,000	16,000	60,500	57,500

[a]U.S. Pharmacopeia (USP) units per tablet or capsule.

Note: The FDA has mandated all enzyme manufacturers to submit New Drug Applications (NDAs) for all pancreatic extract drug products after reviewing data that showed substantial variations among currently marketed products. Numerous manufacturers have investigations under way to seek FDA approval for the treatment of exocrine pancreatic insufficiency due to cystic fibrosis or other conditions under the new guidelines for this class of drugs (*www.fda.gov*).

interfering with the delivery of enzymes into the upper intestine. A recent prospective study reported that pregabalin can improve pain in chronic pancreatitis and lower pain medication requirement.

Endoscopic treatment of chronic pancreatitis pain may involve sphincterotomy, stenting, stone extraction, and drainage of a pancreatic pseudocyst. Therapy directed to the pancreatic duct would seem to be most appropriate in the setting of a dominant stricture, especially if a ductal stone has led to obstruction. The use of endoscopic stenting for patients with chronic pain, but without a dominant stricture, has not been subjected to any controlled trials. It is now appreciated that significant complications can occur from stenting (i.e., bleeding, cholangitis, stent migration, pancreatitis, and stent clogging). In patients with large-duct disease usually from alcohol-induced chronic pancreatitis, ductal decompression with *surgical therapy* has been the therapy of choice. Among such patients, 80% seem to obtain immediate relief; however, at the end of 3 years, one-half of the patients have recurrence of pain. Two randomized prospective trials comparing endoscopic to surgical therapy for chronic pancreatitis demonstrated that surgical therapy was superior to endoscopy at decreasing pain and improving quality of life in selected patients with dilated ducts and abdominal pain. This would suggest that chronic pancreatitis patients with dilated ducts and pain should be considered for surgical intervention. The role of preoperative stenting prior to surgery as a predictor of response has yet to be proven.

A Whipple procedure, total pancreatectomy, and autologous islet cell transplantation have been used in selected patients with chronic pancreatitis and abdominal pain refractory to conventional therapy. The patients who have benefited the most from total pancreatectomy have chronic pancreatitis without prior pancreatic surgery or evidence of islet cell insufficiency. The role of this procedure remains to be fully defined but may be an option in lieu of ductal decompression surgery or pancreatic resection in patients with intractable, painful small-duct disease, hereditary pancreatitis and particularly as the standard surgical procedures tend to decrease islet cell yield. Celiac plexus block has not resulted in long-lasting pain relief.

HEREDITARY PANCREATITIS

Hereditary pancreatitis is a rare disease that is similar to chronic pancreatitis except for an early age of onset and evidence of hereditary factors. A genomewide search using genetic linkage analysis identified the hereditary pancreatitis gene on chromosome 7. Mutations in ion codons 29 (exon 2) and 122 (exon 3) of the cationic trypsinogen gene cause autosomal dominant forms of hereditary pancreatitis. The codon 122 mutations lead to a substitution of the corresponding arginine with another amino acid, usually histidine. This substitution, when it occurs, eliminates a fail-safe trypsin self-destruction site necessary to eliminate trypsin that is prematurely activated within the acinar cell. These patients have recurring attacks of severe abdominal pain that may last from a few days to a few weeks. The serum amylase and lipase levels may be elevated during acute attacks but are usually normal. Patients frequently develop pancreatic calcification, diabetes mellitus, and steatorrhea; in addition, they have an increased incidence of pancreatic carcinoma, with the cumulative incidence being as high as 40% by

age 70 years. A recent natural history study of hereditary pancreatitis in >200 patients from France reported that abdominal pain started in childhood at age 10 years, steatorrhea developed at age 29 years, diabetes at age 38 years, and pancreatic carcinoma at age 55 years. Such patients often require surgical ductal decompression for pain relief. Abdominal complaints in relatives of patients with hereditary pancreatitis should raise the question of pancreatic disease.

PSTI, or SPINK1, is a 56-amino-acid peptide that specifically inhibits trypsin by physically blocking its active site. SPINK1 acts as the first line of defense against prematurely activated trypsinogen in the acinar cell. Recently, it has been shown that the frequency of SPINK1 mutations in patients with idiopathic chronic pancreatitis is markedly increased, suggesting that these mutations may be associated with pancreatitis.

PANCREATIC ENDOCRINE TUMORS

Pancreatic endocrine tumors are discussed in Chap. 80.

OTHER CONDITIONS

ANNULAR PANCREAS

When the ventral pancreatic anlage fails to migrate correctly to make contact with the dorsal anlage, the result may be a ring of pancreatic tissue encircling the duodenum. Such an annular pancreas may cause intestinal obstruction in the neonate or the adult. Symptoms of postprandial fullness, epigastric pain, nausea, and vomiting may be present for years before the diagnosis is entertained. The radiographic findings are symmetric dilation of the proximal duodenum with bulging of the recesses on either side of the annular band, effacement but not destruction of the duodenal mucosa, accentuation of the findings in the right anterior oblique position, and lack of change on repeated examinations. The differential diagnosis should include duodenal webs, tumors of the pancreas or duodenum, postbulbar peptic ulcer, regional enteritis, and adhesions. Patients with annular pancreas have an increased incidence of pancreatitis and peptic ulcer. Because of these and other potential complications, the treatment is surgical even if the condition has been present for years. Retrocolic duodenojejeunostomy is the procedure of choice, although some surgeons advocate Billroth II gastrectomy, gastroenterostomy, and vagotomy.

PANCREAS DIVISUM

Pancreas divisum is present in 7–10% of the population and occurs when the embryologic ventral and dorsal pancreatic anlagen fail to fuse, so that pancreatic drainage is accomplished mainly through the accessory papilla. Pancreas divisum is the most common congenital anatomic variant of the human pancreas. Current evidence indicates that this anomaly does not predispose to the development of pancreatitis in the great majority of patients who harbor it. However, the combination of pancreas divisum and a small accessory orifice could result in dorsal duct obstruction. The challenge is to identify this subset of patients with dorsal duct pathology. Cannulation of the dorsal duct by ERCP is not as easily done as is cannulation of the ventral duct. Patients with pancreatitis and pancreas divisum demonstrated by MRCP or ERCP should be treated with conservative measures. In many of these patients, pancreatitis is idiopathic and unrelated to the pancreas divisum. Endoscopic or surgical intervention is indicated only if pancreatitis recurs and no other cause can be found. If marked dilation of the dorsal duct can be demonstrated, surgical ductal decompression should be performed. It should be stressed that the ERCP/MRCP appearance of pancreas divisum (i.e., a small-caliber ventral duct with an arborizing pattern) may be mistaken as representing an obstructed main pancreatic duct secondary to a mass lesion.

MACROAMYLASEMIA

In macroamylasemia, amylase circulates in the blood in a polymer form too large to be easily excreted by the kidney. Patients with this condition demonstrate an elevated serum amylase value and a low urinary amylase value. The presence of macroamylase can be documented by chromatography of the serum. The prevalence of macroamylasemia is 1.5% of the nonalcoholic general adult hospital population. Usually macroamylasemia is an incidental finding and is not related to disease of the pancreas or other organs. Macrolipasemia has now been documented in a few patients with cirrhosis or non-Hodgkin's lymphoma. In these patients, the pancreas appeared normal on ultrasound and CT examination. Lipase was shown to be complexed with immunoglobulin A. Thus, the possibility of *both* macroamylasemia and macrolipasemia should be considered in patients with elevated blood levels of these enzymes.

Acknowledgment

This chapter represents a revised version of chapters by Drs. Norton J. Greenberger, Phillip P. Toskes, and Bechien Wu that were in previous editions of Harrison's.

FURTHER READING

Duggan SN et al: High prevalence of osteoporosis in patients with chronic pancreatitis: A systematic review and meta-analysis. Clin Gastroenterol Hepatol 12:219, 2014.

Forsmark CE: Management of chronic pancreatitis. Gastroenterology 144:1282, 2013.

Hart PA, Zen Y, Chari ST: Recent advances in autoimmune pancreatitis. Gastroenterology 149:39, 2015.

Tenner S et al: American College of Gastroenterology guideline: Management of acute pancreatitis. Am J Gastroenterol 108:1400, 2013.

Wu B, Banks PA: Clinical Management of patients with acute pancreaitis. Gastroenterology 144:1272, 2013.

Yadav D, Lowenfels AB: The epidemiology of pancreatitis and pancreatic cancer. Gastroenterology 144:1252, 2013.

Section 1 The Immune System in Health and Disease

342 Introduction to the Immune System

Barton F. Haynes, Kelly A. Soderberg, Anthony S. Fauci

DEFINITIONS

- *Adaptive immune system*—recently evolved system of immune responses mediated by T and B lymphocytes. Immune responses by these cells are based on specific antigen recognition by clonotypic receptors that are products of genes that rearrange during development and throughout the life of the organism. Additional cells of the adaptive immune system include various types of antigen-presenting cells (APCs).
- *Antibody*—B cell–produced molecules encoded by genes that re-arrange during B cell development consisting of immunoglobulin heavy and light chains that together form the central component of the B cell receptor (BCR) for antigen. Antibody can exist as B cell–surface antigen-recognition molecules or as secreted molecules in plasma and other body fluids.
- *Antigens*—foreign or self-molecules that are recognized by the adaptive and innate immune systems resulting in immune cell triggering, T cell activation, and/or B cell antibody production.
- *Antimicrobial peptides*—small peptides <100 amino acids in length that are produced by cells of the innate immune system and have anti-infectious agent activity.
- *Apoptosis*—the process of *programmed cell death* whereby signaling through various "death receptors" on the surface of cells (e.g., tumor necrosis factor [TNF] receptors, CD95) leads to a signaling cascade that involves activation of the caspase family of molecules and leads to DNA cleavage and cell death. Apoptosis, which does not lead to induction of inordinate inflammation, is to be contrasted with *cell necrosis*, which does lead to induction of inflammatory responses.
- *Autoimmune diseases*—diseases such as systemic lupus erythematosus and rheumatoid arthritis in which cells of the adaptive immune system such as autoreactive T and B cells become overreactive and produce self-reactive T cell and antibody responses.
- *Autoinflammatory diseases*—hereditary disorders such as hereditary periodic fevers (HPFs) characterized by recurrent episodes of severe inflammation and fever due to mutations in controls of the innate inflammatory response, i.e., the inflammasome (see below and Table 342-6). Patients with HPFs also have rashes and serosal and joint inflammation, and some can have neurologic symptoms. Autoinflammatory diseases are different from autoimmune diseases in that evidence for activation of adaptive immune cells such as autoreactive B cells is not present.
- *Autophagy*—lysosomal degradation pathway mechanism of cells to dispose of intracellular debris and damaged organelles. Autophagy by cells of the innate immune system is used to control intracellular infectious agents such as mycobacteria, in part by initiation of phagosome maturation and enhancing major histocompatibility complex (MHC) class II antigen presentation to CD4 T cells.
- *B cell receptor for antigen*—complex of surface molecules that rearrange during postnatal B cell development, made up of surface immunoglobulin (Ig) and associated Ig αβ chain molecules that recognize nominal antigen via Ig heavy- and light-chain variable regions, and signal the B cell to terminally differentiate to make antigen-specific antibody.
- *B lymphocytes*—bone marrow-derived or bursal-equivalent lymphocytes that express surface immunoglobulin (the BCR for antigen) and secrete specific antibody after interaction with antigen.
- *B regulatory cells*—a population of suppressive B cells that aid in the inhibition of inflammation through the release of cytokines such as IL-10.
- *CD classification of human lymphocyte differentiation antigens*—the development of monoclonal antibody technology led to the discovery of a large number of new leukocyte surface molecules. In 1982, the First International Workshop on Leukocyte Differentiation Antigens was held to establish a nomenclature for cell-surface molecules of human leukocytes. From this and subsequent leukocyte differentiation workshops has come the *cluster of differentiation* (CD) classification of leukocyte antigens.
- *CD4 T cell*—T lymphocyte subset that participates in adaptive immunity and helps B cells make antibody.
- *CD8 T cell*—cytotoxic T lymphocyte subset that destroys tumor cells and cells infected with intracellular pathogens.
- *Chemokines*—soluble molecules that direct and determine immune cell movement and circulation pathways.
- *Complement*—cascading series of plasma enzymes and effector proteins whose function is to lyse pathogens and/or target them to be phagocytized by neutrophils and monocyte/macrophage lineage cells of the reticuloendothelial system.
- *Co-stimulatory molecules*—molecules of APCs (such as B7-1 and B7-2 or CD40) that lead to T cell activation when bound by ligands on activated T cells (such as CD28 or CD40 ligand).
- *Crystallopathies*—nanoparticle or microparticle-sized deposits of crystals, misfolded proteins or airborne particulate matter that can stimulate the inflammasome and initiate inflammation and tissue damage.
- *Cytokines*—soluble proteins that interact with specific cellular receptors that are involved in the regulation of the growth and activation of immune cells and mediate normal and pathologic inflammatory and immune responses.
- *Dendritic cells*—myeloid and/or lymphoid lineage APCs of the adaptive immune system. Immature dendritic cells (DCs), or DC precursors, are key components of the innate immune system by responding to infections with production of high levels of cytokines. DCs are key initiators both of innate immune responses via cytokine production and of adaptive immune responses via presentation of antigen to T lymphocytes.
- *Ig Fc receptors*—receptors found on the surface of certain cells including B cells, natural killer (NK) cells, macrophages, neutrophils, and mast cells. Fc receptors bind to antibodies that have attached to invading pathogen-infected cells. They stimulate cytotoxic cells to destroy microbe-infected cells through a mechanism known as antibody-dependent cell-mediated cytotoxicity (ADCC). Examples of important Fc receptors include CD16 (FcγRIIIa), CD23 (FcεR), CD32 (FcγRII), CD64 (FcγRI), and CD89 (FcαR).
- *Inflammasome*—large cytoplasmic complexes of intracellular proteins that link the sensing of microbial products and cellular stress to the proteolytic activation of interleukin (IL)-1β and IL-18 inflammatory cytokines. Activation of molecules in the inflammasome is a key step in the response of the innate immune system for intracellular recognition of microbial and other danger signals in both health and pathologic states.
- *Innate immune system*—ancient immune recognition system of host cells bearing germline-encoded pattern recognition receptors (PRRs) that recognize pathogens and trigger a variety of mechanisms of pathogen elimination. Cells of the innate immune system include NK cell lymphocytes, monocytes/macrophages, DCs, neutrophils, basophils, eosinophils, tissue mast cells, and epithelial cells.
- *Large granular lymphocytes*—lymphocytes of the innate immune system with azurophilic cytotoxic granules that have NK cell activity capable of killing foreign and host cells with few or no self-MHC class I molecules.

- *Natural killer cells*—large granular lymphocytes (LGLs) that kill target cells expressing few or no human leukocyte antigen (HLA) class I molecules, such as malignantly transformed cells and virally infected cells. NK cells express receptors that inhibit killer cell function when self-MHC class I is present.
- *NK T cells*—innate-like lymphocytes that use an invariant T cell receptor (TCR)-α chain combined with a limited set of TCR-β chains and coexpress receptors commonly found on NK cells. NK T cells recognize lipid antigens of bacterial, viral, fungal, and protozoal infectious agents.
- *Pathogen-associated molecular patterns* (PAMPs)—invariant molecular structures expressed by large groups of microorganisms that are recognized by host cellular PRRs in the mediation of innate immunity.
- *Pattern recognition receptors*—germline-encoded receptors expressed by cells of the innate immune system that recognize PAMPs.
- *Polyreactive natural antibodies*—preexisting low-affinity antibodies produced by B cells that cross-react with multiple antigens and are available at the time of infection to bind to and coat the invading pathogen and harness innate responses to slow the infection until an adaptive high-affinity protective antibody response can be made.
- *T lymphocytes*—thymus-derived lymphocytes that mediate adaptive cellular immune responses including T helper, T regulatory, and cytotoxic T lymphocyte effector cell functions.
- *T cell exhaustion*—state of T cells when the persistence of antigen disrupts memory T cell function, resulting in defects in memory T cell responses. Most frequently occurs in malignancies and in chronic viral infections such as HIV-1 and hepatitis C.
- *T cell receptor (TCR) for antigen*—complex of surface molecules that rearrange during postnatal T cell development made up of clonotypic TCR-α and β chains that are associated with the CD3 complex composed of invariant γ, δ, ε, ζ, and η chains. TCR-α and β chains recognize peptide fragments of protein antigen physically bound in APC MHC class I or II molecules, leading to signaling via the CD3 complex to mediate effector functions.
- *T follicular helper T cells (Tfh)*—CD4 T cells regulated by bcl-6 in B cell follicle germinal centers that produce IL-4 and IL-21 and drive B cell differentiation and affinity maturation in peripheral lymphoid tissues such as lymph node and spleen.
- *T_h1 T cells*—CD4 helper T cell subset regulated by transcription factor T-bet and produces interferon (IFN)-gamma, IL-2 and TNF-beta and participates in cell mediated immunity.
- *T_h2 T cells*—CD4 helper T cell subset regulated by transcription factors STAT6 and GATA3 that produces IL-4, IL-5, IL-6, IL-9, IL-10, and IL-13 and regulates antibody and eosinophil responses.
- *T regulatory cell (Treg)*—CD4 and CD8 T cells regulated by the transcription factor Foxp3 that play roles in modulating the immune system to prevent deleterious immune activation. Expression of Foxp3 is a defining Treg marker.
- *T_H9 T cells*—CD4 T cells regulated by the transcription factor PU.1 that secrete IL-9 and enhance inflammation in atopic disease and inflammatory bowel disease as well as mediate antitumor immunity.
- *T_H17 T cells*—CD4 T cells regulated by the transcription factor RORγt that secrete IL-17, IL-22, and IL-26 and play roles in autoimmune inflammatory disorders as well as defend against bacterial and fungal pathogens.
- *Tolerance*—B and T cell nonresponsiveness to antigens that results from encounter with foreign or self-antigens by B and T lymphocytes in the absence of expression of APC co-stimulatory molecules. Tolerance to antigens may be induced and maintained by multiple mechanisms either centrally (B cell deletion in the thymus for T cells or bone marrow for B cells) or peripherally (by cell deletion or anergy at sites throughout the peripheral immune system).

INTRODUCTION

The human immune system has evolved over millions of years from both invertebrate and vertebrate organisms to develop sophisticated defense mechanisms that protect the host from microbes and their virulence factors. The normal immune system has three key properties: A highly diverse repertoire of antigen receptors that enables recognition of a nearly infinite range of pathogens; immune memory, to mount rapid recall immune responses; and immunologic tolerance, to avoid immune damage to normal self-tissues. From invertebrates, humans have inherited the *innate immune system*, an ancient defense system that uses germline-encoded proteins to recognize pathogens. Cells of the innate immune system, such as macrophages, DCs, and NK lymphocytes, recognize PAMPs that are highly conserved among many microbes and use a diverse set of PRR molecules. Important components of the recognition of microbes by the innate immune system include recognition by germline-encoded host molecules, recognition of key microbe virulence factors but not recognition of self-molecules, and nonrecognition of benign foreign molecules or microbes. Upon contact with pathogens, macrophages and NK cells may kill pathogens directly or, in concert with DCs, may activate a series of events that both slow the infection and recruit the more recently evolved arm of the human immune system, the *adaptive immune system*.

Adaptive immunity is found only in vertebrates and is based on the generation of antigen receptors on T and B lymphocytes by gene rearrangements, such that individual T or B cells express unique antigen receptors on their surface capable of specifically recognizing diverse antigens of the myriad infectious agents in the environment. Coupled with finely tuned specific recognition mechanisms that maintain tolerance (nonreactivity) to self-antigens, T and B lymphocytes bring both *specificity* and *immune memory* to vertebrate host defenses.

This chapter describes the cellular components, key molecules **(Table 342-1)**, and mechanisms that make up the innate and adaptive immune systems and describes how adaptive immunity is recruited to the defense of the host by innate immune responses. An appreciation of the cellular and molecular bases of innate and adaptive immune responses is critical to understanding the pathogenesis of inflammatory, autoimmune, infectious, and immunodeficiency diseases.

THE INNATE IMMUNE SYSTEM

All multicellular organisms, including humans, have developed the use of a limited number of surface and intracellular germline-encoded molecules that recognize pathogens. Because of the myriad of human pathogens, host molecules of the human innate immune system sense "danger signals" and either recognize PAMPs, the common molecular structures shared by many pathogens, or recognize host cell molecules produced in response to infection such as heat shock proteins and fragments of the extracellular matrix. PAMPs must be conserved structures vital to pathogen virulence and survival, such as bacterial endotoxin, so that pathogens cannot mutate molecules of PAMPs to evade human innate immune responses. PRRs are host proteins of the innate immune system that recognize PAMPs as host danger signal molecules **(Tables 342-2 and 342-3)**. Thus, recognition of pathogen molecules by hematopoietic and nonhematopoietic cell types leads to activation/production of the complement cascade, cytokines, and antimicrobial peptides as effector molecules. In addition, pathogen PAMPs as host danger signal molecules activate DCs to mature and to express molecules on the DC surface that optimize antigen presentation to respond to foreign antigens.

PATTERN RECOGNITION

Major PRR families of proteins include transmembrane proteins, such as the Toll-like receptors (TLRs) and C-type lectin receptors (CLRs), and cytoplasmic proteins, such as the retinoic acid–inducible gene (RIG)-1-like receptors (RLRs) and NOD-like receptors (NLRs) (Table 342-3). A major group of PRR collagenous glycoproteins with C-type lectin domains are termed *collectins* and include the serum protein mannose-binding lectin (MBL). MBL and other collectins, as well as two other protein families—the pentraxins (such as C-reactive protein and serum amyloid P) and macrophage scavenger receptors—all have the property of opsonizing (coating) bacteria for phagocytosis by macrophages and can also activate the complement cascade to lyse bacteria. Integrins are cell-surface adhesion molecules that affect attachment between cells and the extracellular matrix and mediate signal transduction that reflects the chemical composition of the cell

TABLE 342-1 Human Leukocyte Surface Antigens—The CD Classification of Leukocyte Differentiation Antigens

SURFACE ANTIGEN (OTHER NAMES)	FAMILY	MOLECULAR MASS, kDa	DISTRIBUTION	LIGAND(s)	FUNCTION
CD1a (T6, HTA-1)	Ig	49	CD, cortical thymocytes, Langerhans type of DCs	TCRγδ T cells	CD1 molecules present lipid antigens of intracellular bacteria such as *Mycobacterium leprae* and *M. tuberculosis* to TCRγδT cells
CD1b	Ig	45	CD, cortical thymocytes, Langerhans type of DCs	TCRγδ T cells	
CD1c	Ig	43	DC, cortical thymocytes, subset of B cells, Langerhans type of DCs	TCRγδ T cells	
CD1d	Ig	37	Cortical thymocytes, intestinal epithelium, Langerhans type of DCs	TCRγδ T cells	
CD2 (T12, LFA-2)	Ig	50	T, NK	CD58, CD48, CD59, CD15	Alternative T cell activation, T cell anergy, T cell cytokine production, T- or NK-mediated cytolysis, T cell apoptosis, cell adhesion
CD3 (T3, Leu-4)	Ig	γ:25–28, δ:21–28, ε:20–25, η:21–22, ζ:16	T	Associates with the TCR	T cell activation and function; ζ is the signal transduction component of the CD3 complex
CD4 (T4, Leu-3)	Ig	55	T, myeloid	MHC-II, HIV, gp120, IL-16, SABP	T cell selection, T cell activation, signal transduction with p56*lck*, primary receptor for HIV
CD7 (3A1, Leu-9)	Ig	40	T, NK	K-12 (CD7L)	T and NK cell signal transduction and regulation of IFN-γ, TNF-α production
CD8 (T8, Leu-2)	Ig	34	T	MHC-I	T cell selection, T cell activation, signal transduction with p56*lck*
CD14 (LPS-receptor)	LRG	53–55	M, G (weak), not by myeloid progenitors	Endotoxin (lipopolysaccharide), lipoteichoic acid, PI	TLR4 mediates with LPS and other PAMP activation of innate immunity
CD16a (FcγRIIIa)	Ig	50–80	NK, macrophages, neutrophils	Fc portion of IgG	Mediates phagocytosis and ADCC
CD19 B4	Ig	95	B (except plasma cells), FDC	Not known	Associates with CD21 and CD81 to form a complex involved in signal transduction in B cell development, activation, and differentiation
CD20 (B1)	Unassigned	33–37	B (except plasma cells)	Not known	Cell signaling, may be important for B cell activation and proliferation
CD21 (B2, CR2, EBV-R, C3dR)	RCA	145	Mature B, FDC, subset of thymocytes	C3d, C3dg, iC3b, CD23, EBV	Associates with CD19 and CD81 to form a complex involved in signal transduction in B cell development, activation, and differentiation; Epstein-Barr virus receptor
CD22 (BL-CAM)	Ig	130–140	Mature B	CDw75	Cell adhesion, signaling through association with p72*sky*, p53/56*lyn*, PI3 kinase, SHP1, fLCγ
CD23 (FcεRII, B6, Leu-20, BLAST-2)	C-type lectin	45	B, M, FDC	IgE, CD21, CD11b, CD11c	Regulates IgE synthesis, cytokine release by monocytes
CD28	Ig	44	T, plasma cells	CD80, CD86	Co-stimulatory for T cell activation; involved in the decision between T cell activation and anergy
CD32a (FcγRIIa)	Ig	40	NK, macrophages, neutrophils	Fc portion of IgG	Mediates phagocytosis and ADCC
CD40	TNFR	48–50	B, DC, EC, thymic epithelium, MP, cancers	CD154	B cell activation, proliferation, and differentiation; formation of GCs; isotype switching; rescue from apoptosis
CD45 (LCA, T200, B220)	PTP	180, 200, 210, 220	All leukocytes	Galectin-1, CD2, CD3, CD4	T and B activation, thymocyte development, signal transduction, apoptosis
CD45RA	PTP	210, 220	Subset T, medullary thymocytes, "naive" T	Galectin-1, CD2, CD3, CD4	Isoforms of CD45 containing exon 4 (A), restricted to a subset of T cells
CD45RB	PTP	200, 210, 220	All leukocytes	Galectin-1, CD2, CD3, CD4	Isoforms of CD45 containing exon 5 (B)
CD45RC	PTP	210, 220	Subset T, medullary thymocytes, "naive" T	Galectin-1, CD2, CD3, CD4	Isoforms of CD45 containing exon 6 (C), restricted to a subset of T cells
CD45RO	PTP	180	Subset T, cortical thymocytes, "memory" T	Galectin-1, CD2, CD3, CD4	Isoforms of CD45 containing no differentially spliced exons, restricted to a subset of T cells

(*Continued*)

TABLE 342-1 Human Leukocyte Surface Antigens—The CD Classification of Leukocyte Differentiation Antigens (Continued)

CD64 (FcγRI)	Ig	45–55	Macrophages and monocytes	Fc portion of IgG	Mediates phagocytosis and ADCC
CD80 (B7-1, BB1)	Ig	60	Activated B and T, MP, DC	CD28, CD152	Co-regulator of T cell activation; signaling through CD28 stimulates and through CD152 inhibits T cell activation
CD86 (B7-2, B70)	Ig	80	Subset B, DC, EC, activated T, thymic epithelium	CD28, CD152	Co-regulator of T cell activation; signaling through CD28 stimulates and through CD152 inhibits T cell activation
CD89 (FCαR)	Ig	55–100	Neutrophils, eosinophils, monocytes, and MP	Fc portion of IgG	Mediates phagocytosis and ADCC of IgA-coated pathogens
CD95 (APO-1, Fas)	TNFR	43	Activated T and B	Fas ligand	Mediates apoptosis
CD152 (CTLA-4)	Ig	30–33	Activated T	CD80, CD86	Inhibits T cell proliferation
CD154 (CD40L)	TNF	33	Activated CD4+ T, subset CD8+ T, NK, M, basophil	CD40	Co-stimulatory for T cell activation, B cell proliferation and differentiation
CD279 (PD-1)	Ig	50–55	B, T, TfH	PD-L1, PD-L2	Inhibits T cell proliferation

Abbreviations: ADCC, antibody-dependent cell-mediated cytotoxicity; CTLA, cytotoxic T lymphocyte–associated protein; DC, dendritic cells; EBV, Epstein-Barr virus; EC, endothelial cells; ECM, extracellular matrix; Fcγ RIII, low-affinity IgG receptor isoform A; FDC, follicular dendritic cells; G, granulocytes; GC, germinal center; GPI, glycosyl phosphatidylinositol; HTA, human thymocyte antigen; Ig, immunoglobulin; IgG, immunoglobulin G; LCA, leukocyte common antigen; LPS, lipopolysaccharide; MHC-I, major histocompatibility complex class I; MP, macrophages; Mr, relative molecular mass; NK, natural killer cells; P, platelets; PBT, peripheral blood T cells; PD-1, programmed cell death-1; PI, phosphatidylinositol; PI3K, phosphatidylinositol 3-kinase; PLC, phospholipase C; PTP, protein tyrosine phosphatase; TCR, T cell receptor; TfH, T follicular helper cells; TNF, tumor necrosis factor; TNFR, tumor necrosis factor receptor. For an expanded list of cluster of differentiation (CD) human antigens, see Harrison's Online at *http://www.accessmedicine.com*; and for a full list of CD human antigens from the most recent Human Workshop on Leukocyte Differentiation Antigens (VII), see *http://mpr.nci.nih.gov/prow/*.

Source: Compiled from T Kishimoto et al (eds): *Leukocyte Typing VI*. New York: Garland Publishing, 1997; R Brines et al: Immunology Today 18S:1, 1997; and S Shaw (ed): Protein reviews on the Web. *http://mpr.nci.nih.gov/prow/*.

environment. For example, integrins signal after cells bind bacterial lipopolysaccharide (LPS) and activate phagocytic cells to ingest pathogens

There are multiple connections between the innate and adaptive immune systems; these include (1) a plasma protein, LPS-binding protein, that binds and transfers LPS to the macrophage LPS receptor, CD14; (2) a human family of proteins called *Toll-like receptor proteins* (TLRs), some of which are associated with CD14, bind LPS, and signal epithelial cells, DCs, and macrophages to produce cytokines and upregulate cell-surface molecules that signal the initiation of adaptive immune responses (**Fig. 342-1**, Tables 342-3 and 342-4), and (3) families of intracellular microbial sensors called NLRs and RLRs. Proteins in the Toll family can be expressed on macrophages, DCs, and B cells as well as on a variety of nonhematopoietic cell types, including respiratory epithelial cells. Eleven TLRs have been identified in humans, and 13 TLRs have been identified in mice **(Tables 342-4 and 342-5)**. Upon ligation, TLRs activate a series of intracellular events that lead to the killing of bacteria- and viral-infected cells as well as to the recruitment and ultimate activation of antigen-specific T and B lymphocytes (Fig. 342-1). Importantly, signaling by massive amounts of LPS through TLR4 leads to the release of large amounts of cytokines that mediate LPS-induced shock. Mutations in TLR4 proteins in mice protect from LPS shock, and TLR mutations in humans protect from LPS-induced inflammatory diseases such as LPS-induced asthma (Fig. 342-1).

TABLE 342-2 Major Components of the Innate Immune System

Pattern recognition receptors (PRRs)	Toll-like receptors (TLRs), C-type lectin receptors (CLRs), retinoic acid–inducible gene (RIG)-1-like receptors (RLRs), and NOD-like receptors (NLRs)
Antimicrobial peptides	α-Defensins, β-defensins, cathelin, protegrin, granulysin, histatin, secretory leukoprotease inhibitor, and probiotics
Cells	Macrophages, dendritic cells, NK cells, NK-T cells, neutrophils, eosinophils, mast cells, basophils, and epithelial cells
Complement components	Classic and alternative complement pathway, and proteins that bind complement components
Cytokines	Autocrine, paracrine, endocrine cytokines that mediate host defense and inflammation, as well as recruit, direct, and regulate adaptive immune responses

Abbreviation: NK, natural killer.

Two other families of cytoplasmic PRRs are the NLRs and the RLRs. These families, unlike the TLRs, are composed primarily of soluble intracellular proteins that scan host cell cytoplasm for intracellular pathogens (Tables 342-2 and 342-3).

The intracellular microbial sensors, NLRs, after triggering, form large cytoplasmic complexes termed *inflammasomes*, which are aggregates of molecules including NOD-like receptor pyrin (NLRP) proteins (Table 342-3). Inflammasomes activate inflammatory caspases and IL-1β in the presence of nonbacterial danger signals (cell stress) and bacterial PAMPs. Mutations in inflammasome proteins can lead to chronic inflammation in a group of periodic febrile diseases called *autoinflammatory syndromes* **(Table 342-6)**. Inflammasomes are activated upon sensing of PAMPs. Crystallopathies are diseases caused by tissue crystal deposition such as monosodium urate that can activate the inflammasome and, in the case of urate deposition, can lead to gout with arthritis or renal disease.

EFFECTOR CELLS OF INNATE IMMUNITY

Cells of the innate immune system and their roles in the first line of host defense are listed in Table 342-5. Equally important as their roles in the mediation of innate immune responses are the roles that each cell type plays in recruiting T and B lymphocytes of the adaptive immune system to engage in specific pathogen responses.

Monocytes-Macrophages Monocytes arise from precursor cells within bone marrow **(Fig. 342-2)** and circulate with a half-life ranging from 1 to 3 days. Monocytes leave the peripheral circulation via capillaries and migration into a vast extravascular cellular pool. Tissue macrophages arise from monocytes that have migrated out of the circulation and by in situ proliferation of macrophage precursors in tissue. Common locations where tissue macrophages (and certain of their specialized forms) are found are lymph node, spleen, bone marrow, perivascular connective tissue, serous cavities such as the peritoneum, pleura, skin connective tissue, lung (alveolar macrophages), liver (Kupffer cells), bone (osteoclasts), central nervous system (microglia cells), and synovium (type A lining cells).

In general, monocytes-macrophages are on the first line of defense associated with innate immunity and ingest and destroy microorganisms through the release of toxic products such as hydrogen peroxide (H_2O_2) and nitric oxide (NO). Inflammatory mediators produced by macrophages attract additional effector cells such as neutrophils to the site of infection. Macrophage mediators include prostaglandins;

TABLE 342-3 Pattern Recognition Receptors (PRRs) and Their Ligands

PRRs	LOCALIZATION	LIGAND	ORIGIN OF THE LIGAND
TLR			
TLR1	Plasma membrane	Triacyl lipoprotein	Bacteria
TLR2	Plasma membrane	Lipoprotein	Bacteria, viruses, parasite, self
TLR3	Endolysosome	dsRNA	Virus
TLR4	Plasma membrane	LPS	Bacteria, viruses, self
TLR5	Plasma membrane	Flagellin	Bacteria
TLR6	Plasma membrane	Diacyl lipoprotein	Bacteria, viruses
TLR7 (human TLR8)	Endolysosome	ssRNA	Virus, bacteria, self
TLR9	Endolysosome	CpG-DNA	Virus, bacteria, protozoa, self
TLR10	Endolysosome	Unknown	Unknown
TLR11	Plasma membrane	Profilin-like molecule	Protozoa
RLR			
RIG-I	Cytoplasm	Short dsRNA, triphosphate dsRNA	RNA viruses, DNA virus
MDA5	Cytoplasm	Long dsRNA	RNA viruses (Picornaviridae)
LGP2	Cytoplasm	Unknown	RNA viruses
NLR			
NOD1	Cytoplasm	iE-DAP	Bacteria
NOD2	Cytoplasm	MDP	Bacteria
CLR			
Dectin-1	Plasma membrane	β^2-Glucan	Fungi
Dectin-2	Plasma membrane	β^2-Glucan	Fungi
MINCLE	Plasma membrane	SAP130	Self, fungi

Abbreviations: CLR, C-type lectin receptors; dsRNA, double-strand RNA; iE-DAP, D-glutamyl-meso-diaminopimelic acid moiety; LGP2, Laboratory of Genetics and Physiology 2 protein encoded by the gene *DHX58*; MDA5, melanoma differentiation-associated protein 5; MDP, MurNAc-L-Ala-D-isoGln, also known as muramyl dipeptide; MINCLE, macrophage-inducible C-type lectin; NLR, NOD-like receptor; NOD, NOTCH protein domain; RIG, retinoic acid–inducible gene; RLR, RIG-like receptors; TLR, Toll-like receptor.

Source: Adapted from O Takeuchi, S Akira: Cell 140:805, 2010, with permission.

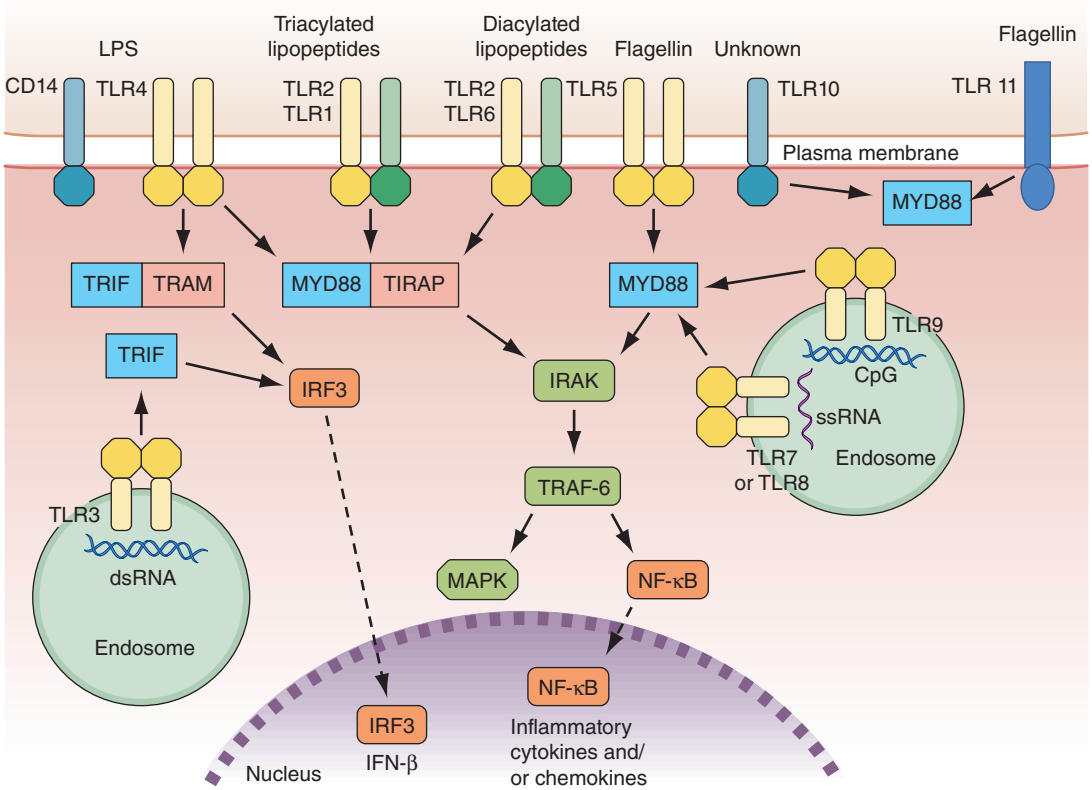


FIGURE 342-1 Overview of major TLR signaling pathways. All TLRs signal through MYD88, with the exception of TLR3. TLR4 and the TLR2 subfamily (TLR1, TLR2, TLR6) also engage TIRAP (Toll-interleukin 1 receptor domain-containing adapter protein). TLR3 signals through TRIF (Toll-interleukin 1 receptor domain-containing adapter-inducing interferon-β). TRIF is also used in conjunction with TRAM (TRIF-related adaptor molecule) in the TLR4-MYD88-independent pathway. *Dashed arrows* indicate translocation into the nucleus. dsRNA, double-strand RNA; IFN, interferon; IRF3, interferon regulatory factor 3; LPS, lipopolysaccharide; MAPK, mitogen-activated protein kinases; NF-κB, nuclear factor-κB; ssRNA, single-strand RNA; TLR, Toll-like receptor. *(Adapted from D van Duin et al: Trends Immunol 27:49, 2006, with permission.)*

TABLE 342-4 The Role of Pattern Recognition Receptors (PRRs) in Modulation of Adaptive Immune Responses

PRR FAMILY	PRRS	LIGAND	DC OR MACROPHAGE CYTOKINE RESPONSE	ADAPTIVE IMMUNE RESPONSE
TLRs	TLR2 (heterodimer with TLR1 or 6)	Lipopeptides	Low IL-12p70	T_H1
		Pam-3-cys (TLR 2/1)	High IL-10	T_H2
		MALP (TLR 2/6)	IL-6	T regulatory
	TLR3	dsRNA	IL-12p70	T_H1
			IFN-α	
			IL-6	
	TLR4	*Escherichia coli* LPS	High IL-12p70	T_H1
			Intermediate IL-10	
			IL-6	
	TLR5	Flagellin	High IL-12p70	T_H1
			Low IL-12p70	T_H2
	TLR7/8	ssRNA	High IL-12p70	T_H1
		Imidazoquinolines	IFN-α	
			IL-6	
	TLR9	CpG DNA	High IL-12p70	T_H1
			Low IL-10	
			IL-6	
			IFN-α	
	TLR10	?	?	?
	TLR11	Profilin-like molecule uropathogenic bacteria	?	?
C-type lectins	DC-SIGN	Env of HIV; core protein of HCV; components of *Mycobacterium tuberculosis*; *Helicobacter pylori*, Lewis Ag	*H. pylori*, Lewis Ag Suppresses IL-12p70 Suppresses TLR signaling in DCs	T_H2 T regulatory
NOD	NOD2	Muramyl dipeptide of peptidoglycan of bacteria	Induces IL-10 in DCs	Weak T cell response (tolerogenic?)
Mannose receptor	Mannose receptor	Mannosylated lipoarabinomannans from bacillus Calmette-Guérin and *M. tuberculosis*	Suppresses IL-12 and TLR signaling in DCs	Weak T cell response? (tolerogenic?)

Abbreviations: CpG, sequences in DNA recognized by TLR-9; DC, dendritic cell; DC-SIGN, DC-specific C-type lectin; dsRNA, double-strand RNA; HCV, hepatitis C virus; HIV, human immunodeficiency virus; LPS, lipopolysaccharide; MALP, macrophage-activating lipopeptide; NOD, NOTCH protein domain; ssRNA, single-strand RNA; T_H1, helper T cell; T_H2, helper T cell; TLR, Toll-like receptor.

Source: B Pulendran: J Immunol 174:2457, 2005. Copyright 2005 The American Association of Immunologists, Inc., with permission.

leukotrienes; platelet activating factor; cytokines such as IL-1, TNF-α, IL-6, and IL-12; and chemokines **(Tables 342-7 to 342-9)**.

Although monocytes-macrophages were originally thought to be the major APCs of the immune system, it is now clear that cell types called *dendritic cells* are the most potent and effective APCs in the body (see below). Monocytes-macrophages mediate innate immune effector functions such as destruction of antibody-coated bacteria, tumor cells, or even normal hematopoietic cells in certain types of autoimmune cytopenias. Monocytes-macrophages ingest bacteria or are infected by viruses, and in doing so, they frequently undergo programmed cell death or *apoptosis*. Macrophages that are infected by intracellular infectious agents are recognized by DCs as infected and apoptotic cells and are phagocytosed by DCs. In this manner, DCs "cross-present" infectious agent antigens of macrophages to T cells. Activated macrophages can also mediate antigen-nonspecific lytic activity and eliminate cell types such as tumor cells in the absence of antibody. This activity is largely mediated by cytokines (i.e., TNF-α and IL-1). Monocytes-macrophages express lineage-specific molecules (e.g., the cell-surface LPS receptor, CD14) as well as surface receptors for a number of molecules, including the Fc region of IgG, activated complement components, and various cytokines (Table 342-7).

Dendritic Cells Human DCs contain several subsets, including myeloid DCs and plasmacytoid DCs. Myeloid DCs can differentiate into either macrophages-monocytes or tissue-specific DCs. In contrast to myeloid DCs, plasmacytoid DCs are inefficient APCs but are potent producers of type I IFN (e.g., IFN-α) in response to viral infections. The maturation of DCs is regulated through cell-to-cell contact and soluble factors, and DCs attract immune effectors through secretion of chemokines. When DCs come in contact with bacterial products, viral proteins, or host proteins released as danger signals from distressed host cells (Fig. 342-2), infectious agent molecules bind to various TLRs and activate DCs to release cytokines and chemokines that drive cells of the innate immune system to become activated to respond to invading organisms, and recruit T and B cells of the adaptive immune system to respond. Plasmacytoid DCs produce antiviral IFN-α that activates NK cell killing of pathogen-infected cells; IFN-α also activates T cells to mature into antipathogen cytotoxic (killer) T cells. Following contact with pathogens, both plasmacytoid and myeloid DCs produce chemokines that attract helper and cytotoxic T cells, B cells, polymorphonuclear cells, and naïve and memory T cells as well as regulatory T cells to ultimately dampen the immune response once the pathogen is controlled. TLR engagement on DCs upregulates MHC class II, B7-1 (CD80), and B7-2 (CD86), which enhance DC-specific antigen presentation and induce cytokine production (Table 342-7). Thus, DCs are important bridges between early (innate) and later (adaptive) immunity. DCs also modulate and determine the types of immune responses induced by pathogens via the TLRs expressed on DCs (TLR7–9 on plasmacytoid DCs, TLR4 on monocytoid DCs) and via the TLR adapter proteins that are induced to associate with TLRs (Fig. 342-1, Table 342-4). In addition, other PRRs, such as C-type lectins, NLRs, and mannose receptors, upon ligation by pathogen products, activate cells of the adaptive immune system and, like TLR stimulation, by a variety of factors, determine the type and quality of the adaptive immune response that is triggered (Table 342-4).

Large Granular Lymphocytes/Natural Killer Cells LGLs or NK cells account for ~5–15% of peripheral blood lymphocytes. NK cells are nonadherent, nonphagocytic cells with large azurophilic

TABLE 342-5 Cells of the Innate Immune System and Their Major Roles in Triggering Adaptive Immunity

CELL TYPE	MAJOR ROLE IN INNATE IMMUNITY	MAJOR ROLE IN ADAPTIVE IMMUNITY
Macrophages	Phagocytose and kill bacteria; produce antimicrobial peptides; bind LPS; produce inflammatory cytokines	Produce IL-1 and TNF-α to upregulate lymphocyte adhesion molecules and chemokines to attract antigen-specific lymphocyte. Produce IL-12 to recruit T_H1 T helper cell responses; upregulate co-stimulatory and MHC molecules to facilitate T and B lymphocyte recognition and activation. Macrophages and dendritic cells, after LPS signaling, upregulate co-stimulatory molecules B7-1 (CD80) and B7-2 (CD86) that are required for activation of pathogen-specific T cells. There are also Toll-like proteins on B cells and dendritic cells that, after LPS ligation, induce CD80 and CD86 on these cells for T cell antigen presentation.
Plasmacytoid dendritic cells (DCs) of lymphoid lineage	Produce large amounts of interferon-α (IFN-α), which has antitumor and antiviral activity, and are found in T cell zones of lymphoid organs; they circulate in blood	IFN-α is a potent activator of macrophage and mature DCs to phagocytose invading pathogens and present pathogen antigens to T and B cells
Myeloid dendritic cells are of two types: interstitial and Langerhans-derived	Interstitial DCs are strong producers of IL-12 and IL-10 and are located in T cell zones of lymphoid organs, circulate in blood, and are present in the interstices of the lung, heart, and kidney; Langerhans DCs are strong producers of IL-12; are located in T cell zones of lymph nodes, skin epithelia, and the thymic medulla; and circulate in blood	Interstitial DCs are potent activators of macrophage and mature DCs to phagocytose invading pathogens and present pathogen antigens to T and B cells
Natural killer (NK) cells	Kill foreign and host cells that have low levels of MHC+ self-peptides. Express NK receptors that inhibit NK function in the presence of high expression of self-MHC.	Produce TNF-α and IFN-γ, which recruit T_H1 helper T cell responses
NK-T cells	Lymphocytes with both T cell and NK surface markers that recognize lipid antigens of intracellular bacteria such as *Mycobacterium tuberculosis* by CD1 molecules and kill host cells infected with intracellular bacteria.	Produce IL-4 to recruit T_H2 helper T cell responses, IgG1 and IgE production
Neutrophils	Phagocytose and kill bacteria, produce antimicrobial peptides	Produce nitric oxide synthase and nitric oxide, which inhibit apoptosis in lymphocytes and can prolong adaptive immune responses
Eosinophils	Kill invading parasites	Produce IL-5, which recruits Ig-specific antibody responses
Mast cells and basophils	Release TNF-α, IL-6, and IFN-γ in response to a variety of bacterial PAMPs	Produce IL-4, which recruits T_H2 helper T cell responses, and recruit IgG1- and IgE-specific antibody responses
Epithelial cells	Produce antimicrobial peptides; tissue-specific epithelia produce mediator of local innate immunity; e.g., lung epithelial cells produce surfactant proteins (proteins within the collectin family) that bind and promote clearance of lung-invading microbes	Produces TGF-β, which triggers IgA-specific antibody responses

Abbreviations: IL-4, IL-5, IL-6, IL-10, and IL-12, interleukin 4, 5, 6, 10, and 12, respectively; MHC, major histocompatibility complex: LPS, lipopolysaccharide; PAMP, pathogen-associated molecular patterns; T_H, helper T cell; TNF-α, tumor necrosis factor-alpha.

Source: Adapted from R Medzhitov, CA Janeway: Curr Opinion Immunol 9:4, 1997. Copyright 1997, with permission from Elsevier.

cytoplasmic granules. NK cells express surface receptors for the Fc portion of IgG (FcR) (CD16) and for NCAM-I (CD56), and many NK cells express T lineage markers, particularly CD8, and proliferate in response to IL-2. NK cells arise in both bone marrow and thymic microenvironments.

Functionally, NK cells share features with both monocytes-macrophages and neutrophils in that they mediate both antibody-dependent cellular cytotoxicity (ADCC) and NK cell activity. ADCC is the binding of an opsonized (antibody-coated) target cell to an Fc receptor-bearing effector cell via the Fc region of antibody, resulting in lysis of the target by the effector cell. NK cell cytotoxicity is the nonimmune (i.e., effector cell never having had previous contact with the target), MHC-unrestricted, non-antibody-mediated killing of target cells, which are usually malignant cell types, transplanted foreign cells, or virus-infected cells. Thus, NK cell cytotoxicity may play an important role in immune surveillance and destruction of malignant and virus-infected host cells. NK cell hyporesponsiveness is also observed in patients with *Chédiak-Higashi syndrome*, an autosomal recessive disease associated with fusion of cytoplasmic granules and defective degranulation of neutrophil lysosomes.

NK cells have a variety of surface receptors that have inhibitory or activating functions and belong to two structural families. These families include the immunoglobulin superfamily and the lectin-like type II transmembrane proteins. NK immunoglobulin superfamily receptors include the killer cell immunoglobulin-like activating or inhibitory receptors (KIRs), many of which have been shown to have HLA class I ligands. The KIRs are made up proteins with either two (KIR2D) or three (KIR3D) extracellular immunoglobulin domains (D). Moreover, their nomenclature designates their function as either inhibitory KIRs with a long (L) cytoplasmic tail and immunoreceptor tyrosine-based inhibitory motif (ITIM) (KIRDL) or activating KIRs with a short (S) cytoplasmic tail (KIRDS). NK cell inactivation by KIRs is a central mechanism to prevent damage to normal host cells. Genetic studies have demonstrated the association of KIRs with viral infection outcome and autoimmune disease **(Table 342-10)**.

In addition to the KIRs, a second set of immunoglobulin superfamily receptors includes the natural cytotoxicity receptors (NCRs), which include NKp46, NKp30, and NKp44. These receptors help to mediate NK cell activation against target cells. The ligands to which NCRs bind on target cells have been recently recognized to be comprised of molecules of pathogens such as influenza, cytomegalovirus, and malaria as well as host molecules expressed on tumor cells.

NK cell signaling is, therefore, a highly coordinated series of inhibiting and activating signals that prevent NK cells from responding to uninfected, nonmalignant self-cells; however, they are activated to attack malignant and virally infected cells **(Fig. 342-3)**. Recent evidence suggests that NK cells, although not possessing rearranging immune recognition genes, may be able to mediate recall for NK cell responses to viruses and for immune responses such as contact hypersensitivity.

Some NK cells express CD3 and invariant TCR-α chains and are termed *NK T cells*. TCRs of NK T cells recognize lipid molecules of intracellular bacteria when presented in the context of CD1d molecules on APCs. Upon activation, NK T cells secrete effector cytokines such as IL-4 and IFN-γ. This mode of recognition of intracellular bacteria such

TABLE 342-6 Diseases Associated with Inflammasome Activity

DISEASE	CLINICAL FEATURES	GENE MUTATED	ETIOLOGIC AGENT	INFLAMMASOME INVOLVEMENT	ANAKINRA RESPONSE[a]
Familial cold autoinflammatory syndrome (FCAS)	Fever, arthralgia, cold-induced urticaria	*NALP3*		Overactive	Yes
Muckle-Wells syndrome (MWS)	Fever, arthralgia, urticaria, sensorineural deafness, amyloidosis	*NLRP3*		Overactive	Yes
Chronic infantile neurologic cutaneous and articular syndrome (CINCA, NOMID)	Fever, severe arthralgia, urticaria, neurologic problems, severe amyloidosis	*NALP3*		Overactive	Yes
Familial Mediterranean fever (FMF)	Fever, peritonitis, pleuritis, amyloidosis	Pyrin		Overactive	Partial
Pyogenic arthritis, pyoderma gangrenosum, and acne syndrome (PAPA)	Pyogenic sterile arthritis	*PSTPIP1*		Overactive	Yes
Hyperimmunoglobulin D syndrome (HIDS)	Arthralgia, abdominal pain, lymphadenopathy	Mevalonate kinase		To be demonstrated	Yes
Tumor necrosis factor receptor-1–associated syndrome (TRAPS)	Fever, abdominal pain, skin lesions	*TNF-R1*		To be demonstrated	Yes
Systemic-onset juvenile idiopathic arthritis (SOJIA)	Chronic joint inflammation		Unknown	To be demonstrated	Yes
Adult-onset Still's disease (AOSD)	Arthralgia, fever		Unknown	To be demonstrated	Yes
Behçet's disease	Arthralgia, uveitis, ulcers		Unknown	To be demonstrated	Yes
Schnitzler's syndrome	Urticaria, fever, arthralgia		Unknown	To be demonstrated	Yes
Gout	Metabolic arthritis, pain		Uric acid (MSU)	Activated	Yes
Pseudogout	Arthritis		CPPD	Activated	Yes
Contact dermatitis	Urticaria		Irritants	Activated	Unknown
Fever syndrome	Fever	*NALP12*		Unknown	Unknown
Hydatidiform mole	Hydatid mole	*NALP7*		Unknown	Unknown
Vitiligo	Skin depigmentation, autoimmunity	*NLRP1*		Overactive	Unknown
Crohn's disease		*NLRP3*		Underactive	Unknown
Multiple Sclerosis		*NLPR3*		Activated	Unknown
Psoriasitic arthritis		*NLRP3*		Activated	Yes

[a]Anakinra is a recombinant interleukin-1 (IL-1) receptor antagonist that functions to block the biologic activity of naturally occurring IL-1.

Abbreviation: CPPD, calcium pyrophosphate dehydrate.

Source: Adapted from F Martinon et al: Ann Rev Immunol 27:229, 2009. Copyright 2009. Reproduced with permission from Annual Reviews Inc.

as *Listeria monocytogenes* and *Mycobacterium tuberculosis* by NK T cells leads to induction of activation of DCs and is thought to be an important innate defense mechanism against these organisms.

The receptors for the Fc portion of IgG (FcγRs) are present on NK cells, B cells, macrophages, neutrophils, and mast cells and mediate interactions of IgG with antibody-coated target cells, such as virally infected cells. Antibody-NK interaction via antibody Fc and NK cell FcR links the adaptive and innate immune systems and regulates the mediation of IgG antibody effector functions such as ADCC. There are both activation and inhibitory FcγRs. Activation FcRs, such as FcγRI (CD64), FcγRIIa (CD32a), and FcγRIIIa (CD16a), are characterized by the presence of an immunoreceptor tyrosine-based activating motif (ITAM) sequence, whereas inhibitory FcRs, such as FcγRIIb (CD32b), contain an ITIM sequence. There is evidence that dysregulation in IgG-FcγR interactions plays roles in arthritis, multiple sclerosis, and systemic lupus erythematosus.

Neutrophils, Eosinophils, and Basophils Granulocytes are present in nearly all forms of inflammation and are amplifiers and effectors of innate immune responses (Fig. 342-2). Unchecked accumulation and activation of granulocytes can lead to host tissue damage, as seen in neutrophil- and eosinophil-mediated *systemic necrotizing vasculitis*. Granulocytes are derived from stem cells in bone marrow. Each type of granulocyte (neutrophil, eosinophil, or basophil) is derived from a different subclass of progenitor cell that is stimulated to proliferate by colony-stimulating factors (Table 342-7). During terminal maturation of granulocytes, class-specific nuclear morphology and cytoplasmic granules appear that allow for histologic identification of granulocyte type.

Neutrophils express Fc receptor IIIa for IgG (CD16a) as well as receptors for activated complement components (C3b or CD35). Upon interaction of neutrophils with antibody-coated (opsonized) bacteria or immune complexes, azurophilic granules (containing myeloperoxidase, lysozyme, elastase, and other enzymes) and specific granules (containing lactoferrin, lysozyme, collagenase, and other enzymes) are released, and microbicidal superoxide radicals (O_2^-) are generated at the neutrophil surface. The generation of superoxide leads to inflammation by direct injury to tissue and by alteration of macromolecules such as collagen and DNA.

Eosinophils are potent cytotoxic effector cells for various parasitic organisms. In *Nippostrongylus brasiliensis* helminth infection, eosinophils are important cytotoxic effector cells for removal of these parasites. Key to regulation of eosinophil cytotoxicity to *N. brasiliensis* worms are antigen-specific T helper cells that produce IL-4, thus providing an example of regulation of innate immune responses by adaptive immunity antigen-specific T cells. Intracytoplasmic contents of eosinophils, such as major basic protein, eosinophil cationic protein, and eosinophil-derived neurotoxin, are capable of directly damaging tissues and may be responsible in part for the organ system dysfunction in the *hypereosinophilic syndromes* **(Chap. 60)**. Because the

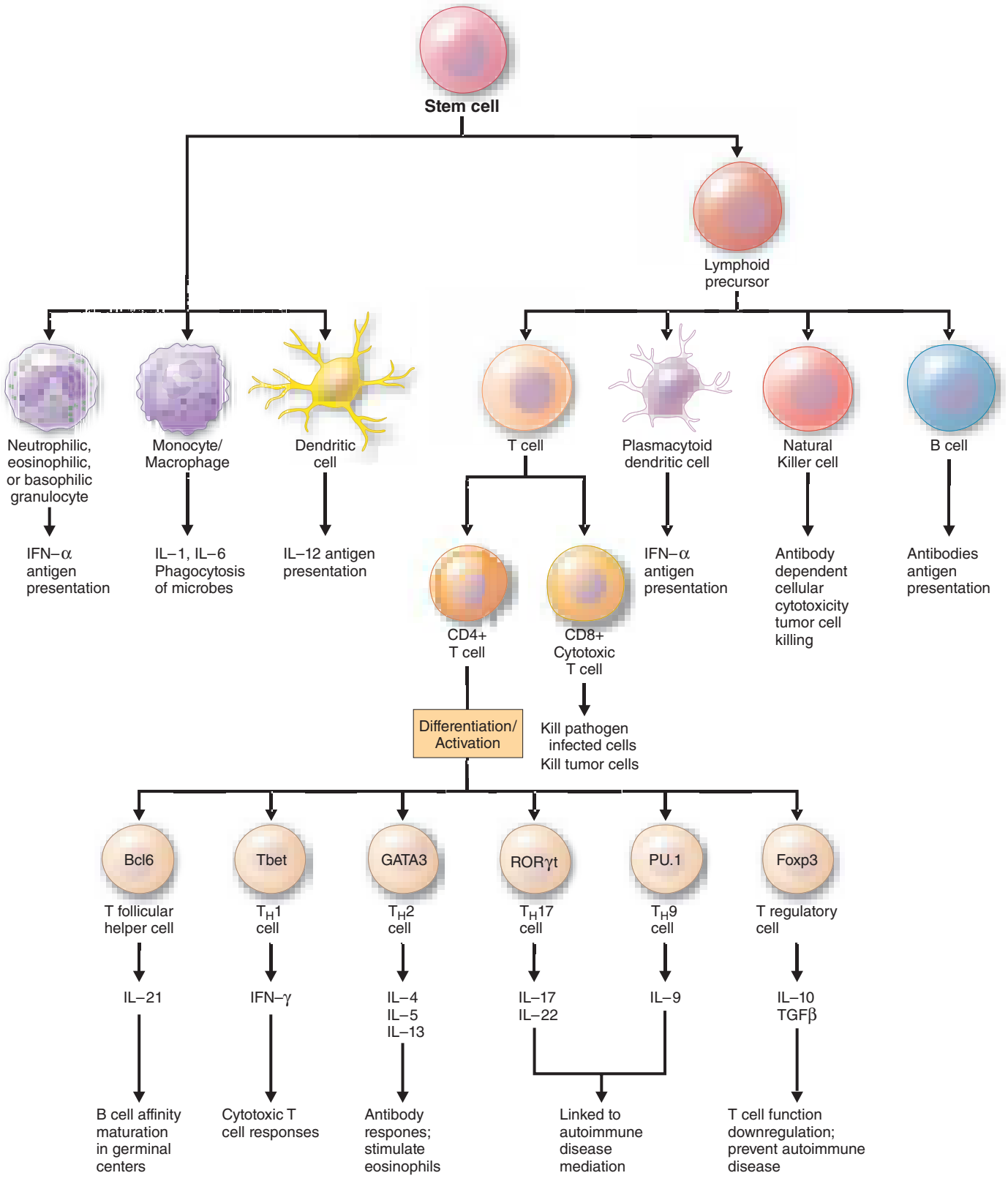


FIGURE 342-2 Model of immune effector cell development. Hematopoietic stem cells differentiate into T cells, antigen-presenting dendritic cells, natural killer cells, macrophages, granulocytes, or B cells. Foreign antigen is processed by dendritic cells, macrophages and B cells, and peptide fragments of foreign antigen are presented to CD4+ and/or CD8+ T cells. CD8+ T cell activation leads to induction of cytotoxic T lymphocyte (CTL) or killer T cell generation, as well as induction of cytokine-producing CD8+ cytotoxic T cells. Granulocytes (neutrophils, eosinophils, or basophils) are effector cells of the innate immune system and mediate anti-infectious agent activity by cytokine production, infectious agent killing or both. T_H1 CD4+ T cells play an important role in defense against intracellular microbes and help in the generation of CD8+ cytotoxic T cells. T_H2 CD4+ T cells producing (IFN) γ or IL-4, IL-5, IL-13 regulate Ig class switching and determine the type of antibody produced. T_H17 cells secrete IL-17 and IL-22, and Th9 cells secrete IL-9. Both are linked to mediation of autoimmune disease. CD4+ T regulatory cells produce IL-10 and TGFβ and downregulate T and B cell responses once the microbe has been eliminated. Each of the types of CD4+ T cells are regulated by different transcription factors and the key transcription factors are shown in the circles above each CD4+ T cell type.

TABLE 342-7 Cytokines and Cytokine Receptors

CYTOKINE	RECEPTOR	CELL SOURCE	CELL TARGET	BIOLOGIC ACTIVITY
IL-1α, β	Type I IL-1r, Type II IL-1r	Monocytes/macrophages, B cells, fibroblasts, most epithelial cells including thymic epithelium, endothelial cells	All cells	Upregulates adhesion molecule expression, neutrophil and macrophage emigration, mimics shock, fever, upregulates hepatic acute-phase protein production, facilitates hematopoiesis
IL-2	IL-2r α, β, common γ	T cells	T cells, B cells, NK cells, monocytes-macrophages	Promotes T cell activation and proliferation, B cell growth, NK cell proliferation and activation, enhanced monocyte/macrophage cytolytic activity
IL-3	IL-3r, common β	T cells, NK cells, mast cells	Monocytes-macrophages, mast cells, eosinophils, bone marrow progenitors	Stimulates hematopoietic progenitors
IL-4	IL-4r α, common γ	T cells, mast cells, basophils	T cells, B cells, NK cells, monocytes-macrophages, neutrophils, eosinophils, endothelial cells, fibroblasts	Stimulates T_H2 helper T cell differentiation and proliferation; stimulates B cell Ig class switch to IgG1 and IgE anti-inflammatory action on T cells, monocytes; produced by T follicular helper cells in B cell germinal centers that stimulate B cell maturation.
IL-5	IL-5r α, common γ	T cells, mast cells, eosinophils	Eosinophils, basophils, murine B cells	Regulates eosinophil migration and activation
IL-6	IL-6r, gp130	Monocytes-macrophages, B cells, fibroblasts, most epithelium including thymic epithelium, endothelial cells	T cells, B cells, epithelial cells, hepatocytes, monocytes-macrophages	Induces acute-phase protein production, T and B cell differentiation and growth, myeloma cell growth, and osteoclast growth and activation
IL-7	IL-7r α, common γ	Bone marrow, thymic epithelial cells	T cells, B cells, bone marrow cells	Differentiates B, T, and NK cell precursors, activates T and NK cells
IL-8	CXCR1, CXCR2	Monocytes-macrophages, T cells, neutrophils, fibroblasts, endothelial cells, epithelial cells	Neutrophils, T cells, monocytes-macrophages, endothelial cells, basophils	Induces neutrophil, monocyte, and T cell migration, induces neutrophil adherence to endothelial cells and histamine release from basophils, and stimulates angiogenesis; suppresses proliferation of hepatic precursors
IL-9	IL-9r α, common γ	T cells	Bone marrow progenitors, B cells, T cells, mast cells	Induces mast cell proliferation and function, synergizes with IL-4 in IgG and IgE production and T cell growth, activation, and differentiation
IL-10	IL-10r	Monocytes-macrophages, T cells, B cells, keratinocytes, mast cells	Monocytes-macrophages, T cells, B cells, NK cells, mast cells	Inhibits macrophage proinflammatory cytokine production, downregulates cytokine class II antigen and B7-1 and B7-2 expression, inhibits differentiation of T_H1 helper T cells, inhibits NK cell function, stimulates mast cell proliferation and function, B cell activation, and differentiation
IL-11	IL-11r α, gp130	Bone marrow stromal cells	Megakaryocytes, B cells, hepatocytes	Induces megakaryocyte colony formation and maturation, enhances antibody responses, stimulates acute-phase protein production
IL-12 (35-kDa and 40-kDa subunits)	IL-12r	Activated macrophages, dendritic cells, neutrophils	T cells, NK cells	Induces T_H1 T helper cell formation and lymphokine-activated killer cell formation; increases CD8+ CTL cytolytic activity; ↓IL-17, ↑IFN-γ
IL-13	IL-13r/IL-4r α	T cells (T_H2)	Monocytes-macrophages, B cells, endothelial cells, keratinocytes	Upregulates VCAM-1 and C-C chemokine expression on endothelial cells and B cell activation and differentiation, and inhibits macrophage proinflammatory cytokine production
IL-14	Unknown	T cells	Normal and malignant B cells	Induces B cell proliferation, inhibits antibody secretion, and expands selected B cell subgroups
IL-15	IL-15r α, common γ, IL2r β	Monocytes-macrophages, epithelial cells, fibroblasts	T cells, NK cells	Promotes T cell activation and proliferation, angiogenesis, and NK cells
IL-16	CD4	Mast cells, eosinophils, CD8+ T cells, respiratory epithelium	CD4+ T cells, monocytes-macrophages, eosinophils	Promotes chemoattraction of CD4+ T cells, monocytes, and eosinophils; inhibits HIV replication; inhibits T cell activation through CD3/T cell receptor
IL-17	IL-17r	CD4+ T cells	Fibroblasts, endothelium, epithelium, macrophages	Enhances cytokine/chemokine secretion; promotes delayed-type reactions
IL-18	IL-18r (IL-1R-related protein)	Keratinocytes, macrophages	T cells, B cells, NK cells	Upregulates IFN-γ production, enhances NK cell cytotoxicity
IL-21	IL-δγ chain/IL-21R	CD4 T cells	NK cells	Downregulates NK cell–activating molecules, NKG2D/DAP10; produced by T follicular helper cells in B cell germinal centers that stimulate B cell maturation.
IL-22	IL-22 R1/IL-10R2	DC, T cells	Epithelial cells	Innate responses against bacterial pathogens; promotes hepatocyte survival
IL-23	IL-12Rb1/IL23R	Macrophages, other cell types	T cells	Opposite effects of IL-12 (↑IL-17, ↑IFN-γ)
IL-24	IL-20 R1/IL-20R2 IL-22R1/IL-20 R2	Macrophages, T_H2 cells	Nonhematopoietic cells such as fibroblasts	Promotes wound healing

(*Continued*)

TABLE 342-7 Cytokines and Cytokine Receptors (Continued)

CYTOKINE	RECEPTOR	CELL SOURCE	CELL TARGET	BIOLOGIC ACTIVITY
IL-25 (also called IL-17E)	IL-17RB	CD4 T cells, mast cells	Fibroblasts, endothelium, epithelium, macrophages	Proinflammatory; induces cytokine production
IL-26	IL-20 R1/IL-10R2	T_H1, T_H17 T cells, synovial cells	Epithelial cells	Proinflammatory; induces cytokine production
IFN-α	Type I interferon receptor	All cells	All cells	Promotes antiviral activity; stimulates T cell, macrophage, and NK cell activity; direct antitumor effects; upregulates MHC class I antigen expression; used therapeutically in viral and autoimmune conditions
IFN-β	Type I interferon receptor	All cells	All cells	Antiviral activity; stimulates T cell, macrophage, and NK cell activity; direct antitumor effects; upregulates MHC class I antigen expression; used therapeutically in viral and autoimmune conditions
IFN-γ	Type II interferon receptor	T cells, NK cells	All cells	Regulates macrophage and NK cell activations; stimulates immunoglobulin secretion by B cells; induction of class II histocompatibility antigens; T_H1 T cell differentiation
TNF-α	TNFrI, TNFrII	Monocytes-macrophages, mast cells, basophils, eosinophils, NK cells, B cells, T cells, keratinocytes, fibroblasts, thymic epithelial cells	All cells except erythrocytes	Fever, anorexia, shock, capillary leak syndrome, enhanced leukocyte cytotoxicity, enhanced NK cell function, acute phase protein synthesis, proinflammatory cytokine induction
TNF-β	TNFrI, TNFrII	T cells, B cells	All cells except erythrocytes	Cell cytotoxicity, lymph node and spleen development
LT-β	LTβR	T cells	All cells except erythrocytes	Cell cytotoxicity, normal lymph node development
G-CSF	G-CSFr; gp130	Monocytes-macrophages, fibroblasts, endothelial cells, thymic epithelial cells, stromal cells	Myeloid cells, endothelial cells	Regulates myelopoiesis; enhances survival and function of neutrophils; clinical use in reversing neutropenia after cytotoxic chemotherapy
GM-CSF	GM-CSFr, common β	T cells, monocytes-macrophages, fibroblasts, endothelial cells, thymic epithelial cells	Monocytes-macrophages, neutrophils, eosinophils, fibroblasts, endothelial cells	Regulates myelopoiesis; enhances macrophage bactericidal and tumoricidal activity; mediator of dendritic cell maturation and function; upregulates NK cell function; clinical use in reversing neutropenia after cytotoxic chemotherapy
M-CSF	M-CSFr (*c-fms* protooncogene)	Fibroblasts, endothelial cells, monocytes-macrophages, T cells, B cells, epithelial cells including thymic epithelium	Monocytes-macrophages	Regulates monocyte-macrophage production and function
LIF	LIFr-α; gp130	Activated T cells, bone marrow stromal cells, thymic epithelium	Megakaryocytes, monocytes, hepatocytes, possibly lymphocyte subpopulations	Induces hepatic acute-phase protein production; stimulates macrophage differentiation; promotes growth of myeloma cells and hematopoietic progenitors; stimulates thrombopoiesis
OSM	OSMr; LIFr; gp130	Activated monocytes-macrophages and T cells, bone marrow stromal cells, some breast carcinoma cell lines, myeloma cells	Neurons, hepatocytes, monocytes-macrophages, adipocytes, alveolar epithelial cells, embryonic stem cells, melanocytes, endothelial cells, fibroblasts, myeloma cells	Induces hepatic acute-phase protein production; stimulates macrophage differentiation; promotes growth of myeloma cells and hematopoietic progenitors; stimulates thrombopoiesis; stimulates growth of Kaposi's sarcoma cells
SCF	SCFr (*c-kit* protooncogene)	Bone marrow stromal cells and fibroblasts	Embryonic stem cells, myeloid and lymphoid precursors, mast cells	Stimulates hematopoietic progenitor cell growth, mast cell growth; promotes embryonic stem cell migration
TGF-β (3 isoforms)	Type I, II, III TGF-β receptor	Most cell types	Most cell types	Downregulates T cell, macrophage, and granulocyte responses; stimulates synthesis of matrix proteins; stimulates angiogenesis
Lymphotactin/ SCM-1	XCR1	NK cells, mast cells, double-negative thymocytes, activated CD8+ T cells	T cells, NK cells	Chemoattractant for lymphocytes; only known chemokine of C class
MCP-1	CCR2	Fibroblasts, smooth-muscle cells, activated PBMCs	Monocytes-macrophages, NK cells, memory T cells, basophils	Chemoattractant for monocytes, activated memory T cells, and NK cells; induces granule release from CD8+ T cells and NK cells; potent histamine-releasing factor for basophils; suppresses proliferation of hematopoietic precursors; regulates monocyte protease production
MCP-2	CCR1, CCR2	Fibroblasts, activated PBMCs	Monocytes-macrophages, T cells, eosinophils, basophils, NK cells	Chemoattractant for monocytes, memory and naïve T cells, eosinophils, ?NK cells; activates basophils and eosinophils; regulates monocyte protease production
MCP-3	CCR1, CCR2	Fibroblasts, activated PBMCs	Monocytes-macrophages, T cells, eosinophils, basophils, NK cells, dendritic cells	Chemoattractant for monocytes, memory and naïve T cells, dendritic cells, eosinophils, ?NK cells; activates basophils and eosinophils; regulates monocyte protease production

(*Continued*)

TABLE 342-7 Cytokines and Cytokine Receptors (Continued)

CYTOKINE	RECEPTOR	CELL SOURCE	CELL TARGET	BIOLOGIC ACTIVITY
MCP-4	CCR2, CCR3	Lung, colon, small intestinal epithelial cells, activated endothelial cells	Monocytes-macrophages, T cells, eosinophils, basophils	Chemoattractant for monocytes, T cells, eosinophils, and basophils
Eotaxin	CCR3	Pulmonary epithelial cells, heart	Eosinophils, basophils	Potent chemoattractant for eosinophils and basophils; induces allergic airways disease; acts in concert with IL-5 to activate eosinophils; antibodies to eotaxin inhibit airway inflammation
TARC	CCR4	Thymus, dendritic cells, activated T cells	T cells, NK cells	Chemoattractant for T and NK cells
MDC	CCR4	Monocytes-macrophages, dendritic cells, thymus	Activated T cells	Chemoattractant for activated T cells; inhibits infection with T cell tropic HIV
MIP-1α	CCR1, CCR5	Monocytes-macrophages, T cells	Monocytes-macrophages, T cells, dendritic cells, NK cells, eosinophils, basophils	Chemoattractant for monocytes, T cells, dendritic cells, and NK cells, and weak chemoattractant for eosinophils and basophils; activates NK cell function; suppresses proliferation of hematopoietic precursors; necessary for myocarditis associated with coxsackievirus infection; inhibits infection with monocytotropic HIV
MIP-1β	CCR5	Monocytes-macrophages, T cells	Monocytes-macrophages, T cells, NK cells, dendritic cells	Chemoattractant for monocytes, T cells, and NK cells; activates NK cell function; inhibits infection with monocytotropic HIV
RANTES	CCR1, CCR2, CCR5	Monocytes-macrophages, T cells, fibroblasts, eosinophils	Monocytes-macrophages, T cells, NK cells, dendritic cells, eosinophils, basophils	Chemoattractant for monocytes-macrophages, CD4+, CD45Ro+ T cells, CD8+ T cells, NK cells, eosinophils, and basophils; induces histamine release from basophils; inhibits infections with monocytotropic HIV
LARC/MIP-3α/ Exodus-1	CCR6	Dendritic cells, fetal liver cells, activated T cells	T cells, B cells	Chemoattractant for lymphocytes
ELC/MIP-3β	CCR7	Thymus, lymph node, appendix	Activated T cells and B cells	Chemoattractant for B and T cells; receptor upregulated on EBV-infected B cells and HSV-infected T cells
I-309/TCA-3	CCR8	Activated T cells	Monocytes-macrophages, T cells	Chemoattractant for monocytes; prevents glucocorticoid-induced apoptosis in some T cell lines
SLC/TCA-4/ Exodus-2	CCR7	Thymic epithelial cells, lymph node, appendix, and spleen	T cells	Chemoattractant for T lymphocytes; inhibits hematopoiesis
DC-CK1/PARC	Unknown	Dendritic cells in secondary lymphoid tissues	Naïve T cells	May have a role in induction of immune responses
TECK	CCR9	Dendritic cells, thymus, liver, small intestine	T cells, monocytes-macrophages, dendritic cells	Thymic dendritic cell-derived cytokine, possibly involved in T cell development
GRO-α/MGSA	CXCR2	Activated granulocytes, monocyte-macrophages, and epithelial cells	Neutrophils, epithelial cells, ?endothelial cells	Neutrophil chemoattractant and activator; mitogenic for some melanoma cell lines; suppresses proliferation of hematopoietic precursors; angiogenic activity
GRO-β/MIP-2α	CXCR2	Activated granulocytes and monocyte-macrophages	Neutrophils and ?endothelial cells	Neutrophil chemoattractant and activator; angiogenic activity
NAP-2	CXCR2	Platelets	Neutrophils, basophils	Derived from platelet basic protein; neutrophil chemoattractant and activator
IP-10	CXCR3	Monocytes-macrophages, T cells, fibroblasts, endothelial cells, epithelial cells	Activated T cells, tumor-infiltrating lymphocytes, ?endothelial cells, ?NK cells	IFN-γ-inducible protein that is a chemoattractant for T cells; suppresses proliferation of hematopoietic precursors
MIG	CXCR3	Monocytes-macrophages, T cells, fibroblasts	Activated T cells, tumor-infiltrating lymphocytes	IFN-γ-inducible protein that is a chemoattractant for T cells; suppresses proliferation of hematopoietic precursors
SDF-1	CXCR4	Fibroblasts	T cells, dendritic cells, ?basophils, ?endothelial cells	Low-potency, high-efficacy T cell chemoattractant; required for B-lymphocyte development; prevents infection of CD4+, CXCR4+ cells by T cell tropic HIV
Fractalkine	CX3CR1	Activated endothelial cells	NK cells, T cells, monocytes-macrophages	Cell-surface chemokine/mucin hybrid molecule that functions as a chemoattractant, leukocyte activator, and cell adhesion molecule
PF-4	Unknown	Platelets, megakaryocytes	Fibroblasts, endothelial cells	Chemoattractant for fibroblasts; suppresses proliferation of hematopoietic precursors; inhibits endothelial cell proliferation and angiogenesis

Abbreviations: B7-1, CD80; B7-2, CD86; CCR, CC-type chemokine receptor; CXCR, CXC-type chemokine receptor; DC-CK, dendritic cell chemokine; EBV, Epstein-Barr virus; ELC, EB11 ligand chemokine (MIP-1b); G-CSF, granulocyte colony-stimulating factor; GM-CSF, granulocyte-macrophage colony-stimulating factor; GRP, growth-related peptide; HSV, herpes simplex virus; IFN, interferon; Ig, immunoglobulin; IL, interleukin; IP-10, IFN-γ-inducible protein-10; LARC, liver- and activation-regulated chemokine; LIF, leukemia inhibitory factor; MCP, monocyte chemotactic protein; M-CSF, macrophage colony-stimulating factor; MDC, macrophage-derived chemokine; MGSA, melanoma growth-stimulating activity; MHC, major histocompatibility complex; MIG, monokine induced by IFN-γ; MIP, macrophage inflammatory protein; NAP, neutrophil-activating protein; NK, natural killer; OSM, oncostatin M; PARC, pulmonary- and activation-regulated chemokine; PBMC, peripheral blood mononuclear cells; PF, platelet factor; RANTES, regulated on activation, normally T cell–expressed and –secreted; SCF, stem cell factor; SDF, stromal cell–derived factor; SLC, secondary lymphoid tissue chemokine; TARC, thymus- and activation-regulated chemokine; TCA, T cell activation protein; TECK, thymus-expressed chemokine; TGF, transforming growth factor; T_H1 and T_H2, helper T cell subsets; TNF, tumor necrosis factor; VCAM, vascular cell adhesion molecule.

Source: Data from JS Sundy et al: Appendix B, in *Inflammation, Basic Principles and Clinical Correlates*, 3rd ed, J Gallin and R Snyderman (eds). Philadelphia, Lippincott Williams and Wilkins, 1999.

TABLE 342-8 CC, CXC_1, CX_3, C_1, and XC Families of Chemokines and Chemokine Receptors

CHEMOKINE RECEPTOR	CHEMOKINE LIGANDS	CELL TYPES	DISEASE CONNECTION
CCR1	CCL3 (MIP-1α), CCL5 (RANTES), CCL7 (MCP-3), CCL14 (HCC1)	T cells, monocytes, eosinophils, basophils	Rheumatoid arthritis, multiple sclerosis
CCR2	CCL2 (MCP-1), CCL8 (MCP-2), CCL7 (MCP-3), CCL13 (MCP-4), CCL16 (HCC4)	Monocytes, dendritic cells (immature), memory T cells	Atherosclerosis, rheumatoid arthritis, multiple sclerosis, resistance to intracellular pathogens, type 2 diabetes mellitus
CCR3	CCL11 (eotaxin), CCL13 (eotaxin-2), CCL7 (MCP-3), CCL5 (RANTES), CCL8 (MCP-2), CCL13 (MCP-4)	Eosinophils, basophils, mast cells, T_H2, platelets	Allergic asthma and rhinitis
CCR4	CCL17 (TARC), CCL22 (MDC)	T cells (T_H2), dendritic cells (mature), basophils, macrophages, platelets	Parasitic infection, graft rejection, T cell homing to skin
CCR5	CCL3 (MIP-1α), CCL4 (MIP-1α), CCL5 (RANTES), CCL11 (eotaxin), CCL14 (HCC1), CCL16 (HCC4)	T cells, monocytes	HIV-1 co-receptor (T cell–tropic strains), transplant rejection
CCR6	CCL20 (MIP-3α, LARC)	T cells (T regulatory and memory), B cells, dendritic cells	Mucosal humoral immunity, allergic asthma, intestinal T cell homing
CCR7	CCL19 (ELC), CCL21 (SLC)	T cells, dendritic cells (mature)	Transport of T cells and dendritic cells to lymph nodes, antigen presentation, and cellular immunity
CCR8	CCL1 (1309)	T cells (T_H2), monocytes, dendritic cells	Dendritic cell migration to lymph node, type 2 cellular immunity, granuloma formation
CCR9	CCL25 (TECK)	T cells, IgA+ plasma cells	Homing of T cells and IgA+ plasma cells to the intestine, inflammatory bowel disease
CCR10	CCL27 (CTACK), CCL28 (MEC)	T cells	T cell homing to intestine and skin
CXCR1	CXCL8 (interleukin-8), CXCL6 (GCP2)	Neutrophils, monocytes	Inflammatory lung disease, COPD
CXCR2	CXCL8, CXCL1 (GROα), CXCL2 (GROα), CXCL3 (GROα), CXCL5 (ENA-78), CXCL6	Neutrophils, monocytes, microvascular endothelial cells	Inflammatory lung disease, COPD, angiogenic for tumor growth
CXCR3-A	CXCL9 (MIG), CXCL10 (IP-10), CXCL11 (I-TAC)	Type 1 helper cells, mast cells, mesangial cells	Inflammatory skin disease, multiple sclerosis, transplant rejection
CXCR3-B	CXCL4 (PF4), CXCL9 (MIG), CXCL10 (IP-10), CXCL11 (I-TAC)	Microvascular endothelial cells, neoplastic cells	Angiostatic for tumor growth
CXCR4	CXCL12 (SDF-1)	Widely expressed	HIV-1 co-receptor (T cell–tropic), tumor metastases, hematopoiesis
CXCR5	CXCL13 (BCA-1)	B cells, follicular helper T cells	Formation of B cell follicles
CXCR6	CXCL16 (SR-PSOX)	CD8+ T cells, natural killer cells, and memory CD4+ T cells	Inflammatory liver disease, atherosclerosis (CXCL16)
CX_3CR1	CX3CL1 (fractalkine)	Macrophages, endothelial cells, smooth-muscle cells	Atherosclerosis
XCR1	XCL1 (lymphotactin), XCL2	T cells, natural killer cells	Rheumatoid arthritis, IgA nephropathy, tumor response

Abbreviations: BCA-1, B-cell chemoattractant 1; COPD, chronic obstructive pulmonary disease; CTACK, cutaneous T cell–attracting chemokine; ELC, Epstein-Barr I1-ligand chemokine; ENA, epithelial cell–derived neutrophil-activating peptide; GCP, granulocyte chemotactic protein; GRO, growth-regulated oncogene; HCC, hemofiltrate chemokine; IP-10, interferon inducible 10; I-TAC, interferon-inducible T cell alpha chemoattractant; LARC, liver- and activation-regulated chemokine; MCP, monocyte chemoattractant protein; MDC, macrophage-derived chemokine; MEC, mammary-enriched chemokine; MIG, monokine induced by interferon-γ; MIP, macrophage inflammatory protein; PF, platelet factor; SDF, stromal cell–derived factor; SLC, secondary lymphoid-tissue chemokine; SR-PSOX, scavenger receptor for phosphatidylserine-containing oxidized lipids; TARC, thymus- and activation-regulated chemokine; TECK, thymus-expressed chemokine; T_H2, type 2 helper T cells.

eosinophil granule contains anti-inflammatory types of enzymes (histaminase, arylsulfatase, phospholipase D), eosinophils may homeostatically downregulate or terminate ongoing inflammatory responses.

Basophils and tissue mast cells are potent reservoirs of cytokines such as IL-4 and can respond to bacteria and viruses with antipathogen cytokine production through multiple TLRs expressed on their surface. Mast cells and basophils can also mediate immunity through the binding of antipathogen antibodies. This is a particularly important host defense mechanism against parasitic diseases. Basophils express high-affinity surface receptors for IgE (FcεRII) (CD23) and, upon cross-linking of basophil-bound IgE by antigen, can release histamine, eosinophil chemotactic factor of anaphylaxis, and neutral protease—all mediators of allergic immediate (anaphylaxis) hypersensitivity responses **(Table 342-11)**. In addition, basophils express surface receptors for activated complement components (C3a, C5a), through which mediator release can be directly affected. Thus, basophils, like most cells of the immune system, can be activated in the service of host defense against pathogens, or they can be activated for mediation release and cause pathogenic responses in allergic and inflammatory diseases. **For further discussion of tissue mast cells, see Chap. 347.**

The Complement System The complement system, an important soluble component of the innate immune system, is a series of plasma enzymes, regulatory proteins, and proteins that are activated in a cascading fashion, resulting in cell lysis. There are four pathways of the complement system: the classic activation pathway activated by antigen/antibody immune complexes, the MBL (a serum collectin; Table 342-3) activation pathway activated by microbes with terminal mannose groups, the alternative activation pathway activated by microbes or tumor cells, and the terminal pathway that is common to the first three pathways and leads to the membrane attack complex that lyses cells **(Fig. 342-4)**. The series of enzymes of the complement system are serine proteases.

Activation of the classic complement pathway via immune complex binding to C1q links the innate and adaptive immune systems via specific antibody in the immune complex. The alternative complement activation pathway is antibody-independent and is activated by binding of C3 directly to pathogens and "altered self" such as tumor cells. In the renal glomerular inflammatory disease *IgA nephropathy*, IgA activates the alternative complement pathway and causes glomerular damage and decreased renal function. Activation of the classic

TABLE 342-9 Cytokine Families Grouped by Structural Similarity

Hematopoietins	IL-2, IL-3, IL-4, IL-5, IL-6, IL-7, IL-9, IL-11, IL-12, IL-15, IL-16, IL-17, IL-21, IL-23, EPO, LIF, GM-CSF, G-CSF, OSM, CNTF, GH, and TPO TNF-α, LT-α, LT-β, CD40L, CD30L, CD27L, 4-1BBL, OX40, OPG, and FasL
IL-1	IL-1α, IL-1β, IL-1ra, IL-18, bFGF, aFGF, and ECGF
PDGF	PDGF A, PDGF B, and M-CSF
TGF-β	TGF-β and BMPs (1, 2, 4, etc.)
C-X-C chemokines	IL-8, Gro-α/β/γ, NAP-2, ENA78, GCP-2, PF4, CTAP-3, MIG, and IP-10
C-C chemokines	MCP-1, MCP-2, MCP-3, MIP-1α, MIP-1β, RANTES

Abbreviations: aFGF, acidic fibroblast growth factor; 4-1BBL, 401 BB ligand; bFGF, basic fibroblast growth factor; BMP, bone marrow morphogenetic proteins; C-C, cysteine-cysteine; CD, cluster of differentiation; CNTF, ciliary neurotrophic factor; CTAP, connective tissue–activating peptide; C-X-C, cysteine-x-cysteine; ECGF, endothelial cell growth factor; EPO, erythropoietin; FasL, Fas ligand; GCP-2, granulocyte chemotactic protein 2; G-CSF, granulocyte colony-stimulating factor; GH, growth hormone; GM-CSF, granulocyte-macrophage colony-stimulating factor; Gro, growth-related gene products; IFN, interferon; IL, interleukin; IP, interferon-γ inducible protein; LIF, leukemia inhibitory factor; LT, lymphotoxin; MCP, monocyte chemoattractant; M-CSF, macrophage colony-stimulating factor; MIG, monokine induced by interferon-γ; MIP, macrophage inflammatory protein; NAP-2, neutrophil activating protein 2; OPG, osteoprotegerin; OSM, oncostatin M; PDGF, platelet-derived growth factor; PF, platelet factor; R, receptor; RANTES, regulated on activation, normal T cell–expressed and –secreted; TGF, transforming growth factor; TNF, tumor necrosis factor; TPO, thyroperoxidase.

complement pathway via C1, C4, and C2 and activation of the alternative pathway via factor D, C3, and factor B both lead to cleavage and activation of C3. C3 activation fragments, when bound to target surfaces such as bacteria and other foreign antigens, are critical for opsonization (coating by antibody and complement) in preparation for phagocytosis. The MBL pathway substitutes MBL-associated serine proteases (MASPs) 1 and 2 for C1q, C1r, and C1s to activate C4. The MBL activation pathway is activated by mannose on the surface of bacteria and viruses.

The three pathways of complement activation all converge on the final common terminal pathway. C3 cleavage by each pathway results in activation of C5, C6, C7, C8, and C9, resulting in the membrane attack complex that physically inserts into the membranes of target cells or bacteria and lyses them.

Thus, complement activation is a critical component of innate immunity for responding to microbial infection. The functional consequences of complement activation by the three initiating pathways and the terminal pathway are shown in Fig. 342-4. In general, the cleavage products of complement components facilitate microbe or damaged cell clearance (C1q, C4, C3), promote activation and enhancement of inflammation (anaphylatoxins, C3a, C5a), and promote microbe or opsonized cell lysis (membrane attack complex).

CYTOKINES

Cytokines are soluble proteins produced by a wide variety of cell types (Tables 342-7 to 342-9). They are critical for both normal innate and adaptive immune responses, and their expression may be perturbed in most immune, inflammatory, and infectious disease states.

Cytokines are involved in the regulation of the growth, development, and activation of immune system cells and in the mediation of the inflammatory response. In general, cytokines are characterized by considerable redundancy; different cytokines have similar functions. In addition, many cytokines are pleiotropic in that they are capable of acting on many different cell types. This pleiotropism results from the expression on multiple cell types of receptors for the same cytokine (see below), leading to the formation of "cytokine networks." The action of cytokines may be (1) autocrine when the target cell is the same cell that secretes the cytokine, (2) paracrine when the target cell is nearby, and (3) endocrine when the cytokine is secreted into the circulation and acts distal to the source.

Cytokines have been named based on presumed targets or based on presumed functions. Those cytokines that are thought to primarily target leukocytes have been named IL-1, 2, 3, etc. Many cytokines that were originally described as having a certain function have retained those

TABLE 342-10 Association of KIRS with Disease

DISEASE	KIR ASSOCIATION	OBSERVATION
Psoriatic arthritis	KIR2DS1/KIR2DS2; HLA-Cw group homozygosity	Susceptibility
Spondylarthritides	Increased KIR3DL2 expression	May contribute to disease pathology
	Interaction HLA-B27 homodimers with KIR3DL1/KIR3DL2; independent of peptide	May contribute to disease pathogenesis
Ankylosing spondylitis	KIR3DL1/3DS1; HLA-B27 genotypes	Susceptibility
Rheumatoid vasculitis	KIR2DS2; HLA-Cw*03	Susceptibility
	Increased KIR2L2/2DS2 in patients with extraarticular manifestations	Clinical manifestations may have different genetic backgrounds with respect to KIR genotype
Rheumatoid arthritis	Decreased KIR2DS1/3DS1 in patients without bone erosions	Susceptibility
	KIR2DS4; HLA-Cw4	Susceptibility
Scleroderma	KIR2DS2+/KIR2DL2–	Susceptibility
Behçet's disease	Altered KIR3DL1 expression	Associated with severe eye disease
Psoriasis vulgaris	2DS1; HLA-Cw*06	Susceptibility
	2DS1; 2DL5; haplotype B	Susceptibility
IDDM	KIR2DS2; HLA-C1	Susceptibility
Type 1 diabetes	KIR2DS2; HLA-C1 and no HLA-C2, no HLA-Bw4	Increased disease progression
Preeclampsia	KIR2DL1 with fewer KIR2DS (mother); HLA-C2 (fetus)	Increased disease progression
AIDS	KIR3DS1; HLA-Bw4Ile80	Decreased disease progression
	KIR3DS1 homozygous; no HLA-Bw4Ile80	Increased disease progression
HCV infection	KIR2DL3 homozygous; HLA-C1 homozygous	Decreased disease progression
Cervical neoplasia (HPV induced)	KIR3DS1; HLA-C1 homozygous and no HLA-Bw4	Increased disease progression
Malignant melanoma	KIR2DL2 and/or KIR2DL3; HLA-C1	Increased disease progression

Abbreviations: HCV, hepatitis C virus; HLA, human leukocyte antigen; HPV, human papillomavirus; IDDM, insulin-dependent diabetes mellitus; KIR, killer cell immunoglobulin-like receptor.

Source: Adapted from R Diaz-Pena et al: Adv Exp Med Biol 649:286, 2009.

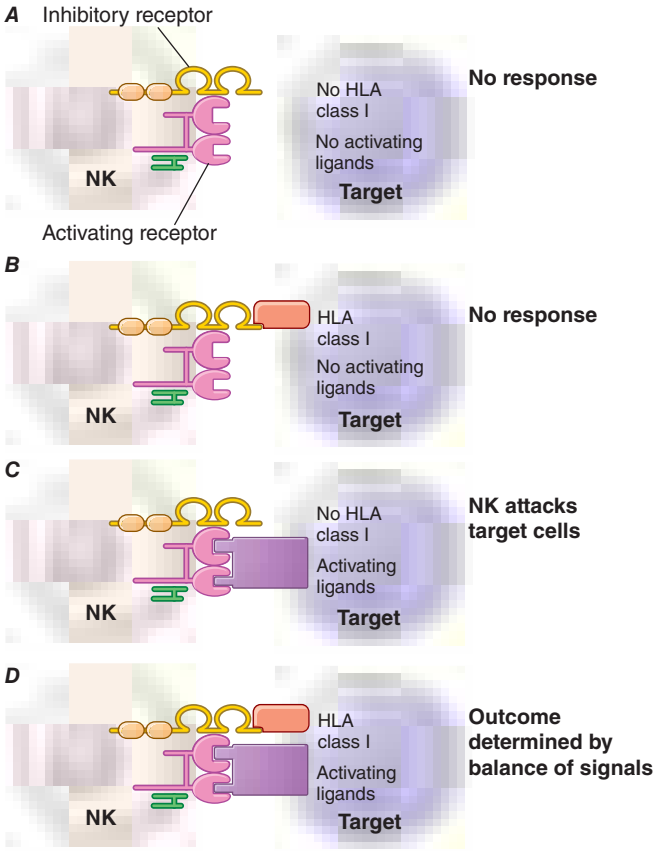


FIGURE 342-3 Encounters between NK cells: Potential targets and possible outcomes. The amount of activating and inhibitory receptors on the NK cells and the amount of ligands on the target cell, as well as the qualitative differences in the signals transduced, determine the extent of the NK response. ***A.*** When target cells have no HLA class I or activating ligands, NK cells cannot kill target cells. ***B.*** When target cells bear self-HLA, NK cells cannot kill targets. ***C.*** When target cells are pathogen-infected and have downregulated HLA and express activating ligands, NK cells kill target cells. ***D.*** When NK cells encounter targets with both self-HLA and activating receptors, then the level of target killing is determined by the balance of inhibitory and activating signals to the NK cell. HLA, human leukocyte antigen; NK, natural killer. *(Adapted from L Lanier: Annu Rev Immunol 23:225, 2005; reproduced with permission from Annual Reviews Inc. Copyright 2011 by Annual Reviews Inc.)*

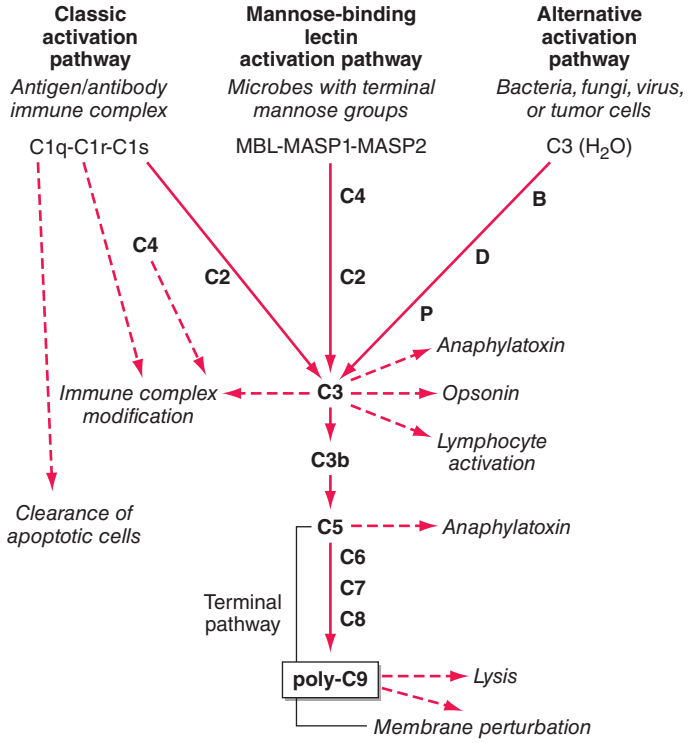


FIGURE 342-4 The four pathways and the effector mechanisms of the complement system. *Dashed arrows* indicate the functions of pathway components. *(After BJ Morley, MJ Walport: The Complement Facts Books. London, Academic Press, 2000, Chap. 2; with permission. Copyright Academic Press, London, 2000.)*

TABLE 342-11 Examples of Mediators Released from Immune Cells and Basophils

MEDIATOR	ACTIONS
Histamine	Smooth-muscle contraction, increased vascular permeability
Slow reacting substance of anaphylaxis (SRSA) (leukotriene C_4, D_4, E_4)	Smooth-muscle contraction
Eosinophil chemotactic factor of anaphylaxis (ECF-A)	Chemotactic attraction of eosinophils
Platelet-activating factor	Activates platelets to secrete serotonin and other mediators: smooth-muscle contraction; induces vascular permeability
Neutrophil chemotactic factor (NCF)	Chemotactic attraction of neutrophils
Leukotactic activity (leukotriene B_4)	Chemotactic attraction of neutrophils
Heparin	Anticoagulant
Basophil kallikrein of anaphylaxis (BK-A)	Cleaves kininogen to form bradykinin

names (e.g., granulocyte colony-stimulating factor [G-CSF]). Cytokines belong in general to three major structural families: the hematopoietin family; the TNF, IL-1, platelet-derived growth factor (PDGF), and transforming growth factor (TGF) β families; and the CXC and C-C chemokine families (Table 342-9). Chemokines are cytokines that regulate cell movement and trafficking; they act through G protein-coupled receptors and have a distinctive three-dimensional structure. IL-8 is the only chemokine that early on was named an IL (Table 342-7).

In general, cytokines exert their effects by influencing gene activation that results in cellular activation, growth, differentiation, functional cell-surface molecule expression, and cellular effector function. In this regard, cytokines can have dramatic effects on the regulation of immune responses and the pathogenesis of a variety of diseases. Indeed, T cells have been categorized on the basis of the pattern of cytokines that they secrete, which results in either humoral immune response (T_H2) or cell-mediated immune response (T_H1). A third type of T helper cell is the T_H17 cell that contributes to host defense against extracellular bacteria and fungi, particularly at mucosal sites (Fig. 342-2).

Cytokine receptors can be grouped into five general families based on similarities in their extracellular amino acid sequences and conserved structural domains. The *immunoglobulin (Ig) superfamily* represents a large number of cell-surface and secreted proteins. The IL-1 receptors (type 1, type 2) are examples of cytokine receptors with extracellular Ig domains.

The hallmark of the *hematopoietic growth factor (type 1) receptor* family is that the extracellular regions of each receptor contain two conserved motifs. One motif, located at the N terminus, is rich in cysteine residues. The other motif is located at the C terminus proximal to the transmembrane region and comprises five amino acid residues, tryptophan-serine-X-tryptophan-serine (WSXWS). This family can be grouped on the basis of the number of receptor subunits they have and on the utilization of shared subunits. A number of cytokine receptors, i.e., IL-6, IL-11, IL-12, and leukemia inhibitory factor, are paired with gp130. There is also a common 150-kDa subunit shared by IL-3, IL-5, and granulocyte-macrophage colony-stimulating factor (GM-CSF) receptors. The gamma chain (γ_c) of the IL-2 receptor is common to the IL-2, IL-4, IL-7, IL-9, and IL-15 receptors. Thus, the specific cytokine receptor is

responsible for ligand-specific binding, whereas the subunits such as gp130, the 150-kDa subunit, and γ_c are important in signal transduction. The γ_c gene is on the X chromosome, and mutations in the γ_c protein result in the *X-linked form of severe combined immune deficiency syndrome (X-SCID)* **(Chap. 344)**.

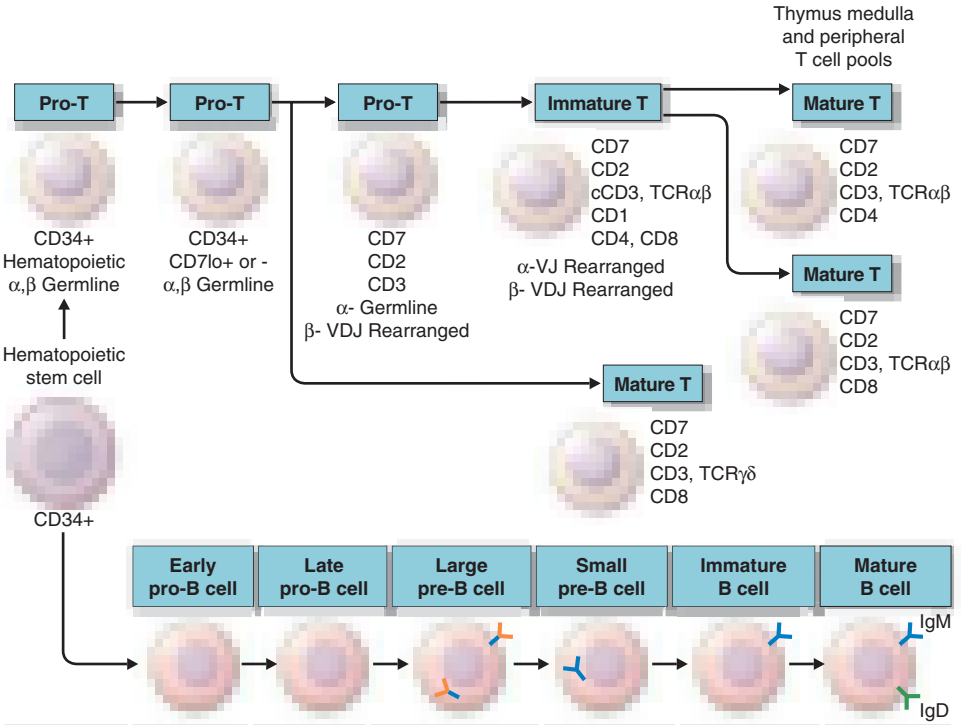
The members of the *interferon (type II) receptor* family include the receptors for IFN-γ and β, which share a similar 210-amino-acid binding domain with conserved cysteine pairs at both the amino and carboxy termini. The members of the *TNF (type III) receptor family* share a common binding domain composed of repeated cysteine-rich regions. Members of this family include the p55 and p75 receptors for TNF (TNF-R1 and TNF-R2, respectively); CD40 antigen, which is an important B cell-surface marker involved in immunoglobulin isotype switching; fas/Apo-1, whose triggering induces apoptosis; CD27 and CD30, which are found on activated T cells and B cells; and nerve growth factor receptor.

The common motif for the *seven transmembrane helix family* was originally found in receptors linked to GTP-binding proteins. This family includes receptors for chemokines (Table 342-8), β-adrenergic receptors, and retinal rhodopsin. It is important to note that two members of the chemokine receptor family, CXC chemokine receptor type 4 (CXCR4) and β chemokine receptor type 5 (CCR5), have been found to serve as the two major co-receptors for binding and entry of HIV into CD4-expressing host cells **(Chap. 197)**.

Significant advances have been made in defining the signaling pathways through which cytokines exert their intracellular effects. The Janus family of protein tyrosine kinases (JAK) is a critical element involved in signaling via the hematopoietin receptors. Four JAK kinases, JAK1, JAK2, JAK3, and Tyk2, preferentially bind different cytokine receptor subunits. Cytokine binding to its receptor brings the cytokine receptor subunits into apposition and allows a pair of JAKs to transphosphorylate and activate one another. The JAKs then phosphorylate the receptor on the tyrosine residues and allow signaling molecules to bind to the receptor, whereby the signaling molecules become phosphorylated. Signaling molecules bind the receptor because they have domains (SH2, or src homology 2 domains) that can bind phosphorylated tyrosine residues. There are a number of these important signaling molecules that bind the receptor, such as the adapter molecule SHC, which can couple the receptor to the activation of the mitogen-activated protein kinase pathway. In addition, an important class of substrate of the JAKs is the signal transducers and activators of transcription (STAT) family of transcription factors. STATs have SH2 domains that enable them to bind to phosphorylated receptors, where they are then phosphorylated by the JAKs. It appears that different STATs have specificity for different receptor subunits. The STATs then dissociate from the receptor and translocate to the nucleus, bind to DNA motifs that they recognize, and regulate gene expression. The STATs preferentially bind DNA motifs that are slightly different from one another and thereby control transcription of specific genes. The importance of this pathway is particularly relevant to lymphoid development. Mutations of JAK3 itself also result in a disorder identical to X-SCID; however, because JAK3 is found on chromosome 19 and not on the X chromosome, JAK3 deficiency occurs in boys and girls **(Chap. 344)**.

THE ADAPTIVE IMMUNE SYSTEM

Adaptive immunity is characterized by antigen-specific responses to a foreign antigen or pathogen. A key feature of adaptive immunity is that following the initial contact with antigen (*immunologic priming*), subsequent antigen exposure leads to more rapid and vigorous immune responses (*immunologic memory*). The adaptive immune system consists of dual limbs of cellular and humoral immunity. The principal effectors of cellular immunity are T lymphocytes, whereas the principal effectors of humoral immunity are B lymphocytes. Both B and T lymphocytes derive from a common stem cell **(Fig. 342-5)**.



Heavy-chain genes	D-J rearranging	VDJ rearranging	VDJ rearranged	VDJ rearranged	VDJ rearranged	VDJ rearranged
Light-chain genes	Germline	Germline	Germline	VJ rearranging	VJ rearranged	VJ rearranged
Surface Ig	Absent	Absent	μ H-chain at surface as part of the Pre-B receptor (orange) containing surrogate light chain (SLC). Receptor is mainly intracellular	μ H-chain in cytoplasm	IgM expressed on cell surface	IgD and IgM made from alternatively spliced H-chain transcripts
Surface marker proteins	CD34 CD10 CD38	CD10 CD19 CD38 CD40	CD19 CD20 CD38	CD19 CD20 CD38	CD19 CD20	CD19 CD20 CD21

FIGURE 342-5 Development stages of T and B cells. Elements of the developing T and B cell receptor for antigen are shown schematically. The classification into the various stages of B cell development is primarily defined by rearrangement of the immunoglobulin (Ig) heavy (H) and light (L) chain genes and by the absence or presence of specific surface markers. The classification of stages of T cell development is primarily defined by cell-surface marker protein expression (sCD3, surface CD3 expression; cCD3, cytoplasmic CD3 expression; TCR, T cell receptor). For B cell development, the pre-B cell receptor is shown as a blue-orange B cell receptor. *(Adapted from CA Janeway et al [eds]: Immunobiology, 9th ed. New York, Garland, 2016; with permission.)*

The proportion and distribution of immunocompetent cells in various tissues reflect cell traffic, homing patterns, and functional capabilities. Bone marrow is the major site of maturation of B cells, monocytes-macrophages, DCs, and granulocytes and contains pluripotent stem cells that, under the influence of various colony-stimulating factors, are capable of giving rise to all hematopoietic cell types. T cell precursors also arise from hematopoietic stem cells and home to the thymus for maturation. Mature T lymphocytes, B lymphocytes, monocytes, and DCs enter the circulation and home to peripheral lymphoid organs (lymph nodes, spleen) and mucosal surface-associated lymphoid tissue (gut, genitourinary, and respiratory tracts) as well as the skin and mucous membranes and await activation by foreign antigen.

T Cells The pool of effector T cells is established in the thymus early in life and is maintained throughout life both by new T cell production in the thymus and by antigen-driven expansion of virgin peripheral T cells into "memory" T cells that reside in peripheral lymphoid organs. The thymus exports ~2% of the total number of thymocytes per day throughout life, with the total number of daily thymic emigrants decreasing by ~3% per year during the first four decades of life.

Mature T lymphocytes constitute 70–80% of normal peripheral blood lymphocytes (only 2% of the total-body lymphocytes are contained in peripheral blood), 90% of thoracic duct lymphocytes, 30–40% of lymph node cells, and 20–30% of spleen lymphoid cells. In lymph nodes, T cells occupy deep paracortical areas around B cell germinal centers, and in the spleen, they are located in periarteriolar areas of white pulp **(Chap. 62)**. T cells are the primary effectors of cell-mediated immunity, with subsets of T cells maturing into CD8+ cytotoxic T cells capable of lysis of virus-infected or foreign cells (short-lived effector T cells) and CD4+ T cells capable of T cell help for CD8+ T cell and B cell development. Two populations of long-lived memory T cells are triggered by infections: effector memory and central memory T cells. Effector memory T cells reside in nonlymphoid organs and respond rapidly to repeated pathogenic infections with cytokine production and cytotoxic functions to kill virus-infected cells. Central memory T cells home to lymphoid organs where they replenish long- and short-lived and effector memory T cells as needed.

In general, CD4+ T cells are the primary regulatory cells of T and B lymphocyte and monocyte function by the production of cytokines and by direct cell contact (Fig. 342-2). In addition, T cells regulate erythroid cell maturation in bone marrow and, through cell contact (CD40 ligand), have an important role in activation of B cells and induction of Ig isotype switching. Considerable evidence now exists that colonization of the gut by commensal bacteria (the gut microbiome) is responsible for expansion of the peripheral CD4+ T cell compartment in normal children and adults.

Human T cells express cell-surface proteins that mark stages of intrathymic T cell maturation or identify specific functional subpopulations of mature T cells. Many of these molecules mediate or participate in important T cell functions (Table 342-1, Fig. 342-5).

The earliest identifiable T cell precursors in bone marrow are CD34+ pro-T cells (i.e., cells in which TCR genes are neither rearranged nor expressed). In the thymus, CD34+ T cell precursors begin cytoplasmic (c) synthesis of components of the CD3 complex of TCR-associated molecules (Fig. 342-5). Within T cell precursors, TCR for antigen gene rearrangement yields two T cell lineages, expressing either TCR-αβ chains or TCR-γδ chains. T cells expressing the TCR-αβ chains constitute the majority of peripheral T cells in blood, lymph node, and spleen and terminally differentiate into either CD4+ or CD8+ cells. Cells expressing TCR-γδ chains circulate as a minor population in blood; their functions, although not fully understood, have been postulated to be those of immune surveillance at epithelial surfaces and cellular defenses against mycobacterial organisms and other intracellular bacteria through recognition of bacterial lipids.

In the thymus, the recognition of self-peptides on thymic epithelial cells, thymic macrophages, and DCs plays an important role in shaping T cell repertoire. As immature cortical thymocytes begin to express surface TCR for antigen, thymocytes with TCRs capable of interacting with self-peptides in the context of self-MHC antigens with low affinity are activated and survive (positive selection). Thymocytes with TCRs that are incapable of binding to self-MHC antigens or bind with high affinity, die of attrition (no selection) or by apoptosis (negative selection). Thymocytes that are positively selected undergo maturation into CD4 or CD8 single positive T cells, and then migrate to the thymus medulla where they interact with self-peptide-self-MHC molecules, where they can again undergo selection. The purpose of negative and positive thymocyte selection is to eliminate potential pathogenic autoreactive T cells, and at the same time, select a repertoire of mature T cells capable of recognizing foreign antigens.

Mature thymocytes that are positively selected are functional MHC class II -restricted CD4+ T cells (Figure 342-2), or they are CD8+ T cells destined to become CD8+ MHC class I -restricted cytotoxic T cells. *MHC class I -* or *class II -restriction* means that T cells recognize antigen peptide fragments only when they are presented in the antigen-recognisite of a class I or class II MHC molecule, respectively **(Chap. 343)**. After thymocyte maturation and selection, CD4 and CD8 thymocytes leave the thymus and migrate to the peripheral immune system. The thymus can continue to be a contributor to the peripheral immune system well into adult life, both normally and when the peripheral T cell pool is damaged, such as occurs in AIDS and cancer chemotherapy.

MOLECULAR BASIS OF T CELL RECOGNITION OF ANTIGEN The TCR for antigen is a complex of molecules consisting of an antigen-binding heterodimer of either αβ or γδ chains noncovalently linked with five CD3 subunits (γ, δ, ε, ζ, and η) **(Fig. 342-6)**. The CD3 ζ chains are either disulfide-linked homodimers (CD3-ζ_2) or disulfide-linked heterodimers composed of one ζ chain and one η chain. TCR-αβ or TCR-γδ molecules must be associated with CD3 molecules to be inserted into the T cell-surface membrane, TCRα being paired with TCR-β and TCR-γ being paired with TCR-δ. Molecules of the CD3 complex mediate transduction of T cell activation signals via TCRs, whereas TCR-α and β or γ and δ molecules combine to form the TCR antigen-binding site.

The α, β, γ, and δ TCR for antigen molecules have amino acid sequence homology and structural similarities to immunoglobulin heavy and light chains and are members of the *immunoglobulin gene superfamily* of molecules. The genes encoding TCR molecules are encoded as clusters of gene segments that rearrange during the course of T cell maturation. This creates an efficient and compact mechanism for housing the diversity requirements of antigen receptor molecules. The TCR-α chain is on chromosome 14 and consists of a series of V (variable), J (joining), and C (constant) regions. The TCR-β chain is on chromosome 7 and consists of multiple V, D (diversity), J, and C TCR-β loci. The TCR-γ chain is on chromosome 7, and the TCR-δ chain is in the middle of the TCR-α locus on chromosome 14. Thus, molecules of the TCR for antigen have constant (framework) and variable regions, and the gene segments encoding the α, β, γ, and δ chains of these molecules are recombined and selected in the thymus, culminating in synthesis of the completed molecule. In both T and B cell precursors (see below), DNA rearrangements of antigen receptor genes involve the same enzymes, recombinase activating gene RAG1 and RAG2, both DNA-dependent protein kinases.

TCR diversity is created by the different V, D, and J segments that are possible for each receptor chain by the many permutations of V, D, and J segment combinations, by "N-region diversification" due to the addition of nucleotides at the junction of rearranged gene segments, and by the pairing of individual chains to form a TCR dimer. As T cells mature in the thymus, the repertoire of antigen-reactive T cells is modified by selection processes that eliminate many autoreactive T cells, enhance the proliferation of cells that function appropriately with self-MHC molecules and antigen, and allow T cells with nonproductive TCR rearrangements to die.

TCR-αβ cells do not recognize native protein or carbohydrate antigens. Instead, T cells recognize only short (~9–13 amino acids) peptide fragments derived from protein antigens taken up or produced in APCs. Foreign antigens may be taken up by endocytosis into acidified intracellular vesicles or by phagocytosis and degraded into small peptides that associate with MHC class II molecules (exogenous antigen-presentation

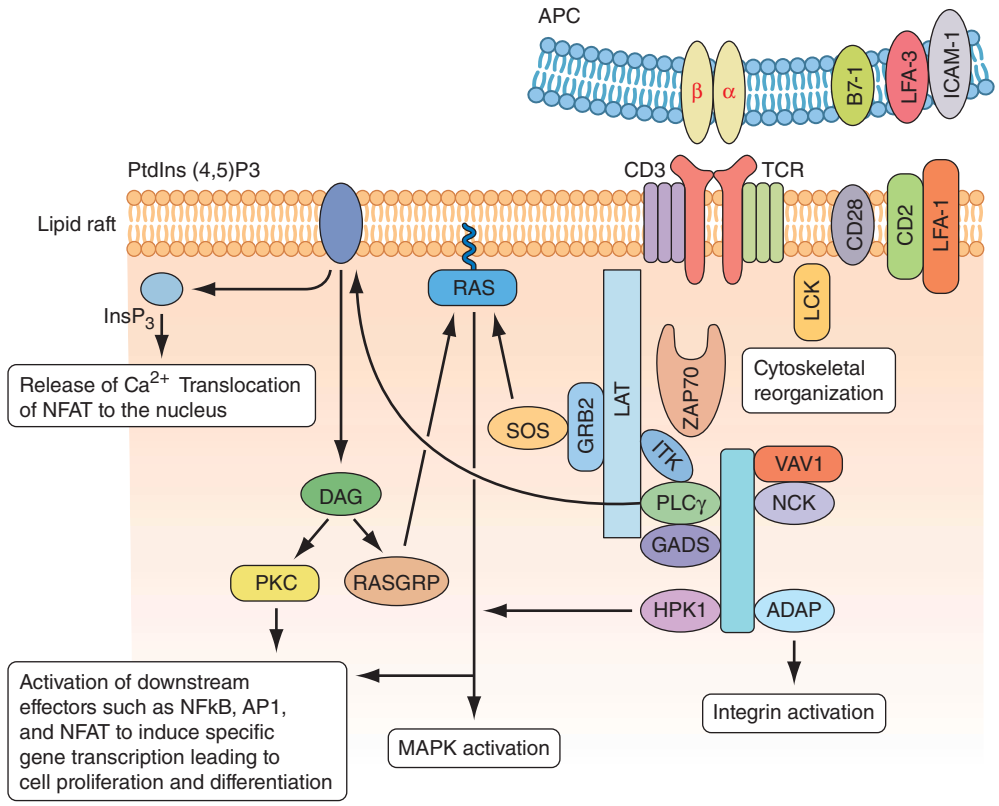


FIGURE 342-6 Signaling through the T cell receptor. Activation signals are mediated via immunoreceptor tyrosine-based activation (ITAM) sequences in LAT and CD3 chains (*blue bars*) that bind to enzymes and transduce activation signals to the nucleus via the indicated intracellular activation pathways. Ligation of the T cell receptor (TCR) by MHC complexed with antigen results in sequential activation of LCK and γ-chain-associated protein kinase of 70 kDa (ZAP70). ZAP70 phosphorylates several downstream targets, including LAT (linker for activation of T cells) and SLP76 (SCR homology 2 [SH2] domain-containing leukocyte protein of 76 kDa). SLP76 is recruited to membrane-bound LAT through its constitutive interaction with GADS (GRB2-related adaptor protein). Together, SLP76 and LAT nucleate a multimolecular signaling complex, which induces a host of downstream responses, including calcium flux, mitogen-activated protein kinase (MAPK) activation, integrin activation, and cytoskeletal reorganization. APC, antigen-presenting cell. *(Adapted from GA Koretzky et al: Nat Rev Immunol 6:67, 2006; with permission from Macmillan Publishers Ltd. Copyright 2006.)*

pathway). Other foreign antigens arise endogenously in the cytosol (such as from replicating viruses) and are broken down into small peptides that associate with MHC class I molecules (endogenous antigen-presenting pathway). Thus, APCs proteolytically degrade foreign proteins and display peptide fragments embedded in the MHC class I or II antigen-recognition site on the MHC molecule surface, where foreign peptide fragments are available to bind to TCR-αβ or TCR-γδ chains of reactive T cells. CD4 molecules act as adhesives and, by direct binding to MHC class II (DR, DQ, or DP) molecules, stabilize the interaction of TCR with peptide antigen (Fig. 342-6). Similarly, CD8 molecules also act as adhesives to stabilize the TCR-antigen interaction by direct CD8 molecule binding to MHC class I (A, B, or C) molecules.

Antigens that arise in the cytosol and are processed via the endogenous antigen-presentation pathway are cleaved into small peptides by a complex of proteases called the *proteasome*. From the proteasome, antigen peptide fragments are transported from the cytosol into the lumen of the endoplasmic reticulum by a heterodimeric complex termed *transporters associated with antigen processing*, or TAP proteins. There, MHC class I molecules in the endoplasmic reticulum membrane physically associate with processed cytosolic peptides. Following peptide association with class I molecules, peptide-class I complexes are exported to the Golgi apparatus, and then to the cell surface, for recognition by CD8+ T cells.

Antigens taken up from the extracellular space via endocytosis into intracellular acidified vesicles are degraded by vesicle proteases into peptide fragments. Intracellular vesicles containing MHC class II molecules fuse with peptide-containing vesicles, thus allowing peptide fragments to physically bind to MHC class II molecules. Peptide-MHC class II complexes are then transported to the cell surface for recognition by CD4+ T cells **(Chap. 343)**.

Whereas it is generally agreed that the TCR-αβ receptor recognizes peptide antigens in the context of MHC class I or class II molecules, lipids in the cell wall of intracellular bacteria such as *M. tuberculosis* can also be presented to a wide variety of T cells, including subsets of TCR-γδ T cells, and a subset of CD8+ TCR-αβ T cells. Importantly, bacterial lipid antigens are not presented in the context of MHC class I or II molecules, but rather are presented in the context of MHC-related CD1 molecules. Some γδ T cells that recognize lipid antigens via CD1 molecules have very restricted TCR usage, do not need antigen priming to respond to bacterial lipids, and may actually be a form of innate rather than acquired immunity to intracellular bacteria.

Just as foreign antigens are degraded and their peptide fragments presented in the context of MHC class I or class II molecules on APCs, endogenous self-proteins also are degraded, and self-peptide fragments are presented to T cells in the context of MHC class I or class II molecules on APCs. In peripheral lymphoid organs, there are T cells that are capable of recognizing self-protein fragments but normally are *anergic* or *tolerant*, i.e., nonresponsive to self-antigenic stimulation, due to lack of self-antigen upregulating APC *co-stimulatory molecules* such as B7-1 (CD80) and B7-2 (CD86) (see below).

Once engagement of mature T cell TCR by foreign peptide occurs in the context of self-MHC class I or class II molecules, binding of non-antigen-specific adhesion ligand pairs such as CD54-CD11/CD18 and CD58-CD2 stabilizes MHC peptide-TCR binding, and the expression of these adhesion molecules is upregulated (Fig. 342-6). Once antigen ligation of the TCR occurs, the T cell membrane is partitioned into *lipid membrane microdomains*, or *lipid rafts*, that coalesce the key signaling molecules TCR/CD3 complex, CD28, CD2, LAT (linker for activation of T cells), intracellular activated (dephosphorylated) src family protein tyrosine kinases (PTKs), and the key CD3ζ-associated protein-70 (ZAP-70) PTK (Fig. 342-6). Importantly, during T cell activation, the CD45 molecule, with protein tyrosine phosphatase activity, is partitioned away from the TCR complex to allow activating phosphorylation events

to occur. The coalescence of signaling molecules of activated T lymphocytes in *microdomains* has suggested that T cell-APC interactions can be considered *immunologic synapses*, analogous in function to neuronal synapses.

After TCR-MHC binding is stabilized, activation signals are transmitted through the cell to the nucleus and lead to the expression of gene products important in mediating the wide diversity of T cell functions such as the secretion of IL-2. The TCR does not have intrinsic signaling activity but is linked to a variety of signaling pathways via ITAMs expressed on the various CD3 chains that bind to proteins that mediate signal transduction. Each of the pathways results in the activation of particular transcription factors that control the expression of cytokine and cytokine receptor genes. Thus, antigen-MHC binding to the TCR induces the activation of the src family of PTKs, Fyn and Lck (Lck is associated with CD4 or CD8 co-stimulatory molecules); phosphorylation of CD3ζ chain; activation of the related tyrosine kinases ZAP-70 and Syk; and downstream activation of the calcium-dependent calcineurin pathway, the ras pathway, and the protein kinase C pathway. Each of these pathways leads to activation of specific families of transcription factors (including *NF-AT*, *fos* and *jun*, and *rel/NF-κB*) that form heteromultimers capable of inducing expression of IL-2, IL-2 receptor, IL-4, TNF-α, and other T cell mediators.

In addition to the signals delivered to the T cell from the TCR complex and CD4 and CD8, molecules on the T cell, such as CD28 and inducible co-stimulator (ICOS), and molecules on DCs, such as B7-1 (CD80) and B7-2 (CD86), also deliver important co-stimulatory signals that upregulate T cell cytokine production and are essential for T cell activation. If signaling through CD28 or ICOS does not occur, or if CD28 is blocked, the T cell becomes anergic rather than activated (see "Immune Tolerance and Autoimmunity" below). CTLA-4 (CD152) is similar to CD28 in its ability to bind CD80 and CD86. Unlike CD28, CTLA-4 transmits an inhibitory signal to T cells, acting as an off switch.

T CELL EXHAUSTION IN VIRAL INFECTIONS AND CANCER In chronic viral infections such as HIV-1, hepatitis C virus, and hepatitis B virus and in chronic malignancies, the persistence of antigen disrupts memory T cell function, resulting in defects in memory T cell responses. This has been defined as *T cell exhaustion* and is associated with T cell programmed cell death protein 1 (PD-1) (CD279) expression. Exhausted T cells have compromised proliferation and lose the ability to produce effector molecules, like IL-2, TNF-α, and IFN-γ. PD-1 downregulates T cell responses and is associated with T cell exhaustion and disease progression. For this reason, inhibition of T cell PD-1 activity to enhance effector T cell function is being explored as a target for immunotherapy in both viral infections and certain malignancies.

T CELL SUPERANTIGENS Conventional antigens bind to MHC class I or II molecules in the groove of the αβ heterodimer and bind to T cells via the V regions of the TCR-α and β chains. In contrast, superantigens bind directly to the lateral portion of the TCR-β chain and MHC class II β chain and stimulate T cells based solely on the Vβ gene segment used independent of the D, J, and Vα sequences present. *Superantigens* are protein molecules capable of activating up to 20% of the peripheral T cell pool, whereas conventional antigens activate <1 in 10,000 T cells. T cell superantigens include staphylococcal enterotoxins and other bacterial products. Superantigen stimulation of human peripheral T cells occurs in the clinical setting of *staphylococcal toxic shock syndrome*, leading to massive overproduction of T cell cytokines that leads to hypotension and shock **(Chap. 142)**.

B CELLS Mature B cells constitute 10–15% of human peripheral blood lymphocytes, 20–30% of lymph node cells, 50% of splenic lymphocytes, and ~10% of bone marrow lymphocytes. B cells express on their surface intramembrane immunoglobulin (Ig) molecules that function as BCRs for antigen in a complex of Ig-associated α and β signaling molecules with properties similar to those described in T cells **(Fig. 342-7)**. Unlike T cells, which recognize only processed peptide fragments of conventional antigens embedded in the notches of MHC class I and class II antigens of APCs, B cells are capable of recognizing and proliferating to whole unprocessed native antigens via antigen binding to B cell–surface Ig (sIg) receptors. B cells also express surface receptors for the Fc region of IgG molecules (CD32) as well as receptors for activated complement components (C3d or CD21, C3b or CD35). The primary function of B cells is to produce antibodies. B cells also serve as APCs and are highly efficient at antigen processing. Their antigen-presenting function is enhanced by a variety of cytokines. Mature B cells are derived from bone marrow precursor cells that arise continuously throughout life (Fig. 342-5).

B lymphocyte development can be separated into antigen-independent and antigen-dependent phases. Antigen-independent B cell development occurs in primary lymphoid organs and includes all stages of B cell maturation up to the sIg+ mature B cell. Antigen-dependent B cell maturation is driven by the interaction of antigen with the mature B cell sIg, leading to memory B cell induction, Ig class switching, and plasma cell formation. Antigen-dependent stages of B cell maturation occur in secondary lymphoid organs, including lymph node, spleen, and gut Peyer's patches. In contrast to the T cell repertoire that is generated intrathymically before contact with foreign antigen, the repertoire of B cells expressing diverse antigen-reactive sites is modified by further alteration of Ig genes after stimulation by antigen—a process called *somatic hypermutation*—that occurs in lymph node germinal centers.

During B cell development, diversity of the antigen-binding variable region of Ig is generated by an ordered set of Ig gene rearrangements that are similar to the rearrangements undergone by TCR α, β, γ, and δ genes. For the heavy chain, there is first a rearrangement of D segments to J segments, followed by a second rearrangement between a V gene segment and the newly formed D-J sequence; the C segment is aligned to the V-D-J complex to yield a functional Ig heavy chain gene (V-D-J-C). During later stages, a functional κ or γ light chain gene is generated by rearrangement of a V segment to a J segment, ultimately yielding an intact Ig molecule composed of heavy and light chains.

The process of Ig gene rearrangement is regulated and results in a single antibody specificity produced by each B cell, with each Ig molecule comprising one type of heavy chain and one type of light chain. Although each B cell contains two copies of Ig light and heavy chain genes, only one gene of each type is productively rearranged and expressed in each B cell, a process termed *allelic exclusion*.

There are ~300 V_κ genes and 5 J_κ genes, resulting in the pairing of V_κ and J_κ genes to create >1500 different kappa light chain combinations. There are ~70 V_λ genes and 4 J_λ genes for >280 different lambda light chain combinations. The number of distinct light chains that can be generated is increased by somatic mutations within the V and J genes, thus creating large numbers of possible specificities from a limited amount of germline genetic information. As noted above, in heavy chain Ig gene rearrangement, the VH domain is created by the joining of three types of germline genes called V_H, D_H, and J_H, thus allowing for even greater diversity in the variable region of heavy chains than of light chains.

The most immature B cell precursors (early pro-B cells) lack cytoplasmic Ig (cIg) and sIg (Fig. 342-5). The large pre-B cell is marked by the acquisition of the surface pre-BCR composed of μ heavy (H) chains and a pre-B light chain, termed V pre-B. V pre-B is a surrogate light chain receptor encoded by the nonrearranged V pre-B and the γ5 light chain locus (the pre-BCR). Pro- and pre-B cells are driven to proliferate and mature by signals from bone marrow stroma—in particular, IL-7. Light chain rearrangement occurs in the small pre-B cell stage such that the full BCR is expressed at the immature B cell stage. Immature B cells have rearranged Ig light chain genes and express sIgM. As immature B cells develop into mature B cells, sIgD is expressed as well as sIgM. At this point, B lineage development in bone marrow is complete, and B cells exit into the peripheral circulation and migrate to secondary lymphoid organs to encounter specific antigens.

Random rearrangements of Ig genes occasionally generate self-reactive antibodies, and mechanisms must be in place to correct these mistakes. One such mechanism is BCR editing, whereby autoreactive BCRs are mutated to not react with self-antigens. If receptor editing is unsuccessful in eliminating autoreactive B cells, then autoreactive B cells undergo negative selection in the bone marrow through induction of apoptosis after BCR engagement of self-antigen.

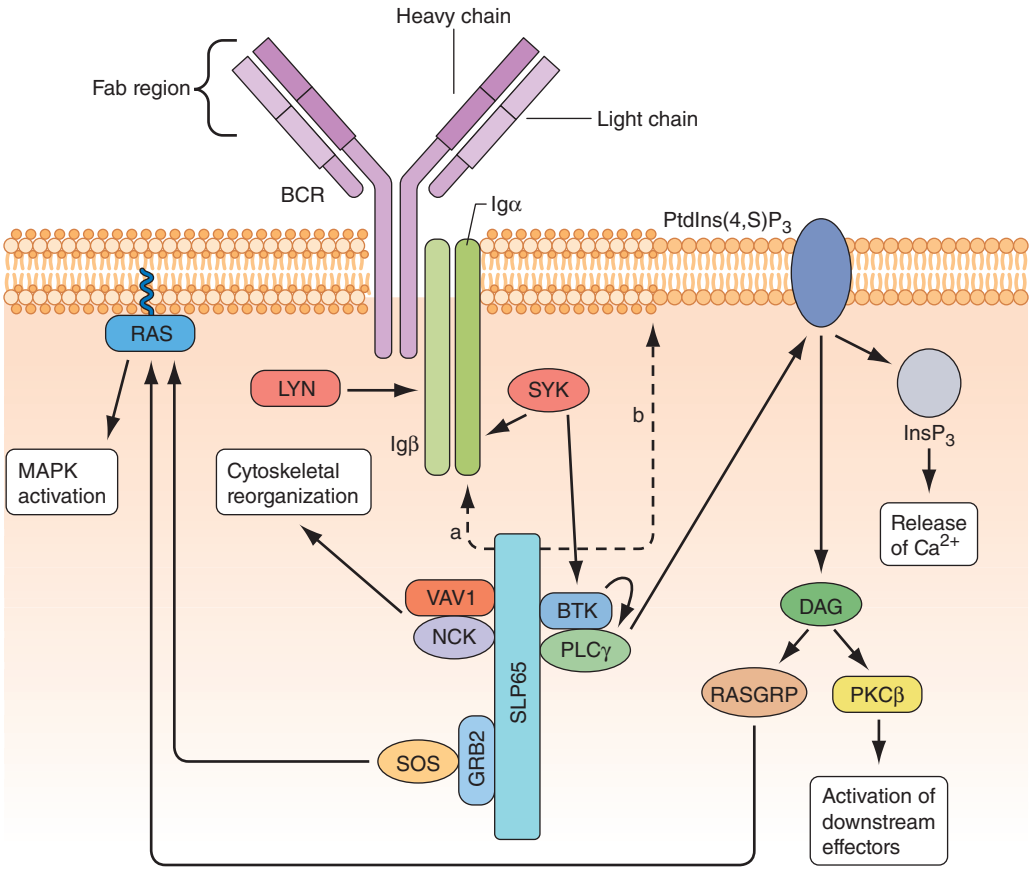


FIGURE 342-7 B cell receptor (BCR) activation results in the sequential activation of protein tyrosine kinases, which results in the formation of a signaling complex and activation of downstream pathways as shown. Whereas SLP76 is recruited to the membrane through GADS and LAT, the mechanism of SLP65 recruitment is unclear. Studies have indicated two mechanisms: (a) direct binding by the SH2 domain of SLP65 to immunoglobulin (Ig) of the BCR complex or (b) membrane recruitment through a leucine zipper in the amino terminus of SLP65 and an unknown binding partner. ADAP, adhesion- and degranulation-promoting adaptor protein; AP1, activator protein 1; BTK, Bruton's tyrosine kinase; DAG, diacylglycerol; GRB2, growth factor receptor-bound protein 2; HPK1, hematopoietic progenitor kinase 1; $InsP_3$, inositol-1,4,5-trisphosphate; ITK, interleukin-2-inducible T cell kinase; NCK, noncatalytic region of tyrosine kinase; NF-B, nuclear factor B; PKC, protein kinase C; PLC, phospholipase C; $PtdIns(4,5)P_2$, phosphatidylinositol-4,5-bisphosphate; RASGRP, RAS guanyl-releasing protein; SOS, son of sevenless homologue; SYK, spleen tyrosine kinase. *(Adapted from GA Koretzky et al: Nat Rev Immunol 6:67, 2006; with permission from Macmillan Publishers Ltd. Copyright 2006.)*

After leaving the bone marrow, B cells populate peripheral B cell sites, such as lymph node and spleen, and await contact with foreign antigens that react with each B cell's clonotypic receptor. Antigen-driven B cell activation occurs through the BCR, and a process known as *somatic hypermutation* takes place whereby point mutations in rearranged H- and L-genes give rise to mutant sIg molecules, some of which bind antigen better than the original sIg molecules. Somatic hypermutation, therefore, is a process whereby memory B cells in peripheral lymph organs have the best binding, or the highest-affinity antibodies. This overall process of generating the best antibodies is called *affinity maturation of antibody*.

Lymphocytes that synthesize IgG, IgA, and IgE are derived from sIgM+, sIgD+ mature B cells. Ig class switching occurs in lymph node and other peripheral lymphoid tissue germinal centers. CD40 on B cells and CD40 ligand on T cells constitute a critical co-stimulatory receptor-ligand pair of immune-stimulatory molecules. Pairs of CD40+ B cells and CD40 ligand+ T cells bind and drive B cell Ig class switching via T cell-produced cytokines such as IL-4 and TGF-β. IL-1, -2, -4, -5, and -6 synergize to drive mature B cells to proliferate and differentiate into Ig-secreting cells.

Humoral Mediators of Adaptive Immunity: Immunoglobulins

Immunoglobulins are the products of differentiated B cells and mediate the humoral arm of the immune response. The primary functions of antibodies are to bind specifically to antigen and bring about the inactivation or removal of the offending toxin, microbe, parasite, or other foreign substance from the body. The structural basis of Ig molecule function and Ig gene organization has provided insight into the role of antibodies in normal protective immunity, pathologic immune-mediated damage by immune complexes, and autoantibody formation against host determinants.

All immunoglobulins have the basic structure of two heavy and two light chains (Fig. 342-7). Immunoglobulin isotype (i.e., G, M, A, D, E) is determined by the type of Ig heavy chain present. IgG and IgA isotypes can be divided further into subclasses (G1, G2, G3, G4, and A1, A2) based on specific antigenic determinants on Ig heavy chains. The characteristics of human immunoglobulins are outlined in **Table 342-12**. The four chains are covalently linked by disulfide bonds. Each chain is made up of a V region and C regions (also called *domains*), themselves made up of units of ~110 amino acids. Light chains have one variable (V_L) and one constant (C_L) unit; heavy chains have one variable unit (V_H) and three or four constant (C_H) units, depending on isotype. As the name suggests, the constant, or C, regions of Ig molecules are made up of homologous sequences and share the same primary structure as all other Ig chains of the same isotype and subclass. Constant regions are involved in biologic functions of Ig molecules. The C_H2 domain of IgG and the C_H4 units of IgM are involved with the binding of the C1q portion of C1 during complement activation. The C_H region at the carboxy-terminal end of the IgG molecule, the Fc region, binds to surface Fc receptors (CD16, CD32, CD64) of macrophages, DCs, NK cells, B cells, neutrophils, and eosinophils. The Fc of IgA binds to FcαR (CD89), and the Fc of IgE binds to FcεR (CD23).

Variable regions (V_L and V_H) constitute the antibody-binding (Fab) region of the molecule. Within the V_L and V_H regions are hypervariable regions (extreme sequence variability) that constitute the antigen-binding site unique to each Ig molecule. The idiotype is defined as the

TABLE 342-12 Physical, Chemical, and Biologic Properties of Human Immunoglobulins

PROPERTY	IgG	IgA	IgM	IgD	IgE
Usual molecular form	Monomer	Monomer, dimer	Pentamer, hexamer	Monomer	Monomer
Other chains	None	J chain, SC	J chain	None	None
Subclasses	G1, G2, G3, G4	A1, A2	None	None	None
Heavy chain allotypes	Gm (=30)	No A1, A2m (2)	None	None	None
Molecular mass, kDa	150	160, 400	950, 1150	175	190
Serum level in average adult, mg/mL	9.5–12.5	1.5–2.6	0.7–1.7	0.04	0.0003
Percentage of total serum Ig	75–85	7–15	5–10	0.3	0.019
Serum half-life, days	23	6	5	3	2.5
Synthesis rate, mg/kg per day	33	65	7	0.4	0.016
Antibody valence	2	2, 4	10, 12	2	2
Classical complement activation	+(G1, 2?, 3)	–	++	–	–
Alternate complement activation	+(G4)	+	–	+	–
Binding cells via Fc	Macrophages, neutrophils, large granular lymphocytes	Lymphocytes	Lymphocytes	None	Mast cells, basophils, B cells
Biologic properties	Placental transfer, secondary Ab for most antipathogen responses	Secretory immunoglobulin	Primary Ab responses	Marker for mature B cells	Allergy, antiparasite responses

Source: After L Carayannopoulos, JD Capra, in WE Paul (ed): *Fundamental Immunology*, 3rd ed. New York, Raven, 1993; with permission.

specific region of the Fab portion of the Ig molecule to which antigen binds. Antibodies against the idiotype portion of an antibody molecule are called *anti-idiotype antibodies*. The formation of such antibodies in vivo during a normal B cell antibody response may generate a negative (or "off") signal to B cells to terminate antibody production.

IgG constitutes ~75–85% of total serum immunoglobulin. The four IgG subclasses are numbered in order of their level in serum, IgG1 being found in greatest amounts and IgG4 the least. IgG subclasses have clinical relevance in their varying ability to bind macrophage and neutrophil Fc receptors and to activate complement (Table 342-12). Moreover, selective deficiencies of certain IgG subclasses give rise to clinical syndromes in which the patient is inordinately susceptible to bacterial infections. IgG antibodies are frequently the predominant antibody made after rechallenge of the host with antigen (secondary antibody response).

IgM antibodies normally circulate as a 950-kDa pentamer with 160-kDa bivalent monomers joined by a molecule called the *J chain*, a 15-kDa nonimmunoglobulin molecule that also effects polymerization of IgA molecules. IgM is the first immunoglobulin to appear in the immune response (primary antibody response) and is the initial type of antibody made by neonates. Membrane IgM in the monomeric form also functions as a major antigen receptor on the surface of mature B cells (Table 342-12). IgM is an important component of immune complexes in autoimmune diseases. For example, IgM antibodies against IgG molecules (rheumatoid factors) are present in high titers in *rheumatoid arthritis*, other collagen diseases, and some infectious diseases (*subacute bacterial endocarditis*).

IgA constitutes only 7–15% of total serum immunoglobulin but is the predominant class of immunoglobulin in secretions. IgA in secretions (tears, saliva, nasal secretions, gastrointestinal tract fluid, and human milk) is in the form of secretory IgA (sIgA), a polymer consisting of two IgA monomers, a joining molecule, again termed the J chain, and a glycoprotein called the *secretory protein*. Of the two IgA subclasses, IgA1 is primarily found in serum, whereas IgA2 is more prevalent in secretions. IgA fixes complement via the alternative complement pathway and has potent antiviral activity in humans by prevention of virus binding to respiratory and gastrointestinal epithelial cells.

IgD is found in minute quantities in serum and, together with IgM, is a major receptor for antigen on the naïve B cell surface. IgE, which is present in serum in very low concentrations, is the major class of immunoglobulin involved in arming mast cells and basophils by binding to these cells via the Fc region. Antigen cross-linking of IgE molecules on basophil and mast cell surfaces results in release of mediators of the immediate hypersensitivity (allergic) response (Table 342-12).

CELLULAR INTERACTIONS IN REGULATION OF NORMAL IMMUNE RESPONSES

The net result of activation of the humoral (B cell) and cellular (T cell) arms of the adaptive immune system by foreign antigen is the elimination of antigen directly by specific effector T cells or in concert with specific antibody. Figure 342-2 is a simplified schematic diagram of the T and B cell responses indicating some of these cellular interactions.

The expression of adaptive immune cell function is the result of a complex series of immunoregulatory events that occur in phases. Both T and B lymphocytes mediate immune functions, and each of these cell types, when given appropriate signals, passes through stages, from activation and induction through proliferation, differentiation, and ultimately effector functions. The effector function expressed may be at the end point of a response, such as secretion of antibody by a differentiated plasma cell, or it might serve a regulatory function that modulates other functions, such as is seen with CD4+ and CD8+ T lymphocytes that modulate both differentiation of B cells and activation of CD8+ cytotoxic T cells.

CD4 helper T cells can be subdivided on the basis of cytokines produced (Fig. 342-2). Activated T_H1-type helper T cells secrete IL-2, IFN-γ, IL-3, TNF-a, GM-CSF, and TNF-β, whereas activated T_H2-type helper T cells secrete IL-3, -4, -5, -6, -10, and -13. T_H1 CD4+ T cells, through elaboration of IFN-γ, have a central role in mediating intracellular killing by a variety of pathogens. T_H1 CD4+ T cells also provide T cell help for generation of cytotoxic T cells and some types of opsonizing antibody, and they generally respond to antigens that lead to delayed hypersensitivity types of immune responses for many intracellular viruses and bacteria (such as HIV or *M. tuberculosis*). In contrast, T_H2 cells have a primary role in regulatory humoral immunity and isotype switching. T_H2 cells, through production of IL-4 and IL-10, have a regulatory role in limiting proinflammatory responses mediated by T_H1 cells (Fig. 342-2). In addition, T_H2 CD4+ T cells provide help to B cells for specific Ig production and respond to antigens that require high antibody levels for foreign antigen elimination (extracellular encapsulated bacteria such as *Streptococcus pneumoniae* and certain parasite infections). Additional subsets of the CD4 T_H cells have been described, one of which is termed T_H17 that secrete cytokines IL-17, -22, and -26. T_H17 cells have been shown to play a role in autoimmune inflammatory disorders in addition to defense against extracellular bacteria and fungi, particularly at mucosal surfaces. T_H9 cells are defined by their secretion of IL-9 and have been shown to play a role in atopic disease, inflammatory bowel disease, and in anti-tumor immunity. Moreover, the Tfh subset of helper T cells is crucial for providing the necessary signals to B cells in germinal centers to undergo affinity maturation.

In summary, the type of T cell response generated in an immune response is determined by the microbe PAMPs presented to the DCs, the TLRs on the DCs that become activated, the types of DCs that are activated, and the cytokines that are produced (Table 342-4). Commonly, myeloid DCs produce IL-12 and activate T_H1 T cell responses that result in IFN-γ and cytotoxic T cell induction, and plasmacytoid DCs produce IFN-α and lead to T_H2 responses that result in IL-4 production and enhanced antibody responses.

As shown in Fig. 342-2, upon activation by DCs, T cell subsets that produce IL-2, IL-3, IFN-γ, and/or IL-4, -5, -6, -10, and -13 are generated and exert positive and negative influences on effector T and B cells. For B cells, trophic effects are mediated by a variety of cytokines, particularly T cell–derived IL-3, -4, -5, and -6, that act at sequential stages of B cell maturation, resulting in B cell proliferation, differentiation, and ultimately antibody secretion. For cytotoxic T cells, trophic factors include inducer T cell secretion of IL-2, IFN-γ, and IL-12.

Important types of immunomodulatory T cells that control immune responses are *CD4+ and CD8+ T regulatory cells*. These cells express the α chain of the IL-2 receptor (CD25), produce IL-10, and suppress both T and B cell responses. T regulatory cells are induced by immature DCs and play key roles in maintaining tolerance to self-antigens. Loss of T regulatory cells is the cause of organ-specific autoimmune disease in mice such as autoimmune thyroiditis, adrenalitis, and oophoritis (see "Immune Tolerance and Autoimmunity" below). T regulatory cells also play key roles in controlling the magnitude and duration of immune responses to microbes. Normally, after the initial immune response to a microbe has eliminated the invader, T regulatory cells are activated to suppress the antimicrobe response and prevent host injury. Some microbes have adapted to induce T regulatory cell activation at the site of infection to promote parasite infection and survival. In *Leishmania* infection, the parasite induces T regulatory cell accumulation at skin infection sites that dampens anti-*Leishmania* T cell responses and prevents parasite elimination. Although B cells recognize native antigen via B cell–surface Ig receptors, B cells require T cell help to produce high-affinity antibody of multiple isotypes that are the most effective in eliminating foreign antigen. T cell–B cell interactions that lead to high-affinity antibody production require (1) processing of native antigen by B cells and expression of peptide fragments on the B cell surface for presentation to T_H cells, (2) the ligation of B cells by both the TCR complex and the CD40 ligand, (3) induction of the process termed *antibody isotype switching* in antigen-specific B cell clones, and (4) induction of the process of affinity maturation of antibody in the germinal centers of B cell follicles of lymph node and spleen.

Naïve B cells express cell-surface IgD and IgM, and initial contact of naïve B cells with antigen is via binding of native antigen to B cell–surface IgM. T cell cytokines, released following T_H2 cell contact with B cells or by a "bystander" effect, induce changes in Ig gene conformation that promote recombination of Ig genes. These events then result in the switching of expression of heavy chain exons in a triggered B cell, leading to the secretion of IgG, IgA, or, in some cases, IgE antibody with the same V region antigen specificity as the original IgM antibody, for response to a wide variety of extracellular bacteria, protozoa, and helminths. CD40 ligand expression by activated T cells is critical for induction of B cell antibody isotype switching and for B cell responsiveness to cytokines. Patients with mutations in T cell CD40 ligand have B cells that are unable to undergo isotype switching, resulting in lack of memory B cell generation and the immunodeficiency syndrome of *X-linked hyper-IgM syndrome* **(Chap. 344)**.

IMMUNE TOLERANCE AND AUTOIMMUNITY

Immune tolerance is defined as the absence of activation of pathogenic autoreactivity to self-antigens. *Autoimmune diseases* are syndromes caused by the activation of T or B cells or both, with no evidence of other causes such as infections or malignancies **(Chap. 348)**. Immune tolerance and autoimmunity are present normally in health; when abnormal, they represent extremes from the normal state. For example, low levels of autoreactivity of T and B cells with self-antigens in the periphery are critical to T and B cell survival. Similarly, low levels of autoreactivity and thymocyte recognition of self-antigens in the thymus are the mechanisms whereby normal T cells are positively selected to survive and leave the thymus to respond to foreign microbes in the periphery and T cells highly reactive to self-antigens are negatively selected and die to prevent overly self-reactive T cells from migrating to the periphery (central tolerance). However, not all self-antigens are expressed in the thymus to delete highly self-reactive T cells, and there are mechanisms for induction of tolerance in peripheral T cells as well. Unlike the presentation of microbial antigens by mature DCs, the presentation of self-antigens by immature DCs neither activates nor matures the DCs to express high levels of co-stimulatory molecules such as B7-1 (CD80) or B7-2 (CD86). When peripheral T cells are stimulated by DCs expressing self-antigens in the context of HLA molecules, sufficient stimulation of T cells occurs to keep them alive, but otherwise they remain anergic, or nonresponsive, until T cells contact a DC with high levels of co-stimulatory molecules expressing microbial antigens and become activated to respond to the microbe. If B cells have high self-reactive BCRs, they normally undergo either deletion in the bone marrow or receptor editing to express a less autoreactive receptor. Although many autoimmune diseases are characterized by abnormal or pathogenic autoantibody production **(Table 342-13)**, most autoimmune diseases are caused by a combination of excess T and B cell reactivity.

Multiple factors contribute to the genesis of autoimmune disease syndromes, including genetic susceptibility (HLAB27 with ankylosing spondylitis) (Table 342-13), environmental immune stimulants such as drugs (e.g., procainamide and phenytoin [Dilantin] with drug-induced systemic lupus erythematosus), infectious agent triggers (such as Epstein-Barr virus and autoantibody production against red blood cells and platelets), and loss of T regulatory cells (leading to thyroiditis, adrenalitis, and oophoritis).

Immunity at Mucosal Surfaces Mucosa covering the respiratory, digestive, and urogenital tracts; the eye conjunctiva; the inner ear; and the ducts of all exocrine glands contain cells of the innate and adaptive mucosal immune system that protect these surfaces against pathogens. In the healthy adult, mucosa-associated lymphoid tissue (MALT) contains 80% of all immune cells within the body and constitutes the largest mammalian lymphoid organ system.

MALT has three main functions: (1) to protect the mucous membranes from invasive pathogens; (2) to prevent uptake of foreign antigens from food, commensal organisms, and airborne pathogens and particulate matter; and (3) to prevent pathologic immune responses from foreign antigens if they do cross the mucosal barriers of the body **(Fig. 342-8)**.

MALT is a compartmentalized system of immune cells that functions independently from systemic immune organs. Whereas the systemic immune organs are essentially sterile under normal conditions and respond vigorously to pathogens, MALT immune cells are continuously bathed in foreign proteins and commensal bacteria, and they must select those pathogenic antigens that must be eliminated. MALT contains anatomically defined foci of immune cells in the intestine, tonsil, appendix, and peribronchial areas that are inductive sites for mucosal immune responses. From these sites, immune T and B cells migrate to effector sites in mucosal parenchyma and exocrine glands where mucosal immune cells eliminate pathogen-infected cells. In addition to mucosal immune responses, all mucosal sites have strong mechanical and chemical barriers and cleansing functions to repel pathogens.

Key components of MALT include specialized epithelial cells called "membrane" or "M" cells that take up antigens and deliver them to DCs or other APCs. Effector cells in MALT include B cells producing antipathogen neutralizing antibodies of secretory IgA as well as IgG isotype, T cells producing similar cytokines as in systemic immune system response, and T helper and cytotoxic T cells that respond to pathogen-infected cells.

Secretory IgA is produced in amounts of >50 mg/kg of body weight per 24 h and functions to inhibit bacterial adhesion, inhibit macromolecule absorption in the gut, neutralize viruses, and enhance antigen elimination in tissue through binding to IgA and receptor-mediated transport of immune complexes through epithelial cells.

TABLE 342-13 Recombinant or Purified Autoantigens Recognized by Autoantibodies Associated with Human Autoimmune Disorders

AUTOANTIGEN	AUTOIMMUNE DISEASES
Cell- or Organ-Specific Autoimmunity	
Acetylcholine receptor	Myasthenia gravis
Actin	Chronic active hepatitis, primary biliary cirrhosis
Adenine nucleotide translator (ANT)	Dilated cardiomyopathy, myocarditis
β-Adrenoreceptor	Dilated cardiomyopathy
Aromatic L-amino acid decarboxylase	Autoimmune polyendocrine syndrome type 1 (APS-1)
Asialoglycoprotein receptor	Autoimmune hepatitis
Bactericidal/permeability-increasing protein (Bpi)	Cystic fibrosis vasculitides
Calcium-sensing receptor	Acquired hypoparathyroidism
Cholesterol side-chain cleavage enzyme (CYP11a)	Autoimmune polyglandular syndrome-1
Collagen type IV-α3-chain	Goodpasture's syndrome
Cytochrome P450 2D6 (CYP2D6)	Autoimmune hepatitis
Desmin	Crohn's disease, coronary artery disease
Desmoglein 1	Pemphigus foliaceus
Desmoglein 3	Pemphigus vulgaris
F-actin	Autoimmune hepatitis
GM gangliosides	Guillain-Barré syndrome
Glutamate decarboxylase (GAD65)	Type 1 diabetes, stiff-person syndrome
Glutamate receptor (GLUR)	Rasmussen encephalitis
H/K ATPase	Autoimmune gastritis
17-α-Hydroxylase (CYP17)	Autoimmune polyglandular syndrome-1
21-Hydroxylase (CYP21)	Addison's disease
IA-2 (ICA512)	Type 1 diabetes
Insulin	Type 1 diabetes, insulin hypoglycemic syndrome (Hirata's disease)
Insulin receptor	Type B insulin resistance, acanthosis, systemic lupus erythematosus (SLE)
Intrinsic factor type 1	Pernicious anemia
Leukocyte function-associated antigen (LFA-1)	Treatment-resistant Lyme arthritis
Myelin-associated glycoprotein (MAG)	Polyneuropathy
Myelin-basic protein	Multiple sclerosis, demyelinating diseases
Myelin oligodendrocyte glycoprotein (MOG)	Multiple sclerosis
Myosin	Rheumatic fever
p-80-Collin	Atopic dermatitis
Pyruvate dehydrogenase complex-E2 (PDC-E2)	Primary biliary cirrhosis
Sodium iodide symporter (NIS)	Graves' disease, autoimmune hypothyroidism
SOX-10	Vitiligo
Thyroid and eye muscle shared protein	Thyroid-associated ophthalmopathy
Thyroglobulin	Autoimmune thyroiditis
Thyroid peroxidase	Autoimmune Hashimoto's thyroiditis
Thyrotropin receptor	Graves' disease
Tissue transglutaminase	Celiac disease
Transcription coactivator p75	Atopic dermatitis
Tryptophan hydroxylase	Autoimmune polyglandular syndrome-1
Tyrosinase	Vitiligo, metastatic melanoma
Tyrosine hydroxylase	Autoimmune polyglandular syndrome-1
Systemic Autoimmunity	
ACTH	ACTH deficiency
Aminoacyl-tRNA histidyl synthetase	Myositis, dermatomyositis
Aminoacyl-tRNA synthetase (several)	Polymyositis, dermatomyositis
Cardiolipin	Systemic lupus erythematosus, antiphospholipid syndrome
Carbonic anhydrase II	Systemic lupus erythematosus, Sjögren's syndrome, systemic sclerosis
Collagen (multiple types)	Rheumatoid arthritis, systemic lupus erythematosus, progressive systemic sclerosis
Centromere-associated proteins	Systemic sclerosis
DNA-dependent nucleoside-stimulated ATPase	Dermatomyositis
Fibrillarin	Scleroderma
Fibronectin	Systemic lupus erythematosus, rheumatoid arthritis, morphea
Glucose-6-phosphate isomerase	Rheumatoid arthritis
β2-Glycoprotein I (B2-GPI)	Primary antiphospholipid syndrome
Golgin (95, 97, 160, 180) Heat shock protein	Sjögren's syndrome, systemic lupus erythematosus, rheumatoid arthritis Various immune-related disorders
Hemidesmosomal protein 180	Bullous pemphigoid, herpes gestationis, cicatricial pemphigoid
Histone H2A-H2B-DNA	Systemic lupus erythematosus
IgE receptor	Chronic idiopathic urticaria
Keratin	Rheumatoid arthritis
Ku-DNA-protein kinase	Systemic lupus erythematosus
Ku-nucleoprotein La phosphoprotein (La 55-B)	Connective tissue syndrome Sjögren's syndrome
Myeloperoxidase	Necrotizing and crescentic glomerulonephritis (NCGN), systemic vasculitis
Proteinase 3 (PR3)	Granulomatosis with polyangiitis (Wegener's), Churg-Strauss syndrome
RNA polymerase I–III (RNP)	Systemic sclerosis, systemic lupus erythematosus
Signal recognition protein (SRP54)	Polymyositis
Topoisomerase-1 (Scl-70)	Scleroderma, Raynaud's syndrome
Tublin	Chronic liver disease, visceral leishmaniasis
Vimentin	Systemic autoimmune disease
Plasma Protein and Cytokine Autoimmunity	
C1 inhibitor	Autoimmune C1 deficiency
C1q	Systemic lupus erythematosus, membrane proliferative glomerulonephritis (MPGN)
Cytokines (IL-1α, IL-1β, IL-6, IL-10, LIF)	Rheumatoid arthritis, systemic sclerosis, normal subjects
Factor II, factor V, factor VII, factor VIII, factor IX, factor X, factor XI, thrombin vWF	Prolonged coagulation time
Glycoprotein IIb/IIIg and Ib/IX	Autoimmune thrombocytopenia purpura
IgA	Immunodeficiency associated with systemic lupus erythematosus, pernicious anemia, thyroiditis, Sjögren's syndrome, and chronic active hepatitis
Oxidized LDL (OxLDL)	Atherosclerosis
Cancer and Paraneoplastic Autoimmunity	
Amphiphysin	Neuropathy, small-cell lung cancer
Cyclin B1	Hepatocellular carcinoma
DNA topoisomerase II	Liver cancer
Desmoplakin	Paraneoplastic pemphigus
Gephyrin	Paraneoplastic stiff-person syndrome

(*Continued*)

TABLE 342-13 Recombinant or Purified Autoantigens Recognized by Autoantibodies Associated with Human Autoimmune Disorders (Continued)

AUTOANTIGEN	AUTOIMMUNE DISEASES
Cancer and Paraneoplastic Autoimmunity (Continued)	
Hu proteins	Paraneoplastic encephalomyelitis
Neuronal nicotinic acetylcholine receptor	Subacute autonomic neuropathy, cancer
p53	Cancer, systemic lupus erythematosus
p62 (IGF-II mRNA-binding protein)	Hepatocellular carcinoma (China)
Recoverin	Cancer-associated retinopathy
Ri protein	Paraneoplastic opsoclonus myoclonus ataxia
βIV spectrin	Lower motor neuron syndrome
Synaptotagmin	Lambert-Eaton myasthenic syndrome
Voltage-gated calcium channels	Lambert-Eaton myasthenic syndrome
Yo protein	Paraneoplastic cerebellar degeneration

Source: From A Lernmark et al: J Clin Invest 108:1091, 2001; with permission.

Recent studies have demonstrated the importance of commensal gut and other mucosal bacteria to the health of the human immune system. Normal commensal flora induces anti-inflammatory events in the gut and protects epithelial cells from pathogens through TLRs and other PRR signaling. When the gut is depleted of normal commensal flora, the immune system becomes abnormal, with loss of T_H1 T cell function. Restoration of the normal gut flora can reestablish the balance in T helper cell ratios characteristic of the normal immune system. Diet also has an impact on the gut microbiome. Altered microbiome composition has been etiologically related to obesity, insulin resistance and diabetes. When the gut barrier is intact, either antigens do not transverse the gut epithelium or, when pathogens are present, a self-limited, protective MALT immune response eliminates the pathogen (Fig. 342-8). However, when the gut barrier breaks down, immune responses to commensal flora antigens can cause inflammatory bowel diseases such as *Crohn's disease* and, perhaps, *ulcerative colitis* (Fig. 342-8) **(Chap. 319)**. Uncontrolled MALT immune responses to food antigens, such as gluten, can cause *celiac disease* **(Chap. 319)**.

THE CELLULAR AND MOLECULAR CONTROL OF PROGRAMMED CELL DEATH

The process of apoptosis (programmed cell death) plays a crucial role in regulating normal immune responses to antigen. In general, a wide variety of stimuli trigger one of several apoptotic pathways to eliminate microbe-infected cells, eliminate cells with damaged DNA, or eliminate activated immune cells that are no longer needed **(Fig. 342-9)**. The largest known family of "death receptors" is the TNF receptor (TNF-R) family (TNF-R1, TNF-R2, Fas [CD95], death receptor 3 [DR3], death receptor 4 [DR4; TNF-related apoptosis-including ligand receptor 1, or TRAIL-R1], and death receptor 5 [DR5, TRAIL-R2]); their ligands are all in the TNF-α family. Binding of ligands to these death receptors leads to a signaling cascade that involves activation of the *caspase* family of molecules that leads to DNA cleavage and cell death. Two other pathways of programmed cell death involve nuclear *p53* in the elimination of cells with abnormal DNA and *mitochondrial cytochrome c* to induce cell death in damaged cells (Fig. 342-9). A number of human diseases have now been described that result from, or are associated with, mutated apoptosis genes **(Table 342-14)**. These include mutations in the Fas and Fas ligand genes in autoimmune and lymphoproliferation syndromes, and multiple associations of mutations in genes in the apoptotic pathway with malignant syndromes.

MECHANISMS OF IMMUNE-MEDIATED DAMAGE TO MICROBES OR HOST TISSUES

Several responses by the host innate and adaptive immune systems to foreign microbes culminate in rapid and efficient elimination of microbes. In these scenarios, the classic weapons of the adaptive immune system (T cells, B cells) interface with cells (macrophages, DCs, NK cells, neutrophils, eosinophils, basophils) and soluble products (microbial peptides, pentraxins, complement and coagulation systems) of the innate immune system **(Chaps. 60 and 345)**.

There are five general phases of host defenses: (1) migration of leukocytes to sites of antigen localization; (2) antigen-nonspecific recognition of pathogens by macrophages and other cells and systems of the innate immune system; (3) specific recognition of foreign antigens mediated by T and B lymphocytes; (4) amplification of the inflammatory response with recruitment of specific and nonspecific effector cells by complement components, cytokines, kinins, arachidonic acid metabolites, and mast cell–basophil products; and (5) macrophage, neutrophil, and lymphocyte participation in destruction of antigen with ultimate removal of antigen particles by phagocytosis (by macrophages or neutrophils) or by direct cytotoxic mechanisms (involving macrophages, neutrophils, DCs, and lymphocytes). Under normal circumstances, orderly progression of host defenses through these phases results in a well-controlled immune and inflammatory response that protects the host from the offending antigen. However, dysfunction of

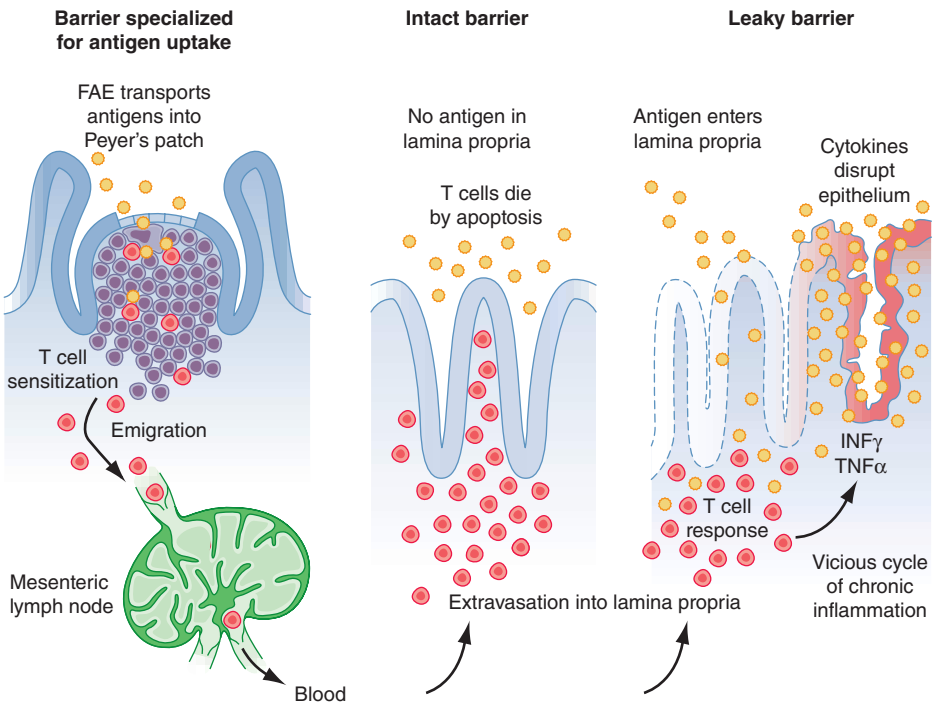


FIGURE 342-8 Increased epithelial permeability may be important in the development of chronic gut T cell–mediated inflammation. CD4 T cells activated by gut antigens in Peyer's patches migrate to the lamina propria (LP). In healthy individuals, these cells die by apoptosis. Increased epithelial permeability may allow sufficient antigen to enter the LP to trigger T cell activation, breaking tolerance mediated by immunosuppressive cytokines and perhaps T regulatory cells. Proinflammatory cytokines then further increase epithelial permeability, setting up a vicious cycle of chronic inflammation. *(From TT MacDonald et al: Science 307:1920, 2005; with permission.)*

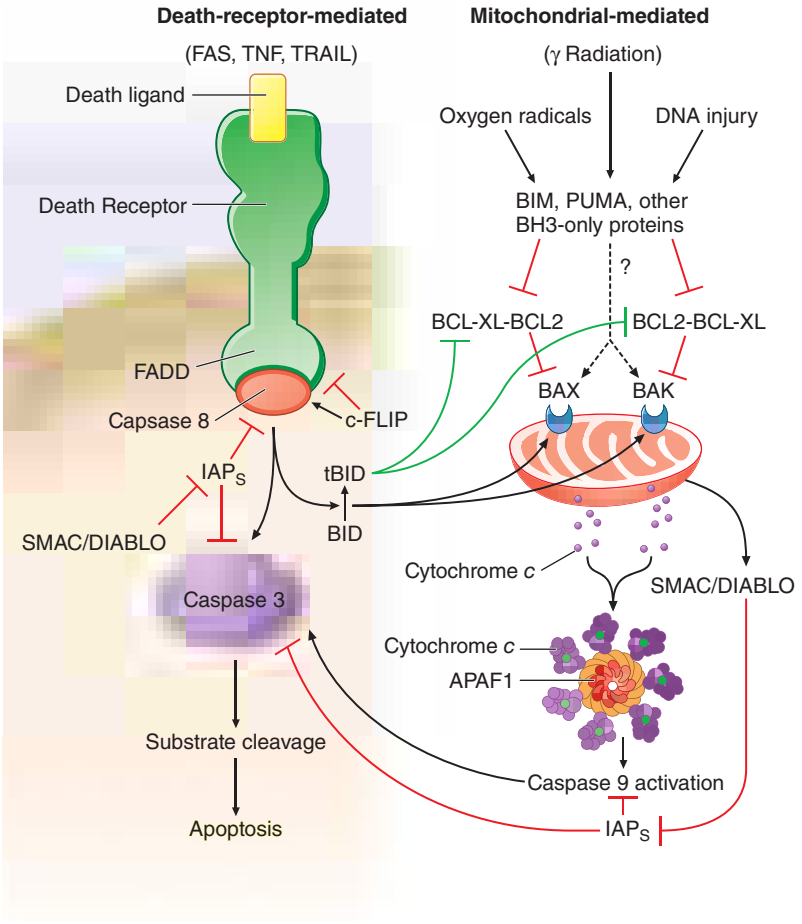


FIGURE 342-9 Pathways of cellular apoptosis. There are two major pathways of apoptosis: the death-receptor pathway, which is mediated by activation of death receptors, and the BCL2-regulated mitochondrial pathway, which is mediated by noxious stimuli that ultimately lead to mitochondrial injury. Ligation of death receptors recruits the adaptor protein FAS-associated death domain (FADD). FADD in turn recruits caspase 8, which ultimately activates caspase 3, the key "executioner" caspase. Cellular FLICE-inhibitory protein (c-FLIP) can either inhibit or potentiate binding of FADD and caspase 8, depending on its concentration. In the intrinsic pathway, proapoptotic BH3 proteins are activated by noxious stimuli, which interact with and inhibit antiapoptotic BCL2 or BCL-XL. Thus, BAX and BAK are free to induce mitochondrial permeabilization with release of cytochrome c, which ultimately results in the activation of caspase 9 through the apoptosome. Caspase 9 then activates caspase 3. SMAC/DIABLO is also released after mitochondrial permeabilization and acts to block the action of inhibitors of apoptosis protein (IAPs), which inhibit caspase activation. There is potential cross-talk between the two pathways, which is mediated by the truncated form of BID (tBID) that is produced by caspase 8–mediated BID cleavage; tBID acts to inhibit the BCL2-BCL-XL pathway and to activate BAX and BAK. There is debate (indicated by the *question mark*) as to whether proapoptotic BH3 molecules (e.g., BIM and PUMA) act directly on BAX and BAK to induce mitochondrial permeability or whether they act only on BCL2-BCL-XL. APAF1, apoptotic protease-activating factor 1; BH3, BCL homologue; TNF, tumor necrosis factor; TRAIL, TNF-related apoptosis-inducing ligand. *(From RS Hotchkiss et al: N Engl J Med 361:1570, 2009; with permission.)*

any of the host defense systems can damage host tissue and produce clinical disease. Furthermore, for certain pathogens or antigens, the normal immune response itself might contribute substantially to the tissue damage. For example, the immune and inflammatory response in the brain to certain pathogens such as *M. tuberculosis* may be responsible for much of the morbidity rate of this disease in that organ system **(Chap. 173)**. In addition, the morbidity rate associated with certain pneumonias such as that caused by *Pneumocystis jiroveci* may be associated more with inflammatory infiltrates than with the tissue-destructive effects of the microorganism itself **(Chap. 215)**.

Molecular Basis of Lymphocyte–Endothelial Cell Interactions The control of lymphocyte circulatory patterns between the bloodstream and peripheral lymphoid organs operates at the level of lymphocyte–endothelial cell interactions to control the specificity of lymphocyte subset entry into organs. Similarly, lymphocyte–endothelial cell interactions regulate the entry of lymphocytes into inflamed tissue. Adhesion molecule expression on lymphocytes and endothelial cells regulates the retention and subsequent egress of lymphocytes within tissue sites of antigenic stimulation, delaying cell exit from tissue and preventing reentry into the circulating lymphocyte pool **(Fig. 342-10)**. All types of lymphocyte migration begin with lymphocyte attachment to specialized regions of vessels, termed *high endothelial venules* (HEVs). An important concept is that adhesion molecules do not generally bind their ligand until a conformational change (ligand activation) occurs in the adhesion molecule that allows ligand binding. Induction of a conformation-dependent determinant on an adhesion molecule can be accomplished by cytokines or via ligation of other adhesion molecules on the cell.

The first stage of lymphocyte–endothelial cell interactions, *attachment and rolling*, occurs when lymphocytes leave the stream of flowing blood cells in a postcapillary venule and roll along venule endothelial cells (Fig. 342-10). Lymphocyte rolling is mediated by the L-selectin molecule (LECAM-1, LAM-1, CD62L) and slows cell transit time through venules, allowing time for activation of adherent cells.

The second stage of lymphocyte–endothelial cell interactions, *firm adhesion with activation-dependent stable arrest*, requires stimulation of lymphocytes by chemoattractants or by endothelial cell–derived cytokines. Cytokines thought to participate in adherent cell activation include members of the IL-8 family, platelet-activation factor, leukotriene B_4, and C5a. In addition, HEVs express chemokines, SLC (CCL21) and ELC (CCL19), which participate in this process. Following activation by chemoattractants, lymphocytes shed L-selectin from the cell surface and upregulate cell CD11b/18 (MAC-1) or CD11a/18 (LFA-1) molecules, resulting in firm attachment of lymphocytes to HEVs.

Lymphocyte homing to peripheral lymph nodes involves adhesion of L-selectin to glycoprotein HEV ligands collectively referred to as *peripheral node addressin* (*PNAd*), whereas homing of lymphocytes to intestine Peyer's patches primarily involves adhesion of the a4β7 integrin to mucosal addressin cell adhesion molecule-1 (MAdCAM-1) on the Peyer's patch HEVs. However, for migration to mucosal Peyer's patch lymphoid aggregates, naïve lymphocytes primarily use L-selectin, whereas memory lymphocytes use a4β7 integrin. a4β1 integrin (CD49d/CD29, VLA-4)–VCAM-1 interactions are important in the initial interaction of memory lymphocytes with HEVs of multiple organs in sites of inflammation **(Table 342-15)**.

The third stage of leukocyte emigration in HEVs is *sticking and arrest*. Sticking of the lymphocyte to endothelial cells and arrest at the site of sticking are mediated predominantly by ligation of a1β2 integrin LFA-1 to the integrin ligand ICAM-1 on HEVs. Whereas the first three stages of lymphocyte attachment to HEVs take only a few seconds, the fourth stage of lymphocyte emigration, *transendothelial migration*, takes ~10 min. Although the molecular mechanisms that control lymphocyte transendothelial migration are not fully characterized, the HEV CD44 molecule and molecules of the HEV glycocalyx (extracellular matrix) are thought to play important regulatory roles in this process (Fig. 342-10). Finally, expression of matrix metalloproteases capable of digesting the subendothelial basement membrane, rich in nonfibrillar collagen, appears to be required for the penetration of lymphoid cells into the extravascular sites.

Abnormal induction of HEV formation and use of the molecules discussed above have been implicated in the induction and maintenance of inflammation in a number of chronic inflammatory diseases. In animal models of type 1 diabetes mellitus, MAdCAM-1 and GlyCAM-1 have been shown to be highly expressed on HEVs in inflamed pancreatic islets, and treatment of these animals with inhibitors of

TABLE 342-14 Immune System Molecule Defects in Animals or Humans That Cause Autoimmune or Malignant Syndromes

PROTEIN	DEFECT	DISEASE OR SYNDROME	OBSERVATION IN ANIMAL MODELS OR HUMANS
Cytokines and Signaling Proteins			
Tumor necrosis factor (TNF) α	Overexpression	Inflammatory bowel disease (IBD), arthritis, vasculitis	Mice
TNF-α	Underexpression	Systemic lupus erythematosus (SLE)	Mice
Interleukin (IL)-1-receptor antagonist	Underexpression	Arthritis	Mice
IL-2	Overexpression	IBD	Mice
IL-7	Overexpression	IBD	Mice
IL-10	Overexpression	IBD	Mice
IL-2 receptor	Overexpression	IBD	Mice
IL-10 receptor	Overexpression	IBD	Mice
IL-3	Overexpression	Demyelinating syndrome	Mice
Interferon-δ	Overexpression in skin	SLE	Mice
STAT-3	Underexpression	IBD	Mice
STAT-4	Overexpression	IBD	Mice
Transforming growth factor (TGF) β	Underexpression	Systemic wasting syndrome and IBD	Mice
TGF-β receptor in T cells	Underexpression	SLE	Mice
Programmed death (CD279, PD-1)	Underexpression	SLE-like syndrome	Mice
Cytotoxic T lymphocyte, antigen-4 (CTLA-4)	Underexpression	Systemic lymphoproliferative disease	Mice
IL-10	Underexpression	IBD (mouse), type 1 diabetes, thyroid disease, primary (human)	Mice and humans
Major Histocompatibility Locus Molecules[a]			
HLA-B27	Allele expression or overexpression	Inflammatory bowel disease	Rats and humans
Complement deficiency of C1, 2, 3 or 4	Underexpression		Humans
LIGHT (TNF superfamily 14)	Overexpression	Systemic lymphoproliferative (mouse) and autoimmunity	Mice
HLA class II DQB10301, DQB10302	Allele expression	Juvenile-onset diabetes	Humans
HLA class II DQB10401, DQB10402	Allele expression	Rheumatoid arthritis	Humans
HLA class I B27	Allele expression	Ankylosing spondylitis, IBD	Rats and humans
Apoptosis Proteins			
TNF receptor 1 (TNF-R1)	Underexpression	Familial periodic fever syndrome	Humans
Fas (CD95; Apo-1)	Underexpression	Autoimmune lymphoproliferative syndrome type 1 (ALPS 1); malignant lymphoma; bladder cancer	Humans
Fas ligand	Underexpression	SLE (only one case identified)	Humans
Perforin	Underexpression	Familial hemophagocytic lymphohistiocytosis (FHL)	Humans
Caspase 10	Underexpression	Autoimmune lymphoproliferative syndrome type II (ALPS II)	Humans
bcl-10	Underexpression	Non-Hodgkin's lymphoma	Humans
P53	Underexpression	Various malignant neoplasms	Humans
Bax	Underexpression	Colon cancer; hematopoietic malignancies	Humans
bcl-2	Underexpression	Non-Hodgkin's lymphoma	Humans
c-IAP2	Underexpression	Low-grade MALT lymphoma	Humans
NAIP1	Underexpression	Spinal muscular atrophy	Humans

[a]Many autoimmune diseases are associated with a myriad of major histocompatibility complex gene allele (HLA) types. They are presented here as examples.

Abbreviation: MALT, mucosa-associated lymphoid tissue.

Source: Adapted from L Mullauer: Mutat Res 488:211, 2001; and A Davidson, B Diamond: N Engl J Med 345:340, 2001.

L-selectin and a4 integrin function blocked the development of type 1 diabetes mellitus **(Chap. 396)**. A similar role for abnormal induction of the adhesion molecules of lymphocyte emigration has been suggested in *rheumatoid arthritis* **(Chap. 351)**, *Hashimoto's thyroiditis* **(Chap. 375)**, *Graves' disease* **(Chap. 375)**, *multiple sclerosis* **(Chap. 436)**, *Crohn's disease* **(Chap. 319)**, and *ulcerative colitis* **(Chap. 319)**.

Immune-Complex Formation Clearance of antigen by immune-complex formation between antigen, complement, and antibody is a highly effective mechanism of host defense. However, depending on the level of immune complexes formed and their physicochemical properties, immune complexes may or may not result in host and foreign cell damage. After antigen exposure, certain types of soluble antigen-antibody complexes freely circulate and, if not cleared by the reticuloendothelial system, can be deposited in blood vessel walls and in other tissues such as renal glomeruli and cause *vasculitis* or *glomerulonephritis* syndromes **(Chaps. 308 and 356)**. Deficiencies of early complement components are associated with inefficient clearance of immune complexes and immune complex mediated tissue damage in autoimmune syndromes, whereas deficiencies of the later complement components are associated with susceptibility to recurrent *Neisseria* infections **(Table 342-16)**.

Immediate-Type Hypersensitivity Helper T cells that drive antiallergen IgE responses are usually T_H2-type inducer T cells that secrete IL-4, IL-5, IL-6, and IL-10. Mast cells and basophils have high-affinity receptors for the Fc portion of IgE (FcRI), and cell-bound antiallergen IgE effectively "arms" basophils and mast cells. Mediator release is triggered by antigen (allergen) interaction with Fc receptor-bound IgE, and the mediators released are responsible for the pathophysiologic changes of *allergic diseases* (Table 342-11). Mediators released from mast cells and

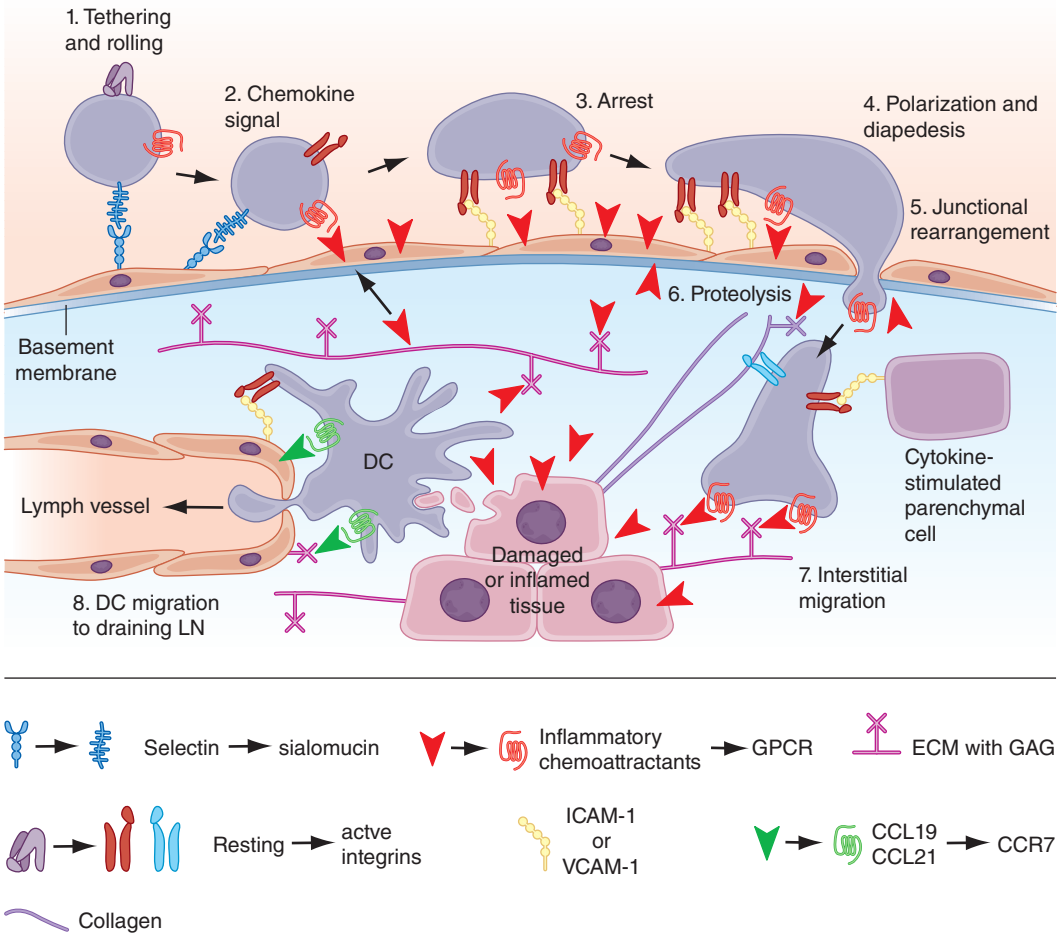


FIGURE 342-10 Key migration steps of immune cells at sites of inflammation. Inflammation due to tissue damage or infection induces the release of cytokines (not shown) and inflammatory chemoattractants (*red arrowheads*) from distressed stromal cells and "professional" sentinels, such as mast cells and macrophages (not shown). The inflammatory signals induce upregulation of endothelial selectins and immunoglobulin "superfamily" members, particularly ICAM-1 and/or VCAM-1. Chemoattractants, particularly chemokines, are produced by or translocated across venular endothelial cells (*red arrow*) and are displayed in the lumen to rolling leukocytes. Those leukocytes that express the appropriate set of trafficking molecules undergo a multistep adhesion cascade (steps 1–3) and then polarize and move by diapedesis across the venular wall (steps 4 and 5). Diapedesis involves transient disassembly of endothelial junctions and penetration through the underlying basement membrane (step 6). Once in the extravascular (interstitial) space, the migrating cell uses different integrins to gain "footholds" on collagen fibers and other ECM molecules, such as laminin and fibronectin, and on inflammation-induced ICAM-1 on the surface of parenchymal cells (step 7). The migrating cell receives guidance cues from distinct sets of chemoattractants, particularly chemokines, which may be immobilized on glycosaminoglycans (GAG) that "decorate" many ECM molecules and stromal cells. Inflammatory signals also induce tissue dendritic cells (DCs) to undergo maturation. Once DCs process material from damaged tissues and invading pathogens, they upregulate CCR7, which allows them to enter draining lymph vessels that express the CCR7 ligand CCL21 (and CCL19). In lymph nodes (LNs), these antigen-loaded mature DCs activate naïve T cells and expand pools of effector lymphocytes, which enter the blood and migrate back to the site of inflammation. T cells in tissue also use this CCR7-dependent route to migrate from peripheral sites to draining lymph nodes through afferent lymphatics. *(Adapted from AD Luster et al: Nat Immunol 6:1182, 2005; with permission from Macmillan Publishers Ltd. Copyright 2005.)*

basophils can be divided into three broad functional types: (1) those that increase vascular permeability and contract smooth muscle (histamine, platelet-activating factor, SRS-A, BK-A), (2) those that are chemotactic for or activate other inflammatory cells (ECF-A, NCF, leukotriene B_4), and (3) those that modulate the release of other mediators (BK-A, platelet-activating factor) **(Chap. 345)**.

Cytotoxic Reactions of Antibody In this type of immunologic injury, complement-fixing (C1-binding) antibodies against normal or foreign cells or tissues (IgM, IgG1, IgG2, IgG3) bind complement via the classic pathway and initiate a sequence of events similar to that initiated by immune-complex deposition, resulting in cell lysis or tissue injury. Examples of antibody-mediated cytotoxic reactions include red cell lysis in *transfusion reactions*, *Goodpasture's syndrome* with anti–glomerular basement membrane antibody formation, and *pemphigus vulgaris* with antiepidermal antibodies inducing blistering skin disease.

Delayed-Type Hypersensitivity Reactions Inflammatory reactions initiated by mononuclear leukocytes and not by antibody alone have been termed *delayed-type hypersensitivity reactions*. The term *delayed* has been used to contrast a secondary cellular response that appears 48–72 h after antigen exposure with an *immediate* hypersensitivity response generally seen within 12 h of antigen challenge and initiated by basophil mediator release or preformed antibody. For example, in an individual previously infected with *M. tuberculosis* organisms, intradermal placement of tuberculin purified protein derivative as a skin test challenge results in an indurated area of skin at 48–72 h, indicating previous exposure to tuberculosis.

The cellular events that result in classic delayed-type hypersensitivity responses are centered on T cells (predominantly, although not exclusively, IFN-γ, IL-2, and TNF-α-secreting T_H1-type helper T cells) and macrophages. Recently, NK cells have been suggested to play a major role in the form of delayed hypersensitivity that occurs following skin contact with immunogens. First, local immune and inflammatory responses at the site of foreign antigen upregulate endothelial cell adhesion molecule expression, promoting the accumulation of lymphocytes at the tissue site. In the general scheme outlined in Fig. 342-2, antigen is processed by DCs and presented to small numbers of CD4+ T cells expressing a TCR specific for the antigen. IL-12 produced by APCs induces T cells to produce IFN-γ (T_H1 response). Macrophages frequently undergo epithelioid cell transformation and fuse to form

TABLE 342-15 Trafficking Molecules Involved In Inflammatory Disease Processes

		PROPOSED LEUKOCYTE RECEPTORS FOR ENDOTHELIAL TRAFFIC SIGNALS		
DISEASE	**KEY EFFECTOR CELL**	**L-SELECTIN, LIGAND**	**GPCR**	**INTEGRIN[a]**
Acute Inflammation				
Myocardial infarction	Neutrophil	PSGL-1	CXCR1, CXCR2, PAFR, BLT1	LFA-1, Mac-1
Stroke	Neutrophil	L-Selectin, PSGL-1	CXCR1, CXCR2, PAFR, BLT1	LFA-1, Mac-1
Ischemia-reperfusion	Neutrophil	PSGL-1	CXCR1, CXCR2, PAFR, BLT1	LFA-1, Mac-1
T_H1 Inflammation				
Atherosclerosis	Monocyte	PSGL-1	CCR1, CCR2, BLT1, CXCR2, CX3CR1	VLA-4
	T_H1	PSGL-1	CXCR3, CCR5	VLA-4
Multiple sclerosis	T_H1	PSGL-1 (?)	CXCR3, CXCR6	VLA-4, LFA-1
	Monocyte	PSGL-1 (?)	CCR2, CCR1	VLA-4, LFA-1
Rheumatoid arthritis	Monocyte	PSGL-1	CCR1, CCR2	VLA-1, VLA-2, VLA-4, LFA-1
	T_H1	PSGL-1	CXCR3, CXCR6	VLA-1, VLA-2, VLA-4, LFA-1
	Neutrophil	L-Selectin, PSGL-1	CXCR2, BLT1	LFA-1[b]
Psoriasis	Skin-homing T_H1	CLA	CCR4, CCR10, CXCR3	VLA-4[c], LFA-1
Crohn's disease	Gut-homing T_H1	PSGL-1	CCR9, CXCR3	α4, β7, LFA-1
Type 1 diabetes	T_H1	PSGL-1 (?)	CCR4, CCR5	VLA-4, LFA-1
	CD8	L-Selectin (?), PSGL-1 (?)	CXCR3	VLA-4, LFA-1
Allograft rejection	CD8	PSGL-1	CXCR3, CX3CR1, BLT1	VLA-4, LFA-1
	B cell	L-Selectin, PSGL-1	CXCR5, CXCR4	VLA-4, LFA-1
Hepatitis	CD8	PSGL-1	CXCR3, CCR5, CXCR6	VLA-4
Lupus	T_H1	None	CXCR6	VLA-4[d]
	Plasmacytoid DC	L-Selectin, CLA	CCR7, CXCR3, ChemR23	LFA-1, Mac-1
	B cell	CLA (?)	CXCR5, CXCR4	LFA-1
T_H2 Inflammation				
Asthma	T_H2	PSGL-1	CCR4, CCR8, BLT1	LFA-1
	Eosinophil	PSGL-1	CCR3, PAFR, BLT1	VLA-4, LFA-1
	Mast cells	PSGL-1	CCR2, CCR3, BLT1	VLA-4, LFA-1
Atopic dermatitis	Skin-homing T_H2	CLA	CCR4, CCR10	VLA-4, LFA-1

[a]Various β_1 integrins have been linked in different ways in basal lamina and interstitial migration of distinct cell types and inflammatory settings. [b]In some settings, Mac-1 has been linked to transmigration. [c]CD44 can act in concert with VLA-4 in particular models of leukocyte arrest. [d]T_H2 cells require VAP-1 to traffic to inflamed liver.

Source: From AD Luster et al: Nat Immunol 6:1182, 2005; with permission from Macmillan Publishers Ltd. Copyright 2005.

multinucleated giant cells in response to IFN-γ. This type of mononuclear cell infiltrate is termed *granulomatous inflammation*. Examples of diseases in which delayed-type hypersensitivity plays a major role are fungal infections (*histoplasmosis*; **Chap. 207**), mycobacterial infections (*tuberculosis*, *leprosy*; **Chaps. 173 and 174**), chlamydial infections (*lymphogranuloma venereum*; **Chap. 184**), helminth infections (*schistosomiasis*; **Chap. 229**), reactions to toxins (*berylliosis*; **Chap. 283**), and hypersensitivity reactions to organic dusts (*hypersensitivity pneumonitis*; **Chap. 282**). In addition, delayed-type hypersensitivity responses play important roles in tissue damage in autoimmune diseases such as *rheumatoid arthritis*, *temporal arteritis*, and granulomatosis with polyangiitis (Wegener's) **(Chaps. 351 and 356)**.

TABLE 342-16 Complement Deficiencies and Associated Diseases

COMPONENT	ASSOCIATED DISEASES
Classic Pathway	
Clq, Clr, Cls, C4	Immune-complex syndromes,[a] pyogenic infections
C2	Immune-complex syndromes,[a] few with pyogenic infections
C1 inhibitor	Rare immune-complex disease, few with pyogenic infections
C3 and Alternative Pathway C3	
C3	Immune-complex syndromes,[a] pyogenic infections
D	Pyogenic infections
Properdin	*Neisseria* infections
I	Pyogenic infections
H	Hemolytic-uremic syndrome
Membrane Attack Complex	
C5, C6, C7, C8	Recurrent *Neisseria* infections, immune-complex disease
C9	Rare *Neisseria* infections

[a]Immune-complex syndromes include systemic lupus erythematosus (SLE) and SLE-like syndromes, glomerulonephritis, and vasculitis syndromes.

Source: After JA Schifferli, DK Peters: Lancet 322:957, 1983. Copyright 1983, with permission from Elsevier.

Autophagy Autophagy is a process that involves a lysosomal degradation pathway mechanism of cells to dispose of intracellular debris and damaged organelles. Autophagy by cells of the innate immune system is used to control intracellular infectious agents such as *Mycobacterium tuberculosis*, in part by initiation of phagosome maturation and enhancing MHC class II antigen presentation to CD4 T cells.

CLINICAL EVALUATION OF IMMUNE FUNCTION

Clinical assessment of immunity requires investigation of the four major components of the immune system that participate in host defense and in the pathogenesis of autoimmune diseases: (1) humoral immunity (B cells); (2) cell-mediated immunity (T cells, monocytes); (3) phagocytic cells of the reticuloendothelial system (macrophages), as well as polymorphonuclear leukocytes; and (4) complement. Clinical problems that require an evaluation of immunity include chronic infections, recurrent infections, unusual infecting agents, and certain autoimmune syndromes. The type of clinical syndrome under evaluation can provide information regarding possible immune defects **(Chap. 344)**. Defects in cellular immunity generally result in viral, mycobacterial, and fungal infections. An extreme example of deficiency in cellular immunity is AIDS **(Chap. 197)**. Antibody deficiencies result in recurrent bacterial infections, frequently with organisms such as *S. pneumoniae* and *Haemophilus influenzae* **(Chap. 344)**. Disorders of phagocyte function are frequently manifested by recurrent skin infections, often due to *Staphylococcus aureus* **(Chap. 60)**. Finally, deficiencies

of early and late complement components are associated with autoimmune phenomena and recurrent *Neisseria* infections (Table 342-16). **For further discussion of useful initial screening tests of immune function, see Chap. 344.**

IMMUNOTHERAPY

Many therapies for autoimmune and inflammatory diseases involve the use of nonspecific immune-modulating or immunosuppressive agents such as glucocorticoids or cytotoxic drugs. The goal of development of new treatments for immune-mediated diseases is to design ways to specifically interrupt pathologic immune responses, leaving nonpathologic immune responses intact. Novel ways to interrupt pathologic immune responses that are under investigation include the use of anti-inflammatory cytokines or specific cytokine inhibitors as anti-inflammatory agents, the use of monoclonal antibodies against T or B lymphocytes as therapeutic agents, the use of intravenous Ig for certain infections and immune complex–mediated diseases, the use of specific cytokines to reconstitute components of the immune system, and bone marrow transplantation to replace the pathogenic immune system with a more normal immune system **(Chaps. 60, 344, and 197)**. In particular, the use of a monoclonal antibody to B cells (rituximab, anti-CD20 MAb) is approved in the United States for the treatment of non-Hodgkin's lymphoma **(Chap. 104)** and, in combination with methotrexate, for treatment of adult patients with severe rheumatoid arthritis resistant to TNF-α inhibitors **(Chap. 351)**. CTLA-4 inhibitors such as Ipilimumab and Tremelimumab and anti-PD-1 antibodies such as Nivolumab have been shown to reverse CD8 T cell exhaustion in melanoma and other solid tumors and induce immune cell control of tumor growth. CTLA4 and PD-1 inhibitors are currently being studied in HCV or HIV-1 infection to reverse anti-viral CD8 T cell dysfunction and promote the reduction of virus infected cells. A new technique that engineers autologous T cells to express antibody receptors that target leukemic cells, termed T cells with chimeric antigen receptors (T CARs), is currently showing promising results in clinical trials for the treatment of certain types of leukemias and lymphomas.

Cell-based therapies have been studied for many years, including ex vivo activation of NK cells for reinfusion into patients with malignancies, and DC therapy of ex vivo priming of DCs for enhanced presentation of cancer antigens, with reinfusion of primed DCs into the patient. One such strategy for DC therapy has been approved by the FDA for treatment of advanced prostate cancer.

Cytokines and Cytokine Inhibitors Several TNF inhibitors are used as biological therapies in the treatment of rheumatoid arthritis; these include monoclonal antibodies, TNF-R Fc fusion proteins, and Fab fragments. Use of anti-TNF-α antibody therapies such as adalimumab, infliximab, and golimumab has resulted in clinical improvement in patients with these diseases and has opened the way for targeting TNF-α to treat other severe forms of autoimmune and/or inflammatory disease. Blockage of TNF-α has been effective in *rheumatoid arthritis*, *psoriasis*, *Crohn's disease*, and *ankylosing spondylitis*. Other cytokine inhibitors are recombinant soluble TNF-α receptor (R) fused to human Ig and anakinra (soluble *IL-1 receptor antagonist*, or IL-1ra). The treatment of autoinflammatory syndromes (Table 342-6) with recombinant IL-1 receptor antagonist can prevent symptoms in these syndromes, because the overproduction of IL-1β is a hallmark of these diseases.

TNF-αR-Fc fusion protein (etanercept) and IL-1ra act to inhibit the activity of pathogenic cytokines in rheumatoid arthritis, i.e., TNF-α and IL-1, respectively. Similarly, anti-IL-6, IFN-β, and IL-11 act to inhibit pathogenic proinflammatory cytokines. Anti-IL-6 (tocilizumab) inhibits IL-6 activity, whereas IFN-β and IL-11 decrease IL-1 and TNF-α production.

Of particular note has been the successful use of IFN-γ in the treatment of the phagocytic cell defect in *chronic granulomatous disease* **(Chap. 60)**.

Th17 CD4 T cells have been implicated in the pathogenesis of psoriasis, ulcerative colitis, and other autoimmune diseases. Monoclonal antibodies have now been developed that target cytokines (IL-12, IL-23) that induce Th17 T cell differentiation, and are licensed by the FDA for treatment of psoriasis. Monoclonal antibodies that directly target IL-17 have also recently been licensed for psoriasis and psoriatic arthritis treatment.

Monoclonal Antibodies to T and B Cells The OKT3 MAb against human T cells has been used for several years as a T cell–specific immunosuppressive agent that can substitute for horse antithymocyte globulin (ATG) in the treatment of solid organ transplant rejection. OKT3 produces fewer allergic reactions than ATG but does induce human anti-mouse Ig antibody—thus limiting its use. Anti-CD4 MAb therapy has been used in trials to treat patients with rheumatoid arthritis. While inducing profound immunosuppression, anti-CD4 MAb treatment also induces susceptibility to severe infections. Treatment of patients with a MAb against the T cell molecule CD40 ligand (CD154) is under investigation to induce tolerance to organ transplants, with promising results reported in animal studies. Monoclonal antibodies to the CD25 (IL-2a) receptor (basiliximab) are being used for treatment of graft-versus-host disease in bone marrow transplantation, and anti-CD20 MAb (rituximab) is used to treat hematologic neoplasms, autoimmune diseases, kidney transplant rejection, and rheumatoid arthritis. The anti-IgE monoclonal antibody (omalizumab) is used for blocking antigen-specific IgE that causes *hay fever* and *allergic rhinitis* **(Chap. 345)**; however, side effects of anti-IgE include increased risk of anaphylaxis. Studies have shown that T_H17 cells, in addition to T_H1, are mediators of inflammation in Crohn's disease, and anti–IL-12/IL-23p40 antibody therapy has been studied as a treatment.

It is important to realize the potential risks of these immunosuppressive monoclonal antibodies. Natalizumab is a humanized IgG antibody against an a4 integrin that inhibits leukocyte migration into tissues and has been approved for treatment of multiple sclerosis in the United States. Both it and anti-CD20 (rituximab) have been associated with the onset of progressive multifocal leukoencephalopathy (PML)—a serious and usually fatal CNS infection caused by JC polyomavirus. Efalizumab, a humanized IgG monoclonal antibody previously approved for treatment of plaque psoriasis, has now been taken off the market due to reactivation of JC virus leading to fatal PML. Thus, use of any currently approved immunosuppressant immunotherapies should be undertaken with caution and with careful monitoring of patients according to FDA guidelines.

Intravenous Immunoglobulin (IVIg) IVIg has been used successfully to block reticuloendothelial cell function and immune complex clearance in various immune cytopenias such as immune thrombocytopenia **(Chap. 111)**. In addition, IVIg is useful for prevention of tissue damage in certain inflammatory syndromes such as Kawasaki disease **(Chap. 356)** and as Ig replacement therapy for certain types of immunoglobulin deficiencies **(Chap. 344)**. In addition, controlled clinical trials support the use of IVIg in selected patients with graft-versus-host disease, multiple sclerosis, myasthenia gravis, Guillain-Barré syndrome, and chronic demyelinating polyneuropathy.

Stem Cell Transplantation Hematopoietic stem cell transplantation (SCT) is now being comprehensively studied to treat several autoimmune diseases including systemic lupus erythematosus, multiple sclerosis, and scleroderma. The goal of immune reconstitution in autoimmune disease syndromes is to replace a dysfunctional immune system with a normally reactive immune cell repertoire. Preliminary results in patients with scleroderma and lupus have showed encouraging results. Controlled clinical trials in these three diseases are now being launched in the United States and Europe to compare the toxicity and efficacy of conventional immunosuppression therapy with that of myeloablative autologous SCT. Recently, SCT was used in the setting of HIV-1 infection. HIV-1 infection of CD4+ T cells requires the presence of surface CD4 receptor and the chemokine receptor 5 (CCR5) co-receptor. Studies have demonstrated that patients who are homozygous for a 32-bp deletion in the CCR5 allele do not express CD4+ T cell CCR5 and thus are resistant to HIV-1 infection with HIV-1 strains that use this co-receptor. Stem cells from a homozygous CCR5 delta32 donor were transplanted to an HIV-infected patient following standard conditioning for such

transplants, and the patient has maintained long-term control of the virus without antiretrovirals. Thus, a number of recent insights into immune system function have spawned a new field of interventional immunotherapy and have enhanced the prospect for development of more specific and nontoxic therapies for immune and inflammatory diseases.

FURTHER READING

Crotty S: T follicular helper cell differentiation, function, and roles in disease. Immunity 41:529, 2014.

Dupage M, Bluestone JA: Harnessing the plasticity of CD4+T cells to treat immune-mediated disease. Nat Rev Immunol 3:149, 2016.

Kim BS et al: Targeting IL-17 in autoimmunity and inflammation. Arch Pharm Res 39:1537, 2016.

Levine B et al: Autophagy in immunity and inflammation. Nature 469:323, 2011.

Mulay SR, Anders HJ: Crystallopathies. N Engl J Med 374:25, 2016.

Pedersen HK et al: Human gut microbes impact host serum metabolome and insulin sensitivity. Nature 535:376, 2016.

Pfeiffer JK, Virgin HW: Transkingdom control of viral infection and immunity in the mammalian intestine. Science 351:239, 2016.

Ratner D et al: Bacterial secretion systems and regulation of inflammasome activation. J Leuk Bio 101(1):165, 2017.

Yang F et al: The diverse biological functions of neutrophils, beyond the defense against infections. Inflammation 40:311, 2017.

343 The Major Histocompatibility Complex

Gerald T. Nepom

THE HLA COMPLEX AND ITS PRODUCTS

The human major histocompatibility complex (MHC), commonly called the human leukocyte antigen (HLA) complex, is a 4-megabase (Mb) region on chromosome 6 (6p21.3) that is densely packed with expressed genes. The best known of these genes are the *HLA class I* and *class II genes*, whose products are critical for immunologic specificity and transplantation histocompatibility, and they play a major role in susceptibility to a number of autoimmune diseases and some forms of drug hypersensitivity. Many other genes in the HLA region are also essential to the innate and antigen-specific functioning of the immune system. The HLA region shows extensive conservation with the MHC of other mammals in terms of genomic organization, gene sequence, and protein structure and function.

The *HLA class I genes* are located in a 2-Mb stretch of DNA at the telomeric end of the HLA region (Fig. 343-1). The classic (MHC class Ia) HLA-A, B, and C loci, the products of which are integral participants in the immune response to intracellular infections, tumors, and allografts, are expressed in all nucleated cells and are highly polymorphic in the population. *Polymorphism* refers to a high degree of allelic variation within a genetic locus that leads to extensive variation between different individuals expressing different alleles. More than 3400 alleles at HLA-A, 4300 alleles at HLA-B, and 3100 at HLA-C have been identified in different human populations, making this the most highly polymorphic segment known within the human genome. Each of the alleles at these loci encodes a *heavy chain* (also called an α *chain*) that associates noncovalently with the nonpolymorphic light chain β_2*-microglobulin*, encoded on chromosome 15.

The designation of HLA genes and their products is based on a World Health Organization (WHO) nomenclature, in which alleles are given a single designation that indicates locus, allotype, and sequence-based subtype. For example, *HLA-A*02:01* indicates subtype 1 of a group of alleles that encode HLA-A2 molecules. Subtypes that differ from each other at the nucleotide but not the amino acid sequence level are designated by an extra numeral (e.g., *HLA-B*07:02:01* and *HLA-B*07:02:02* are two variants of *HLA-B*07:02*, both encoding the same HLA-B7 molecule). The nomenclature of class II genes, discussed below, is made more complicated by the fact that both chains of a class II molecule are encoded by closely linked HLA-encoded loci, each of which may be polymorphic, and by the presence of differing numbers of isotypic DRB loci in different individuals. It has become clear that accurate HLA genotyping requires DNA sequence analysis, and the identification of alleles at the DNA sequence level has contributed greatly to the understanding of the role of HLA molecules as peptide-binding ligands, to the analysis of associations of HLA alleles with certain diseases, to the study of the population genetics of HLA, and to a clearer understanding of the contribution of HLA differences to allograft rejection and graft-versus-host disease. Current databases of HLA class I and class II sequences can be accessed by the Internet (e.g., from the IMGT/HLA Database, *http://www.ebi.ac.uk/imgt/hla*), and frequent updates of HLA gene lists are published in several journals. It is also possible to predict HLA genotypes by virtue of their linkage with single nucleotide polymorphisms (SNPs) prevalent in the genome. Imputation of HLA alleles using this technique is not as precise as targeted sequencing; however, the technology is much simpler and cheaper.

The biologic significance of this MHC genetic diversity, resulting in extreme variation in the human population, is evident from the perspective of the structure of MHC molecules. As shown in Fig. 343-2, the MHC class I and class II genes encode MHC molecules that bind small peptides, and together this complex (pMHC; peptide-MHC) forms the ligand for recognition by T lymphocytes, through the antigen-specific T cell receptor (TCR). There is a direct link between the genetic variation and this structural interaction: The allelic changes in genetic sequence result in diversification of the peptide-binding capabilities of each MHC molecule and in differences for specific TCR binding. Thus, different pMHC complexes bind different antigens and are targets for recognition by different T cells.

The class I MHC and class II MHC structures, shown in Fig. 343-2*B,C*, are structurally closely related; however, there are a few key differences. While both bind peptides and present them to T cells, the binding pockets have different shapes, which influence the types of immune responses that result (discussed below). In addition, there are structural contact sites for T cell molecules known as CD8 and CD4, expressed on the class I or class II membrane-proximal domains,

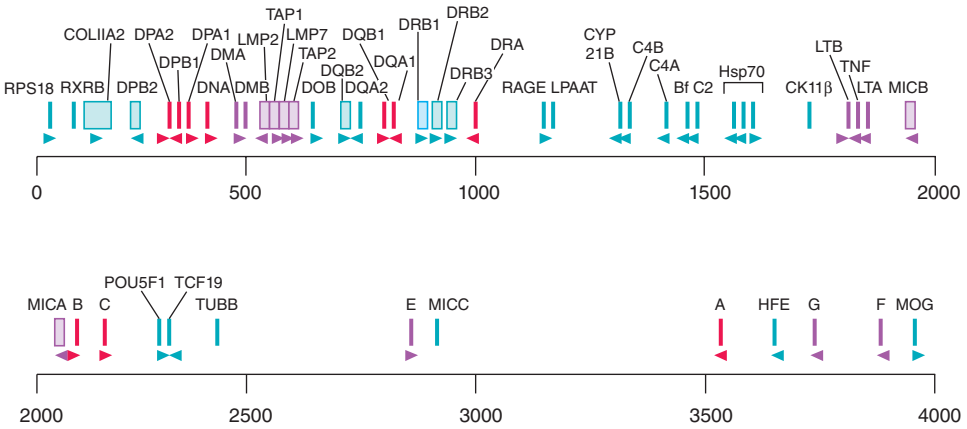


FIGURE 343-1 Physical map of the HLA region, showing the class I and class II loci, other immunologically important loci, and a sampling of other genes mapped to this region. Gene orientation is indicated by *arrowheads*. Scale is in kilobase (kb). The approximate genetic distance from DP to A is 3.2 cM. This includes 0.8 cM between A and B (including 0.2 cM between C and B), 0.4–0.8 cM between B and DR-DQ, and 1.6–2.0 cM between DR-DQ and DP.

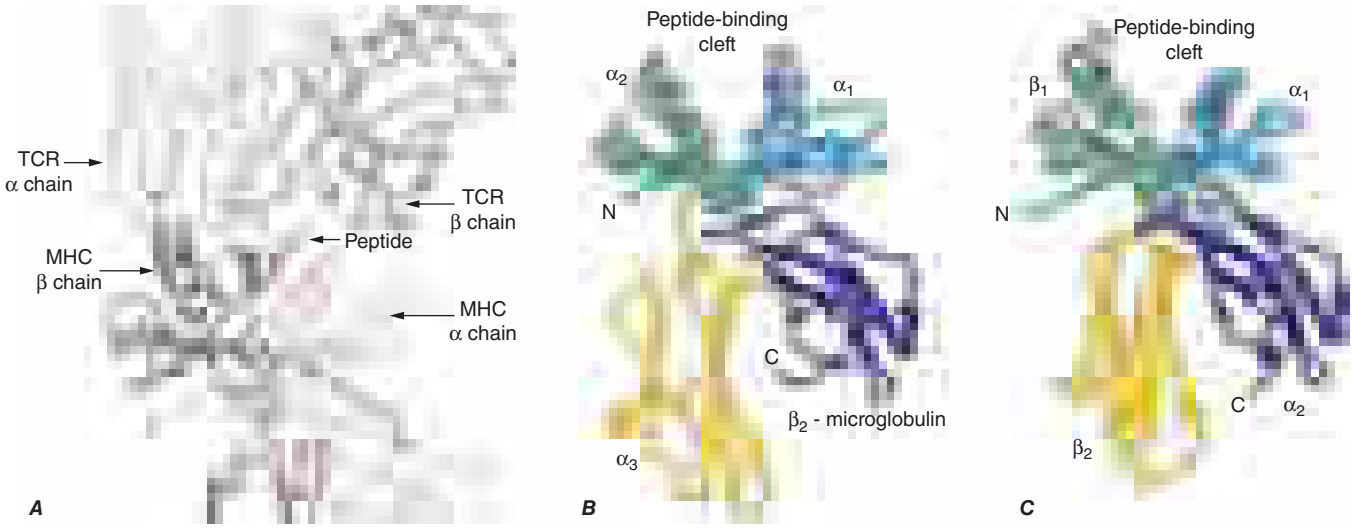


FIGURE 343-2 ***A.*** The trimolecular complex of TCR (*top*), MHC molecule (*bottom*), and a bound peptide form the structural determinants of specific antigen recognition. Other panels (***B*** and ***C***) show the domain structure of MHC class I (***B***) and class II (***C***) molecules. The α_1 and α_2 domains of class I and the α_1 and β_1 domains of class II form a β-sheet platform that forms the floor of the peptide-binding groove, and α helices that form the sides of the groove. The α_3 (***B***) and β_2 domains (***C***) project from the cell surface and form the contact sites for CD8 and CD4, respectively. *(Adapted from EL Reinherz et al: Science 286:1913, 1999; and C Janeway et al: Immunobiology Bookshelf, 2nd ed. Garland Publishing, New York, 1997; with permission.)*

respectively. This ensures that when peptide antigens are presented by class I molecules, the responding T cells are predominantly of the CD8 class, and similarly, that T cells responding to class II pMHC complexes are predominantly CD4.

The nonclassic, or class Ib, MHC molecules, HLA-E, F, and G, are much less polymorphic than MHC Ia and appear to have distinct functions. The HLA-E molecule has a peptide repertoire displaying signal peptides cleaved from classic MHC class I molecules and is the major self-recognition target for the natural killer (NK) cell–inhibitory receptors NKG2A or NKG2C paired with CD94 (see below and **Chap. 342**). This appears to be a function of immune surveillance, because loss of MHC class I signal peptides serves as a surrogate marker for injured or infected cells, leading to release of the inhibitory signal and subsequent activation of NK cells. HLA-E can also bind and present peptides to CD8 T cells, albeit with a limited scope, as eight allelic HLA-E molecules are known. HLA-G was originally described in stem cells and in extravillous trophoblasts, where it is implicated in regulation of maternal-fetal tolerance in pregnancy. It is now recognized as a widely expressed regulatory molecule that is expressed in multiple alternatively spliced forms, and provides inhibitory signals in both cell-bound and soluble forms; induction of expression is associated with downregulatory immunomodulation at sites of inflammation or malignancy. Eighteen allelic HLA-G molecules have been identified, interacting with receptors on NK, T cell, and dendritic cells. HLA-F occurs in four allelic forms, and is expressed on proliferating lymphoid and monocyte cells; its function is largely unknown, although it has been shown to form complexes that interact with specific NK receptors, sometimes together with other class I molecules in the absence of bound peptides. In general, the emerging view of non-classical class Ib molecules is a complex regulatory network for engaging immunomodulatory responses in the absence of traditional forms of antigen recognition attributed to classical class Ia molecules.

Additional class I–like genes have been identified, some HLA-linked and some encoded on other chromosomes, that show only distant homology to the class Ia and Ib molecules but share the three-dimensional class I structure. Those on chromosome 6p21 include MIC-A and MIC-B, which are encoded centromeric to HLA-B, and HLA-HFE, located 3 to 4 cM (centi-Morgan) telomeric of HLA-F. MIC-A and MIC-B do not bind peptide but are expressed on gut and other epithelium in a stress-inducible manner and serve as activation signals for certain γδ T cells, NK cells, CD8 T cells, and activated macrophages, acting through the activating NKG2D receptors. Over 100 MIC-A and 40 MIC-B alleles are known, and additional diversification comes from variable alanine repeat sequences in the transmembrane domain. Due to this structural diversity, MIC-A can be recognized as a foreign tissue target during organ transplantation, contributing to graft failure. HLA-HFE encodes the gene defective in hereditary hemochromatosis **(Chap. 407)**. Among the non-HLA, class I–like genes, CD1 refers to a family of molecules that present glycolipids or other nonpeptide ligands to certain T cells, including T cells with NK activity; FcRn binds IgG within lysosomes and protects it from catabolism **(Chap. 342)**; and Zn-α_2-glycoprotein 1 binds a nonpeptide ligand and promotes catabolism of triglycerides in adipose tissue. Like the HLA-A, B, C, E, and G heavy chains, each of which forms a heterodimer with β2-microglobulin (Fig. 343-2), the class I–like molecules, HLA-HFE, FcRn, and CD1 also bind to β_2-microglobulin, but MIC-A, MIC-B, and Zn-α_2-glycoprotein 1 do not.

The HLA class II region is also illustrated in Fig. 343-1. Multiple class II genes are arrayed within the centromeric 1 Mb of the HLA region, forming distinct haplotypes. A *haplotype* refers to an array of alleles at polymorphic loci along a chromosomal segment. Multiple class II genes are present on a single haplotype, clustered into three major subregions: HLA-DR, DQ, and DP. Each of these subregions contains at least one functional alpha (A) locus and one functional beta (B) locus. Together these encode proteins that form the α and β polypeptide chains of a mature class II HLA molecule. Thus, the DRA and DRB genes encode an HLA-DR molecule; DQA and DQB genes encode HLA-DQ molecules; and DPA and DPB genes encode HLA-DP molecules. There are several DRB genes (*DRB1*, *DRB2*, *DRB3*, etc.), so that two expressed DR molecules are encoded on most haplotypes by combining the α-chain product of the DRA gene with separate β chains. Nearly 2000 alleles have been identified at the HLA-DRB1 locus, with most of the variation occurring within limited segments encoding residues that interact with antigens. Detailed analysis of sequences and population distribution of these alleles strongly suggest that this diversity is actively selected by environmental pressures associated with pathogen diversity. In the DQ region, both DQA1 and DQB1 are polymorphic, with over 70 DQA1 alleles and 900 DQB1 alleles. The current nomenclature is largely analogous to that discussed above for class I, using the convention "locus * allele."

In addition to allelic polymorphism, products of different DQA alleles can, with some limitations, pair with products of different DQB alleles through both *cis* and *trans* pairing to create combinatorial complexity and expand the number of expressed class II molecules. Because of the enormous allelic diversity in the general population, most individuals are heterozygous at all of the class I and class II loci. Thus, most individuals express six classic class I molecules (two each of HLA-A, -B, and -C) and many class II molecules—two DP, two to four DR, and multiple DQ (both *cis* and *trans* dimers).

OTHER GENES IN THE MHC

In addition to the class I and class II genes themselves, there are numerous genes interspersed among the HLA loci that have interesting and important immunologic functions. Our current concept of the function of MHC genes now encompasses many of these additional genes, some of which are also highly polymorphic. Indeed, direct comparison of the complete DNA sequences for eight of the entire 4-Mb MHC regions from different haplotypes shows >44,000 nucleotide variations, encoding an extremely high potential for biologic diversity, and at least 97 genes located in this region are known to have coding region sequence variation. Specific examples include the TAP and LMP genes, as discussed in more detail below, which encode molecules that participate in intermediate steps in the HLA class I biosynthetic pathway. Another set of HLA genes, DMA and DMB, performs an analogous function for the class II pathway. These genes encode an intracellular molecule that facilitates the proper complexing of HLA class II molecules with antigen (see below). The *HLA class III region* is a name given to a cluster of genes between the class I and class II complexes, which includes genes for the two closely related cytokines tumor necrosis factor (TNF)-α and lymphotoxin (TNF-β); the complement components C2, C4, and Bf; heat shock protein (HSP) 70; and the enzyme 21-hydroxylase.

The class I genes HLA-A, B, and C are expressed in all nucleated cells, although generally to a higher degree on leukocytes than on nonleukocytes. In contrast, the class II genes show a more restricted distribution: HLA-DR and HLA-DP genes are constitutively expressed on most cells of the myeloid cell lineage, whereas all three class II gene families (HLA-DR, -DQ, and -DP) are inducible by certain stimuli provided by inflammatory cytokines such as interferon γ. Within the lymphoid lineage, expression of these class II genes is constitutive on B cells and inducible on human T cells. Most endothelial and epithelial cells in the body, including the vascular endothelium and the intestinal epithelium, are also inducible for class II gene expression, and some cells show specialized expression, such as HLA-DQA2 and HLA-DQB2 on Langerhans cells. While somatic tissues normally express only class I and not class II genes, during times of local inflammation, they are recruited by cytokine stimuli to express class II genes as well, thereby becoming active participants in ongoing immune responses. Class II expression is controlled largely at the transcriptional level through a conserved set of promoter elements that interact with a protein known as *CIITA*. Cytokine-mediated induction of CIITA is a principal method by which tissue-specific expression of HLA gene expression is controlled. Other HLA genes involved in the immune response, such as TAP and LMP, are also susceptible to upregulation by signals such as interferon γ.

LINKAGE DISEQUILIBRIUM

In addition to extensive polymorphism at the class I and class II loci, another characteristic feature of the HLA complex is *linkage disequilibrium*. This is formally defined as a deviation from Hardy-Weinberg equilibrium for alleles at linked loci. This is reflected in the very low recombination rates between certain loci within the HLA complex. For example, recombination between DR and DQ loci is almost never observed in family studies, and characteristic haplotypes with particular arrays of DR and DQ alleles are found in every population. Similarly, the complement components C2, C4, and Bf are almost invariably inherited together, and the alleles at these loci are found in characteristic haplotypes. In contrast, there is a recombinational hotspot between DQ and DP, which are separated by 1–2 cM of genetic distance, despite their close physical proximity. Certain extended haplotypes encompassing the interval from DQ into the class I region are commonly found, the most notable being the haplotype DR3-B8-A1, which is found, in whole or in part, in 10–30% of northern European whites. As discussed below under HLA and immunologic disease, one consequence of the phenomenon of linkage disequilibrium has been the resulting difficulty in assigning HLA-disease associations to a single allele at a single locus.

MHC STRUCTURE AND FUNCTION

Class I and class II molecules display a distinctive structural architecture, which contains specialized functional domains responsible for the unique genetic and immunologic properties of the HLA complex. The principal known function of both class I and class II HLA molecules is to bind antigenic peptides in order to present antigen to an appropriate T cell. The ability of a particular peptide to satisfactorily bind to an individual HLA molecule is a direct function of the molecular fit between the amino acid residues on the peptide with respect to the amino acid residues of the HLA molecule. The bound peptide forms a tertiary structure called the *MHC-peptide complex*, which communicates with T lymphocytes through binding to the TCR molecule. The first site of TCR-MHC-peptide interaction in the life of a T cell occurs in the thymus, where self-peptides are presented to developing thymocytes by MHC molecules expressed on thymic epithelium and hematopoietically derived antigen-presenting cells, which are primarily responsible for positive and negative selection, respectively **(Chap. 342)**. Thus, the population of MHC–T cell complexes expressed in the thymus shapes the TCR repertoire. Mature T cells encounter MHC molecules in the periphery both in the maintenance of tolerance **(Chap. 348)** and in the initiation of immune responses. The TCR-MHC-peptide interaction is the central event in the initiation of most antigen-specific immune responses, since it is the structural determinant of the specificity. For potentially immunogenic peptides, the ability of a given peptide to be generated and bound by an HLA molecule is a primary feature of whether or not an immune response to that peptide can be generated, and the repertoire of peptides that a particular individual's HLA molecules can bind exerts a major influence over the specificity of that individual's immune response.

When a TCR molecule binds to an HLA-peptide complex, it forms intermolecular contacts with both the antigenic peptide and with the HLA molecule itself. The outcome of this recognition event depends on the density and duration of the binding interaction, accounting for a dual specificity requirement for activation of the T cell. That is, the TCR must be specific both for the antigenic peptide and for the HLA molecule. The polymorphic nature of the presenting molecules, and the influence that this exerts on the peptide repertoire of each molecule, results in the phenomenon of *MHC restriction* of the T cell specificity for a given peptide. The binding of CD8 or CD4 molecules to the class I or class II molecule, respectively, also contributes to the interaction between T cell and the HLA-peptide complex, by providing for the selective activation of the appropriate T cell.

CLASS I STRUCTURE

(Fig. 343-2B) As noted above, MHC class I molecules provide a cell-surface display of peptides derived from intracellular proteins, and they also provide the signal for self-recognition by NK cells. Surface-expressed class I molecules consist of an MHC-encoded 44-kD glycoprotein heavy chain, a non-MHC-encoded 12-kD light chain β_2-microglobulin, and an antigenic peptide, typically 8–11 amino acids in length and derived from intracellularly produced protein. The heavy chain displays a prominent peptide-binding groove. In HLA-A and B molecules, the groove is ~3 nm in length by 1.2 nm in maximum width (30 Å × 12 Å), whereas it is apparently somewhat wider in HLA-C. Antigenic peptides are noncovalently bound in an extended conformation within the peptide-binding groove, with both N- and C-terminal ends anchored in pockets within the groove (A and F pockets, respectively) and, in many cases, with a prominent kink, or arch, approximately one-third of the way from the N-terminus that elevates the peptide main chain off the floor of the groove.

A remarkable property of peptide binding by MHC molecules is the ability to form highly stable complexes with a wide array of peptide sequences. This is accomplished by a combination of peptide sequence–independent and peptide sequence–dependent bonding. The former consists of hydrogen bond and van der Waals interactions between conserved residues in the peptide-binding groove and charged or polar atoms along the peptide backbone. The latter is dependent upon the six side pockets that are formed by the irregular surface produced by protrusion of amino acid side chains from within the binding groove. The side chains lining the pockets interact with some of the peptide side chains. The sequence polymorphism among different class I alleles and isotypes predominantly affects the residues that line these pockets, and the interactions of these residues with

peptide residues constitute the sequence-dependent bonding that confers a particular sequence "motif" on the range of peptides that can bind each MHC molecule.

CLASS I BIOSYNTHESIS

(Fig. 343-3*A*) The biosynthesis of the classic MHC class I molecules reflects their role in presenting endogenous peptides. The heavy chain is cotranslationally inserted into the membrane of the endoplasmic reticulum (ER), where it becomes glycosylated and associates sequentially with the chaperone proteins calnexin and ERp57. It then forms a complex with β_2-microglobulin, and this complex associates with the chaperone calreticulin and the MHC-encoded molecule tapasin, which physically links the class I complex to TAP, the MHC-encoded transporter associated with antigen processing. Meanwhile, peptides generated within the cytosol from intracellular proteins by the multisubunit, multicatalytic proteasome complex are actively transported into the ER by TAP, where they are trimmed by enzymes known as ER aminopeptidases. At this point, peptides with appropriate sequence complementarity bind specific class I molecules to form complete, folded heavy chain–β_2-microglobulin–peptide trimer complexes. These are transported rapidly from the ER, through the *cis*- and *trans*-Golgi where the N-linked oligosaccharide is further processed, and thence to the cell surface.

Most of the peptides transported by TAP are produced in the cytosol by proteolytic cleavage of intracellular proteins by the multisubunit, multicatalytic proteasome, and inhibitors of the proteasome dramatically reduce expression of class I–presented antigenic peptides. A thiol-dependent oxidoreductase ERp57, which mediates disulfide bond rearrangements, also appears to play an important role in folding the class I–peptide complex into a stable multicomponent molecule. The MHC-encoded proteasome subunits LMP2 and LMP7 may influence the spectrum of peptides produced but are not essential for proteasome function.

CLASS I FUNCTION

Peptide Antigen Presentation On any given cell, a class I molecule occurs in 100,000–200,000 copies and binds several hundred to several thousand distinct peptide species. The vast majority of these peptides are self-peptides to which the host immune system is tolerant by one or more of the mechanisms that maintain tolerance (e.g., clonal deletion in the thymus or clonal anergy or clonal ignorance in the periphery [Chaps. 342 and 348]). However, class I molecules bearing foreign peptides expressed in a permissive immunologic context activate CD8 T cells, which, if naïve, will then differentiate into cytolytic T lymphocytes (CTLs). These T cells and their progeny, through their αβ TCRs, are then capable of Fas/CD95- and/or perforin-mediated cytotoxicity and/or cytokine secretion (Chap. 342) upon further encounter with the class I–peptide combination that originally activated it, or other structurally related class I–peptide complexes. As alluded to above, this phenomenon by which T cells recognize foreign antigens in the context of specific MHC alleles is termed *MHC restriction*, and the specific MHC molecule is termed the *restriction element*. The most common source of foreign peptides presented by class I molecules is viral infection, in the course of which peptides from viral proteins enter the class I pathway. The generation of a strong CTL response that destroys virally infected cells represents an important antigen-specific defense against many viral infections (Chap. 342). In the case of some viral infections—hepatitis B, for example—CTL-induced target cell apoptosis is thought to be a more important mechanism of tissue damage than

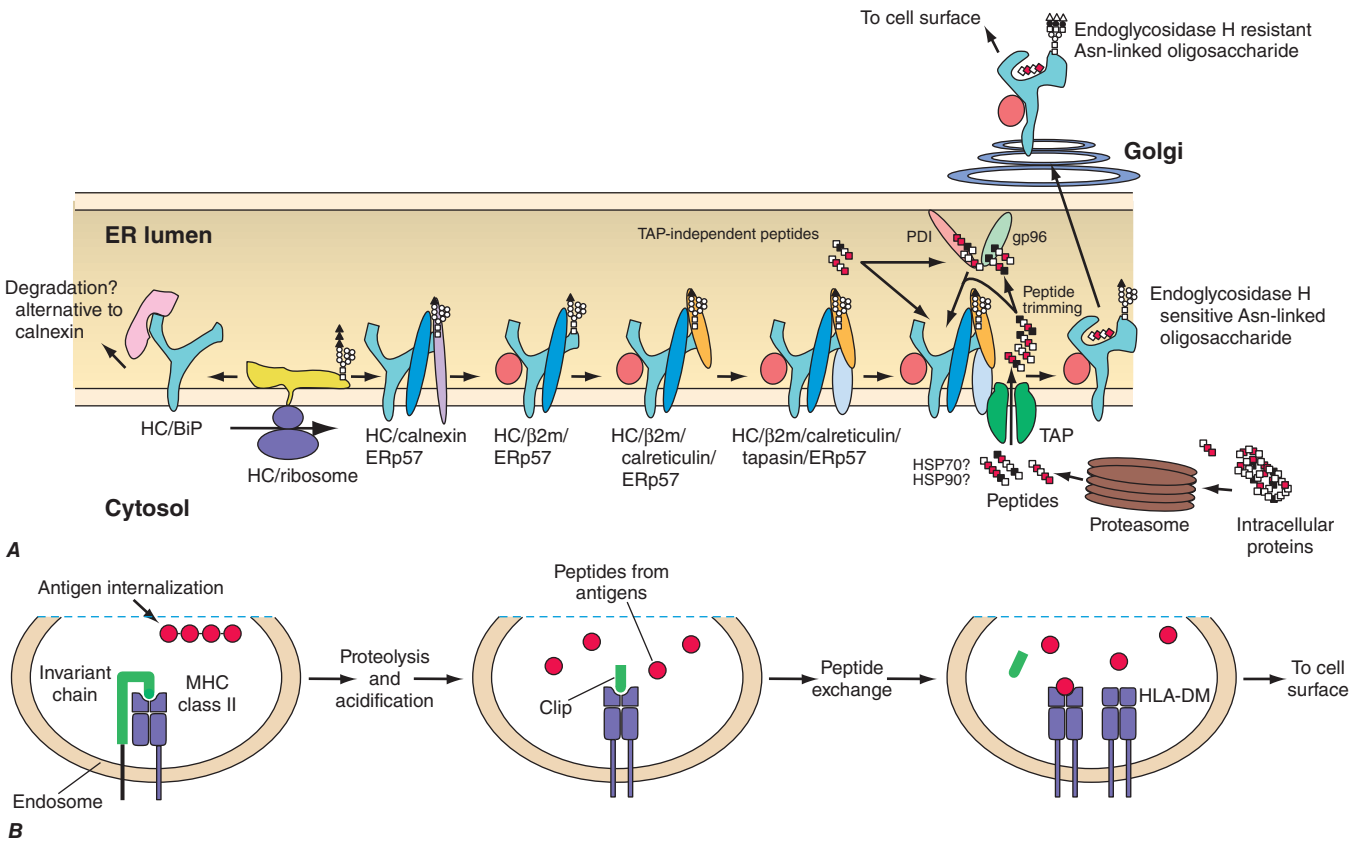


FIGURE 343-3 Biosynthesis of class I (*A*) and class II (*B*) molecules. ***A.*** Nascent heavy chain (HC) becomes associated with β_2-microglobulin (β2m) and peptide through interactions with a series of chaperones. Peptides generated by the proteasome are transported into the endoplasmic reticulum (ER) by TAP. Peptides undergo N-terminal trimming in the ER and become associated with chaperones, including gp96 and PDI. Once peptide binds to HC-β2m, the HC-β2m-peptide trimeric complex exits the ER and is transported by the secretory pathway to the cell surface. In the Golgi, the N-linked oligosaccharide undergoes maturation, with addition of sialic acid residues. Molecules are not necessarily drawn to scale. ***B.*** Pathway of HLA class II molecule assembly and antigen processing. After transport through the Golgi and post-Golgi compartment, the class II–invariant chain complex moves to an acidic endosome, where the invariant chain is proteolytically cleaved into fragments and displaced by antigenic peptides, facilitated by interactions with the DMA-DMB chaperone protein. This class II molecule–peptide complex is then transported to the cell surface.

any direct cytopathic effect of the virus itself. The importance of the class I pathway in the defense against viral infection is underscored by the identification of a number of viral products that interfere with the normal class I biosynthetic pathway and thus block the immunogenic expression of viral antigens.

Other examples of intracellularly generated peptides that can be presented by class I molecules in an immunogenic manner include peptides derived from nonviral intracellular infectious agents (e.g., *Listeria*, *Plasmodium*), tumor antigens, minor histocompatibility antigens, and certain autoantigens. There are also situations in which cell surface–expressed class I molecules are thought to acquire and present exogenously derived peptides.

HLA Class I Receptors and NK Cell Recognition (Chap. 342)
NK cells, which play an important role in innate immune responses, are activated to cytotoxicity and cytokine secretion by contact with cells that lack MHC class I expression, and NK cell activation is inhibited by cells that express MHC class I. In humans, the recognition of class I molecules by NK cells is carried out by three classes of receptor families, the killer cell–inhibitory cell receptor (KIR) family, the leukocyte Ig-like receptor (LIR) family, and the CD94/NKG2 family. The KIR family, also called CD158, is encoded on chromosome 19q13.4. KIR gene nomenclature is based on the number of domains (2D or 3D) and the presence of long (L) or short (S) cytoplasmic domains. The KIR2DL1 and S1 molecules primarily recognize alleles of HLA-C, which possess a lysine at position 80 (HLA-Cw2, -4, -5, and -6), whereas the KIR2DL2/S2 and KIR2DL3/S3 families primarily recognize alleles of HLA-C with asparagine at this position (HLA-Cw1, -3, -7, and -8). The KIR3DL1 and S1 molecules predominantly recognize HLA-B alleles that fall into the HLA-Bw4 class determined by residues 77–83 in the α_1 domain of the heavy chain, whereas the KIR3DL2 molecule is an inhibitory receptor for HLA-A*03. One of the KIR products, KIR2DL4, is known to be an activating receptor for HLA-G, and KIR3DL2 and KIR2DS4 have been described as immunoregulatory ligands interacting with HLA-F. The most common KIR haplotype in whites contains one activating KIR and six inhibitory KIR genes, although there is a great deal of diversity in the population, with >100 different combinations. It appears that most individuals have at least one inhibitory KIR for a self-HLA class I molecule, providing a structural basis for NK cell target specificity, which helps prevent NK cells from attacking normal cells. The importance of KIR-HLA interactions to many immune responses is illustrated by studies associating KIR3DL1 or S1 with multiple sclerosis (Chap. 436), an autoimmune disease, but also with partial protection against HIV (Chap. 197), in both cases consistent with a role for HLA-KIR–mediated NK activation. Studies also show an association of KIR2DS1 with protection from relapse following allogeneic bone marrow transplantation in acute myeloid leukemia when these inhibitory receptors in the donors do not recognize the recipient HLA-C.

The LIR gene family (CD85, also called ILT) is encoded centromeric of the KIR locus on 19q13.4, and it encodes a variety of inhibitory immunoglobulin-like receptors expressed on many lymphocyte and other hematopoietic lineages. Interaction of LIR-1 (ILT2) with NK or T cells inhibits activation and cytotoxicity, mediated by many different HLA class I molecules, including HLA-G. HLA-F also appears to interact with LIR molecules, although the functional context for this is not understood.

The third family of NK receptors for HLA is encoded in the NK complex on chromosome 12p12.3-13.1 and consists of CD94 and five NKG2 genes, A/B, C, E/H, D, and F. These molecules are C-type (calcium-binding) lectins, and most of these function as disulfide-bonded heterodimers between CD94 and one of the NKG2 glycoproteins. The principal ligand of CD94/NKG2A receptors is the HLA-E molecule, complexed to a peptide derived from the signal sequence of classic HLA class I molecules and HLA-G. Thus, analogous to the way in which KIR receptors recognize HLA-C, the NKG2 receptor monitors self–class I expression, albeit indirectly through peptide recognition in the context of HLA-E. NKG2C, E, and H appear to have similar specificities but act as activating receptors. NKG2D is expressed as a homodimer and functions as an activating receptor expressed on NK cells, γδ TCR T cells, and activated CD8 T cells. When complexed with an adaptor called DAP10, NKG2D recognizes MIC-A and MIC-B molecules and activates the cytolytic response. NKG2D also binds a class of molecules known as *ULBP*, structurally related to class I molecules but not encoded in the MHC. **The function of NK cells in immune responses is discussed in Chap. 342.**

CLASS II STRUCTURE

(Fig. 343-2*C*) A specialized functional architecture similar to that of the class I molecules can be seen in the example of a class II molecule depicted in Fig. 343-2*C*, with an antigen-binding cleft arrayed above a supporting scaffold that extends the cleft toward the external cellular environment. However, in contrast to the HLA class I molecular structure, β_2-microglobulin is not associated with class II molecules. Rather, the class II molecule is a heterodimer, composed of a 29-kD α chain and a 34-kD β chain. The amino-terminal domains of each chain form the antigen-binding elements that, like the class I molecule, cradle a peptide in a groove bounded by extended α-helical loops, one encoded by the A (α chain) gene and one by the B (β chain) gene. Like the class I groove, the class II antigen-binding groove is punctuated by pockets that contact the side chains of amino acid residues of the bound peptide, but unlike the class I groove, it is open at both ends. Therefore, peptides bound by class II molecules vary greatly in length, since both the N- and C-terminal ends of the peptides can extend through the open ends of this groove. Approximately 11 amino acids within the bound peptide form intimate contacts with the class II molecule itself, with backbone hydrogen bonds and specific side chain interactions combining to provide, respectively, stability and specificity to the binding (Fig. 343-4).

The genetic polymorphisms that distinguish different class II genes correspond to changes in the amino acid composition of the class II molecule, and these variable sites are clustered predominantly around the pocket structures within the antigen-binding groove. As with class I, this is a critically important feature of the class II molecule, which explains how genetically different individuals have functionally different HLA molecules.

BIOSYNTHESIS AND FUNCTION OF CLASS II MOLECULES

(Fig. 343-3*B*) The intracellular assembly of class II molecules occurs within a specialized compartmentalized pathway that differs dramatically from the class I pathway described above. As illustrated in Fig. 343-3*B*, the class II molecule assembles in the ER in association with a chaperone molecule, known as the *invariant chain*. The invariant chain performs at least two roles. First, it binds to the class II molecule and blocks the peptide-binding groove, thus preventing antigenic peptides from binding. This role of the invariant chain appears to account for one of the important differences between class I and class II MHC pathways, since it can explain why class I molecules present endogenous peptides from proteins newly synthesized in the ER but class II molecules generally do not. Second, the invariant chain contains molecular localization signals that direct the class II molecule to traffic into post-Golgi compartments known as *endosomes*, which develop into specialized acidic compartments where proteases cleave the invariant chain, and antigenic peptides can now occupy the class II groove. The specificity and tissue distribution of these proteases appear to be an important way in which the immune system regulates access to the peptide-binding groove and T cells become exposed to specific self-antigens. Differences in protease expression in the thymus and in the periphery may in part determine which specific peptide sequences comprise the peripheral repertoire for T cell recognition. It is at this stage in the intracellular pathway, after cleavage of the invariant chain, that the MHC-encoded DM molecule catalytically facilitates the exchange of peptides within the class II groove to help optimize the specificity and stability of the MHC-peptide complex.

Once this MHC-peptide complex is deposited in the outer cell membrane, it becomes the target for T cell recognition via a specific TCR expressed on lymphocytes. Because the endosome environment contains internalized proteins retrieved from the extracellular environment, the class II–peptide complex often contains bound antigens that

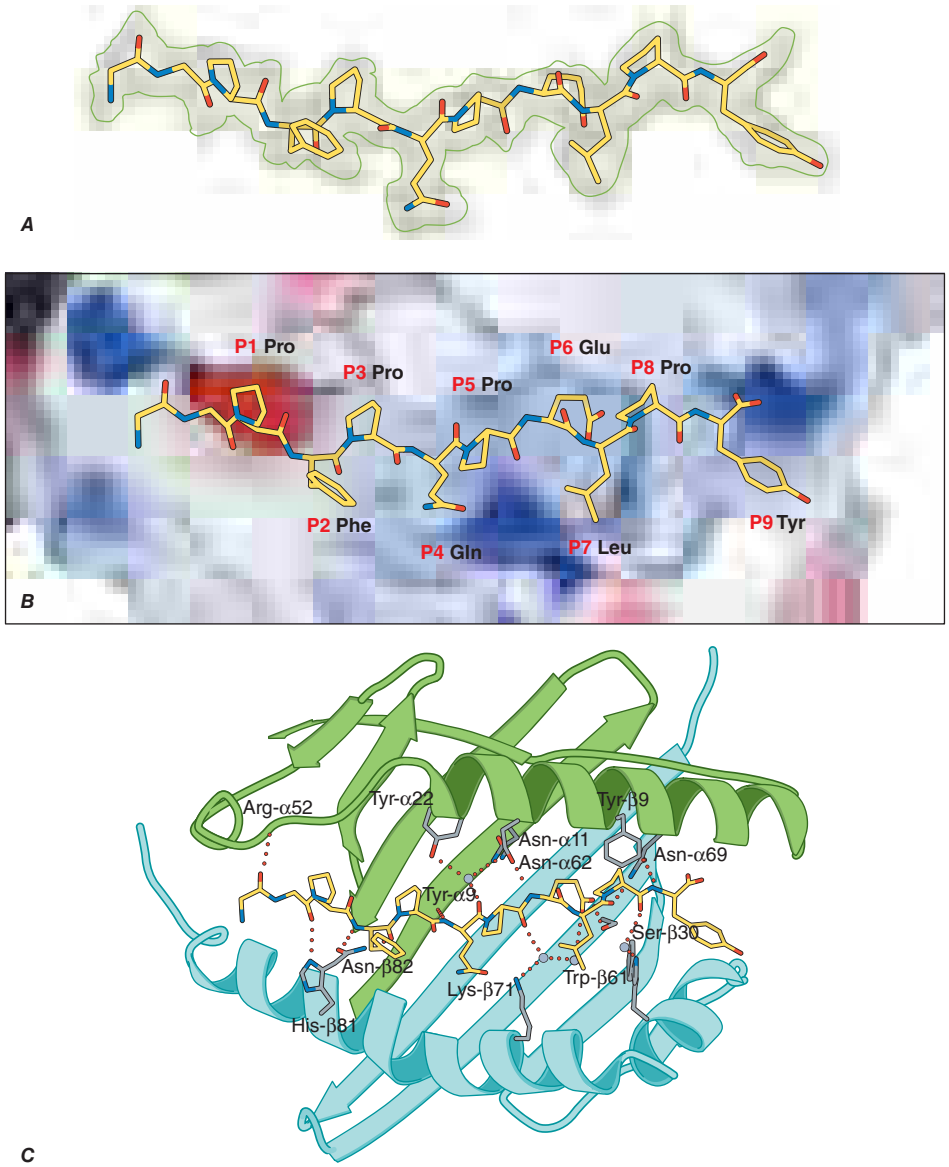


FIGURE 343-4 Specific intermolecular interactions determine peptide binding to MHC class II molecules. A short peptide sequence derived from alpha-gliadin (***A***) is accommodated within the MHC class II binding groove by specific interactions between peptide side chains (the P1–P9 residues illustrated in ***B***) and corresponding pockets in the MHC class II structure. The latter are determined by the genetic polymorphisms of the MHC gene, in this case encoding an HLA-DQ2 molecule (***C***). This shows the extensive hydrogen bond and salt bridge network, which tightly constrains the pMHC complex and presents the complex of antigen and restriction element for CD4 T cell recognition. *(From C Kim et al: Structural basis for HLA-DQ2-mediated presentation of gluten epitopes in celiac disease. Proc Natl Acad Sci USA 101:4175, 2004.)*

were originally derived from extracellular proteins. In this way, the class II peptide–loading pathway provides a mechanism for immune surveillance of the extracellular space. This appears to be an important feature that permits the class II molecule to bind foreign peptides, distinct from the endogenous pathway of class I–mediated presentation.

ROLE OF HLA IN TRANSPLANTATION

The development of modern clinical transplantation in the decades since the 1950s provided a major impetus for elucidation of the HLA system, as allograft survival is highest when donor and recipient are HLA-identical. Although many molecular events participate in transplantation rejection, allogeneic differences at class I and class II loci play a major role. Class I molecules can promote T cell responses in several different ways. In the cases of allografts in which the host and donor are mismatched at one or more class I loci, host T cells can be activated by classic *direct alloreactivity*, in which the antigen receptors on the host T cells react with the foreign class I molecule expressed on the allograft. In this situation, the response of any given TCR may be dominated by the allogeneic MHC molecule, the peptide bound to it, or some combination of the two. Another type of host anti-graft T cell response involves the uptake and processing of donor MHC antigens by host antigen-presenting cells and the subsequent presentation of the resulting peptides by host MHC molecules. This mechanism is termed *indirect alloreactivity*.

In the case of class I molecules on allografts that are shared by the host and the donor, a host T cell response may still be triggered because of peptides that are presented by the class I molecules of the graft but not of the host. The most common basis for the existence of these endogenous antigen peptides, called *minor histocompatibility antigens*, is a genetic difference between donor and host at a non-MHC locus encoding the structural gene for the protein from which the peptide is derived. These loci are termed *minor histocompatibility loci*, and nonidentical individuals typically differ at many such loci. CD4 T cells react to analogous class II variation, both direct and indirect, and class II differences alone are sufficient to drive allograft rejection.

ASSOCIATION OF HLA ALLELES WITH SUSCEPTIBILITY TO DISEASE

It has long been postulated that infectious agents provide the driving force for the allelic diversification seen in the HLA system. An important corollary of this hypothesis is that resistance to specific pathogens may differ between individuals, based on HLA genotype. Observations of specific HLA genes associated with resistance to malaria or dengue fever, persistence of hepatitis B, and to disease progression in HIV infection are consistent with this model. For example, failure to clear persistent hepatitis B or C viral infection may reflect the inability of particular HLA molecules to present viral antigens effectively to T cells. Similarly, both protective and susceptible HLA allelic associations have been described for human papilloma virus–associated cervical neoplasia, implicating the MHC as an influence in mediating viral clearance in this form of cancer.

Pathogen diversity is probably also the major selective pressure favoring HLA heterozygosity. The extraordinary scope of HLA allelic diversity increases the likelihood that most new pathogens will be recognized by some HLA molecules, helping to ensure immune fitness to the host. However, another consequence of diversification is that some alleles may become capable of recognition of "innocent bystander" molecules, including drugs, environmental molecules, and tissue-derived self-antigens. In a few instances, single HLA alleles display a strong selectivity for binding of a particular agent that accounts for a genetically determined response: Hypersensitivity to abacavir, an antiretroviral therapeutic, is directly linked to binding of abacavir in the antigen-binding pockets of *HLA-B*57:01*, where it is buried underneath antigenic peptides and distorts the landscape, changing T cell recognition specificity; an adverse drug reaction to abacavir is >500 times more likely to occur in persons with *HLA-B*57:01* than in individuals without this HLA allele. Other examples include chronic beryllium toxicity, which is linked to binding of beryllium by HLA-DP molecules

with a specific glutamic acid polymorphic residue on the class II beta chain, clindamycin-related cutaneous drug reactions which are more common in individuals with *HLA-B*51:01*, and dapsone hypersensitivity in patients with leprosy who express *HLA-B*13:01*. Even in the case of more complex diseases, particular HLA alleles are strongly associated with certain inappropriate immune-mediated disease states, particularly for some common autoimmune disorders (Chap. 348). By comparing allele frequencies in patients with any particular disease and in control populations, >100 such associations have been identified, some of which are listed in Table 343-1. The strength of genetic association is reflected in the term *relative risk*, which is a statistical odds ratio representing the risk of disease for an individual carrying a particular genetic marker compared with the risk for individuals in that population without that marker. The nomenclature shown in Table 343-1 reflects both the HLA serotype (e.g., DR3, DR4) and the HLA genotype (e.g., *DRB1*03:01*, *DRB1*04:01*). It is very likely that the class I and class II alleles themselves are the true susceptibility alleles for most of these associations. However, because of the extremely strong linkage disequilibrium between the DR and DQ loci, in some cases it has been difficult to determine the specific locus or combination of class II loci involved. In some cases, the susceptibility gene may be one of the HLA-linked genes located near the class I or class II region, but not the HLA gene itself, and in other cases, the susceptibility gene may be a non-HLA gene such as TNF-α, which is nearby. Indeed, since linkage disequilibrium of some haplotypes extends across large segments of the MHC region, it is quite possible that combinations of genes may account for the particular associations of HLA haplotypes with disease. For example, on some haplotypes associated with rheumatoid arthritis (RA), both HLA-DRB1 alleles and a particular polymorphism associated with the TNF locus may be contributory to disease risk. Other candidates for similar epistatic effects include the IKBL gene and the MICA locus, potentially in combination with classic HLA class II risk alleles.

As might be predicted from the known function of the class I and class II gene products, almost all of the diseases associated with specific HLA alleles have an immunologic component to their pathogenesis. The recent development of soluble HLA-peptide recombinant molecules as biological probes of T cell function, often in multivalent complexes referred to as "MHC tetramers," represents an opportunity to use HLA genetic associations to develop biomarkers for detection of early disease progression. However, it should be stressed that even the strong HLA associations with disease (those associations with relative risk of ≥10) implicate normal, rather than defective, alleles. Most individuals who carry these susceptibility genes do not express the associated disease; in this way, the particular HLA gene is permissive for disease but requires other environmental (e.g., the presence of specific antigens) or genetic factors for full penetrance. In each case studied, even in diseases with very strong HLA associations, the concordance of disease in monozygotic twins is higher than in HLA-identical dizygotic twins or other sibling pairs, indicating that non-HLA genes contribute to susceptibility and can significantly modify the risk attributable to HLA.

TABLE 343-1 Significant HLA Class I and Class II Associations with Disease

	MARKER	GENE	STRENGTH OF ASSOCIATION
Spondyloarthropathies			
Ankylosing spondylitis	B27	*B*27:02,–04, –05*	++++
Reactive arthritis (Reiter's)	B27		++++
Acute anterior uveitis	B27		+++
Reactive arthritis (*Yersinia, Salmonella, Shigella, Chlamydia*)	B27		+++
Psoriatic spondylitis	B27		+++
Collagen-Vascular Diseases			
Juvenile arthritis, pauciarticular	DR8		++
	DR5		++
Rheumatoid arthritis	DR4	*DRB1*04:01, –04, –05*	+++
Sjögren's syndrome	DR3		++
Systemic lupus erythematosus			+
White	DR3		++
Japanese	DR2		
Autoimmune Gut and Skin			
Gluten-sensitive enteropathy (celiac disease)	DQ2	*DQA1*05:01 DQB1*02:01*	+++
Chronic active hepatitis	DR3		++
Dermatitis herpetiformis	DR3		+++
Psoriasis vulgaris	Cw6		++
Pemphigus vulgaris	DR4	*DRB1*04:02*	+++
	DQ1	*DQB1*05:03*	+
Bullous pemphigoid variant	DQ7	*DQB1*03:01*	
Autoimmune Endocrine			
Type 1 diabetes mellitus	DQ8	*DQB1*03:02*	+++
	DR4	*DRB1*04:01,–04*	++
	DR3		—
	DR2	*DQB1*06:02*	+[a]
Hyperthyroidism (Graves')	B8		+
	DR3		+
Hyperthyroidism (Japanese)	B35		++
Adrenal insufficiency	DR3		
Autoimmune Neurologic			
Myasthenia gravis	B8		+
Multiple sclerosis	DR2	*DRB1*15:01*	+
	DR2	*DRB5*01:01*	++
Other			
Behçet's disease	B51		++
Congenital adrenal hyperplasia	B47	*21·OH (Cyp21B)*	+++
Narcolepsy	DR2	*DQB1*06:02*	++++
Goodpasture's syndrome (anti-GBM)	DR2		++
Abacavir hypersensitivity	B57	*B*57:01*	++++

[a]Strong negative association, that is, genetic association with protection from diabetes.

Abbreviation: GBM, glomerular basement membrane.

Another group of diseases is genetically linked to HLA, not because of the immunologic function of HLA alleles but rather because they are caused by autosomal dominant or recessive abnormal alleles at loci that happen to reside in or near the HLA region. Examples of these are 21-hydroxylase deficiency (Chap. 379), hemochromatosis (Chap. 407), and spinocerebellar ataxia (Chap. 429).

CLASS I ASSOCIATIONS WITH DISEASE

Although the associations of human disease with particular HLA alleles or haplotypes predominantly involve the class II region, there are also several prominent disease associations with class I alleles. These include the association of Behçet's disease (Chap. 357) with HLA-B51, psoriasis vulgaris (Chap. 53) with HLA-Cw6, and, most notably, the spondyloarthritides (Chap. 355) with HLA-B27. More than 150 HLA-B locus alleles, designated *HLA-B*27:01–B*27:154*, encode the family of

B27 class I molecules. All of the subtypes share a common B pocket in the peptide-binding groove—a deep, negatively charged pocket that shows a strong preference for binding the arginine side chain. In addition, B27 is among the most negatively charged of HLA class I heavy chains, and the overall preference is for positively charged peptides. *HLA-B*27:05* is the predominant subtype in whites and most other non-Asian populations, and this subtype is very highly associated with ankylosing spondylitis (AS) **(Chap. 355)**, both in its idiopathic form and in association with chronic inflammatory bowel disease or psoriasis vulgaris. It is also associated with reactive arthritis (ReA) **(Chap. 355)**, with other idiopathic forms of peripheral arthritis (undifferentiated spondyloarthropathy), and with recurrent acute anterior uveitis. B27 is found in 50–90% of individuals with these conditions, compared with a prevalence of ~7% in North American whites.

Evidence that the B27 molecule itself is involved in disease pathogenesis comes both from clinical epidemiology and on the occurrence of a spondyloarthropathy-like disease in HLA-B27 transgenic rats. The association of B27 with these diseases may derive from the specificity of a particular peptide or family of peptides bound to B27 or through another mechanism that is independent of the peptide specificity of B27. In particular, HLA-B27 has been shown to form heavy chain homodimers, utilizing the cysteine residue at position 67 of the B57 α chain, in the absence of β_2-microglobulin. These homodimers are expressed on the surface of lymphocytes and monocytes from patients with AS, and receptors including KIR3DL1, KIR3DL2, and ILT4 (LILRB2) are capable of binding to them, promoting the activation and survival of cells expressing these receptors. Alternatively, this dimerization "misfolding" of B27 may initiate an intracellular stress signaling response, called the unfolded protein response (UPR), capable of modulating immune cell function, possibly in enthesial-resident T cells that act as sensors of damage and environmental stress.

CLASS II DISEASE ASSOCIATIONS

As can be seen in Table 343-1, the majority of associations of HLA and disease are with class II alleles. Several diseases have complex HLA genetic associations.

Celiac Disease In the case of celiac disease **(Chap. 318)**, it is probable that the HLA-DQ genes are the primary basis for the disease association. HLA-DQ genes present on both the celiac-associated DR3 and DR7 haplotypes include the *DQB1*02:01* gene, and further detailed studies have documented a specific class II αβ dimer encoded by the *DQA1*05:01* and *DQB1*02:01* genes, which appears to account for most of the HLA genetic contribution to celiac disease susceptibility. This specific HLA association with celiac disease may have a straightforward explanation: Peptides derived from the wheat gluten component gliaden are bound to the molecule encoded by *DQA1*05:01* and *DQB1*02:01* and presented to T cells. Gliaden-derived peptides that are implicated in this immune activation bind the DQ class II dimer best when the peptide contains a glutamine to glutamic acid substitution. It has been proposed that tissue transglutaminase, an enzyme present at increased levels in the intestinal cells of celiac patients, converts glutamine to glutamic acid in gliadin, creating peptides that are capable of being bound by the DQ2 molecule and presented to T cells.

Pemphigus Vulgaris In the case of pemphigus vulgaris **(Chap. 55)**, there are two HLA genes associated with disease, *DRB1*04:02* and *DQB1*05:03*. Peptides derived from desmoglein-3, an epidermal autoantigen, bind to the *DRB1*04:02*- and *DQB1*05:03*-encoded HLA molecules, and this combination of specific peptide binding and disease-associated class II molecule is sufficient to stimulate desmoglein-specific T cells. A bullous pemphigoid clinical variant, not involving desmoglein recognition, has been found to be associated with *HLA-DQB1*03:01*.

Juvenile Arthritis Pauciarticular juvenile arthritis **(Chap. 351)** is an autoimmune disease associated with genes at the DRB1 locus and also with genes at the DPB1 locus. Patients with both *DPB1*02:01* and a DRB1 susceptibility allele (usually *DRB1*08* or **05*) have a higher relative risk than expected from the additive effect of those genes alone. In juvenile patients with rheumatoid factor–positive polyarticular disease, heterozygotes carrying both *DRB1*04:01* and **04:04* have a relative risk >100, reflecting an apparent synergy in individuals inheriting both of these susceptibility genes.

Type 1 Diabetes Mellitus Type 1 (autoimmune) diabetes mellitus **(Chap. 396)** is associated with MHC genes on more than one haplotype. The presence of both the DR3 and DR4 haplotypes in one individual confers a twentyfold increased risk for type 1 diabetes; the strongest single association is with *DQB1*03:02*, and all haplotypes that carry a *DQB1*03:02* gene are associated with type 1 diabetes, whereas related haplotypes that carry a different DQB1 gene are not. However, the relative risk associated with inheritance of this gene can be modified, depending on other HLA genes present either on the same or a second haplotype. For example, the presence of a DR2-positive haplotype containing a *DQB1*06:02* gene is associated with decreased risk. This gene, *DQB1*06:02*, is considered "protective" for type 1 diabetes. Even some DRB1 genes that can occur on the same haplotype as *DQB1*03:02* may modulate risk, so that individuals with the DR4 haplotype that contains *DRB1*04:03* are less susceptible to type 1 diabetes than individuals with other DR4-DQB1*03:02 haplotypes. There are some characteristic structural features of the diabetes-associated DQ molecule encoded by *DQB1*03:02*, particularly the capability for binding peptides that have negatively charged amino acids near their C-termini. This may indicate a role for specific antigenic peptides or T cell interactions in the immune response to islet-associated proteins.

Although the presence of a DR3 haplotype in combination with the DR4-DQB1*0302 haplotype is a very high-risk combination for diabetes susceptibility, the specific gene on the DR3 haplotype that is responsible for this synergy is not yet identified.

Rheumatoid Arthritis The HLA genes associated with RA **(Chap. 351)** encode a distinctive sequence of amino acids from codons 67 to 74 of the DRβ molecule: RA-associated class II molecules carry the sequence LeuLeuGluGlnArgArgAlaAla or LeuLeuGluGlnLysArgAlaAla in this region, whereas non-RA-associated genes carry one or more differences in this region. These residues form a portion of the molecule that lies in the middle of the α-helical portion of the DRB1-encoded class II molecule, termed the *shared epitope*.

The highest risk for susceptibility to RA comes in individuals who carry both a *DRB1*04:01* and *DRB1*04:04* gene. These DR4-positive RA-associated alleles with the *shared epitope* are most frequent among patients with more severe, erosive disease. Several mechanisms have been proposed that link the shared epitope to immune reactivity in RA. This portion of the class II molecule may allow preferential binding of an arthritogenic peptide, it may favor the expansion of a type of self-reactive T lymphocyte, or it may itself form part of the pMHC ligand recognized by TCR that initiates synovial tissue recognition.

MOLECULAR MECHANISMS FOR HLA-DISEASE ASSOCIATIONS

As noted above, HLA molecules play a key role in the selection and establishment of the antigen-specific T cell repertoire and a major role in the subsequent activation of those T cells during the initiation of an immune response. Precise genetic polymorphisms characteristic of individual alleles dictate the specificity of these interactions and thereby instruct and guide antigen-specific immune events. These same genetically determined pathways are therefore implicated in disease pathogenesis when specific HLA genes are responsible for autoimmune disease susceptibility.

The fate of developing T cells within the thymus is determined by the affinity of interaction between TCR and HLA molecules bearing self-peptides, and thus the particular HLA types of each individual control the precise specificity of the T cell repertoire **(Chap. 342)**. The primary basis for HLA-associated disease susceptibility may well lie within this thymic maturation pathway. The positive selection of potentially autoreactive T cells, based on the presence of specific HLA susceptibility genes, may establish the threshold for disease risk in a particular individual.

At the time of onset of a subsequent immune response, the primary role of the HLA molecule is to bind peptide and present it to antigen-specific T cells. The HLA complex can therefore be viewed as encoding genetic determinants of precise immunologic activation events. Antigenic peptides that bind particular HLA molecules are capable of stimulating T cell immune responses; peptides that do not bind are not presented to T cells and are not immunogenic. This genetic control of the immune response is mediated by the polymorphic sites within the HLA antigen–binding groove that interact with the bound peptides. In autoimmune and immune-mediated diseases, it is likely that specific tissue antigens that are targets for pathogenic lymphocytes are complexed with the HLA molecules encoded by specific susceptibility alleles. In autoimmune diseases with an infectious etiology, it is likely that immune responses to peptides derived from the initiating pathogen are bound and presented by particular HLA molecules to activate T lymphocytes that play a triggering or contributory role in disease pathogenesis. The concept that early events in disease initiation are triggered by specific HLA-peptide complexes offers some prospects for therapeutic intervention, since it may be possible to design compounds that interfere with the formation or function of specific HLA-peptide–TCR interactions.

When considering mechanisms of HLA associations with immune response and disease, it is well to remember that immune-mediated disease is a multistep process in which initial HLA-peptide recognition helps establish a repertoire of potentially reactive T cells, whereas subsequent HLA-associated antigen presentation provides the essential peptide-binding specificity for peripheral T cell recognition leading to activation. These deterministic events can occur long before clinical evidence of autoimmunity, as exemplified by the HLA genetic associations with detection of specific autoantibodies in type 1 diabetes and in rheumatoid arthritis that are present for several years before disease diagnosis.

FURTHER READING

Karlin E, Phillips E: Genotyping for severe drug hypersensitivity. Curr Allergy Asthma Rep 14:418, 2014.

Liu J et al: Phenome-wide association study maps new diseases to the human major histocompatibility complex region. J Med Genet 53:681, 2016.

Robinson J et al: The IPD and IPD-IMGT/HLA Database: Allele variant databases. Nucleic Acids Res 43:D423, 2015.

Zheng X et al: HIBAG—HLA genotype imputation with attribute bagging. Pharmacogenomics J 14:192, 2014.

344 Primary Immune Deficiency Diseases

Alain Fischer

Immunity is intrinsic to life and an important tool in the fight for survival against pathogenic microorganisms. The human immune system can be divided into two major components: the innate immune system and the adaptive immune system **(Chap. 342)**. The innate immune system provides the rapid triggering of inflammatory responses based on the recognition (at the cell surface or within cells) of either molecules expressed by microorganisms or molecules that serve as "danger signals" released by cells under attack. These receptor/ligand interactions trigger signaling events that ultimately lead to inflammation. Virtually all cell lineages (not just immune cells) are involved in innate immune responses; however, myeloid cells (i.e., neutrophils and macrophages) play a major role because of their phagocytic capacity. The adaptive immune system operates by clonal recognition of antigens followed by a dramatic expansion of antigen-reactive cells and execution of an immune effector program. Most of the effector cells die off rapidly, whereas memory cells persist. Although both T and B lymphocytes recognize distinct chemical moieties and execute distinct adaptive immune responses, the latter is largely dependent on the former in generating long-lived humoral immunity. Adaptive responses utilize components of the innate immune system; for example, the antigen-presentation capabilities of dendritic cells help to determine the type of effector response. Not surprisingly, immune responses are controlled by a series of regulatory mechanisms.

Hundreds of gene products have been characterized as effectors or mediators of the immune system **(Chap. 342)**. Whenever the expression or function of one of these products is genetically impaired (provided the function is nonredundant), a primary immunodeficiency (PID) occurs.

PIDs are genetic diseases with primarily Mendelian inheritance. More than 350 conditions have now been described, and deleterious mutations in ~346 genes have been identified. The overall prevalence of PIDs has been estimated in various countries at 5–10 per 100,000 individuals; however, given the difficulty in diagnosing these rare and complex diseases, this figure is probably an underestimate. PIDs can involve all possible aspects of immune responses, from innate through adaptive, cell differentiation, and effector function and regulation. For the sake of clarity, PIDs should be classified according to (1) the arm of the immune system that is defective and (2) the mechanism of the defect (when known). **Table 344-1** classifies the most prevalent PIDs according to this manner of classification; however, one should bear in mind that the classification of PIDs sometimes involves arbitrary decisions because of overlap and, in some cases, lack of data.

The consequences of PIDs vary widely as a function of the molecules that are defective. This concept translates into multiple levels of vulnerability to infection by pathogenic and opportunistic microorganisms, ranging from extremely broad (as in severe combined immunodeficiency [SCID]) to narrowly restricted to a single microorganism (as in Mendelian susceptibility to mycobacterial disease [MSMD]). The locations of the sites of infection and the causal microorganisms involved will thus help physicians arrive at proper diagnoses. PIDs can also lead to immunopathologic responses such as allergy (as in Wiskott-Aldrich syndrome [WAS]), lymphoproliferation, and autoimmunity. A combination of recurrent infections, inflammation, and autoimmunity can be observed in a number of PIDs, thus creating obvious therapeutic challenges. Finally, some PIDs increase the risk of cancer, notably but not exclusively lymphocytic cancers, for example, lymphoma.

DIAGNOSIS OF PRIMARY IMMUNODEFICIENCIES

The most frequent symptom prompting the diagnosis of a PID is the presence of recurrent or unusually severe infections. As mentioned above, recurrent allergic or autoimmune manifestations may also alert the physician to a possible diagnosis of PID. In such cases, a detailed account of the subject's personal and family medical history should be obtained. It is of the utmost importance to gather as much medical information as possible on relatives and up to several generations of ancestors. In addition to the obvious focus on primary symptoms, the clinical examination should evaluate the size of lymphoid organs and, when appropriate, look for the characteristic signs of a number of complex syndromes that may be associated with a PID.

The performance of laboratory tests should be guided to some extent by the clinical findings. Infections of the respiratory tract (bronchi, sinuses) mostly suggest a defective antibody response. In general, invasive bacterial infections can result from complement deficiencies, signaling defects of innate immune responses, asplenia, or defective antibody responses. Viral infections, recurrent *Candida* infections, and opportunistic infections are generally suggestive of impaired T cell immunity. Skin infections and deep-seated abscesses primarily reflect innate immune defects (such as chronic granulomatous disease); however, they may also appear in the autosomal dominant hyper-IgE syndrome. **Table 344-2** summarizes the laboratory tests that are most frequently used to diagnose a PID. More specific tests (notably genetic tests) are then used to make a definitive diagnosis. Genomic tools now allow us to more efficiently track genetic defects through usage of gene panel resequencing and/or whole exome sequencing.

TABLE 344-1 Classification of Primary Immune Deficiency Diseases

Deficiencies of the Innate Immune System	
• Phagocytic cells:	
- Impaired production: severe congenital neutropenia (SCN)	
- Asplenia	
- Impaired adhesion: leukocyte adhesion deficiency (LAD)	
- Impaired killing: chronic granulomatous disease (CGD)	
• Innate immunity receptors and signal transduction:	
- Defects in Toll-like receptor signaling	
- Mendelian susceptibility to mycobacterial disease	
• Complement deficiencies:	
- Classical, alternative, and lectin pathways	
- Lytic phase	
Deficiencies of the Adaptive Immune System	
• T lymphocytes:	
- Impaired development	Severe combined immune deficiencies (SCIDs)
	DiGeorge's syndrome
- Impaired survival, migration, function	Combined immunodeficiencies
	Hyper-IgE syndrome (autosomal dominant)
	DOCK8 deficiency
	CD40 ligand deficiency
	Wiskott-Aldrich syndrome
	Ataxia-telangiectasia and other DNA repair deficiencies
• B lymphocytes:	
- Impaired development	XL and AR agammaglobulinemia
- Impaired function	Hyper-IgM syndrome
	Common variable immunodeficiency (CVID)
	IgA deficiency
Regulatory Defects	
• Innate immunity	Autoinflammatory syndromes (outside the scope of this chapter)
	Severe colitis
	Hemophagocytic lymphohistiocytosis (HLH)
• Adaptive immunity	Autoimmune lymphoproliferation syndrome (ALPS)
	Autoimmunity and inflammatory diseases (IPEX, APECED)

Abbreviations: APECED, autoimmune polyendocrinopathy candidiasis ectodermal dysplasia; AR, autosomal recessive; IPEX, immunodysregulation polyendocrinopathy enteropathy X-linked syndrome; XL, X-linked.

TABLE 344-2 Tests Most Frequently Used to Diagnose a Primary Immune Deficiency (PID)

TEST	INFORMATION	PID DISEASE
• Blood cell counts and cell morphology	Neutrophil counts[a]	↓ Severe congenital neutropenia, ↑↑ LAD
	Lymphocyte counts[a]	T cell ID
	Eosinophilia	WAS, hyper-IgE syndrome
	Howell-Jolly bodies	Asplenia
• Chest x-ray	Thymic shadow	SCID, DiGeorge's syndrome
	Costochondral junctions	Adenosine deaminase deficiency
• Bone x-ray	Metaphyseal ends	Cartilage hair hypoplasia
• Immunoglobulin serum levels	IgG, IgA, IgM	B cell ID
	IgE	Hyper-IgE syndrome, WAS, T cell ID
• Lymphocyte phenotype	T, B lymphocyte counts	T cell ID, agammaglobulinemia
• Dihydrorhodamine fluorescence (DHR) assay	Reactive oxygen species production by PMNs	Chronic granulomatous disease
Nitroblue tetrazolium (NBT) assay		
• CH50, AP50	Classic and alternative complement pathways	Complement deficiencies
• Ultrasonography of the abdomen	Spleen size	Asplenia

[a]Normal counts vary with age. For example, the lymphocyte count is between 3000 and 9000/μL of blood below the age of 3 months and between 1500 and 2500/μL in adults.

Abbreviations: ID, immunodeficiency; LAD, leukocyte adhesion deficiency; PMNs, polymorphonuclear leukocytes; SCID, severe combined immunodeficiency; WAS, Wiskott-Aldrich syndrome.

The PIDs discussed below have been grouped together according to the affected cells and the mechanisms involved (Table 344-1, Fig. 344-1).

PRIMARY IMMUNODEFICIENCIES OF THE INNATE IMMUNE SYSTEM

PIDs of the innate immune system are relatively rare and account for ~10% of all PIDs.

SEVERE CONGENITAL NEUTROPENIA

Severe congenital neutropenia (SCN) consists of a group of inherited diseases that are characterized by severely impaired neutrophil counts (<500 polymorphonuclear leukocytes [PMN]/μL of blood). The condition is usually manifested from birth. SCN may also be cyclic (with a 3-week periodicity), and other neutropenia syndromes can also be intermittent. Although the most frequent inheritance pattern for SCN is autosomal dominant, autosomal recessive and X-linked recessive conditions also exist. Bacterial infections at the interface between the body and the external milieu (e.g., the orifices, wounds, and the respiratory tract) are common manifestations. Bacterial infections can rapidly progress through soft tissue and are followed by dissemination in the bloodstream. Severe visceral fungal infections can also ensue. The absence of pus is a hallmark of this condition.

Diagnosis of SCN requires examination of the bone marrow. Most SCNs are associated with a block in granulopoiesis at the promyelocytic stage (Fig. 344-1). SCN has multiple etiologies, and to date, mutations in 16 different genes have been identified. Most of these mutations result in isolated SCN, whereas others are syndromic (Chap. 60). The most frequent forms of SCN are caused by the premature cell death of granulocyte precursors, as observed in deficiencies of GFI1, HAX1, and elastase 2 (*ELANE*), with the latter accounting for 50% of SCN sufferers. Certain *ELANE* mutations cause cyclic neutropenia syndrome. A gain-of-function mutation in the *WASP* gene (see the section on "Wiskott-Aldrich Syndrome" below) causes X-linked SCN, which is also associated with monocytopenia.

As mentioned above, SCN exposes the patient to life-threatening, disseminated bacterial and fungal infections. Treatment requires careful hygiene measures, notably in infants. Later in life, special oral and dental care is essential, along with the prevention of bacterial infection by prophylactic administration of trimethoprim/sulfamethoxazole. Subcutaneous injection of the cytokine granulocyte colony-stimulating factor (G-CSF) usually improves neutrophil development and thus prevents infection in most SCN diseases. However, there are two caveats: (1) a few cases of SCN with *ELANE* mutation are refractory to G-CSF and may require curative treatment via allogeneic hematopoietic stem cell transplantation (HSCT); and (2) a subset of G-CSF-treated patients carrying *ELANE* mutations are at a greater risk of developing acute myelogenous leukemia associated (in most cases) with somatic gain-of-function mutations of the G-CSF receptor gene.

A few SCN conditions are associated with additional immune defects involving leukocyte migration as observed in the WHIM syndrome (gain of function mutation of the chemokine CXCR4) or in moesin deficiency.

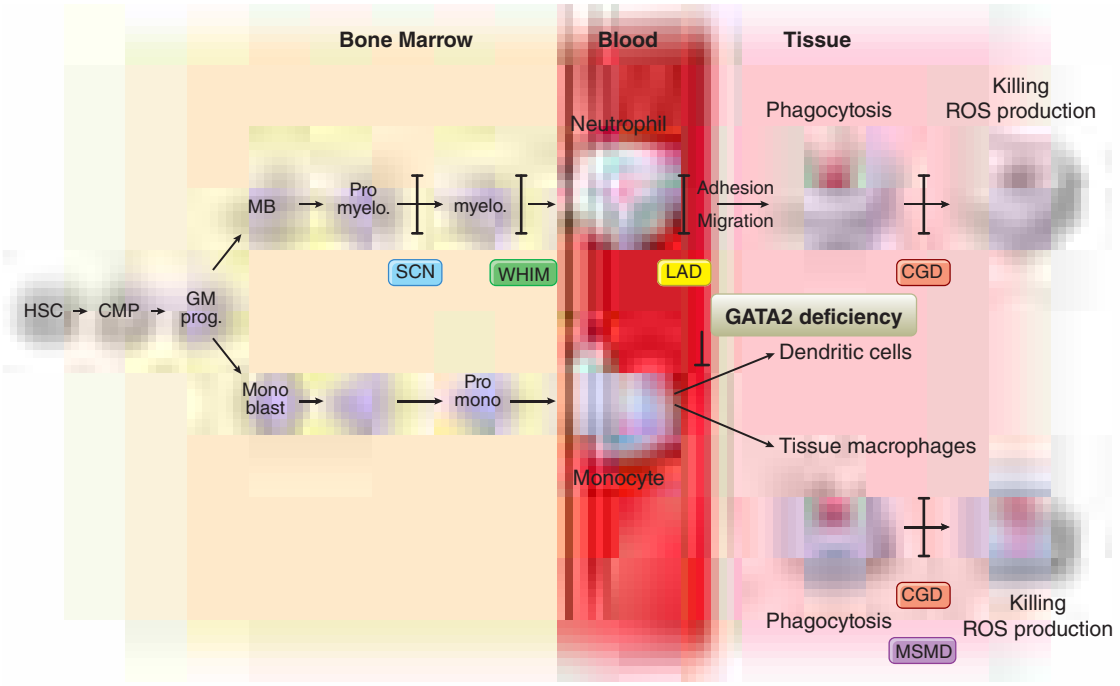


FIGURE 344-1 Differentiation of phagocytic cells and related primary immunodeficiencies (PIDs). Hematopoietic stem cells (HSCs) differentiate into common myeloid progenitors (CMPs) and then granulocyte-monocyte progenitors (GM-prog.), which, in turn, differentiate into neutrophils (MB: myeloblasts; Promyelo: promyelocytes; myelo: myelocytes) or monocytes (monoblasts and promonocytes). Upon activation, neutrophils adhere to the vascular endothelium, transmigrate, and phagocytose the targets. Reactive oxygen species (ROS) are delivered to the microorganism-containing phagosomes. Macrophages in tissues kill using the same mechanism. Following activation by interferon γ (not shown here), macrophages can be armed to kill intracellular pathogens such as mycobacteria. For sake of simplicity, not all cell differentiation stages are shown. The abbreviations for PIDs are contained in boxes placed at corresponding stages of the pathway. CGD, chronic granulomatous diseases; GATA2, zinc finger transcription factor; LAD, leukocyte adhesion deficiencies; MSMD, Mendelian susceptibility to mycobacterial disease; SCN, severe congenital neutropenia; WHIM, warts, hypogammaglobulinemia, infections, and myelokathexis.

ASPLENIA

Primary failure of the development of a spleen is an extremely rare disease that can be either syndromic (in Ivemark syndrome) or isolated with an autosomal dominant expression; in the latter case, mutations in the ribosomal protein SA and the NKX2.5 genes were recently found. Due to the absence of natural filtration of microbes in the blood, asplenia predisposes affected individuals to fulminant infections by encapsulated bacteria. Although most infections occur in the first years of life, cases may also arise in adulthood. The diagnosis is confirmed by abdominal ultrasonography and the detection of Howell-Jolly bodies in red blood cells. Effective prophylactic measures (twice-daily oral penicillin and appropriate vaccination programs) usually prevent fatal outcomes.

GATA2 DEFICIENCY

Recently an immunodeficiency combining monocytopenia and dendritic and lymphoid (B and natural killer [NK]) cell deficiency (DCML), also called monocytopenia with nontuberculous mycobacterial infections (mono-MAC), has been described as a consequence of a dominant mutation in the gene *GATA2*, a transcription factor involved in hematopoiesis. This condition also predisposes to lymphedema, myelodysplasia, and acute myeloid leukemia. Infections (bacterial and viral) are life-threatening, thus indicating, together with the malignant risk, HSCT.

LEUKOCYTE ADHESION DEFICIENCY

Leukocyte adhesion deficiency (LAD) consists of three autosomal recessive conditions (LAD I, II, and III) **(Chap. 60)**. The most frequent condition (LAD I) is caused by mutations in the β2 integrin gene; following leukocyte activation, β2 integrins mediate adhesion to inflamed endothelium expressing cognate ligands. LAD III results from a defect in a regulatory protein (kindlin, also known as Fermt 3) involved in activating the ligand affinity of β2 integrins. The extremely rare LAD II condition is the end result of a defect in selectin-mediated leukocyte rolling that occurs prior to β2 integrin binding. There is a primary defect in fucose transporter such that oligosaccharide selectin ligands are missing in this syndromic condition.

Given that neutrophils are not able to reach infected tissues, LAD renders the individual susceptible to bacterial and fungal infections in a way that is similar to that of patients with SCN. LAD also causes impaired wound healing and delayed loss of the umbilical cord. A diagnosis can be suspected in cases of pus-free skin/tissue infections and massive hyperleukocytosis (>30,000/μL) in the blood (mostly granulocytes). Patients with LAD III also develop bleeding because the β2 integrin in platelets is not functional. Use of immunofluorescence and functional assays to detect β2 integrin can help form a diagnosis. Severe forms of LAD may require HSCT, although gene therapy is also now being considered. Neutrophil-specific granule deficiency (a very rare condition caused by a mutation in the gene for transcription factor C/EBPα) results in a condition that is clinically similar to LAD.

CHRONIC GRANULOMATOUS DISEASES

Chronic granulomatous diseases (CGDs) are characterized by impaired phagocytic killing of microorganisms by neutrophils and macrophages **(Chap. 60)**. The incidence is ~1 per 200,000 live births. About 70% of cases are associated with X-linked recessive inheritance versus autosomal inheritance in the remaining 30%. CGD causes deep-tissue bacterial and fungal abscesses in macrophage-rich organs such as the lymph nodes, liver, and lungs. Recurrent skin infections (such as folliculitis) are common and can prompt an early diagnosis of CGD. The infectious agents are typically catalase-positive bacteria (such as *Staphylococcus aureus* and *Serratia marcescens*) but also include *Burkholderia cepacia*, pathogenic mycobacteria (in certain regions of the world), and fungi (mainly filamentous molds, such as *Aspergillus*).

CGD is caused by defective production of reactive oxygen species (ROS) in the phagolysosome membrane following phagocytosis of microorganisms. It results from the lack of a component of NADPH oxidase (gp91phox or p22phox) or of the associated adapter/activating proteins (p47phox, p67phox, or p40phox) that mediate the transport of electrons into the phagolysosome for creating ROS by interaction

with O_2. Under normal circumstances, these ROS either directly kill engulfed microorganisms or enable the rise in pH needed to activate the phagosomal proteases that contribute to microbial killing. Diagnosis of CGD is based on assays of ROS production in neutrophils and monocytes (Table 344-2). As its name suggests, CGD is also a granulomatous disease. Macrophage-rich granulomas can often arise in the liver, spleen, and other organs. These are sterile granulomas that cause disease by obstruction (bladder, pylorus, etc.) or inflammation (colitis, restrictive lung disease).

The management of infections in patients with CGD can be a complex process. The treatment of bacterial infections is generally based on combination therapy with antibiotics that are able to penetrate into cells. The treatment of fungal infections requires aggressive, long-term use of antifungals. Inflammatory/granulomatous lesions are usually steroid-sensitive; however, glucocorticoids often contribute to the spread of infections. Hence, there is strong need for new therapeutic options in what is still a poorly understood disease.

The treatment of CGD mostly relies on preventing infections. It has been unambiguously demonstrated that prophylactic usage of trimethoprim/sulfamethoxazole is both well tolerated and highly effective in reducing the risk of bacterial infection. Daily administration of azole derivatives (notably itraconazole) also reduces the frequency of fungal complications. It has long been suggested that interferon γ administration is helpful, although medical experts continue to disagree over this controversial issue. Patients may do reasonably well with prophylaxis and careful management. However, other patients develop lifelong severe and persistent fungal infections and/or chronic inflammatory complications, leading to consideration of performing HSCT. Due to increase in reported successes, HSCT is now an established curative approach for CGD; however, the risk-versus-benefit ratio must be carefully assessed on a case-by-case basis. Gene therapy approaches are also being evaluated.

MENDELIAN SUSCEPTIBILITY TO MYCOBACTERIAL DISEASE

This group of diseases is characterized by a defect in the interleukin-12 (IL-12)–interferon (IFN) γ axis (including IL-12p40, IL-12 receptor [R] β_1, IFN-γ R_1 and R_2, STAT1, IRF8, and ISG515 deficiencies), which ultimately leads to impaired IFN-γ-dependent macrophage activation. Both recessive and dominant inheritance modes have been observed. The hallmark of this PID is a specific and narrow vulnerability to tuberculous and nontuberculous mycobacteria. The most severe phenotype (as observed in complete IFN-γ receptor deficiency) is characterized by disseminated infection that can be fatal even when aggressive and appropriate antimycobacterial therapy is applied. In addition to mycobacterial infections, MSMD patients (and particularly those with an IL-12/IL-12 R deficiency) are prone to developing *Salmonella* infections. Although MSMDs are very rare, they should be considered in any patient with persistent mycobacterial infection. Treatment with IFN-γ may efficiently bypass an IL-12/IL-12R deficiency.

TOLL-LIKE RECEPTOR (TLR) PATHWAY DEFICIENCIES

In a certain group of patients with early-onset, invasive *Streptococcus pneumoniae* infections or (less frequently) *Staphylococcus aureus* or other pyogenic infections, conventional screening for PIDs does not identify the cause of the defect in host defense. It has been established that these patients carry recessive mutations in genes that encode essential adaptor molecules (IRAK4 and MYD88) involved in the signaling pathways of the majority of known TLRs **(Chap. 342)**. Remarkably, susceptibility to infection appears to decrease after the first few years of life—perhaps an indication that adaptive immunity (once triggered by an initial microbial challenge) is then able to prevent recurrent infections.

Certain TLRs (TLR-3, 7, 8, and 9) are involved in the recognition of RNA and DNA and usually become engaged during viral infections. Very specific susceptibility to herpes simplex encephalitis has been described in patients with a deficiency in Unc93b (a molecule associated with TLR-3, 7, 8, and 9 required for correct subcellular localization), TLR-3, or associated signaling molecules TRIF, TBK1, and TRAF3, resulting in defective type I IFN production. The fact that no other TLR deficiencies have been found—despite extensive screening of patients with unexplained, recurrent infections—strongly suggests that these receptors are functionally redundant. Hypomorphic mutations in NEMO/IKK-γ (a member of the NF-κB complex, which is activated downstream of TLR receptors) lead to a complex, variable immunodeficiency, and a number of associated features. Susceptibility to both invasive, pyogenic infections and mycobacteria may be observed in this particular setting.

COMPLEMENT DEFICIENCY

The complement system is composed of a complex cascade of plasma proteins **(Chap. 342)** that leads to the deposition of C3b fragments on the surface of particles and the formation of immune complexes that can culminate in the activation of a lytic complex at the bacterial surface. C3 cleavage can be mediated via three pathways: the classic, alternate, and lectin pathways. C3b coats particles as part of the opsonization process that facilitates phagocytosis following binding to cognate receptors. A deficiency in any component of the classic pathway (C1q, C1r, C1s, C4, and C2) can predispose an individual to bacterial infections that are tissue-invasive or that occur in the respiratory tract. Likewise, a C3 deficiency or a deficiency in factor I (a protein that regulates C3 consumption, thus leading to a C3 deficiency due to its absence) also results in the same type of vulnerability to infection. It has recently been reported that a very rare deficiency in ficolin-3 predisposes affected individuals to bacterial infections. Deficiencies in the alternative pathway (factors D and properdin) are associated with the occurrence of invasive *Neisseria* infections.

Lastly, deficiencies of any complement component involved in the lytic phase (C5, C6, C7, C8, and, to a lesser extent, C9) predispose affected individuals to systemic infection by *Neisseria*. This is explained by the critical role of complement in the lysis of the thick cell wall possessed by this class of bacteria.

Diagnosis of a complement deficiency relies primarily on testing the status of the classic and alternate pathway via functional assays, that is, the CH50 and AP50 tests, respectively. When either pathway is profoundly impaired, determination of the status of the relevant components in that pathway enables a precise diagnosis. Appropriate vaccinations and daily administration of oral penicillin are efficient means of preventing recurrent infections. It is noteworthy that several complement deficiencies (in the classic pathway and the lytic phase) may also predispose affected individuals to autoimmune diseases (notably systemic lupus erythematosus; **Chap. 349**).

PRIMARY IMMUNODEFICIENCIES OF THE ADAPTIVE IMMUNE SYSTEM

T LYMPHOCYTE DEFICIENCIES (TABLE 344-1, FIGS. 344-2 AND 344-3)

Given the central role of T lymphocytes in adaptive immune responses **(Chap. 342)**, PIDs involving T cells generally have severe pathologic consequences; this explains the poor overall prognosis and the need for early diagnosis and the early intervention with appropriate therapy. Several differentiation pathways of T cell effectors have been described, one or all of which may be affected by a given PID (Fig. 344-2). Follicular helper CD4+ T cells in germinal centers are required for T-dependent antibody production, including the generation of Ig class-switched, high-affinity antibodies. CD4+ T_H1 cells provide cytokine-dependent (mostly IFN-γ-dependent) help to macrophages for intracellular killing of various microorganisms, including mycobacteria and *Salmonella*. CD4+ T_H2 cells produce IL-4, IL-5, and IL-13 and thus recruit and activate eosinophils and other cells required to fight helminth infections. CD4+ T_H17 cells produce IL-17 and IL-22 cytokines that recruit neutrophils to the skin and lungs to fight bacterial and fungal infections. Cytotoxic CD8+ T cells can kill infected cells, notably in the context of viral infections. In addition, certain T cell deficiencies predispose affected individuals to *Pneumocystis jiroveci* lung infections early in life and to chronic gut/biliary duct/liver infections by *Cryptosporidium* and related genera later on in life.

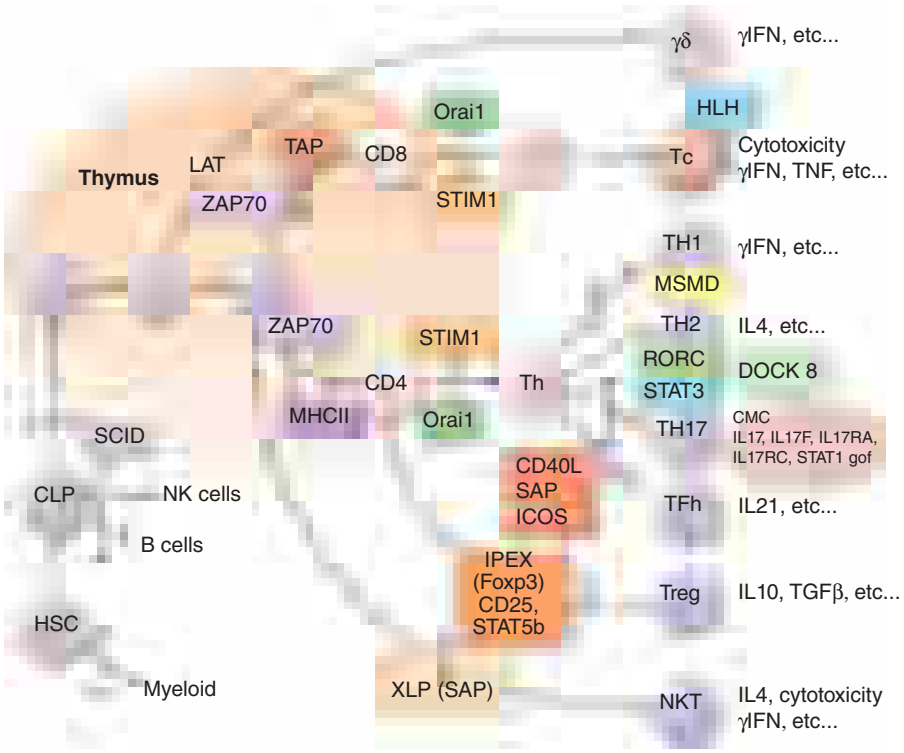


FIGURE 344-2 T cell differentiation, effector pathways, and related primary immunodeficiencies (PIDs). Hematopoietic stem cells (HSCs) differentiate into common lymphoid progenitors (CLPs), which, in turn, give rise to the T cell precursors that migrate to the thymus. The development of CD4+ and CD8+ T cells is shown. Known T cell effector pathways are indicated, that is, γδ cells, cytotoxic T cells (Tc), T_H1, T_H2, T_H17, TFh (follicular helper) CD4 effector T cells, regulatory T cells (Treg), and natural killer T cells (NKTs); abbreviations for PIDs are contained in boxes. Vertical bars indicate a complete deficiency; broken bars a partial deficiency. DOCK8, autosomal recessive form of hyper-IgE syndrome; HLH, hematopoietic lymphohistiocytosis; IL17F, IL17RA, STAT1 (gof: gain of function), CMC (chronic mucocutaneous candidiasis), CD40L, ICOS, SAP deficiencies; IPEX, immunodysregulation polyendocrinopathy enteropathy X-linked syndrome; LAT, linker for activation of T cells; MHCII, major histocompatibility complex class II deficiency; MSMD, Mendelian susceptibility to mycobacterial disease; Orai1, STIM1 deficiencies; RORC, RAR related orphan receptor C; SCID, severe combined immunodeficiency; STAT3, autosomal dominant form of hyper-IgE syndrome; TAP, TAP1 and TAP2 deficiencies; XLP, X-linked proliferative syndromes; ZAP70, zeta-associated protein deficiency.

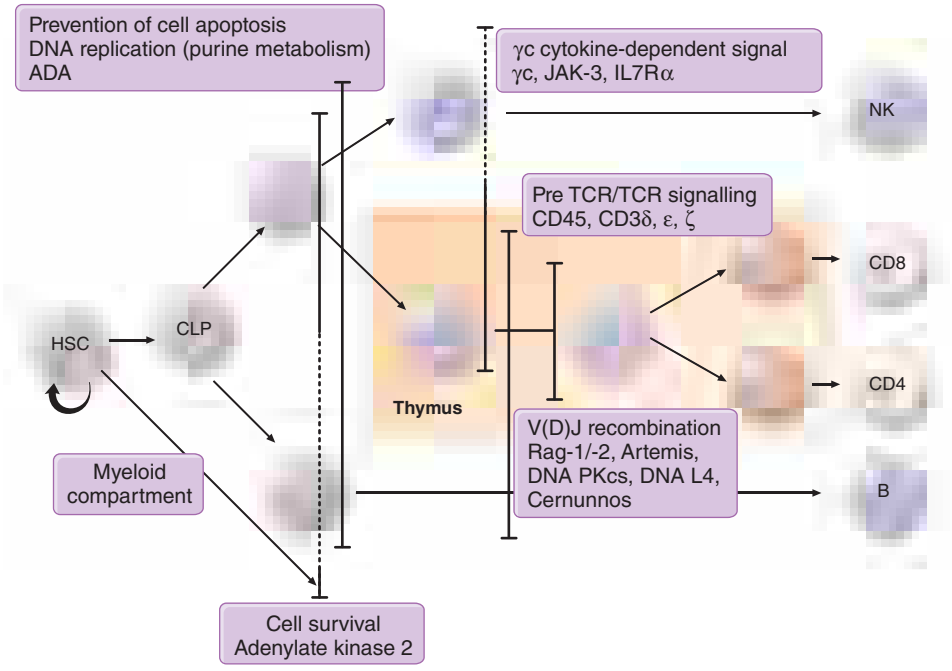


FIGURE 344-3 T cell differentiation and severe combined immunodeficiencies (SCIDs). The vertical bars indicate the five mechanisms currently known to lead to SCID. The names of deficient proteins are indicated in the boxes adjacent to the vertical bars. A broken line means that deficiency is partial or involves only some of the indicated immunodeficiencies. ADA, adenosine deaminase deficiency; CLPs, common lymphoid progenitors; DNAL4, DNA ligase 4; HSCs, hematopoietic stem cells; NKs, natural killer cells; TCR, T cell receptor.

Lastly, naturally occurring or induced regulatory T cells are essential for controlling inflammation (notably reactivity to commensal bacteria in the gut) and autoimmunity. The role of other T cell subsets with limited T cell receptor (TCR) diversity (such as γδTCR T cells or natural killer T [NKT] cells) in PIDs is less well known; however, these subsets can be defective in certain PIDs, and this finding can sometimes contribute to the diagnosis (e.g., NKT cell deficiency in X-linked proliferative syndrome [XLP]). T cell deficiencies account for ~20% of all cases of PID.

Severe Combined Immunodeficiencies SCIDs constitute a group of rare PIDs characterized by a profound block in T cell development and thus the complete absence of these cells. The developmental block is always the consequence of an intrinsic deficiency. The incidence of SCID is estimated to be 1 in 50,000 live births. Given the severity of the T cell deficiency, clinical consequences occur early in life (usually within 3 to 6 months of birth). The most frequent clinical manifestations are recurrent oral candidiasis, failure to thrive, and protracted diarrhea and/or acute interstitial pneumonitis caused by *Pneumocystis jiroveci* (although the latter can also be observed in the first year of life in children with B cell deficiencies). Severe viral infections or invasive bacterial infections can also occur. Patients may also experience complications related to infections caused by live vaccines (notably bacille Calmette-Guérin [BCG]) that may lead not only to local and regional infection but also to disseminated infection manifested by fever, splenomegaly, and skin and lytic bone lesions. A scaly skin eruption can be observed in a context of maternal T cell engraftment (see below). A diagnosis of SCID can be suspected based on the patient's clinical history and, possibly, a family history of deaths in very young children (suggestive of either X-linked or recessive inheritance). Lymphocytopenia is strongly suggestive of SCID in >90% of cases (Table 344-2). The absence of a thymic shadow on a chest x-ray can also be suggestive of SCID. An accurate diagnosis relies on precise determination of the number of circulating T, B, and NK lymphocytes and their subsets. T cell lymphopenia may be masked in some patients by the presence of maternal T cells (derived from maternal-fetal blood transfers) that cannot be eliminated. Although counts are usually low (<500/μL of blood), higher maternal T cell counts may, under some circumstances, initially mask the presence of SCID. Thus, screening for maternal cells by using adequate genetic markers should be performed whenever necessary. Inheritance pattern analysis and lymphocyte phenotyping can discriminate between various forms of SCID and provide guidance in the choice of accurate molecular diagnostic tests (see below). To date, five distinct causative

mechanisms for SCID (Fig. 344-3) have been identified. T cell quantification of receptor excision circles (TREC) by using the Guthrie card is a reliable diagnostic test for newborn screening. It is now operational in most of the United States and is being evaluated elsewhere. Its more widespread use will lead to the provision of therapy (see below) to uninfected patients resulting in a maximal chance of cure.

SEVERE COMBINED IMMUNODEFICIENCY CAUSED BY A CYTOKINE-SIGNALING DEFICIENCY The most frequent SCID phenotype (accounting for 40–50% of all cases) is the absence of both T and NK cells. This outcome results from a deficiency in either the common γ chain (γc) receptor that is shared by several cytokine receptors (the IL-2, 4, 7, 9, 15, and 21 receptors) or Jak-associated kinase (JAK) 3 that binds to the cytoplasmic portion of the γc chain receptor and induces signal transduction following cytokine binding. The former form of SCID (γc deficiency) has an X-linked inheritance mode, while the second is autosomal recessive. A lack of the IL-7Rα chain (which, together with γc, forms the IL-7 receptor) induces a selective T cell deficiency.

PURINE METABOLISM DEFICIENCY Ten to 20% of SCID patients exhibit a deficiency in adenosine deaminase (ADA), an enzyme of purine metabolism that deaminates adenosine (ado) and deoxyadenosine (dAdo). An ADA deficiency results in the accumulation of ado and dAdo metabolites that induce premature cell death of lymphocyte progenitors. The condition results in the absence of B and NK lymphocytes as well as T cells. The clinical expression of complete ADA deficiency typically occurs very early in life. Since ADA is a ubiquitous enzyme, its deficiency can also cause bone dysplasia with abnormal costochondral junctions and metaphyses (found in 50% of cases) and neurologic defects. The very rare purine nucleoside phosphorylase (PNP) deficiency causes a profound although incomplete T cell deficiency that is often associated with severe neurologic impairments.

DEFECTIVE REARRANGEMENTS OF T AND B CELL RECEPTORS A series of SCID conditions are characterized by a selective deficiency in T and B lymphocytes with autosomal recessive inheritance. These conditions account for 20–30% of SCID cases and result from mutations in genes encoding proteins that mediate the recombination of V(D)J gene elements in T and B cell antigen receptor genes (required for the generation of diversity in antigen recognition). The main deficiencies involve RAG1, RAG2, DNA-dependent protein kinase, and Artemis. A less severe (albeit variable) immunologic phenotype can result from other deficiencies in the same pathway, that is, DNA ligase 4 and Cernunnos deficiencies. Given that these latter factors are involved in DNA repair, these deficiencies also cause develop-mental defects.

DEFECTIVE (PRE-)T CELL RECEPTOR SIGNALING IN THE THYMUS A selective T cell defect can be caused by a series of rare deficiencies in molecules involved in signaling via the pre-TCR or the TCR. These include deficiencies in CD3 subunits associated with the (pre-)TCR (i.e., CD3δ, ε, and ζ) and CD45.

RETICULAR DYSGENESIS Reticular dysgenesis is an extremely rare form of SCID that causes T and NK deficiencies with severe neutropenia and sensorineural deafness. It results from an adenylate kinase 2 deficiency.

Patients with SCID require appropriate care with aggressive anti-infective therapies, immunoglobulin replacement, and (when necessary) parenteral nutrition support. In most cases, curative treatment relies on HSCT. Today, HSCT provides a very high curative potential for SCID patients who are otherwise in reasonably good condition. In this regard, neonatal screening, based on quantification of TRECs on a Guthrie card sample, is being developed. Gene therapy has been found to be successful for cases of X-linked SCID (γc deficiency) and SCID caused by an ADA deficiency, although toxicity has become an issue in the treatment of the former disease that may now be overcome by use of newly generated vectors. Lastly, a third option for the treatment of ADA deficiency consists of enzyme substitution with a pegylated enzyme.

Thymic Defects A profound T cell defect can also result from faulty development of the thymus, as is most often observed in rare cases of DiGeorge's syndrome—a relatively common condition leading to a constellation of developmental defects. In ~1% of such cases, the thymus is completely absent, leading to virtually no mature T cells. However, expansion of oligoclonal T cells can occur and is associated with skin lesions. Diagnosis (using immunofluorescence in situ hybridization) is based on the identification of a hemizygous deletion in the long arm of chromosome 22. To recover the capability for T cell differentiation, these cases require a thymic graft. CHARGE (*c*oloboma of the eye, *h*eart anomaly, choanal *a*tresia, *r*etardation, *g*enital, and *e*ar anomalies) syndrome (CHD7 deficiency) is a less frequent cause of impaired thymus development. Lastly, the very rare "nude" defect is characterized by the absence of both hair and the thymus.

Omenn Syndrome *Omenn syndrome* consists of a subset of T cell deficiencies that present with a unique phenotype, including early-onset erythrodermia, alopecia, hepatosplenomegaly, and failure to thrive. These patients usually display T cell lymphocytosis, eosinophilia, and low B cell counts. It has been found that the T cells of these patients exhibit a low TCR heterogeneity. This peculiar syndrome is the consequence of hypomorphic mutations in genes usually associated with SCID, that is, *RAG-1*, *RAG-2*, or (less frequently) *ARTEMIS* or *IL-7Rα*. The impaired homeostasis of differentiating T cells thus causes this immune system–associated disease. These patients are very fragile, requiring simultaneous anti-infective therapy, nutritional support, and immunosuppression. HSCT provides a curative approach.

Functional T Cell Defects (Fig. 344-2) A subset of T cell PIDs with autosomal inheritance is characterized by partially preserved T cell differentiation but defective activation resulting in abnormal effector function. There are many causes of these defects, but all lead to susceptibility to viral and opportunistic infections, chronic diarrhea, and failure to thrive, with onset during childhood. Careful phenotyping and in vitro functional assays are required to identify these diseases, the best characterized of which are the following.

ZETA-ASSOCIATED PROTEIN 70 (ZAP70) DEFICIENCY Zeta-associated protein 70 (ZAP70) is recruited to the TCR following antigen recognition. A ZAP70 deficiency leads typically to an almost complete absence of CD8+ T cells; CD4+ T cells are present but cannot be activated in vitro by TCR stimulation.

CALCIUM SIGNALING DEFECTS A small number of patients have been reported who exhibit a profound defect in in vitro T and B cell activation as a result of defective antigen receptor-mediated Ca^{2+} influx. This defect is caused by a mutation in the calcium channel gene (*ORAI1*) or its activator (*STIM-1*). It is noteworthy that these patients are also prone to autoimmune manifestations (blood cytopenias) and exhibit a nonprogressive muscle disease.

HUMAN LEUKOCYTE ANTIGEN (HLA) CLASS II DEFICIENCY Defective expression of HLA class II molecules is the hallmark of a group of four recessive genetic defects all of which affect molecules (RFX5, RFXAP, RFXANK, and CIITA) involved in the transactivation of the genes coding for HLA class II. As a result, low but variable CD4+ T cell counts are observed in addition to defective antigen-specific T and B cell responses. These patients are particularly susceptible to herpesvirus, adenovirus, and enterovirus infections and chronic gut/liver *Cryptosporidium* infections.

HLA CLASS I DEFICIENCY Defective expression of molecules involved in antigen presentation by HLA class I molecules (i.e., TAP-1, TAP-2, and Tapasin) leads to reduced CD8+ T cell counts, loss of HLA class I antigen expression, and a particular phenotype consisting of chronic obstructive pulmonary disease and severe vasculitis.

OTHER DEFECTS A variety of other T cell PIDs have been described, some of which are associated with a precise molecular defect (e.g., IL-2-inducible T cell kinase [ITK] deficiency, IL-21 and IL21 receptor deficiencies, CARD11 deficiency, DOCK2 deficiency, RORC deficiency). These conditions are also characterized by profound vulnerability to infections, such as severe Epstein-Barr virus (EBV)–induced B cell proliferation and autoimmune disorders in ITK deficiency. Milder phenotypes are associated with CD8 and CD3γ deficiencies.

HSCT is indicated for most of these diseases, although the prognosis is worse than in SCID because many patients are chronically infected at the time of diagnosis. Fairly aggressive immunosuppression and myeloablation may be necessary to achieve engraftment of allogeneic stem cells.

T Cell Primary Immunodeficiencies with DNA Repair Defects This is a group of PIDs characterized by a combination of T and B cell defects of variable intensity, together with a number of nonimmunologic features resulting from DNA fragility. The autosomal recessive disorder *ataxia-telangiectasia* (AT) is the most frequently encountered condition in this group. It has an incidence of 1:40,000 live births and causes B cell defects (low IgA, IgG2 deficiency, and low antibody production), which often require immunoglobulin replacement. AT is associated with a progressive T cell immunodeficiency. As the name suggests, the hallmark features of AT are telangiectasia and cerebellar ataxia. The latter manifestations may not be detectable before the age of 3–4 years, so that AT should be considered in young children with IgA deficiency and recurrent and problematic infections. Diagnosis is based on a cytogenetic analysis showing excessive chromosomal rearrangements (mostly affecting chromosomes 7 and 14) in lymphocytes. AT is caused by a mutation in the gene encoding the ATM protein—a kinase that plays an important role in the detection and repair of DNA lesions (or cell death if the lesions are too numerous) by triggering several different pathways. Overall, AT is a progressive disease that carries a very high risk of lymphoma, leukemia, and (during adulthood) carcinomas. A variant of AT ("AT-like disease") is caused by mutation in the *MRE11* gene.

Nijmegen breakage syndrome (NBS) is a less common condition that also results from chromosome instability (with the same cytogenetic abnormalities as in AT). NBS is characterized by a severe T and B cell combined immune deficiency with autosomal recessive inheritance. Individuals with NBS exhibit microcephaly and a bird-like face, but have neither ataxia nor telangiectasia. The risk of malignancies is very high. NBS results from a deficiency in nibrin (NBSI, a protein associated with MRE11 and Rad50 that is involved in checking DNA lesions) caused by hypomorphic mutations.

Severe forms of *dyskeratosis congenita* (also known as Hoyeraal-Hreidarsson syndrome) combine a progressive immunodeficiency that can also include an absence of B and NK lymphocytes, progressive bone marrow failure, microcephaly, in utero growth retardation, and gastrointestinal disease. The disease can be X-linked or, more rarely, autosomal recessive. It is caused by the mutation of genes encoding telomere maintenance proteins, including dyskerin (DKC1).

Finally, *immunodeficiency with centromeric and facial anomalies* (ICF) is a complex syndrome of autosomal recessive inheritance that variably combines a mild T cell immune deficiency with a more severe B cell immune deficiency, coarse face, digestive disease, and mild mental retardation. A diagnostic feature is the detection by cytogenetic analysis of multiradial aspects in multiple chromosomes (most frequently 1, 9, and 16) corresponding to an abnormal DNA structure secondary to defective DNA methylation. It is the consequence of a deficiency in most cases in the DNA methyltransferase DNMT3B, ZBTB24, CDCA7, or HELLS.

T Cell Primary Immunodeficiencies with Hyper-IgE Several T cell PIDs are associated with elevated serum IgE levels (as in Omenn syndrome). A condition sometimes referred to as *autosomal recessive hyper-IgE syndrome* is notably characterized by recurrent bacterial infections in the skin and respiratory tract and severe skin and mucosal infections by pox viruses and human papillomaviruses, together with severe allergic manifestations. T and B lymphocyte counts are low. Mutations in the *DOCK8* gene have been found in many of these patients. This condition is an indication for HSCT.

A very rare, related condition with autosomal recessive inheritance that causes a similar susceptibility to infection with various microbes (see above), including mycobacteria, reportedly results from a deficiency in Tyk-2, a JAK family kinase involved in the signaling of many different cytokine receptors.

Autosomal Dominant Hyper-IgE Syndrome This unique condition, the *autosomal dominant hyper-IgE syndrome*, is usually diagnosed by the combination of recurrent skin and lung infections that can be complicated by pneumatoceles. Infections are caused by pyogenic bacteria and fungi. Several other manifestations characterize hyper-IgE syndrome, including facial dysmorphy, defective loss of primary teeth, hyperextensibility, scoliosis, and osteoporosis. Elevated serum IgE levels are typical of this syndrome. Defective T_H17 effector responses have been shown to account at least in part for the specific patterns of susceptibility to particular microbes. This condition is caused by a heterozygous (dominant) mutation in the gene encoding the transcription factor STAT3 that is required in a number of signaling pathways following binding of cytokine to cytokine receptors (such as that of IL-6 and the IL-6 receptor). It also results in partially defective antibody production because of defective IL-21R signaling. Hence, immunoglobulin substitution can be considered as prophylaxis of bacterial infections.

Cartilage Hair Hypoplasia The autosomal recessive *cartilage hair hypoplasia* (CHH) disease is characterized by short-limb dwarfism, metaphyseal dysostosis, and sparse hair, together with a combined T and B cell PID of extremely variable intensity (ranging from quasi-SCID to no clinically significant immune defects). The condition can predispose to erythroblastopenia, autoimmunity, and tumors. It is caused by mutations in the *RMRP* gene for a noncoding ribosome-associated RNA.

CD40 Ligand and CD40 Deficiencies *Hyper-IgM syndrome* (HIGM) is a well-known PID that is usually classified as a B cell immune deficiency (see **Fig. 344-4** and below). It results from defective immunoglobulin class switch recombination (CSR) in germinal centers and leads to profound deficiency in production of IgG, IgA, and IgE (although IgM production is maintained). Approximately half of HIGM sufferers are also prone to opportunistic infections, for example, interstitial pneumonitis caused by *Pneumocystis jiroveci* (in young children), protracted diarrhea and cholangitis caused by *Cryptosporidium*, and infection of the brain with *Toxoplasma gondii*.

In the majority of cases, this condition has an X-linked inheritance and is caused by a deficiency in CD40 ligand (L). CD40L induces signaling events in B cells that are necessary for both CSR and adequate activation of other CD40-expressing cells that are involved in innate immune responses against the above-mentioned microorganisms. More rarely, the condition is caused by a deficiency in CD40 itself. The poorer prognosis of CD40L and CD40 deficiencies (relative to most other HIGM conditions) implies that (1) thorough investigations have to be performed in all cases of HIGM and (2) potentially curative HSCT should be discussed on a case-by-case basis for this group of patients.

Wiskott-Aldrich Syndrome WAS is a complex, recessive, X-linked disease with an incidence of ~1 in 200,000 live births. It is caused by mutations in the *WASP* gene that affect not only T lymphocytes but also the other lymphocyte subsets, dendritic cells, and platelets. WAS is typically characterized by the following clinical manifestations: recurrent bacterial infections, eczema, and bleeding caused by thrombocytopenia. However, these manifestations are highly variable—mostly as a consequence of the many different *WASP* mutations that have been observed. Null mutations predispose affected individuals to invasive and bronchopulmonary infections, viral infections, severe eczema, and autoimmune manifestations. The latter include autoantibody-mediated blood cytopenia, glomerulonephritis, skin and visceral vasculitis (including brain vasculitis), erythema nodosum, and arthritis. Another possible consequence of WAS is lymphoma, which may be virally induced (e.g., by EBV or Kaposi's sarcoma–associated herpesvirus). Thrombocytopenia can be severe and compounded by the peripheral destruction of platelets associated with autoimmune disorders. Hypomorphic mutations usually lead to milder outcomes that are generally limited to thrombocytopenia. It is noteworthy that even patients with "isolated" X-linked thrombocytopenia can develop severe autoimmune disease or lymphoma later in life. The immunologic workup is not very informative; there can be a relative CD8+ T cell deficiency, frequently accompanied by low

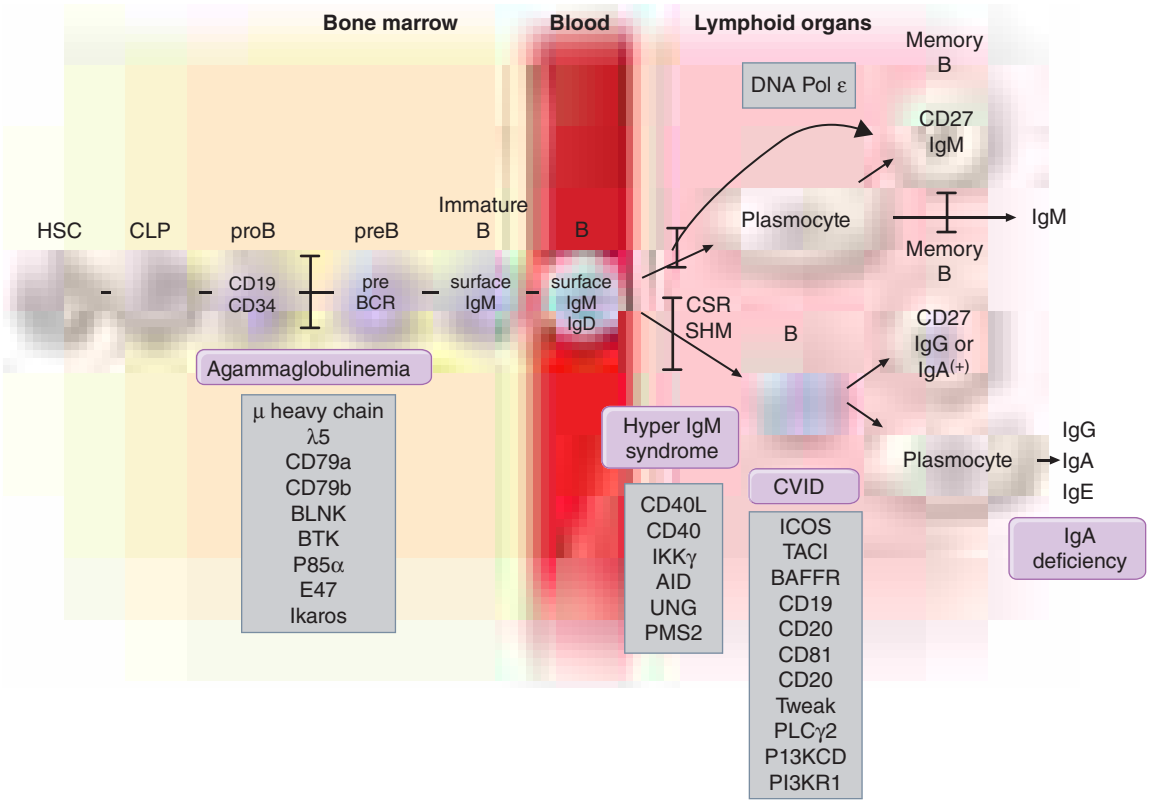


FIGURE 344-4 B cell differentiation and related primary immunodeficiencies (PIDs). Hematopoietic stem cells (HSCs) differentiate into common lymphoid progenitors (CLPs), which give rise to pre-B cells. The B cell differentiation pathway goes through the pre–B cell stage (expression of the μ heavy chain and surrogate light chain), the immature B cell stage (expression of surface IgM), and the mature B cell stage (expression of surface IgM and IgD). The main phenotypic characteristics of these cells are indicated. In lymphoid organs, B cells can differentiate into plasma cells and produce IgM or undergo (in germinal centers) Ig class switch recombination (CSR) and somatic mutation of the variable region of V genes (SHM) that enable selection of high-affinity antibodies. These B cells produce antibodies of various isotypes and generate memory B cells. PIDs are indicated in the purple boxes. CVID, common variable immunodeficiency.

serum IgM levels and decreased antigen-specific antibody responses. A typical feature is reduced-sized platelets on a blood smear. Diagnosis is based on intracellular immunofluorescence analysis of WAS protein (WASp) expression in blood cells. WASp regulates the actin cytoskeleton and thus plays an important role in many lymphocyte functions, including cell adhesion and migration and the formation of synapses between antigen-presenting and target cells. Predisposition to autoimmune disorders is in part related to defective regulatory T cells. The treatment of WAS should match the severity of disease expression. Prophylactic antibiotics, immunoglobulin G (IgG) supplementation, and careful topical treatment of eczema are indicated. Although splenectomy improves platelet count in a majority of cases, this intervention is associated with a significant risk of infection (both before and after HSCT). Allogeneic HSCT is curative, with fairly good results overall. Gene therapy trials are also under way. A similar condition has been reported in a girl with a deficiency in the Wiskott-Aldrich interacting protein (WIP).

A few other complex PIDs are worth mentioning. *Sp110 deficiency* causes a T cell PID with liver venoocclusive disease and hypogammaglobulinemia. *Chronic mucocutaneous candidiasis* (CMC) is a heterogeneous disease, considering the different inheritance patterns that have been observed. In some cases, chronic candidiasis is associated with late-onset bronchopulmonary infections, bronchiectasis, and brain aneurysms. Moderate forms of CMC are related to autoimmunity and AIRE deficiency (see below). In this setting, predisposition to *Candida* infection is associated with the detection of autoantibodies to T_H17 cytokines. Recently, deficiencies in IL-17A, IL-17F, and IL-17 receptor A and C and in the associated protein Act1, and above all, gain-of-function mutations in *STAT1* have been found to be associated with CMC. In all cases, CMC is related to defective T_H17 function. Innate immunodeficiency in CARD9 also predisposes to chronic invasive fungal infection.

B LYMPHOCYTE DEFICIENCIES (TABLE 344-1, FIG. 344-4)

Deficiencies that predominantly affect B lymphocytes are the most frequent PIDs and account for 60–70% of all cases. B lymphocytes make antibodies. Pentameric IgMs are found in the vascular compartment and are also secreted at mucosal surfaces. IgG antibodies diffuse freely into extravascular spaces, whereas IgA antibodies are produced and secreted predominantly from mucosa-associated lymphoid tissues. Although Ig isotypes have distinct effector functions, including Fc receptor–mediated and (indirectly) C_3 receptor–dependent phagocytosis of microorganisms, they share the ability to recognize and neutralize a given pathogen. Defective antibody production therefore allows the establishment of invasive, pyogenic bacterial infections as well as recurrent sinus and pulmonary infections (mostly caused by *Streptococcus pneumoniae*, *Haemophilus influenzae*, *Moraxella catarrhalis*, and, less frequently, gram-negative bacteria). If left untreated, recurrent bronchial infections lead to bronchiectasis and, ultimately, cor pulmonale and death. Parasitic infections such as caused by *Giardia lamblia* and bacterial infections caused by *Helicobacter* and *Campylobacter* of the gut are also observed. A complete lack of antibody production (namely agammaglobulinemia) can also predispose affected individuals to severe, chronic, disseminated enteroviral infections causing meningoencephalitis, hepatitis, and a dermatomyositis-like disease.

Even with the most profound of B cell deficiencies, infections rarely occur before the age of 6 months; this is because of transient protection provided by the transplacental passage of immunoglobulins during the last trimester of pregnancy. Conversely, a genetically nonimmunodeficient child born to a mother with hypogammaglobulinemia is, in the absence of maternal Ig substitution, usually prone to severe bacterial infections in utero and for several months after birth.

Diagnosis of B cell PIDs relies on the determination of serum Ig levels (Table 344-2). Determination of antibody production following

immunization with tetanus toxoid vaccine or nonconjugated pneumococcal polysaccharide antigens can also help diagnose more subtle deficiencies. Another useful test is B cell phenotype determination in switched μ−δ− CD27+ and nonswitched memory B cells (μ+δ+ CD27+). In agammaglobulinemic patients, examination of bone marrow B cell precursors (Fig. 344-4) can help obtain a precise diagnosis and guide the choice of genetic tests.

Agammaglobulinemia Agammaglobulinemia is characterized by a profound defect in B cell development (<1% of the normal B cell blood count). In most patients, very low residual Ig isotypes can be detected in the serum. In 85% of cases, agammaglobulinemia is caused by a mutation in the *BTK* gene that is located on the X chromosome. The *BTK* gene product is a kinase that participates in (pre) B cell receptor signaling. When the kinase is defective, there is a block (albeit a leaky one) at the pre-B to B cell stage (Fig. 344-4). Detection of *BTK* by intracellular immunofluorescence of monocytes, and lack thereof in patients with X-linked agammaglobulinemia (XLA), is a useful diagnostic test. Not all of the mutations in *BTK* result in agammaglobulinemia, since some patients have a milder form of hypogammaglobulinemia and low but detectable B cell counts. These cases should not be confused with common variable immunodeficiency (CVID, see below). About 10% of agammaglobulinemia cases are caused by alterations in genes encoding elements of the pre-B cell receptor, i.e., the μ heavy chain, the λ5 surrogate light chain, Igα or Igβ, the scaffold protein BLNK, and the p85 α subunit of phosphatidylinositol 3 phosphate kinase (P13K) and the Ikaros transcription factor. In 5% of cases, the defect is unknown. It is noteworthy that agammaglobulinemia can be observed in patients with ICF syndrome, despite the presence of normal peripheral B cell counts. Lastly, agammaglobulinemia can be a manifestation of a myelodysplastic syndrome (associated or not with neutropenia). Treatment of agammaglobulinemic patients is based on immunoglobulin replacement (see below). Profound hypogammaglobulinemia is also observed in adults, in association with thymoma.

Hyper-IgM (HIGM) Syndromes *HIGM* is a rare B cell PID characterized by defective Ig CSR. It results in very low serum levels of IgG and IgA and elevated or normal serum IgM levels. The clinical severity is similar to that seen in agammaglobulinemia, although chronic lung disease and sinusitis are less frequent and enteroviral infections are uncommon. As discussed above, a diagnosis of HIGM involves screening for an X-linked CD40L deficiency and an autosomal recessive CD40 deficiency, which affect both B and T cells. In 50% of cases affecting only B cells, these isolated HIGM syndromes result from mutations in the gene encoding activation-induced deaminase, the protein that induces CSR in B cell germinal centers. These patients usually have enlarged lymphoid organs. In the other 50% of cases, the etiology is unknown (except for rare UNG and PMS2 deficiencies). Furthermore, IgM-mediated autoimmunity and lymphomas can occur in HIGM syndrome. It is noteworthy that HIGM can result from fetal rubella syndrome or can be a predominant immunologic feature of other PIDs, such as the immunodeficiency associated with ectodermic anhydrotic hypoplasia X-linked NEMO deficiency and the combined T and B cell PIDs caused by DNA repair defects such as AT and Cernunnos deficiency.

Common Variable Immunodeficiency CVID is an ill-defined condition characterized by low serum levels of one or more Ig isotypes. Its prevalence is estimated to be 1 in 20,000. The condition is recognized predominantly in adults, although clinical manifestations can occur earlier in life. Hypogammaglobulinemia is associated with at least partially defective antibody production in response to vaccine antigens. B lymphocyte counts are often normal but can be low. Besides infections, CVID patients may develop lymphoproliferation (splenomegaly), granulomatous lesions, colitis, antibody-mediated autoimmune disease, and lymphomas. A family history is found in 10% of cases. A clear-cut dominant inheritance pattern is found in some families, whereas recessive inheritance is observed more rarely. In most cases, no molecular cause can be identified. A small number of patients in Germany were found to carry mutations in the *ICOS* gene encoding a T cell-membrane protein that contributes to B cell activation and survival. In 10% of patients with CVID, monoallelic or biallelic mutations of the gene encoding TACI (a member of the tumor necrosis factor [TNF] receptor family that is expressed on B cells) have been found. In fact, heterozygous TACI mutations correspond to a genetic susceptibility factor, since similar heterozygous mutations are found in 1% of controls. The B-cell activating factor (BAFF) receptor was found to be defective in a kindred with CVID, although not all individuals carrying the mutation have CVID. Recently a group of patients with hypogammaglobulinemia and lymphoproliferation was shown to exhibit dominant gain of function mutations in the *PIK3CD* gene encoding the p110δ form of P13 kinase or in the *PI3KR1* gene encoding the regulatory p85α subunit of PI3 kinase. Rare cases of hypogammaglobulinemia were found to be associated with CD19 and CD81 deficiencies. These patients have B cells that can be identified by typing for other B cell markers.

A diagnosis of CVID should be made after excluding the presence of hypomorphic mutations associated with agammaglobulinemia or more subtle T cell defects; this is particularly the case in children. It is possible that many cases of CVID result from a constellation of factors, rather than a single genetic defect. Hypogammaglobulinemia can be associated with neutropenia and lymphopenia in the warts, hypogammaglobulinemia, infections, and myelokathexis syndrome (WHIM) caused by dominant gain-of-function mutation of *CXCR4*, resulting in cell retention in the bone marrow.

Selective Ig Isotype Deficiencies *IgA deficiency* and CVID represent polar ends of a clinical spectrum due to the same underlying gene defect(s) in a large subset of these patients. IgA deficiency is the most common PID; it can be found in 1 in every 600 individuals. It is asymptomatic in most cases; however, individuals may present with increased numbers of acute and chronic respiratory infections that may lead to bronchiectasis. In addition, over their lifetime, these patients experience an increased susceptibility to drug allergies, atopic disorders, and autoimmune diseases. Symptomatic IgA deficiency is probably related to CVID, since it can be found in relatives of patients with CVID. Furthermore, IgA deficiency may progress to CVID. It is thus important to assess serum Ig levels in IgA-deficient patients (especially when infections occur frequently) in order to detect changes that should prompt the initiation of immunoglobulin replacement. Selective IgG2 (+G4) deficiency (which in some cases may be associated with IgA deficiency) can also result in recurrent sinopulmonary infections and should thus be specifically sought in this clinical setting. These conditions are ill-defined and often transient during childhood. A pathophysiologic explanation has not been found.

Selective Antibody Deficiency to Polysaccharide Antigens Some patients with normal serum Ig levels are prone to *S. pneumoniae* and *H. influenzae* infections of the respiratory tract. Defective production of antibodies against polysaccharide antigens (such as those in the *S. pneumoniae* cell wall) can be observed and is probably causative. This condition may correspond to a defect in marginal zone B cells, a B cell subpopulation involved in T-independent antibody responses.

Immunoglobulin Replacement IgG antibodies have a half-life of 21–28 days. Thus, injection of plasma-derived polyclonal IgG containing a myriad of high-affinity antibodies can provide protection against disease-causing microorganisms in patients with defective IgG antibody production. This form of therapy should not be based on laboratory data alone (i.e., IgG and/or antibody deficiency) but should be guided by the occurrence or not of infections; otherwise, patients might be subjected to unjustified IgG infusions. Immunoglobulin replacement can be performed by IV or subcutaneous routes. In the former case, injections have to be repeated every 3–4 weeks, with a residual target level of 800 mg/mL in patients who had very low IgG levels prior to therapy. Subcutaneous injections are typically performed once a week, although the frequency can be adjusted on a case-by-case basis. A trough level of 800 mg/mL is desirable. Whatever the mode of administration, the main goal is to reduce the frequency of the respiratory

tract infections and prevent chronic lung and sinus disease. The two routes appear to be equally safe and efficacious, and so the choice should be left to the preference of the patient.

In patients with chronic lung disease, chest physical therapy with good pulmonary toilet and the cyclic use of antibiotics are also needed. Immunoglobulin replacement is well tolerated by most patients, although the selection of the best-tolerated Ig preparation may be necessary in certain cases. Since IgG preparations contain a small proportion of IgAs, caution should be taken in patients with residual antibody production capacity and a complete IgA deficiency, as these subjects may develop anti-IgA antibodies that can trigger anaphylactic shock. These patients should be treated with IgA-free IgG preparations. Immunoglobulin replacement is a lifelong therapy; its rationale and procedures have to be fully understood and mastered by the patient and his or her family in order to guarantee the strict observance required for efficacy.

PRIMARY IMMUNODEFICIENCIES AFFECTING REGULATORY PATHWAYS (TABLE 344-1)

An increasing number of PIDs have been found to cause homeostatic dysregulation of the immune system, either alone or in association with increased vulnerability to infections. Defects of this type affecting the innate immune system and autoinflammatory syndromes will not be covered in this chapter. However, three specific entities (hemophagocytic lymphohistiocytosis [HLH], lymphoproliferation, and autoimmunity) will be described below.

HEMOPHAGOCYTIC LYMPHOHISTIOCYTOSIS

HLH is characterized by an unremitting activation of CD8+ T lymphocytes and macrophages that leads to organ damage (notably in the liver, bone marrow, and central nervous system). This syndrome results from a broad set of inherited diseases, most of which impair T and NK lymphocyte cytotoxicity. The manifestations of HLH are often induced by a viral infection. EBV is the most frequent trigger. In severe forms of HLH, disease onset may start during the first year of life or even (in rare cases) at birth.

Diagnosis relies on the identification of the characteristic symptoms of HLH (fever, hepatosplenomegaly, edema, neurologic diseases, blood cytopenia, increased liver enzymes, hypofibrinogenemia, high triglyceride levels, elevated markers of T cell activation, and hemophagocytic features in the bone marrow or cerebrospinal fluid). Functional assays of postactivation cytotoxic granule exocytosis (CD107 fluorescence at the cell membrane) can suggest genetically determined HLH. The conditions can be classified into three subsets:

1. Familial HLH with autosomal recessive inheritance, including perforin deficiency (30% of cases) that can be recognized by assessing intracellular perforin expression; Munc13-4 deficiency (30% of cases); syntaxin 11 deficiency (10% of cases); Munc18-2 deficiency (20% of cases); and a few residual cases that lack a known molecular defect.
2. HLH with partial albinism. Three conditions combine HLH and abnormal pigmentation, where hair examination can help in the diagnosis: Chédiak-Higashi syndrome, Griscelli syndrome, and Hermansky-Pudlak syndrome type II. Chédiak-Higashi syndrome is also characterized by the presence of giant lysosomes within leukocytes **(Chap. 60)**, in addition to a primary neurologic disorder with slow progression of symptoms over time.
3. XLP is characterized in most patients by the induction of HLH following EBV infection, while other patients develop progressive hypogammaglobulinemia similar to what is observed in CVID and/or certain lymphomas. XLP is caused by a mutation in the *SH2DIA* gene that encodes the adaptor protein SAP (associated with a SLAM family receptor). Several immunologic abnormalities have been described, including low 2B4-mediated NK cell cytotoxicity, impaired differentiation of NKT cells, defective antigen-induced T cell death, and defective T cell helper activity for B cells. A related disorder (XLP2) has recently been described. It is also X-linked and induces HLH (frequently after EBV infection), although the clinical manifestation may be less pronounced. The condition is associated with a deficiency of the antiapoptotic molecule XIAP. The pathophysiology of XLP2 remains unclear; however, it may be related to control of inflammation in macrophages as there is a functional link between XIAP and NLRC4, an inflammasome component, in which gain of function can also induce HLH. XLP2 is also frequently associated with colitis.

HLH is a life-threatening complication. The treatment of this condition requires aggressive immunosuppression with either the cytotoxic agent etoposide or anti–T cell antibodies; specific therapy targeting interferon γ, which is critical in causing HLH, is an additional option to consider. Once remission has been achieved, HSCT should be performed, since it provides the only curative form of therapy.

AUTOIMMUNE LYMPHOPROLIFERATIVE SYNDROME

Autoimmune lymphoproliferative syndrome (ALPS) is characterized by nonmalignant T and B lymphoproliferation causing splenomegaly and enlarged lymph nodes; 70% of patients also display autoimmune manifestations such as autoimmune cytopenias, Guillain-Barré syndrome, uveitis, and hepatitis **(Chaps. 62 and 342)**. A hallmark of ALPS is the presence of CD4–CD8– TCRαβ+ T cells (2–50%) in the blood of affected individuals. Hypergammaglobulinemia involving IgG and IgA is also frequently observed. The syndrome is caused by a defect in Fas-mediated apoptosis of lymphocytes, which can thus accumulate and mediate autoimmunity. Furthermore, ALPS can lead to malignancies.

Most patients carry a heterozygous mutation in the gene encoding Fas that is characterized by dominant inheritance and variable penetrance, depending on the nature of the mutation. A rare and severe form of the disease with early onset can be observed in patients carrying a biallelic mutation of Fas, which profoundly impairs the protein's expression and/or function. Fas-ligand, caspase 10, caspase 8, and somatic neuroblastoma RAS viral oncogene homologue (NRAS) mutations have also been reported in a few cases of ALPS. Many cases of ALPS have not been precisely delineated at the molecular level. A B cell–predominant ALPS has recently been found associated with a protein kinase Cδ gene mutation. Treatment of ALPS is essentially based on the use of proapoptotic drugs, which need to be carefully administered in order to avoid toxicity.

COLITIS, AUTOIMMUNITY, AND PRIMARY IMMUNODEFICIENCIES

Several PIDs (most of which are T cell–related) can cause severe gut inflammation. The prototypic example is *immunodysregulation polyendocrinopathy enteropathy X-linked syndrome* (IPEX), characterized by a widespread inflammatory enteropathy, food intolerance, skin rashes, autoimmune cytopenias, and diabetes. The syndrome is caused by loss-of-function mutations in the gene encoding the transcription factor FOXP3, which is required for the acquisition of effector function by regulatory T cells. In most cases of IPEX, CD4+CD25+ regulatory T cells are absent from the blood. This condition has a poor prognosis and requires aggressive immunosuppression. The only possible curative approach is allogeneic HSCT. IPEX-like syndromes that lack a FOXP3 mutation have also been described. In some cases, a CD25 deficiency has been found. Defective CD25 expression also impairs regulatory cell expansion/function. This functional T cell deficiency means that CD25-deficient patients are also at increased risk of opportunistic infections. It is noteworthy that abnormalities in regulatory T cells have been described in other PID settings, such as in Omenn syndrome, STAT5b deficiency, STIM1 (Ca flux) deficiency, and WAS; these abnormalities may account (at least in part) for the occurrence of inflammation and autoimmunity. The autoimmune features observed in a small fraction of patients with DiGeorge's syndrome may have the same cause. Recently, severe inflammatory gut disease has been described in patients with a deficiency in the IL-10 receptor or IL-10.

Dominant mutations in genes encoding the regulatory molecule CTLA4, recessive mutations in the gene encoding LRBA (a molecule involved in recycling of CTLA4) as well as dominant gain of function

mutation of STAT3 cause a multifaceted lymphoproliferative and autoimmune syndrome, frequently involving inflammatory bowel disease that can be associated with hypogammaglobulinemia. Molecular diagnosis is required before adapted targeted therapies are undertaken.

A distinct autoimmune entity is observed in *autoimmune polyendocrinopathy candidiasis ectodermal dysplasia* (APECED) syndrome, which is characterized by autosomal recessive inheritance. It consists of multiple autoimmune manifestations that can affect solid organs in general and endocrine glands in particular. Mild, chronic *Candida* infection is often associated with this syndrome. The condition is due to mutations in the autoimmune regulator (*AIRE*) gene and results in impaired thymic expression of self-antigens by medullary epithelial cells and impaired negative selection of self-reactive T cells that leads to autoimmune manifestations.

A combination of hypogammaglobulinemia, autoantibody production, cold-induced urticaria or skin granulomas, or autoinflammation has been reported, and has been termed the *PLCγ2-associated antibody deficiency and immune dysregulation* (PLAID or APLAID).

CONCLUSION

The variety and complexity of the clinical manifestations of the many different PIDs strongly indicate that it is important to raise awareness of these diseases. Indeed, early diagnosis is essential for establishing an appropriate therapeutic regimen. Hence, patients with suspected PIDs must always be referred to experienced clinical centers that are able to perform appropriate molecular and genetic tests. A precise molecular diagnosis is not only necessary for initiating the most suitable treatment, but is also important for genetic counseling and prenatal diagnosis.

One pitfall that may hamper diagnosis is the high variability that is associated with many PIDs. Variable disease expression can result from the differing consequences of various mutations associated with a given condition, as exemplified by WAS and, to a lesser extent, XLA. There can also be effects of modifier genes (as also suspected in XLA) and environmental factors such as EBV infection that can be the main trigger of disease in XLP conditions. Furthermore, it has recently been established that somatic mutations in an affected gene can attenuate the phenotype of a number of T cell PIDs. This has been described for ADA deficiency, X-linked SCID, RAG deficiencies, NF-κB essential modulator (NEMO) deficiency, and, most frequently, WAS. In contrast, somatic mutations can create disease states analogous to PID, as reported for ALPS. Lastly, cytokine-neutralizing autoantibodies can mimic a PID, as shown for IFN-γ.

Many aspects of the pathophysiology of PIDs are still unknown, and the disease-causing gene mutations have not been identified in all cases (as illustrated by CVID and IgA deficiency). However, our medical understanding of PIDs has now reached the stage where scientifically based approaches to the diagnosis and treatment of these diseases can be implemented. A genetic diagnosis has become a milestone step in the care of PID patients.

FURTHER READING

Casanova JL et al: Guidelines for genetic studies in single patients: Lessons from primary immunodeficiencies. J Exp Med 211:2137, 2014.

Fischer A et al: Gene therapy for primary immunodeficiencies. Clin Genet 88:507, 2015.

Fischer A et al: Severe combined immunodeficiencies and related disorders. Nat Rev Dis Primers 1:15061, 2015.

Holland SM: Chronic granulomatous disease. Hematol Oncol Clin North Am 2013;27:89, viii.

Kwan A et al: Newborn screening for severe combined immunodeficiency in 11 screening programs in the United States. JAMA 312:729, 2014.

Notarangelo LD: Functional T cell immunodeficiencies (with T cells present). Annu Rev Immunol 31:195, 2013.

Ochs HD et al (eds): *Primary Immunodeficiencies: A Molecular and Genetic Approach*. New York, Oxford University Press, 2013.

Picard C et al: International Union of Immunological Societies: 2017 Primary Immunodeficiency Diseases Committee Report on Inborn Errors of Immunity. J Clin Immunol 38:96, 2018.

Picard C, Fischer A: Contribution of high-throughput DNA sequencing to the study of primary immunodeficiencies. Eur J Immunol 44:2854, 2014.

Resnick ES, Cunningham-Rundles C: The many faces of the clinical picture of common variable immune deficiency. Curr Opin Allergy Clin Immunol 12:595, 2012.

Section 2 Disorders of Immune-Mediated Injury

345 Urticaria, Angioedema, and Allergic Rhinitis

Katherine N. Cahill, Joshua A. Boyce

INTRODUCTION

The term *atopy* implies a tendency to manifest asthma, rhinitis, urticaria, and atopic dermatitis alone or in combination, in association with the presence of allergen-specific IgE. However, individuals without an atopic background may also develop hypersensitivity reactions, particularly urticaria and anaphylaxis, associated with the presence of IgE. Since mast cells are key effector cells in allergic rhinitis and asthma, and the dominant effector in urticaria, anaphylaxis, and systemic mastocytosis, its developmental biology, activation pathway, product profile, and target tissues will be considered in the introduction to these clinical disorders. Dysregulation of mast cell development seen in mastocytosis will be covered in a separate chapter.

The binding of IgE to human mast cells and basophils, a process termed *sensitization*, prepares these cells for subsequent antigen-specific activation. The high-affinity Fc receptor for IgE, designated FcεRI, is composed of one α, one β, and two disulfide-linked γ chains, which together cross the plasma membrane seven times. The α chain is responsible for IgE binding, and the β and γ chains provide for signal transduction that follows the aggregation of the sensitized tetrameric receptors by polymeric antigen. The binding of IgE stabilizes the α chain at the plasma membrane, thus increasing the density of FcεRI receptors at the cell surface while sensitizing the cell for effector responses. This accounts for the correlation between serum IgE levels and the numbers of FcεRI receptors detected on circulating basophils. Signal transduction is initiated through the action of a Src family–related tyrosine kinase termed Lyn that is constitutively associated with the β chain. Lyn transphosphorylates the canonical immunoreceptor tyrosine-based activation motifs (ITAMs) of the β and γ chains of the receptor, resulting in recruitment of more active Lyn to the β chain and of Syk tyrosine kinase. The phosphorylated tyrosines in the ITAMs function as binding sites for the tandem *src* homology two (SH2) domains within Syk. Syk activates not only phospholipase Cγ, which associates with the linker of activated T cells at the plasma membrane, but also phosphatidylinositol 3-kinase to provide phosphatidylinositol-3,4,5-trisphosphate, which allows membrane targeting of the Tec family kinase Btk and its activation by Lyn. In addition, the Src family tyrosine kinase Fyn becomes activated after aggregation of IgE receptors and phosphorylates the adapter protein Gab2 that enhances activation of phosphatidylinositol 3-kinase. Indeed, this additional input is essential for mast cell activation, but it can be partially inhibited by Lyn, indicating that the extent of mast cell activation is in part regulated by the interplay between these Src family kinases. Activated phospholipase Cγ cleaves phospholipid membrane substrates to provide inositol-1,4,5-trisphosphate (IP_3) and 1,2-diacylglycerols (1,2-DAGs) so as to mobilize intracellular calcium and activate protein kinase C, respectively. The subsequent opening of calcium-regulated activated channels provides the sustained elevations

of intracellular calcium required to recruit the mitogen-activated protein kinases, ERK, JNK, and p38 (serine/threonine kinases), which provide cascades to augment arachidonic acid release and to mediate nuclear translocation of transcription factors for various cytokines. The calcium ion–dependent activation of phospholipases cleaves membrane phospholipids to generate lysophospholipids, which, like 1,2-DAG, may facilitate the fusion of the secretory granule perigranular membrane with the cell membrane, a step that releases the membrane-free granules containing the preformed mast cell mediators.

The secretory granule of the human mast cell has a crystalline structure, unlike mast cells of lower species. IgE-dependent cell activation results in solubilization and swelling of the granule contents within the first minute of receptor perturbation; this reaction is followed by the ordering of intermediate filaments about the swollen granule, movement of the granule toward the cell surface, and fusion of the perigranular membrane with that of other granules and with the plasmalemma to form extracellular channels for mediator release while maintaining cell viability.

In addition to exocytosis, aggregation of FcεRI initiates two other pathways for generation of bioactive products, namely, lipid mediators and cytokines. The biochemical steps involved in expression of such cytokines as tumor necrosis factor α (TNF-α), interleukin (IL) 1, IL-6, IL-4, IL-5, IL-13, granulocyte-macrophage colony-stimulating factor (GM-CSF), and others, including an array of chemokines, have not been specifically defined for mast cells. Inhibition studies of cytokine production (IL-1β, TNF-α, and IL-6) in mouse mast cells with cyclosporine or FK506 reveal binding to the ligand-specific immunophilin and attenuation of the calcium ion- and calmodulin-dependent serine/threonine phosphatase, calcineurin.

Lipid mediator generation **(Fig. 345-1)** involves translocation of calcium ion–dependent cytosolic phospholipase A_2 to the outer nuclear membrane, with subsequent release of arachidonic acid for metabolic processing by the distinct prostanoid and leukotriene pathways. The constitutive prostaglandin endoperoxide synthase-1 (PGHS-1/cyclooxygenase-1) and the de novo inducible PGHS-2 (cyclooxygenase-2) convert released arachidonic acid to the sequential intermediates, prostaglandins G_2 and H_2. The glutathione-dependent

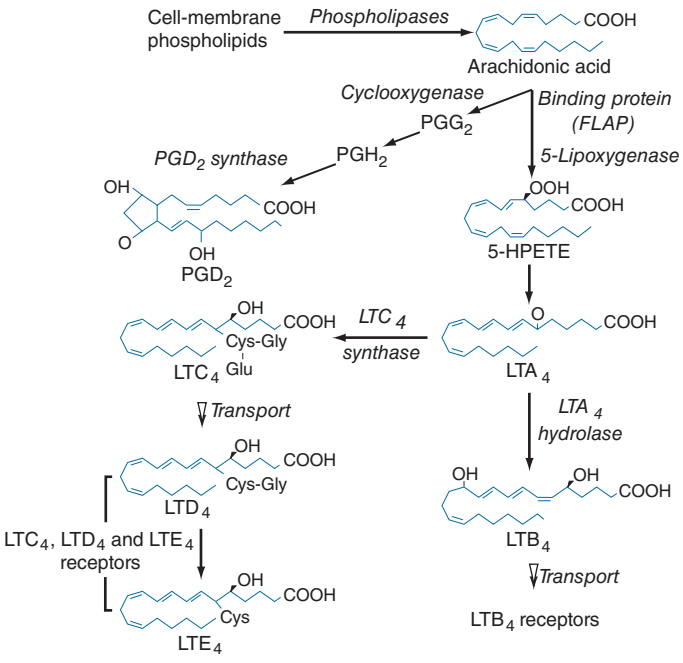


FIGURE 345-1 Pathways for biosynthesis and release of membrane-derived lipid mediators from mast cells. In the 5-lipoxygenase pathway, leukotriene A_4 (LTA_4) is the intermediate from which the terminal-pathway enzymes generate the distinct final products, leukotriene C_4 (LTC_4) and leukotriene B_4 (LTB_4), which leave the cell by separate saturable transport systems. Gamma glutamyl transpeptidase and a dipeptidase then cleave glutamic acid and glycine from LTC_4 to form LTD_4 and LTE_4, respectively. The major mast cell product of the cyclooxygenase system is PGD_2.

hematopoietic prostaglandin D_2 (PGD_2) synthase then converts PGH_2 to PGD_2, the predominant mast cell prostanoid. The PGD_2 receptor DP_1 is expressed by platelets, natural killer cells, dendritic cells, and epithelial cells, whereas DP_2 is expressed by T_H2 lymphocytes, innate lymphoid type 2 cells, eosinophils, and basophils. Mast cells also generate thromboxane A_2 (TXA_2), a short lived but powerful mediator that induces bronchoconstriction and platelet activation through the T prostanoid (TP) receptor.

For leukotriene biosynthesis, the released arachidonic acid is metabolized by 5-lipoxygenase (5-LO) in the presence of an integral nuclear membrane protein, 5-LO activating protein (FLAP). The calcium ion–dependent translocation of 5-LO to the nuclear membrane converts the arachidonic acid to the sequential intermediates, 5-hydroperoxyeicosatetraenoic acid (5-HPETE) and leukotriene (LT) A_4. LTA_4 is conjugated with reduced glutathione by LTC_4 synthase, an integral nuclear membrane protein homologous to FLAP. Intracellular LTC_4 is released by a carrier-specific export step for extracellular metabolism to the additional cysteinyl leukotrienes, LTD_4 and LTE_4, by the sequential removal of glutamic acid and glycine. Alternatively, cytosolic LTA_4 hydrolase converts some LTA_4 to the dihydroxy leukotriene LTB_4, which also undergoes specific export. Two receptors for LTB_4, BLT_1 and BLT_2, mediate chemotaxis of human neutrophils. Two receptors for the cysteinyl leukotrienes, $CysLT_1$ and $CysLT_2$, are present on smooth muscle of the airways and the microvasculature and on hematopoietic cells such as macrophages, eosinophils, and mast cells. Whereas the $CysLT_1$ receptor has a preference for LTD_4 and is blocked by the receptor antagonists in clinical use, the $CysLT_2$ receptor is equally responsive to LTD_4 and LTC_4, is unaffected by these antagonists, and is a negative regulator of the function of the $CysLT_1$ receptor. LTD_4, acting at $CysLT_1$ receptors, is the most potent known bronchoconstrictor, whereas LTE_4 induces a vascular leak and mediates the recruitment of eosinophils to the bronchial mucosa. Recently, GPR99, CysLT3 receptor, was identified as an LTE4 receptor. The lysophospholipid formed during the release of arachidonic acid from 1-*O*-alkyl-2-acyl-*sn*-glyceryl-3-phosphorylcholine can be acetylated in the second position to form platelet-activating factor (PAF). Serum levels of PAF correlated positively with the severity of anaphylaxis to peanut in a recent study, whereas the levels of PAF acetyl hydrolase (a PAF-degrading enzyme) were inversely related to the same outcome.

Unlike most other cells of bone marrow origin, mast cells circulate as committed progenitors lacking their characteristic secretory granules. These committed progenitors express c-*kit*, the receptor for stem cell factor (SCF). Unlike most other lineages, they retain and increase c-*kit* expression with maturation. The SCF interaction with c-*kit* is an absolute requirement for the development of constitutive tissue mast cells residing in skin and connective tissue sites and for the accumulation of mast cells at mucosal surfaces during T_H2-type immune responses. Several T cell-derived cytokines (IL-3, IL-4, IL-5, and IL-9) can potentiate SCF-dependent mast cell proliferation and/or survival in vitro in mice and humans. Indeed, mast cells are absent from the intestinal mucosa in clinical T cell deficiencies, but are present in the submucosa. Based on the immunodetection of secretory granule neutral proteases, mast cells in the lung parenchyma and intestinal mucosa selectively express tryptase, and those in the intestinal and airway submucosa, perivascular spaces, skin, lymph nodes, and breast parenchyma express tryptase, chymase, and carboxypeptidase A (CPA). In the mucosal epithelium of severe asthmatics, mast cells can express tryptase and CPA without chymase. The secretory granules of mast cells selectively positive for tryptase exhibit closed scrolls with a periodicity suggestive of a crystalline structure by electron microscopy, whereas the secretory granules of mast cells with multiple proteases are scroll-poor, with an amorphous or lattice-like appearance.

Mast cells are distributed at cutaneous and mucosal surfaces and in submucosal tissues about venules and could influence the entry of foreign substances by their rapid response capability **(Fig. 345-2)**. Upon stimulus-specific activation and secretory granule exocytosis, histamine and acid hydrolases are solubilized, whereas the neutral proteases, which are cationic, remain largely bound to the anionic proteoglycans, heparin and chondroitin sulfate E, with which they function as a complex. Histamine and the various lipid mediators

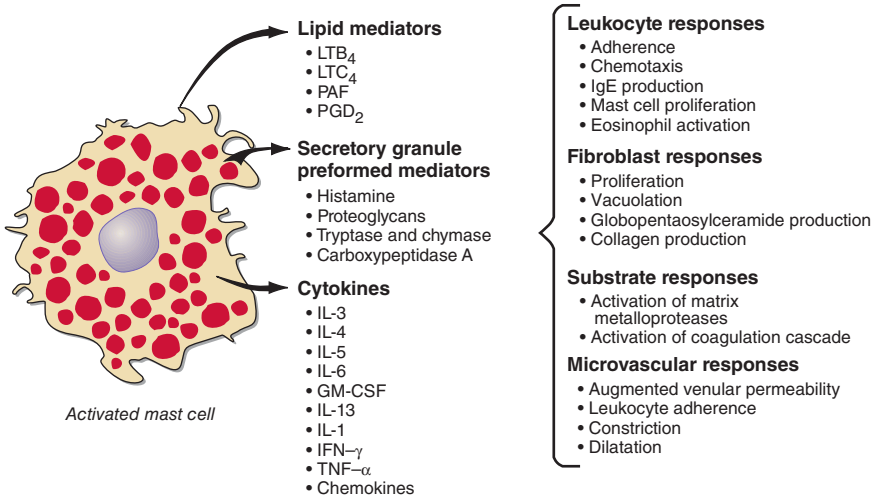


FIGURE 345-2 Bioactive mediators of three categories generated by IgE-dependent activation of murine mast cells can elicit common but sequential target cell effects leading to acute and sustained inflammatory responses. GM-CSF, granulocyte-macrophage colony-stimulating factor; IL, interleukin; IFN, interferon; LT, leukotriene; PAF, platelet-activating factor; PGD_2, prostaglandin D_2; TNF, tumor necrosis factor.

(PGD_2, $LTC_4/D_4/E_4$, PAF) alter venular permeability, thereby allowing influx of plasma proteins such as complement and immunoglobulins, whereas LTB_4 mediates leukocyte–endothelial cell adhesion and subsequent directed migration (chemotaxis). The accumulation of leukocytes and plasma opsonins facilitates defense of the microenvironment. The inflammatory response can also be detrimental, as in asthma, where the smooth-muscle constrictor activity of the cysteinyl leukotrienes is evident and much more potent than that of histamine.

The cellular component of the mast cell–mediated inflammatory response is augmented and sustained by cytokines and chemokines. IgE-dependent activation of human skin mast cells in situ elicits TNF-α production and release, which in turn induces endothelial cell responses favoring leukocyte adhesion. Similarly, activation of purified human lung mast cells or cord blood–derived cultured mast cells in vitro results in substantial production of proinflammatory (TNF-α) and immunomodulatory cytokines (IL-4, IL-5, IL-13) and chemokines. Bronchial biopsy specimens from patients with asthma reveal that mast cells are immunohistochemically positive for IL-4 and IL-5, but that the predominant localization of IL-4, IL-5, and GM-CSF is to T cells, defined as T_H2 by this profile. IL-4 modulates the T cell phenotype to the T_H2 subtype, determines the isotype switch to IgE (as does IL-13), and upregulates FcεRI-mediated expression of cytokines by mast cells based on in vitro studies.

An immediate and late cellular phase of allergic inflammation can be induced in the skin, nose, or lung of some allergic humans with local allergen challenge. The immediate phase in the nose involves pruritus and watery discharge; in the lung, it involves bronchospasm and mucus secretion; and in the skin, it involves a wheal-and-flare response with pruritus. The reduced nasal patency, reduced pulmonary function, or erythema with swelling at the skin site in a late-phase response at 6–8 h is associated with biopsy findings of infiltrating and activated T_H2 cells, eosinophils, basophils, and some neutrophils. The progression from early mast cell activation to late cellular infiltration has been used as an experimental surrogate of rhinitis or asthma. However, in asthma, there is an intrinsic hyperreactivity of the airways independent of the associated inflammation. Moreover, early- and late-phase responses (at least in the lung) are far more sensitive to blockade of IgE-dependent mast cell activation (or actions of histamine and cysteinyl leukotrienes) than are spontaneous or virally induced asthma exacerbations.

Consideration of the mechanism of immediate-type hypersensitivity diseases in the human has focused largely on the IgE-dependent recognition of otherwise innocuous substances. A region of chromosome 5 (5q23-31) contains genes implicated in the control of IgE levels including IL-4 and IL-13, as well as IL-3 and IL-9, which are involved in mucosal mast cell hyperplasia, and IL-5 and GM-CSF, which are central to eosinophil development and their enhanced tissue viability. Genes with linkage to the specific IgE response to particular allergens include those encoding the major histocompatibility complex (MHC) and certain chains of the T cell receptor (TCR-αδ). The complexity of atopy and the associated diseases includes susceptibility, severity, and therapeutic responses, each of which is among the separate variables modulated by both innate and adaptive immune stimuli.

The induction of allergic disease requires sensitization of a predisposed individual to a specific allergen. The greatest propensity for the development of atopic allergy occurs in childhood and early adolescence. The allergen is processed by antigen-presenting cells of the monocytic lineage (particularly dendritic cells) located throughout the body at surfaces that contact the outside environment, such as the nose, lungs, eyes, skin, and intestine. These antigen-presenting cells present the epitope-bearing peptides via their MHC to T helper cells and their subsets. The T cell response depends both on cognate recognition and on the cytokine microenvironment provided by the antigen-presenting dendritic cells, with IL-4 directing a T_H2 subset, interferon (IFN) γ a T_H1 profile, and IL-6 with transforming growth factor β (TGF-β) a T_H17 subset. Allergens not only present antigenic epitopes via dendritic cells but also contain pattern recognition ligands that facilitate the immune response by direct initiation of cytokine generation from innate cell types such as basophils, mast cells, eosinophils, and others. The T_H2 response is associated with activation of specific B cells that can also present allergens or that transform into plasma cells for antibody production. Synthesis and release into the plasma of allergen-specific IgE results in sensitization of FcεR1-bearing cells such as mast cells and basophils, which become activated on exposure to the specific allergen. In certain diseases, including those associated with atopy, the monocyte and eosinophil populations can express a trimeric FcεR1, which lacks the β chain, and yet respond to its aggregation. An additional recently recognized class of c-*kit*-expressing innate cells (termed group 2 innate lymphoid cells of ILC2) can generate large quantities of IL-5 and IL-13 during antihelminth responses, are prominent in nasal polyps from humans, and contribute to inflammation in allergic diseases.

URTICARIA AND ANGIOEDEMA

DEFINITION

Urticaria and angioedema represent the same pathophysiologic process occurring at different levels of the skin. Urticaria involves dilation of vascular structures in the superficial dermis, while angioedema originates from the deeper dermis and subcutaneous tissues. Not surprisingly they often appear together, with roughly 40% of patients reporting both, and affect >20% of the population at sometime during their lifespan. Urticaria can occur on any area of the body as well-circumscribed wheals with erythematous raised serpiginous borders and blanched centers that may coalesce to become giant wheals. Urticarial lesions last for <24 h, frequently migrate around the body, leave no bruising or scarring and are intensely pruritic. Angioedema is marked by dramatic swelling with more pain than pruritus and minimal erythema, which may develop with a pruritic prodrome and takes hours to days to resolve. Acute urticaria and/or angioedema are episodes that occur for <6 weeks' duration, whereas attacks persisting >6 weeks are designated chronic.

PREDISPOSING FACTORS AND ETIOLOGY

Acute or chronic urticaria and/or angioedema can occur at any point in the lifespan with the third to fifth decade the most common for chronic. Women are affected more often than men with a slight predominance for those with a history of atopy. Acute urticaria is most often the result

TABLE 345-1 Classification of Urticaria and/or Angioedema

ACUTE	CHRONIC
Drug Reactions	Idiopathic—a subset with autoimmune component
NSAIDS, IV contrast, angiotensin-converting enzyme (ACE) inhibitors, etc.	Collagen vascular disease—urticarial vasculitis
Foods	Physical stimuli
Inhalation or contact with environmental allergens	Dermatographism
Transfusion reactions	Cholinergic urticaria
Insects	Vibration, cold, pressure, water (aquagenic)
Infections—viral, bacterial, parasitic	Sun (solar)
	Mastocytosis/Urticaria pigmentosa
	Hereditary
	Hereditary angioedema (HAE)
	Familial cold urticaria
	C3b inhibitor deficiency
	Muckle-Wells syndrome
	Schnitzler syndrome
	Hypereosinophilic syndrome
	Gleich syndrome

of exposure to a food, environmental or drug allergen or viral infection while chronic urticaria is often idiopathic.

The classification of urticaria-angioedema presented in **Table 345-1** focuses on the different mechanisms for eliciting clinical disease and can be useful for differential diagnosis.

Additional etiologies include physical stimuli such as cold, heat, solar rays, exercise, and mechanical irritation. The physical urticarias can be distinguished by the precipitating event and other aspects of the clinical presentation. *Dermographism*, which occurs in 1–4% of the population, is defined by the appearance of a linear wheal with surrounding erythema at the site of a brisk stroke with a firm object **(Fig. 345-3)**. Dermographism has a prevalence that peaks in the second to third decades. It is not influenced by atopy and has a duration generally of <5 years. *Pressure urticaria*, which often accompanies chronic idiopathic urticaria, presents in response to a sustained stimulus such as a shoulder strap or belt, running (feet), or manual labor (hands). *Cholinergic urticaria* is distinctive in that the pruritic wheals are of small size (1–2 mm) and are surrounded by a large area of erythema; attacks are precipitated by fever, a hot bath or shower, or exercise and are presumptively attributed to a rise in core body temperature. *Exercise-induced anaphylaxis* can be precipitated by exertion alone or can be dependent on prior food ingestion. There is an association with the presence of IgE specific for α-5 gliadin, a component of wheat. The clinical presentation can be limited to flushing, erythema, and pruritic urticaria but may progress to angioedema of the face, oropharynx, larynx, or intestine or to vascular collapse; it is distinguished from cholinergic urticaria by presenting with wheals of conventional size and by not occurring with fever or a hot bath. *Cold urticaria* is local at body areas exposed to low ambient temperature or cold objects but can progress to vascular collapse with immersion in cold water (swimming). *Solar urticaria* is subdivided into six groups by the response to specific portions of the light spectrum. *Vibratory angioedema* may occur after years of occupational exposure or can be idiopathic; it may be accompanied by cholinergic urticaria. Other rare forms of physical allergy, always defined by stimulus-specific elicitation, include *local heat urticaria*, *aquagenic urticaria* from contact with water of any temperature (sometimes associated with polycythemia vera), and *contact urticaria* from direct interaction with some chemical substance (such as latex).

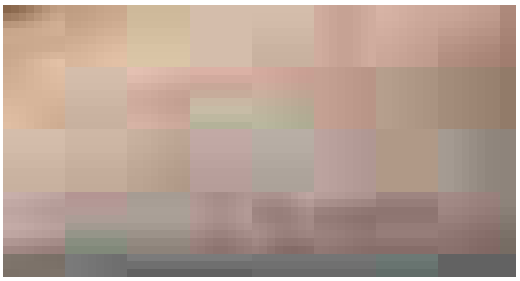


FIGURE 345-3 Dermographic urticarial lesion induced by stroking the forearm lightly with the edge of a tongue blade. The photograph, taken after 10 min, demonstrates a prominent wheal-and-flare reaction in the shape of a *hashtag*. *(Photograph provided by Katherine N. Cahill, MD, Harvard Medical School.)*

Isolated Angioedema Angioedema without urticaria can be idiopathic or due to the generation of bradykinin in the setting of C1 inhibitor (C1INH) deficiency that may be inborn as an autosomal dominant characteristic or may be acquired through the appearance of an autoantibody in the setting of malignancy. The angiotensin-converting enzyme (ACE) inhibitors can provoke a similar clinical presentation in 0.2–0.7% of exposed patients due to delayed degradation of bradykinin. Black race, organ transplant, female gender, smoking, and increasing age are known risk factors for ACE-inhibitor related angioedema.

CLINICAL PRESENTATION AND PATHOPHYSIOLOGY

Urticarial eruptions are distinctly pruritic, may involve any area of the body from the scalp to the soles of the feet, and appear in crops of 12- to 36-h duration, with old lesions fading as new ones appear. Most of the physical urticarias (cold, cholinergic, dermatographism) are an exception, with individual lesions lasting <2 h. Neither urticaria nor angioedema lesions are symmetric or dependent in distribution. The most common sites for angioedema are often periorbital and perioral. Angioedema of the upper respiratory tract may be life-threatening due to transient laryngeal obstruction, whereas gastrointestinal involvement may present with abdominal colic, with or without nausea and vomiting, and can result in unnecessary surgical intervention. No residual scaring occurs with either urticaria or angioedema unless there is an underlying vasculitic process.

The pathology is characterized by edema of the superficial dermis in urticaria and of the subcutaneous tissue and deep dermis in angioedema. Collagen bundles in affected areas are widely separated, and the venules are sometimes dilated. Any perivenular infiltrate consists of lymphocytes, monocytes, eosinophils, and neutrophils that are present in varying combination and numbers.

The best evidence for IgE- and mast cell–involvement in urticaria and angioedema is *cold urticaria*. Cryoglobulins or cold agglutinins are present in up to 5% of these patients. Immersion of an extremity in an ice bath precipitates angioedema of the distal portion with urticaria at the air interface within minutes of the challenge. Histologic studies reveal marked mast cell degranulation with associated edema of the dermis and subcutaneous tissues. Elevated levels of histamine have been found in the plasma of venous effluent and in the fluid of suction blisters at experimentally induced lesional sites in patients with cold urticaria, dermographism, pressure urticaria, vibratory angioedema, light urticaria, and heat urticaria. By ultrastructural analysis, the pattern of mast cell degranulation in cold urticaria resembles an IgE-mediated response with solubilization of granule contents, fusion of the perigranular and cell membranes, and discharge of granule contents, whereas in a dermatographic lesion, there is additional superimposed zonal (piecemeal) degranulation. Elevations of plasma histamine levels with biopsy-proven mast cell degranulation have also been demonstrated with generalized attacks of *cholinergic urticaria*.

Up to 45% of patients with chronic urticaria have an autoimmune cause for their disease including autoantibodies to IgE or to the α chain of FcεRI. In some patients, autologous serum injected into their own skin can induce a wheal-and-flare reaction involving mast cell activation. The presence of these antibodies can also be recognized by their capacity to release histamine or induce activation markers such as CD63 or CD203 on basophils. An association with antibodies to microsomal peroxidase and/or thyroglobulin has been observed with both clinically significant Hashimoto's thyroiditis as well as a euthyroid state. In vitro studies reveal that these autoantibodies can mediate basophil degranulation with enhancement by serum as a source of the anaphylatoxic fragment, C5a.

The urticaria and angioedema associated with classic serum sickness or with hypocomplementemic cutaneous necrotizing angiitis (urticarial vasculitis) are believed to be immune-complex diseases. Reactions to mast cell granule–releasing agents (opioids, contrast media) and to nonsteroidal anti-inflammatory drugs are most often limited to urticaria and/or angioedema, but may be systemic.

Hereditary angioedema (HAE) is a fully penetrant, autosomal dominant disease due to a mutation in the *SERPING1* gene leading to a deficiency of C1INH (type 1) in about 85% of patients or to a dysfunctional protein (type 2) in the remainder affecting 1:30,000–80,000 in the general population. A third less common type of HAE has been described in which C1INH function is normal, and the causal lesion is a mutant form of factor XII, which leads to generation of excessive bradykinin. C1INH deficiency can also develop in a sporadic acquired form as a result of excessive consumption of C1INH due either to formation of immune complexes or to the generation of an autoantibody directed to C1INH in the setting of lymphoproliferative disease. C1INH blocks the catalytic function of activated factor XII (Hageman factor) and of kallikrein, as well as the C1r/C1s components of C1, with the common result of degrading bradykinin. During clinical attacks of angioedema, C1INH function or levels fall, patients develop elevated plasma levels of bradykinin leading to angioedema and excessive activation of C1 results in a decline in C4 and C2 levels.

The use of ACE inhibitors results in impaired bradykinin degradation and explains the angioedema that occurs idiosyncratically in ACE inhibitor-exposed patients with a normal C1INH. Bradykinin-mediated angioedema, whether caused by ACE inhibitors or by C1INH deficiency, is noteworthy for the conspicuous absence of concomitant urticaria or pruritus, the frequent involvement of the gastrointestinal tract, and the duration of symptoms >24 h.

DIAGNOSIS

The classification of urticarial and angioedematous states as presented in Table 345-1 in terms of duration can facilitate identification of possible mechanisms. History alone of self-limited urticarial and/or angioedema episodes can be sufficient to make a diagnosis in the setting of acute disease triggered by drug, environmental or food allergen with history-directed confirmatory skin testing or assay for serum allergen-specific IgE. Direct reproduction of the lesion in physical urticarias is particularly valuable because it so often establishes the cause of the lesion. Even with chronic urticaria/angioedema, initial diagnostic testing should be limited and expanded testing guided by history. Complete blood count with assessment for eosinophilia, erythrocyte sedimentation rate and thyroid stimulation hormone level are recommended by consensus guidelines even though the vast majority of chronic urticaria is associated with no laboratory abnormality. Urticarial lesions that last longer than 36 h result in scarring and are reported as painful and not pruritic warrant biopsy to evaluate for cellular infiltration, nuclear debris, and fibrinoid necrosis of the venules consistent with urticarial vasculitis. Chronic angioedema without urticaria warrants assessment of complement levels. Concomitant flushing and hyperpigmented papules that urticate with stroking in the absence of angioedema raise the question of mastocytosis. An appropriate travel history should trigger an evaluation for parasites.

The diagnosis of HAE is suggested not only by family history but also by the lack of pruritus and of urticarial lesions, the prominence of recurrent gastrointestinal attacks of colic, and episodes of laryngeal edema. Laboratory diagnosis depends on demonstrating a deficiency of C1INH antigen (type 1) or a nonfunctional protein (type 2) by a catalytic inhibition assay. While levels of C1 are normal, its substrates, C4 and C2, are chronically depleted and fall further during attacks due to the activation of additional C1. Patients with the acquired forms of C1INH deficiency have the same clinical manifestations but differ in the lack of a familial element. Furthermore, their sera exhibit a reduction of C1 function and C1q protein as well as C1INH, C4, and C2. Inborn C1INH deficiency and ACE inhibitor–elicited angioedema are associated with elevated levels of bradykinin. Lastly, type 3 HAE is associated with normal levels of complement proteins and a factor XII gene mutation.

TREATMENT

Urticaria and Angioedema

For most forms of urticaria, H_1 antihistamines such as chlorpheniramine or diphenhydramine effectively attenuate both urtication and pruritus, but because of their side effects and short half-life, long-acting, non-sedating agents such as loratadine, desloratadine, and fexofenadine, or low-sedating agents such as cetirizine or levocetirizine generally are used first and increased to four times daily (QID) dosing. The addition of an H_2 antagonist such as cimetidine, ranitidine, or famotidine in conventional dosages may add benefit when H_1 antihistamines are inadequate. A $CysLT_1$ receptor antagonist such as montelukast, 10 mg daily, or zafirlukast, 20 mg twice a day, can be an important add-on therapy. For chronic urticaria which has failed to respond to a combination of long-acting H_1 antihistamines QID and a $CysLT_1$ receptor antagonist or cold urticaria, monoclonal anti-IgE antibodies such as omalizumab are now the next line of therapy. Older agents with antihistamine properties such as doxepin, cyproheptadine, and hydroxyzine have proven effective when H_1 antihistamines fail but are less effective than omalizumab and are sedating.

Topical glucocorticoids are of no value, and systemic glucocorticoids are generally avoided in idiopathic, allergen-induced, or physical urticarias due to their long-term toxicity. Systemic glucocorticoids are useful in the management of patients with pressure urticaria, vasculitic urticaria (especially with eosinophil prominence), idiopathic angioedema with or without urticaria, or chronic urticaria that responds poorly to conventional treatment and should be considered in any patient with debilitating disease. With persistent vasculitic urticaria, hydroxychloroquine, dapsone, or colchicine may be added to the regimen after hydroxyzine and before or along with systemic glucocorticoids. Cyclosporine is efficacious for patients with chronic idiopathic urticaria that is severe and poorly responsive to other modalities and/or where glucocorticoids are a requirement.

Infusion of isolated or recombinant C1INH protein is approved for prophylaxis of and acute HAE attacks while administration of a bradykinin 2 receptor antagonist (Icatibant) or a kallikrein inhibitor (Ecallantide) may be used for treatment of an acute attack of HAE. Older, less expensive preventative options include attenuated androgens, which stimulate production by the normal gene of an amount of functional C1INH sufficient to control the spontaneous activation of C1. The antifibrinolytic agent ε-aminocaproic acid may be used for preoperative prophylaxis, but is contraindicated in patients with thrombotic tendencies or ischemia due to arterial atherosclerosis. Fresh frozen plasma infusion can be used for acute attacks in a setting which lacks access to the newer agents. Bradykinin 2 receptor antagonist and C1INH protein are being studied for ACE inhibitor-induced angioedema. Treatment of the underlying hematologic malignancy is indicated for acquired C1INH deficiency.

ALLERGIC RHINITIS

DEFINITION

Rhinitis is characterized by sneezing; rhinorrhea; obstruction of the nasal passages; conjunctival, nasal, and pharyngeal itching; and lacrimation and can be classified as allergic or non-allergic. A clinical history of rhinitis symptoms occurring in a temporal relationship to allergen exposure and documentation of sensitization to an environmental allergen are required for a diagnosis of allergic rhinitis. Although commonly seasonal due to elicitation by airborne pollens, it can be perennial in an environment of chronic exposure to house dust mites, animal danders, or insect (cockroach) products. The overall prevalence in North America has increased in the past 20 years and is 10–30%, with the peak prevalence of >30% occurring in the fifth decade.

PREDISPOSING FACTORS AND ETIOLOGY

Allergic rhinitis generally occurs in atopic individuals, often in association with atopic dermatitis, food allergy, urticaria, and/or asthma

(Chap. 281). Up to 50% of patients with allergic rhinitis manifest asthma, whereas 70–80% of individuals with asthma and 80% of individuals with chronic bilateral sinusitis experience allergic rhinitis. Female sex, particulate air pollution exposure, and maternal tobacco smoking increase the risk of developing allergic rhinitis over the life span.

Trees, grasses, and weeds that depend on wind rather than insects for pollination produce sufficient quantities of pollen suitable for wide distribution by air currents to elicit seasonal allergic rhinitis. The dates of pollination of these species historically varied little from year to year in a particular locale, but may be quite different in another climate. In the temperate areas of North America, trees typically pollinate from March through May, grasses in June and early July, and ragweed from mid-August to early October. Molds, which are widespread in nature because they occur in soil or decaying organic matter, propagate spores in a pattern that depends on climatic conditions. Climate change is impacting these patterns with early tree pollination and prolonged ragweed season with the delay of the first frost. Perennial allergic rhinitis occurs in response to allergens that are present throughout the year, including animal dander, cockroach-derived proteins, mold spores, or dust mites such as *Dermatophagoides farinae* and *Dermatophagoides pteronyssinus*. Dust mites are scavengers of human skin and excrete cysteine protease allergens in their feces. In up to 40% of patients with perennial rhinitis, no clear-cut allergen can be demonstrated as causative.

PATHOPHYSIOLOGY AND MANIFESTATIONS

Episodic rhinorrhea, sneezing, obstruction of the nasal passages with lacrimation, and pruritus of the conjunctiva, nasal mucosa, and oropharynx are the hallmarks of allergic rhinitis. The nasal mucosa is pale and boggy, the conjunctiva congested and edematous, and the pharynx generally unremarkable. Swelling of the turbinates and mucous membranes with obstruction of the sinus ostia and eustachian tubes precipitates secondary infections of the sinuses and middle ear, respectively. A growing number of patients with seasonal allergic rhinitis demonstrate pollen-associated food allergen syndrome characterized by oropharyngeal pruritus and/or mild swelling following the ingestion of raw plant-based foods which contain cross-reacting pollen-related allergens.

Nasal polyps, representing mucosal protrusions containing edema fluid with variable numbers of eosinophils and degranulated mast cells, can increase obstructive symptoms with anosmia as a defining feature and can concurrently arise within the nasopharynx or sinuses. Atopy is not a risk factor for nasal polyps, which instead may occur in the setting of cystic fibrosis, aspirin-exacerbated respiratory disease characterized by the triad of asthma, rhinosinusitis, and respiratory reactions to all cyclooxygenase-1 inhibitors, and in patients with chronic staphylococcal colonization, which produces superantigens leading to an intense T_H2 inflammatory response.

The nose presents a large mucosal surface area through the folds of the turbinates and serves to adjust the temperature and moisture content of inhaled air and to filter out particulate materials >10 μm in size by impingement in a mucous blanket; ciliary action moves the entrapped particles toward the pharynx. Entrapment of pollen and digestion of the outer coat by mucosal enzymes such as lysozymes release protein allergens. The initial interaction occurs between the allergen and intraepithelial mast cells and then proceeds to involve deeper perivenular mast cells, both of which are sensitized with specific IgE. During the symptomatic season when the mucosae are already swollen and hyperemic, there is enhanced adverse reactivity to the seasonal pollen as well as irritants such as tobacco smoke and fragrances. Biopsy specimens of nasal mucosa during seasonal rhinitis show submucosal edema with infiltration by eosinophils, along with some basophils and neutrophils.

The mucosal surface fluid contains IgA that is present because of its secretory piece and also IgE, which apparently arrives by diffusion from plasma cells in proximity to mucosal surfaces. IgE fixes to mucosal and submucosal mast cells, and the intensity of the clinical response to inhaled allergens is quantitatively related to the naturally occurring pollen dose. In sensitive individuals, the introduction of allergen into the nose is associated with sneezing, nasal obstruction, and discharge, and the fluid contains histamine, PGD_2, and leukotrienes. Thus the mast cells of the nasal mucosa and submucosa generate and release mediators through IgE-dependent reactions that are capable of producing tissue edema and eosinophilic infiltration.

DIAGNOSIS

The diagnosis of seasonal allergic rhinitis depends largely on an accurate history of occurrence coincident with the pollination of the offending weeds, grasses, or trees. The continuous character of perennial allergic rhinitis due to contamination of the home or place of work makes historic analysis difficult, but there may be variability in symptoms that can be related to exposure to animal dander, dust mite and/or cockroach allergens, fungal spores, or work-related allergens such as latex. Patients with perennial rhinitis commonly develop the problem in adult life, and manifest nasal congestion and a postnasal discharge, often associated with thickening of the sinus membranes demonstrated by radiography. Perennial nonallergic rhinitis with eosinophilia syndrome (NARES) occurs in the middle decades of life and is characterized by nasal obstruction, anosmia, chronic sinusitis, and prominent eosinophilic nasal discharge in the absence of allergen sensitization. The term *vasomotor rhinitis* or *perennial nonallergic rhinitis* designates a condition of enhanced reactivity of the nasopharynx in which a symptom complex resembling perennial allergic rhinitis occurs with nonspecific stimuli, including chemical odors, temperature and humidity variations, and position changes but occurs without tissue eosinophilia or an allergic etiology. Other entities to be excluded are structural abnormalities of the nasopharynx; exposure to irritants; gustatory rhinitis associated with cholinergic activation that occurs while eating or ingesting alcohol; hypothyroidism; upper respiratory tract infection; pregnancy with prominent nasal mucosal edema; prolonged topical use of α-adrenergic agents in the form of nasal sprays (rhinitis medicamentosa); and the use of certain systemic agents such as β-adrenergic antagonists, ACE inhibitors, direct vasodilators (hydralazine), α1-adrenergic receptor antagonists, estrogens, progesterone, NSAIDS, gabapentin, phosphodiesterase-5 inhibitors, and psychotropics (Risperidone, chlorpromazine, amitriptyline).

The nasal secretions of allergic patients are rich in eosinophils, and a modest peripheral eosinophilia can be observed. Local or systemic neutrophilia implies infection. Total serum IgE is frequently elevated, but the demonstration of immunologic specificity for IgE is critical to an etiologic diagnosis. A skin test by the intracutaneous route (puncture or prick) with the allergens of interest provides a rapid and reliable approach to identifying allergen-specific IgE that has sensitized cutaneous mast cells. A positive intracutaneous skin test with 1:10–1:20 weight/volume of extract has a high predictive value for the presence of allergy. An intradermal test with a 1:500–1:1000 dilution of 0.05 mL may follow if indicated by history when the intracutaneous test is negative, but while more sensitive, it is less reliable due to the reactivity of some asymptomatic individuals at the test dose.

Newer methodology for detecting total IgE, including the development of enzyme-linked immunosorbent assays (ELISA) employing anti-IgE bound to either a solid-phase or a liquid-phase particle, provides rapid and cost-effective determinations. Measurements of specific anti-IgE in serum are obtained by its binding to an allergen and quantitation by subsequent uptake of labeled anti-IgE. As compared to the skin test, the assay of specific IgE in serum is less sensitive but has high specificity.

TREATMENT

Allergic Rhinitis

Although allergen avoidance is the most cost-effective means of managing allergic rhinitis, only in the case of animal dander and possibly dust mites is it really feasible. Treatment with pharmacologic agents represents the standard approach to seasonal or perennial allergic rhinitis. Oral long-acting H_1 antihistamines are effective for nasopharyngeal itching, sneezing, and watery

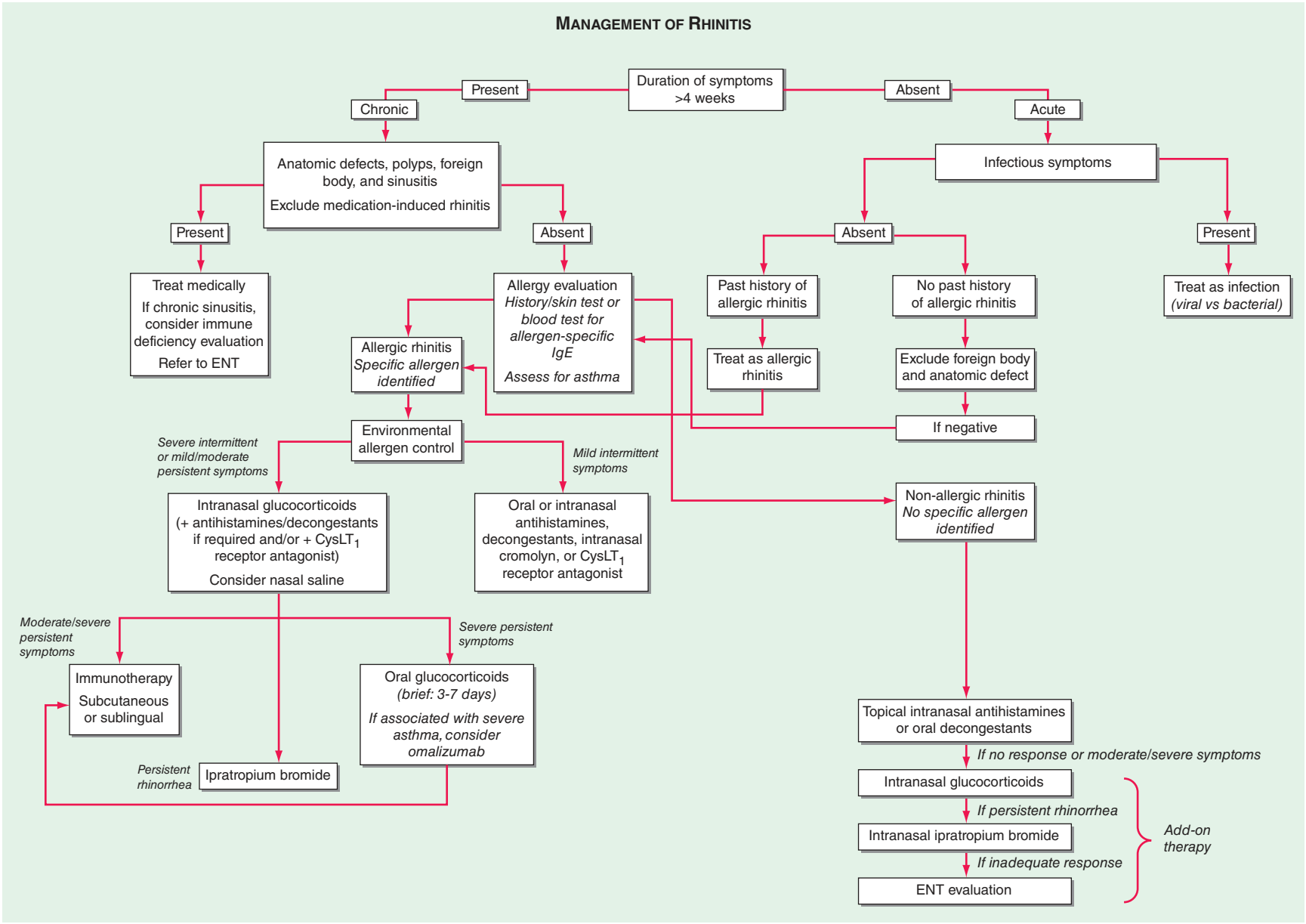


FIGURE 345-4 Algorithm for the diagnosis and management of rhinitis. Persistent defined as >4 days/week for >4 weeks. Moderate/severe defined as abnormal sleep, impaired daily activities (school, work, sport, leisure) and/or troublesome symptoms. CysLT, cysteinyl leukotriene; ENT, ear, nose, and throat; IgE, immunoglobulin E.

rhinorrhea and for such ocular manifestations as itching, tearing, and erythema, but they are less efficacious for the nasal congestion. The older antihistamines are sedating, and they induce psychomotor impairment, including reduced eye-hand coordination and impaired automobile driving skills. Their anticholinergic (muscarinic) effects include visual disturbance, urinary retention, and constipation. Because the newer H_1 antihistamines such as fexofenadine, loratadine, desloratadine, cetirizine, levocetirizine, olopatadine, bilastine, and azelastine are less lipophilic and more H_1 selective, their ability to cross the blood-brain barrier is reduced, and thus their sedating and anticholinergic side effects are minimized. These newer antihistamines do not differ appreciably in efficacy for relief of rhinitis and/or sneezing. Intranasal high-potency glucocorticoids are the most potent drugs available for the relief of established rhinitis, seasonal or perennial, and are effective in relieving nasal congestion as well as ocular symptoms. They provide efficacy with substantially reduced side effects as compared with this same class of agent administered orally. Their most frequent side effect is local irritation, with *Candida* overgrowth being a rare occurrence. The currently available intranasal glucocorticoids—beclomethasone, flunisolide, triamcinolone, budesonide, fluticasone propionate, fluticasone furoate, ciclosenide, and mometasone furoate—are equally effective for nasal symptom relief, including nasal congestion; these agents all achieve up to 70% overall symptom relief with some variation in the time period for onset of benefit. Azelastine nasal spray may benefit individuals with nonallergic vasomotor rhinitis as well as additive benefit to intranasal steroids in allergic rhinitis, but it has an adverse effect of dysgeusia (taste perversion) in some patients. Alternative nasal decongestants include α-adrenergic agents such as phenylephrine or oxymetazoline; however, the duration of their efficacy is limited because of rebound rhinitis (i.e., 7- to 14-day use can lead to rhinitis medicamentosa) and such systemic responses as hypertension. Oral α-adrenergic agonist decongestants containing pseudoephedrine are standard for the management of nasal congestion, generally in combination with an antihistamine. While oral antihistamines typically reduce nasal and ocular symptoms by about one-third, pseudoephedrine must be added to achieve a similar reduction in nasal congestion. These pseudoephedrine combination products can cause insomnia and are precluded from use in patients with narrow angle glaucoma, urinary retention, severe hypertension, marked coronary artery disease, or a first-trimester pregnancy. The $CysLT_1$ blocker montelukast is approved for treatment of both seasonal and perennial rhinitis, and it reduces both nasal and ocular symptoms by about 20%. Cromolyn sodium nasal spray inhibits mast cell degranulation, and can be used prophylactically on a continuous basis during the season. Topical ipratropium is an anticholinergic agent effective in reducing rhinorrhea, including that of patients with perennial non-allergic symptoms, and it can be additionally efficacious when combined with intranasal glucocorticoids. For concomitant allergic conjunctivitis, topical treatment with cromolyn sodium is effective in treating mild allergic symptoms and topical antihistamines such as olopatadine, azelastine, ketotifen, or epinastine administered to the eye provide rapid relief of itching and redness and are more effective than oral antihistamines.

Immunotherapy Immunotherapy consists of repeated exposure to gradually increasing concentrations of the allergen(s) considered to be specifically responsible for the symptom complex. Two forms of immunotherapy, subcutaneous (SCIT) and sublingual (SLIT), are currently available. Controlled studies of ragweed, grass, dust mite, and cat dander allergens administered via SCIT for treatment of allergic rhinitis have demonstrated improved symptom control over medications alone with the advantage of providing a durable benefit. The duration of SCIT is 3–5 years, with discontinuation being based on minimal symptoms over two consecutive seasons of exposure to the allergen. Clinical benefit appears related to the administration of a high dose of relevant allergen, advancing from weekly to monthly intervals. Patients should remain at the treatment site for at least 30 min after allergen administration so that any systemic reactions including anaphylaxis can be managed. Two to three percent of SCIT patients experience a systemic reaction over a 12-month period. Local reactions with erythema and induration are not uncommon and may persist for 1–3 days. SLIT is prepared as a tablet to be dissolved under the tongue at home after the first dose. The efficacy of SLIT is comparable to SCIT but only for the three allergens formulations available, dust mite, timothy/northern grasses and ragweed. Systemic reactions are less frequent with SLIT but transient oral pruritus is common. Immunotherapy is contraindicated in patients with significant cardiovascular disease or unstable asthma and should be conducted with particular caution in any patient requiring β-adrenergic blocking therapy because of the difficulty in managing an anaphylactic complication. The response to immunotherapy is associated with a complex of cellular and humoral effects that includes a modulation in T cell cytokine production and allergen-specific IgG_4 expansion. Immunotherapy should be reserved for clearly documented seasonal or perennial rhinitis that is clinically related to defined allergen exposure with confirmation by the presence of allergen-specific IgE through skin or in vitro specific IgE testing. Systemic treatment with a monoclonal antibody to IgE (omalizumab) that blocks mast cell and basophil sensitization has efficacy for allergic rhinitis and can be used with immunotherapy to enhance safety and efficacy. However, current approval is only for treatment of patients with persistent allergic asthma not controlled by inhaled glucocorticoid therapy. A sequence for the management of allergic or perennial rhinitis based on an allergen-specific diagnosis and stepwise management as required for symptom control would include the following: (1) identification of the offending allergen(s) by history with confirmation of the presence of allergen-specific IgE by skin test and/or serum assay; (2) avoidance of the offending allergen; and (3) medical management in a stepwise fashion **(Fig. 345-4)**. Mild intermittent symptoms of allergic rhinitis are treated with oral antihistamines, oral $CysLT_1$ receptor antagonists, intranasal antihistamines, or intranasal cromolyn prophylaxis. Moderate to more severe allergic rhinitis is managed with intranasal glucocorticoids plus oral antihistamines, oral $CysLT_1$ receptor antagonists, or antihistamine-decongestant combinations. Persistent or seasonal allergic rhinitis, rhinoconjunctivitis, or asthma which remains uncontrolled with maximal medical therapy merit consideration of allergen-specific immunotherapy.

FURTHER READING

Bernstein DI et al: Allergic rhinitis: Mechanisms and treatment. Immunol Allergy Clin North Am 36:261, 2016.

Cho SH et al: Chronic rhinosinusitis without nasal polyps. J Allergy Clin Immunol Pract 4:575, 2016.

Cicardi M et al: Classification, diagnosis, and approach to treatment for angioedema: Consensus report from the Hereditary Angioedema International Working Group. Allergy 69:602, 2014.

Corren J et al: Allergic and nonallergic rhinitis, in *Middleton's Allergy: Principles and Practice*, 8th ed. NF Adkinson et al (eds). Philadelphia, Saunders, 2014, pp 664–685.

Jutel M et al: International consensus on allergen immunotherapy II: Mechanisms, standardization, and pharmacoeconomics. J Allergy Clin Immunol 137:358, 2016.

Maurer M et al: Omalizumab for the treatment of chronic idiopathic or spontaneous urticaria. N Engl J Med 368:924, 2013.

Saini SS: Urticaria and angioedema, in *Middleton's Allergy: Principles and Practice*, 8th ed. NF Adkinson et al (eds). Philadelphia, Saunders, 2014, pp 575–587.

Stevens WW et al: Chronic rhinosinusitis with nasal polyps. J Allergy Clin Immunol Pract 4:565, 2016.

346 Anaphylaxis

David Hong, Joshua A. Boyce

DEFINITION

Anaphylaxis is a potentially life-threatening systemic allergic reaction involving one or more organ systems that typically occurs within seconds to minutes of exposure to the anaphylactic trigger, most often a drug, food, or hymenoptera sting. Other triggers of anaphylaxis include radiocontrast administration or latex exposure. The term "anaphylaxis" was first described in 1902 by Charles Richet and Paul Portier who attempted to immunize dogs against sea anemone toxin in the same way Pasteur was able to vaccinate individuals against the smallpox virus. To their surprise, repeated administration of small, sub-lethal doses of sea anemone toxin reliably induced acute-onset death when re-administered 2–3 weeks after initial "vaccination" to the toxin. The phenomenon was termed ana (anti)-phylaxis ("protection or guarding") because vaccination with anemone toxin resulted in the opposite intended immune effect. Charles Richet was awarded the Nobel Prize in Physiology or Medicine in 1913 for this work which led to further insights into hypersensitivity and mast cell biology.

Clinical Manifestations While 80–90% of anaphylactic episodes are uniphasic, about 10–20% of cases are biphasic in which anaphylactic symptoms return about an hour or longer after resolution of initial symptoms. Anaphylactic reactions are particularly dangerous when hypotension or hypoxia occurs, leading potentially to cardiovascular collapse or respiratory failure, respectively. There may be upper or lower airway obstruction or both. Laryngeal edema may be experienced as a "lump" in the throat, hoarseness, or stridor, whereas bronchial obstruction is associated with a feeling of tightness in the chest and/or audible wheezing. Patients with underlying asthma are predisposed to severe involvement of the lower airways and increased mortality associated with anaphylaxis. In fatal cases with clinical bronchial obstruction, the lungs show marked hyperinflation on gross and microscopic examination. The microscopic findings in the bronchi, however, are limited to luminal secretions, peribronchial congestion, submucosal edema, and eosinophilic infiltration, and the acute emphysema is attributed to intractable bronchospasm that subsides with death. Angioedema resulting in death by mechanical obstruction occurs in the epiglottis and larynx; however, the process also is evident in the hypopharynx and to some extent in the trachea. On microscopic examination, there is wide separation of the collagen fibers and the glandular elements; vascular congestion and eosinophilic infiltration also are present. Patients dying of vascular collapse without antecedent hypoxia from respiratory insufficiency have visceral congestion with a presumptive loss of intravascular fluid volume. The associated electrocardiographic abnormalities, with or without infarction, in some patients may reflect a primary cardiac event mediated by mast cells (which are prominent near the coronary vessels) or may be secondary to a critical reduction in blood volume.

Gastrointestinal manifestations represent another severe presentation of anaphylaxis, and include nausea, vomiting, crampy abdominal pain, and/or fecal incontinence. Angioedema of the bowel wall may also cause sufficient intravascular volume depletion to precipitate cardiovascular collapse.

Cutaneous manifestations are among the most common presentations of anaphylaxis (>90% of cases). Symptoms include urticarial eruptions, flushing with diffuse erythema, and/or a feeling of generalized warmth. Urticarial eruptions are intensely pruritic and may be localized or disseminated. They may coalesce to form giant hives but seldom persist beyond 48 h.

PREDISPOSING FACTORS AND ETIOLOGY

Because the most dangerous manifestations of anaphylaxis involve the cardiovascular and/or respiratory systems, preexisting asthma and underlying cardiovascular disease could lead to more rapid decompensation from anaphylaxis. Atopy is not generally thought to be a risk factor for anaphylaxis from drug reactions or hymenoptera stings, but is associated with radiocontrast sensitivity, exercise-induced anaphylaxis, idiopathic anaphylaxis, and allergy to foods or latex. Severe hymenoptera-induced anaphylaxis (generally with prominent hypotension) can be a presenting feature of underlying systemic mastocytosis. Hymenoptera allergy is also more likely in patients whose occupations (i.e., beekeepers, trash haulers, and landscape workers) place them in regular proximity to stinging insects. Most commonly, allergen-induced cross-linking of IgE-bound FcεRI receptors on mast cells and basophils initiates the signal transduction events leading to hypersensitivity syndromes including anaphylaxis. The generation of allergen-specific IgE is the end result of sensitization via the adaptive immune system. The mechanisms underlying sensitization are beyond the scope of this topic; however, environmental factors, innate immune responses, and cytokines are among the many variables leading to antigen-specific IgE production by B cells and plasma cells. IgE-mediated drug allergies are most common with antibiotics and certain chemotherapy drugs, though theoretically, they can occur with almost any medication. As is the case with environmental allergies, repeated exposure to the allergy-causing antigen is an important risk factor to keep in mind when evaluating patients with anaphylaxis. In the case of allergy to carboplatin, the incidence of hypersensitivity is 27% in patients who have had ≥7 lifetime infusions and as high as 46% in patients who have had ≥15 lifetime infusions. Similarly, patients with cystic fibrosis have a relatively high incidence of allergic reactions to IV antibiotics that they receive periodically to treat exacerbations of bronchiectasis. Drugs can also function as haptens that form immunogenic conjugates with host proteins. The conjugating hapten may be the parent compound, a nonenzymatically derived storage product, or a metabolite formed in the host. Recombinant biologics can also induce the formation of IgE against the proteins or against glycosylated structures that serve as immunogens. More recently, outbreaks of anaphylaxis to the EGFR antibody, cetuximab, were reported in association with elevated titers of serum IgE to alpha-1,3-galactose (alpha-gal), an oligosaccharide found in non-primate mammals. Cetuximab is derived from a mouse cell line expressing a transferase that tags the Fab' portion of the cetuximab heavy chain with alpha-gal. Interestingly, patients with a history of multiple bites from *Amblyomma americanum* ticks commonly found in the Carolinas, Arkansas, and Tennessee are more likely to have anti-alpha-gal IgE as compared to control patients living outside those states. Such individuals who become sensitized to alpha-gal can develop episodes of delayed anaphylaxis to beef, lamb, and pork.

PATHOPHYSIOLOGY

Many of the important early mediators of anaphylaxis are derived from mast cells, basophils, and eosinophils. Mast cells and basophils contain preformed granules comprised of histamine, proteases (tryptase, chymase), proteoglycans (heparin, chondroitin sulfate), and TNF-α, which are rapidly released into surrounding tissue upon cell activation, a process known as degranulation. Mast cells, basophils, and eosinophils are also sources of arachidonic acid-derived products which include cysteinyl leukotrienes, prostaglandins, and platelet activating factor (PAF). Histamine release results in flushing, urticaria, pruritus, and, in high concentrations, hypotension and tachycardia. Cysteinyl leukotrienes and prostaglandin D_2 cause bronchoconstriction and increased microvascular permeability. Prostaglandin D_2 causes cutaneous flushing, and attracts eosinophils and basophils to the site of mast cell activation. Serum PAF levels correlate with anaphylaxis severity and are inversely proportional to the constitutive level of PAF acetylhydrolase, which is necessary for PAF inactivation. Tryptase and chymase can activate complement and coagulation pathways. Activation of these pathways results in production of the anaphylotoxins, C3a and C5a, and activation of the kallikrein-kinin system which regulates blood pressure and vascular permeability. The actions of these anaphylactic mediators are likely additive or synergistic at the target tissues.

Non-IgE-mediated reactions to certain drugs (which may occur upon the first exposure) can mimic the pathophysiology of IgE-dependent anaphylaxis due to a similar profile of mediators. For example, paclitaxel

is a chemotherapy agent derived from yew tree bark and needles that requires polyethoxylated castor oil (Cremophor) to be solubilized into aqueous solution. Cremophor directly activates the complement cascade, resulting in complement-dependent induced histamine release from mast cells and basophils. A version of paclitaxel that is solubilized by being bound to albumin nanoparticles, Abraxane, has a far lower rate of hypersensitivity, especially for patients who have had infusion reactions to Cremophor-solubilized paclitaxel. Reactions to radiocontrast and vancomycin are other examples of non-IgE-mediated hypersensitivity. Opiates and NSAIDs are other drug categories that can have similar adverse reactions.

DIAGNOSIS

The diagnosis of an anaphylactic reaction depends primarily on a history revealing the onset of symptoms and signs within seconds to minutes after the putative trigger is encountered. An exception is delayed anaphylaxis to meats in alpha-gal sensitized patients. Every attempt to identify the specific cause or causes should be made so as to minimize the risk of recurrent anaphylaxis. If a particular drug or food is suspected, skin or serum specific IgE testing is useful to confirm clinical suspicions. If a specific trigger cannot be identified, a workup of underlying atopic diatheses may be useful to identify risk factors that could play a potential contributory role. In the acute setting, laboratory biomarkers of mast cell degranulation may be useful to document the severity of an anaphylactic episode. The most obvious serum biomarker to assay, histamine, has an extremely short half-life with a measurable time-window that expires <1 h from the onset of anaphylaxis. A more practical and useful biomarker is serum tryptase which peaks 60–90 min after the onset of anaphylaxis and can be measured as long as 5 h after the onset of anaphylaxis. It may be useful to follow-up an elevated tryptase measurement in the acute setting with another measurement when the patient is clinically stable to establish a baseline reference. An elevated baseline tryptase level may warrant further workup for mastocytosis, especially if the presenting reaction occurred in the setting of hymenoptera sting.

TREATMENT

Early recognition of an anaphylactic reaction is mandatory since severe, even fatal, complications, can occur within minutes after symptoms first appear. The treatment of first choice is intramuscular administration of 0.3–0.5 mL of 1:1000 (1 mg/mL) epinephrine, with repeated doses at 5–20 min intervals as needed for a severe reaction. The failure to use epinephrine within the first 20 min of symptoms is a risk factor for poor clinical outcomes in various studies of anaphylaxis. Another important variable that may affect anaphylaxis survival is body posture, as an upright or sitting posture may lead to the "empty heart syndrome" in which there is insufficient venous return to the heart from sudden onset hypotension secondary to intravascular volume depletion. Epinephrine can further accelerate empty heart syndrome due to its chronotropic effects. For this reason, it is recommended that patients who suffer from anaphylaxis be placed in the supine position before receiving epinephrine. IV fluids and vasopressor agents may be administered in the acute medical setting if intractable hypotension occurs. Epinephrine provides both α- and β-adrenergic effects, resulting in vasoconstriction, bronchial smooth-muscle relaxation, and attenuation of enhanced venular permeability. Beta blockers may attenuate this response; therefore, an alternative anti-hypertensive may be considered in patients at high risk of needing emergency epinephrine. Oxygen alone via a nasal catheter or with nebulized albuterol may be helpful; however, either endotracheal intubation or a tracheostomy is mandatory for oxygen delivery if progressive hypoxia develops. Ancillary agents such as antihistamines, glucocorticoids, and bronchodilators are also useful therapeutics to treat urticaria/angioedema and bronchospasm once the patient is hemodynamically stable.

PREVENTION

Avoidance The simplest, most straightforward approach to the long-term management of a patient with a history of anaphylaxis is strict avoidance of known anaphylactic triggers and education on acute management, that is, instructing the patient on the proper use and indications for use of self-administered epinephrine. Lifelong avoidance is not easy if the trigger is an occupational exposure, hymenoptera sting, a common food (i.e., peanut), or a drug representing the sole or best therapeutic option for the patient. Special management options may exist for these patients.

Specific Immunotherapy Patients with large local reactions to hymenoptera stings are unlikely to have anaphylaxis with subsequent stings; however, patients of any age who have had documented anaphylaxis should be formally evaluated and started on venom immunotherapy (VIT) if skin or serologic IgE testing confirms the history. Immunotherapy is a means of "tolerizing" patients to allergen by means of serial subcutaneous administration of escalating doses of extract containing relevant allergen until a target maintenance dose is achieved. As in the case of Richet's unfortunate dogs, anaphylaxis can sometimes occur during the course of administering immunotherapy extracts, so formulating extracts and administering them is typically done under the care of a specialist familiar with this type of treatment. In the case of hymenoptera allergy, patients receive VIT extracts containing actual hymenoptera venom with a maintenance dose equivalent to 2–5 stings. The recommended duration of treatment is 3–5 years; however, patients who have experienced severe respiratory of cardiovascular anaphylaxis are often on lifelong therapy.

Tolerance Induction IgE sensitization to foods occurs most frequently in infants and young children, especially those with atopic dermatitis, and is a risk factor for anaphylaxis (although detection of specific IgE through skin or serum testing has relatively poor predictive value). While most allergy to egg, milk, soy, and/or wheat resolves spontaneously during childhood, ~80% of children with peanut allergy remain sensitive for life. A sharp rise in the prevalence of peanut allergy was also observed in the late 1990s–early 2000s especially in countries with Western diets where the average age of peanut introduction was age ≥3 years. Curiously, in cultures where peanut was introduced much earlier into children's diets, the prevalence of peanut allergy remained low. The landmark "Learning Early About Peanut Allergy" (LEAP) study demonstrated that early introduction of peanut protein to the diet of high risk infants (4–11 months of age with atopic dermatitis and/or egg allergy) can prevent the development of most (80% or more) peanut allergy compared with children who did not consume peanuts (avoidance group), even when IgE sensitization (based on positive skin test) had already developed at the time of study entry. While the induction of tolerance at an early age seems to be key to preventing clinical reactivity later in life, it is not yet clear if this principle holds true for other foods commonly associated with hypersensitivity reactions.

Desensitization For patients who have suffered anaphylaxis from drug allergy and whose treatment regimen requires the administration of the offending drug, desensitization may be a short-term treatment option to prevent reactions. Desensitization elicits a temporary state of tolerance to the drug in sensitized, clinically reactive patients. While it has been a proven technique for penicillin-allergic patients for decades, desensitization has more recently been proven to be effective for certain chemotherapy agents, especially platin-based chemotherapy agents which can induce IgE-mediated sensitization with repeated exposures. The exact mechanisms underlying desensitization are not fully understood; however, temporary tolerance can be achieved through the serial administration of gradually escalating doses of drug, starting from extremely low doses, over the course of hours. So long as the patient continues to receive the drug in question at regular intervals based on drug half-life, a "desensitized" state can also be maintained until the drug is no longer needed. Drug desensitization works best for IgE-mediated reactions; however, it has been performed in cases of non-IgE-mediated anaphylaxis from Cremophor-solubilized paclitaxel as described earlier in this chapter. Other non-IgE-mediated anaphylactic reactions can often be prevented with premedication regimens. A typical premedication regimen for radiocontrast, for example, will have the patient receive prednisone 0.5 mg/kg at 13, 6, and 1 h prior to

contrast administration. Diphenhydramine 25 mg is also given 1 h prior to contrast. Flushing reactions from vancomycin can often be alleviated with antihistamine premedication and down titrating the infusion rate.

FURTHER READING

CASTELLS MC et al: Hypersensitivity reactions to chemotherapy: Outcomes and safety of rapid desensitization in 413 cases. J Allergy Clin Immunol 122:574, 2008.

CHUNG CH et al: Cetuximab-induced anaphylaxis and IgE specific for galactose-alpha-1,3-galactose. N Engl J Med 358:1109, 2008.

DU TOIT G et al: LEAP Study Team. Randomized trial of peanut consumption in infants at risk for peanut allergy. N Engl J Med 373:803, 2015.

DU TOIT G et al: Immune Tolerance Network LEAP-On Study Team. Effect of avoidance on peanut allergy after early peanut consumption. N Engl J Med 374:1435, 2016.

LIEBERMAN P et al: Anaphylaxis—A practice parameter update 2015. Ann Allergy Asthma Immunol 115:341, 2015.

347 Mastocytosis

Cem Akin, Joshua A. Boyce

DEFINITION

Mastocytosis is defined by accumulation of clonally expanded mast cells in tissues such as skin, bone marrow, liver, spleen, and gut. The mast cell expansion is generally recognized in skin and/or bone marrow. Mastocytosis occurs at any age and has a slight preponderance in males. Mastocytosis is a rare disorder and its exact prevalence is not known; however, it is estimated to occur in ~1 in 20,000 people. Familial occurrence is rare, and atopy is not increased compared to the general population.

CLASSIFICATION AND PATHOPHYSIOLOGY

A consensus classification for mastocytosis recognizes cutaneous mastocytosis with variants, five systemic forms, and rare mast cell sarcoma **(Table 347-1)**.

Cutaneous mastocytosis is the most common diagnosis of mastocytosis in children and indicates disease limited to skin with absence of pathologic infiltrates in internal organs. It is usually diagnosed within the first year of life with demonstration of fixed, maculopapular, and hyperpigmented lesions (maculopapular cutaneous mastocytosis [MPCM], formerly known as urticaria pigmentosa), mastocytoma(s) or diffuse cutaneous mastocytosis. Systemic mastocytosis (SM) refers to involvement of a non-cutaneous site (usually bone marrow). There are five distinct variants of SM; the form designated as *indolent systemic mastocytosis* (ISM) accounts for the majority of adult patients. ISM is diagnosed when there is no evidence of an associated hematologic disorder, mast cell leukemia or tissue dysfunction due to mast cell infiltration and is not known to alter life expectancy. Systemic smoldering mastocytosis (formerly considered a subvariant of ISM) is characterized by high mast cell burden as evidenced by a bone marrow infiltration of >30% and a baseline serum tryptase >200 ng/ml (B-findings), but absence of SM-AHNMD or ASM **(Table 347-2)**. In *systemic mastocytosis associated with clonal hematologic non–mast cell lineage disease* (SM-AHNMD, or SM-AHN for short), the prognosis is determined by the nature of the associated disorder, which can range from dysmyelopoiesis to leukemias usually of myeloid origin. In *aggressive systemic mastocytosis* (ASM), mast cell infiltration/proliferation in multiple organs such as liver, spleen, gut, bone, and bone marrow resulting in 1 or more C findings and a poor prognosis (Table 347-2). *Mast cell leukemia* (MCL) is the rarest form of SM and is invariably fatal at present; the peripheral blood contains circulating, metachromatically staining, and atypical mast cells. An aleukemic form of MCL is recognized without circulating mast cells when the percentage of high-grade immature mast cells in bone marrow smears exceeds 20% in a nonspicular area. Mast cell sarcoma is a rare solid mast cell tumor with malignant invasive features.

A point mutation of A to T at codon 816 of *KIT* that causes an aspartic acid to valine substitution, resulting in a somatic gain-of-function mutation, is found in mast cells and sometimes in multiple other cell lineages in patients with mastocytosis. This substitution, as well as other rare mutations of *KIT*, is characteristic of patients with all forms of SM, but is also present in some children with cutaneous mastocytosis in lesional skin, as might be anticipated because mast cells are of bone marrow lineage. Additional mutations in genes such as *TET2, SRSF2, ASLX1*, and *RUNX1* known to be associated with other hematologic neoplastic disorders can be detected in patients usually with advanced (non-ISM) forms of SM. The prognosis for patients with cutaneous mastocytosis and for almost all patients with ISM is a normal life expectancy, whereas that for patients with SM-AHNMD is determined by the non–mast cell component. ASM and MCL carry a poorer prognosis, while patients with SSM carry an intermediate prognosis. Progression from ISM to a more advanced form is rare (approximately 3% overall); however, patients should be monitored for emergence of hematologic disease and end organ manifestations of ASM. In infants and children with cutaneous manifestations, namely, maculopapular cutaneous mastocytosis, mastocytoma(s), or bullous lesions, visceral involvement is usually lacking, and spontaneous resolution is common prior to adolescence. Progression from CM to ISM may occur in ~10% of children, especially in those with high mast cell burden (diffuse cutaneous mastocytosis), hematologic abnormalities and those who present with smaller uniform lesions with diameters measuring <2 cm.

TABLE 347-1 Classification of Mastocytosis

Cutaneous mastocytosis (CM)
Maculopapular cutaneous mastocytosis (MPCM)
Solitary mastocytoma of skin
Diffuse cutaneous mastocytosis
Indolent systemic mastocytosis (ISM)
Smoldering systemic mastocytosis
Systemic mastocytosis with an associated clonal hematologic non–mast cell lineage disease (SM-AHNMD)
Aggressive systemic mastocytosis (ASM)
Mast cell leukemia (MCL)
Mast cell sarcoma (MCS)

Source: Modified from H-P Horny et al: Mastocytosis. In: *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*, revised 4th ed. SH Swerdlow et al (eds). Lyon, France, IARC Press, 2017, pp 61–69.

TABLE 347-2 B and C Findings for Diagnosis of SSM and ASM

B-Findings (2 or more in the absence of any C findings are required for a diagnosis of SSM):

1. MC infiltration in bone marrow biopsy of >30% and the basal serum tryptase level >200 ng/mL
2. Hypercellular bone marrow with signs of dysmyelopoiesis but without cytopenias meeting C criteria or WHO criteria for an MDS or MPN
3. Palpable hepatomegaly, palpable splenomegaly, or lymphadenopathy (on CT or ultrasound: >2 cm) without impaired liver function or hypersplenism

C-Findings: (1 or more required for a diagnosis of ASM). C finding should be reasonably attributable to high tissue mast cell infiltration.

1. Cytopenia(s): ANC<1,000/μL or Hb<10 g/dL or PLT<100,000/μL
2. Hepatomegaly with ascites and impaired liver function
3. Palpable splenomegaly with associated hypersplenism
4. Malabsorption with hypoalbuminemia and weight loss
5. Skeletal lesions: large area(s) of osteolyses with pathologic fractures (presence of osteoporosis alone without osteolytic lesions does not satisfy this criterion)

Abbreviations: ANC, absolute neutrophil count; ASM, aggressive systemic mastocytosis; CT, computed tomography; Hb, hemoglobin; MC, mast cells; MDS, myelodysplastic syndromes; MPN, myeloproliferative disorders; PLT, platelets; SSM, smoldering systemic mastocytosis; WHO, World Health Organization.

CLINICAL MANIFESTATIONS

The clinical manifestations of SM, distinct from a leukemic complication, are due to the release of bioactive substances acting at both local and distal sites, tissue occupancy by the mast cell mass, and the tissue response to that mass. The pharmacologically induced manifestations are intermittent flushing, tachycardia and vascular collapse, gastric distress, lower abdominal crampy pain, and diarrhea. The increase in local cell burden is evidenced by the lesions of MPCM (urticaria pigmentosa) at skin sites and internal organ biopsies such as bone marrow and gastrointestinal tract and may be a direct local cause of bone pain and/or malabsorption. Mast cell–mediated fibrotic changes may occur in liver, spleen, and bone marrow but not in gastrointestinal tissue or skin. Immunofluorescent analysis of bone marrow and skin lesions in ISM and of spleen, lymph node, and skin in ASM has revealed only one mast cell phenotype, namely, scroll-poor cells expressing tryptase, chymase, and CPA.

The cutaneous lesions of MPCM (formerly known as urticaria pigmentosa) are reddish-brown macules, papules, or plaques that respond to trauma with urtication and erythema (Darier's sign). Children with CM may present with MPCM, mastocytomas, or diffuse cutaneous mastocytosis (DCM). Mastocytomas are generally solitary elevated lesions that are yellow, brown, or red in color. Their size may vary from a few millimeters to several centimeters. Rubbing or irritation of the mastocytoma lesion may lead to systemic symptoms such as flushing and urticaria. Children with DCM present without distinct lesions, but rather a generalized thickening of skin (pachydermia) due to diffuse mast cell infiltration. DCM may be associated with bullae formation and more severe systemic symptoms including upper GI irritation and vascular collapse in the first few years of life. Maculopapular skin lesions of mastocytosis may be present in patients with adult-onset systemic disease. The apparent incidence of cutaneous lesions is ≥80% in patients with ISM and <50% in those with SM-AHNMD or ASM. In the upper gastrointestinal tract, gastritis and peptic ulcer are significant problems. In the lower intestinal tract, the occurrence of diarrhea and abdominal pain is attributed to increased motility due to mast cell mediators; this problem can be aggravated by malabsorption, which can also cause secondary nutritional insufficiency and osteomalacia. The periportal fibrosis associated with mast cell infiltration and a prominence of eosinophils may lead to portal hypertension and ascites. In some patients, anaphylaxis with rapid and life threatening vascular collapse can be induced by hymenoptera stings. These patients often have evidence of venom specific IgE. The neuropsychiatric disturbances are clinically most evident as impaired recent memory, decreased attention span, and "migraine-like" headaches. Patients may experience exacerbation of a specific clinical sign or symptom variably with alcohol ingestion, temperature changes, stress, use of mast cell–interactive opioids, or ingestion of NSAIDs.

DIAGNOSIS

Cutaneous mastocytosis is diagnosed by observing the characteristic lesions of MPCM or mastocytoma(s). A skin biopsy can be obtained to confirm these subvariants of CM, whereas patients with suspected DCM and bullous mastocytosis usually require a skin biopsy to confirm the diagnosis. Although the diagnosis of SM is generally suspected on the basis of the clinical history and physical findings, and can be supported by laboratory procedures, it can be established only by a tissue diagnosis. By convention, the diagnosis of SM depends heavily on bone marrow biopsy to meet the criteria of one major plus one minor or three minor findings **(Table 347-3)**. The bone marrow provides the major criterion by revealing aggregates of mast cells, often in paratrabecular and perivascular locations with lymphocytes and eosinophils, as well as the minor criteria of abnormal mast cell morphology, aberrant mast cell membrane immunophenotype, or a codon 816 mutation in an extracutaneous tissue. A basal serum total tryptase level is a noninvasive approach to consider before bone marrow biopsy. The pro-β and α forms of tryptase are elevated in more than one-half of patients with SM and provide a minor criterion; the fully processed ("mature") β form is increased in patients undergoing an anaphylactic reaction. A rare histolopathologic subvariant called "well differentiated systemic mastocytosis" (WDSM) is characterized by clusters of mature appearing fully granulated and round mast cells, lack of aberrant CD25 and CD2 expression, and lack of D816V KIT mutation in most patients. These patients often have a history of childhood onset cutaneous disease and their mast cells may display aberrant CD30 expression and other markers of clonality such as atypical (non-D816V) KIT mutations. Additional studies directed by the presentation include a bone densitometry, bone scan, or skeletal survey; computed tomography scan, or endoscopy; and a neuropsychiatric evaluation. Osteoporosis is increased in mastocytosis and may lead to pathologic fractures.

TABLE 347-3 Diagnostic Criteria for Systemic Mastocytosis[a]

Major:
Multifocal dense infiltrates of mast cells (>15 mast cells per aggregate) in bone marrow or other extracutaneous tissues
Minor:
Abnormal mast cell morphology (spindle shape, bi- or multi-lobed or eccentric nucleus, hypogranulated cytoplasm)
Aberrant mast cell surface phenotype with expression of CD25 (IL-2 receptor alpha chain) and/or CD2
Detection of codon 816 mutation in peripheral blood cells, bone marrow cells, or an extracutaneous lesional tissue
Total serum tryptase >20 ng/mL

[a]Diagnosis requires either the major criterion and one minor criterion or three minor criteria.

Some patients presenting with recurrent mast cell activation symptoms (particularly hypotensive syncopal anaphylactic episodes) have been found to have underlying mastocytosis. A subset of these patients may be found to have the D816V *KIT* mutation or aberrant mast cells displaying CD25, but lack other diagnostic criteria for SM. Such patients are termed to have "monoclonal mast cell activation syndrome."

The differential diagnosis requires the exclusion of other flushing disorders. The 24-h urine assessment of 5-hydroxy-indoleacetic acid and metanephrines should exclude a carcinoid tumor or a pheochromocytoma, respectively. Some patients presenting with recurrent mast cell activation symptoms without an obvious increase in mast cell burden in skin or bone marrow have been shown to carry aberrant mast cells with clonality markers of D816V *KIT* mutation or surface CD25 expression. Most patients with recurrent IgE-induced or idiopathic anaphylaxis present with urticaria, angioedema, and/or wheezing, which are not manifestations of SM.

TREATMENT

Mastocytosis

The management of SM uses a stepwise and symptom/sign–directed approach that includes an H_1 antihistamine for flushing and pruritus, an H_2 antihistamine or proton pump inhibitor for gastric acid hypersecretion, oral cromolyn sodium for diarrhea and abdominal pain, and occasionally aspirin (in those who are known to be tolerant of NSAIDs) for severe flushing with or without associated vascular collapse, despite use of H_1 and H_2 antihistamines, to block biosynthesis of PGD_2. Systemic glucocorticoids appear to alleviate the malabsorption. Mast cell cytoreductive therapy consisting of midostaurin, IFN-α or cladribine is generally reserved for advanced, nonindolent variants of SM. Midostaurin is a multi-kinase inhibitor with activity against D816V mutated and wild type KIT, and was recently approved by Food and Drug Administration for treatment of advanced systemic mastocytosis (SM-AHNMD, ASM, and MCL), and should be considered as a first-line therapy of these disease variants. The efficacy of cytoreductive therapy in mastocytosis is variable, perhaps because of dosage limitations due to side effects. Imatinib is not effective in most cases as D816V *KIT* mutation provides resistance against it. Combination chemotherapy is appropriate for the frank leukemias. Stem cell transplantation has shown to be effective in a small subset of patients with advanced mastocytosis.

A self-injectable epinephrine prescription is recommended for most patients due to increased incidence of anaphylaxis. Patients with a history of systemic hymenoptera venom reaction should be tested for venom specific IgE and placed on lifelong venom immunotherapy. Other investigational tyrosine kinase inhibitors with a capacity to inhibit D816V *KIT* mutation are currently in clinical trials.

FURTHER READING

Arock M et al: Current treatment options in patients with mastocytosis: Status in 2015 and future perspectives. Eur J Haematol 94:474, 2014.

Gotlib J et al: Efficacy and safety of midostaurin in advanced systemic mastocytosis. N Engl J Med 374:2530, 2016.

Hartmann K et al: Cutaneous manifestations in patients with mastocytosis: Consensus report of the European Competence Network on Mastocytosis; the American Academy of Allergy, Asthma & Immunology; and the European Academy of Allergology and Clinical Immunology. J Allergy Clin Immunol 137:35, 2016.

Horny H-P et al: Mastocytosis (mast cell disease). In: *WHO Classification of Tumours. Pathology & Genetics. Tumours of Haematopoietic and Lymphoid Tissues*. SH Swerdlow, et al (eds). Lyon, France, IARC Press, 2008, pp 54–63.

Theoharides TC et al: Mast cells, mastocytosis, and related disorders. N Engl J Med 373:163, 2015.

Ustun C et al: Consensus opinion on allogeneic hematopoietic cell transplantation in advanced systemic Mastocytosis. Biol Blood Marrow Transplant 22:1348, 2016.

Valent P et al: Standards and standardization in mastocytosis: Consensus statements on diagnostics, treatment recommendations and response criteria. Eur J Clin Invest 37:435, 2007.

Valent P et al: European Competence Network on Mastocytosis. Proposed diagnostic algorithm for patients with suspected mastocytosis: A proposal of the European Competence Network on Mastocytosis. Allergy 69:1267, 2014.

348 Autoimmunity and Autoimmune Diseases

Betty Diamond, Peter E. Lipsky

One of the central features of the immune system is the capacity to mount an inflammatory response to potentially harmful foreign materials while avoiding damage to self-tissues. Whereas recognition of self plays an important role in shaping the repertoires of immune receptors on both T and B cells and in clearing apoptotic and other tissue debris from sites throughout the body, the development of potentially harmful immune responses to self-antigens is, in general, prohibited. The essential feature of an *autoimmune disease* is that tissue injury is caused by the immunologic reaction of the organism against its own tissues. *Autoimmunity*, on the other hand, refers merely to the presence of antibodies or T lymphocytes that react with self-antigens and does not necessarily imply that the self-reactivity has pathogenic consequences. Autoimmunity is present in all individuals and increases with age; however, autoimmune disease occurs only in those individuals in whom the breakdown of one or more of the basic mechanisms regulating immune tolerance results in self-reactivity that can cause tissue damage.

Polyreactive autoantibodies that recognize many host antigens are present throughout life. These antibodies are usually of the IgM heavy chain isotype and are encoded by nonmutated germline immunoglobulin variable region genes. These antibodies are essential, as they remove apoptotic debris through non inflammatory pathways. Expression of these autoantibodies may be increased after some inciting events. When autoimmunity is induced by an inciting event, such as infection or tissue damage from trauma or ischemia, the autoreactivity is in general self-limited. When such autoimmunity does persist, however, pathology may or may not result. Even in the presence of organ pathology, it may be difficult to determine whether the damage is mediated by autoreactivity. After an inciting event, the development of self-reactivity may be the consequence of an ongoing pathologic process, may be nonpathogenic, or may exacerbate tissue inflammation and damage. Individuals with autoimmune disease may have numerous autoantibodies, only some or even none of which may be pathogenic. For example, patients with systemic sclerosis may have a wide array of antinuclear antibodies that are important in disease classification but are not clearly pathogenic; in contrast, patients with pemphigus may also exhibit a wide array of autoantibodies, one of which (antibody to desmoglein 1 and 3) is known to be pathogenic.

MECHANISMS OF AUTOIMMUNITY

Since Ehrlich first postulated the existence of mechanisms to prevent the generation of self-reactivity in the early 1900s, there has been a progressive increase in understanding of this prohibition in parallel with a progressive increase in understanding of the immune system. Burnet's clonal selection theory included the idea that interaction of lymphoid cells with their specific antigens during fetal or early postnatal life would lead to elimination of such "forbidden clones." This idea was refuted, however, when it was shown that autoimmune diseases could be induced in experimental animals by simple immunization procedures, that autoantigen-binding cells could be demonstrated easily in the circulation of normal individuals, and that self-limited autoimmune phenomena frequently developed after tissue damage from infection or trauma. These observations indicated that clones of cells capable of responding to autoantigens were present in the repertoire of antigen-reactive cells in normal adults and suggested that mechanisms in addition to clonal deletion were responsible for preventing their activation.

Currently, three general processes are thought to be involved in the maintenance of selective unresponsiveness to autoantigens **(Table 348-1)**: (1) sequestration of self-antigens, rendering them inaccessible to the immune system; (2) specific unresponsiveness (tolerance or anergy) of relevant T or B cells; and (3) limitation of potential reactivity by regulatory mechanisms. Derangements of these normal processes may predispose to the development of autoimmunity **(Table 348-2)**. In general, these abnormal responses require both an exogenous trigger, such as bacterial or viral infection or cigarette smoking, and the presence of endogenous abnormalities in the cells of the immune system. Microbial superantigens, such as staphylococcal protein A and staphylococcal enterotoxins, are substances that can stimulate a broad range of T and B cells through specific interactions with selected families of immune receptors, irrespective of their antigen specificity. If autoantigen-reactive T and/or B cells express these receptors, autoimmunity may develop. Alternatively, molecular mimicry or cross-reactivity between a microbial product and a self-antigen may lead to activation of autoreactive lymphocytes. One of the best examples of autoreactivity and autoimmune disease resulting from molecular mimicry is rheumatic fever,

TABLE 348-1 Mechanisms Preventing Autoimmunity

1. Sequestration of self-antigens
2. Generation and maintenance of tolerance
   a. Central deletion of autoreactive lymphocytes
   b. Peripheral anergy of autoreactive lymphocytes
   c. Receptor replacement in autoreactive lymphocytes
3. Regulatory mechanisms
   a. Regulatory T cells
   b. Regulatory B cells
   c. Regulatory mesenchymal cells
   d. Regulatory cytokines
   e. Idiotype network

TABLE 348-2 Mechanisms of Autoimmunity

- I. Exogenous
  - A. Molecular mimicry
  - B. Superantigenic stimulation
  - C. Microbial and tissue damage–associated adjuvanticity
- II. Endogenous
  - A. Altered antigen presentation
    - 1. Loss of immunologic privilege
    - 2. Presentation of novel or cryptic epitopes (epitope spreading)
    - 3. Alteration of self-antigen
    - 4. Enhanced function of antigen-presenting cells
      - a. Costimulatory molecule expression
      - b. Cytokine production
  - B. Increased T cell help
    - 1. Cytokine production
    - 2. Costimulatory molecules
  - C. Increased B cell function
    - 1. B cell activating factor
    - 2. Costimulatory molecules
  - D. Apoptotic defects or defects in clearance of apoptotic material
  - E. Cytokine imbalance
  - F. Altered immunoregulation

in which antibodies to the M protein of streptococci cross-react with myosin, laminin, and other matrix proteins as well as with neuronal antigens. Deposition of these autoantibodies in the heart initiates an inflammatory response, whereas their penetration into the brain can result in Sydenham's chorea. Molecular mimicry between microbial proteins and host tissues has been reported in type 1 diabetes mellitus, rheumatoid arthritis, systemic lupus erythematosus (SLE), celiac disease, and multiple sclerosis. It is presumed that infectious agents may be able to overcome self-tolerance because they possess *pathogen-associated molecular patterns* (PAMPs). These molecules (e.g., bacterial endotoxin, RNA, or DNA) exert adjuvant-like effects on the immune system by interacting with *Toll-like receptors* (TLRs) and other *pattern recognition receptors* (PRRs) that increase the immunogenicity and immunostimulatory capacity of the microbial material. The adjuvants activate dendritic cells, which in turn stimulate the activation of previously quiescent lymphocytes that recognize both microbial antigens and self-antigens. Similarly, cellular and tissue damage due to the release of *damage-associated molecular patterns* (DAMPs), including DNA, RNA nucleosomes, and other tissue debris, may activate cells of the inflammatory and immune systems through engagement of the same array of PRRs. This pathway may lead to autoimmune disease in individuals who have impairments in mechanisms for clearance of tissue debris.

Although previous work focused on the role of pathogenic microorganisms in triggering autoimmunity, more recent studies have focus on the role of the microbiome, the collection of non-pathogenic microorganisms that reside on various body surfaces. It has become clear that the interaction between specific constituents of these microbiota and the immune system can shape the nature of the immune response to either favor or discourage immune/inflammatory responses. Thus, some genera within the microbiome may favor a nonresponsive state dominated by regulatory T cells whereas others may favor the development of T effector cells and a proinflammatory state. Gender bias in autoimmune conditions may also be favored by differences in the dominant organisms within the microbiome.

Endogenous derangements of the immune system may also contribute to the loss of immunologic tolerance to self-antigens and the development of autoimmunity (Table 348-2). Some autoantigens reside in immunologically privileged sites, such as the brain or the anterior chamber of the eye. These sites are characterized by the inability of engrafted tissue to elicit immune responses. Immunologic privilege results from a number of events, including the limited entry of proteins from those sites into lymphatics, the local production of immunosuppressive cytokines such as transforming growth factor β, and the local expression of molecules (including Fas ligand) that can induce apoptosis of activated T cells. Lymphoid cells remain in a state of immunologic ignorance (neither activated nor anergized) with regard to proteins expressed uniquely in immunologically privileged sites. If the privileged site is damaged by trauma or inflammation or if T cells are activated elsewhere, proteins expressed at this site can become immunogenic and also be the targets of immunologic assault. In multiple sclerosis and sympathetic ophthalmia, for example, antigens uniquely expressed in the brain and eye, respectively, become the target of activated T cells.

Alterations in antigen presentation may also contribute to autoimmunity. Peptide determinants (*epitopes*) of a self-antigen that are not routinely presented to lymphocytes may be recognized as a result of altered proteolytic processing of the molecule and the ensuing presentation of novel peptides (*cryptic epitopes*). When B cells rather than dendritic cells present self-antigen, they may also present cryptic epitopes that can activate autoreactive T cells. These cryptic epitopes will not previously have been available to effect the silencing of autoreactive lymphocytes. Furthermore, once there is immunologic recognition of one protein component of a multimolecular complex, reactivity may be induced to other components of the complex after internalization and presentation of all molecules within the complex (epitope spreading). Finally, inflammation, environmental agents, drug exposure, or normal senescence may cause a post-translational alteration in proteins, resulting in the generation of immune responses that cross-react with normal self-proteins. For example, the induction and/or release of protein arginine deiminase enzymes results in the conversion of arginine residues to citrullines in a variety of proteins, thereby altering their capacity to induce immune responses. Production of antibodies to citrullinated proteins has been observed in rheumatoid arthritis and chronic lung disease as well as in normal smokers. These antibodies may be the target of autoantibodies that contribute to organ pathology. Alterations in the availability and presentation of autoantigens may be important components of immunoreactivity in certain models of organ-specific autoimmune diseases. In addition, these factors may be relevant to an understanding of the pathogenesis of various drug-induced autoimmune conditions. However, the diversity of autoreactivity manifesting in non-organ-specific systemic autoimmune diseases suggests that these conditions may result from a more general activation of the immune system rather than from an alteration in individual self-antigens.

Many autoimmune diseases are characterized by the presence of antibodies that react with antigens present in apoptotic material. Defects in the clearance of apoptotic material have been shown to elicit autoimmunity and autoimmune disease in a number of animal models. Moreover, such defects have been found in patients with SLE. Apoptotic debris that is not cleared quickly by the immune system can function as endogenous ligands for a number of PRRs on dendritic cells and B cells. Under such circumstances, dendritic cells and/or B cells are activated, and an immune response to apoptotic debris can develop. In addition, the presence of uncleared extracellular apoptotic material within germinal centers of secondary lymphoid organs in patients with SLE may facilitate the direct activation of autoimmune B cell clones or may function to select such clones during immune responses.

Deficiency in C1q, likewise, can predispose or exacerbate autoimmunity. C1q assists in the clearance of apoptotic debris binding to IgM autoantibodies and to inhibitory receptors on monocytes and dendritic cells. If C1q is not present, a mechanism of immune suppression is lost. Moreover, if antibodies have undergone class switch recombination to IgG, the apoptotic debris containing immune complexes will engage activating Fc receptors on myeloid cells to induce an inflammatory response. Studies in a number of experimental models have suggested that intense stimulation of T lymphocytes can produce nonspecific signals that bypass the need for antigen-specific helper T cells and lead to polyclonal B cell activation with the formation of multiple autoantibodies. For example, antinuclear, antierythrocyte, and antilymphocyte antibodies are produced during the chronic graft-versus-host reaction. In addition, true autoimmune diseases, including autoimmune hemolytic anemia and immune complex–mediated glomerulonephritis, can

be induced in this manner. While such diffuse activation of helper T cell activity clearly can cause autoimmunity, nonspecific stimulation of B lymphocytes can also lead to the production of autoantibodies. Thus, the administration of polyclonal B cell activators, such as bacterial endotoxin, to normal mice leads to the production of a number of autoantibodies, including those to DNA and IgG (rheumatoid factor). A variety of genetic modifications resulting in hyperresponsiveness of B cells also can lead to the production of autoantibodies and, in animals of appropriate genetic background, a lupus-like syndrome. Moreover, excess B cell activating factor (BAFF), a B cell survival promoting cytokine, can impair B cell tolerance, cause T cell–independent B cell activation, and lead to the development of autoimmunity. SLE can also be induced in mice through exuberant dendritic cell activation, through a redundancy of TLR7 on the Y chromosome (as in BXSB-Yaa mice), or through exposure to CpG, a ligand for TLR9. The ensuing induction of inflammatory mediators can cause a switch from the production of nonpathogenic IgM autoantibodies to the production of pathogenic IgG autoantibodies in the absence of antigen-specific T cell help. Aberrant selection of the B or T cell repertoire at the time of antigen receptor expression can also predispose to autoimmunity. For example, B cell immunodeficiency caused by an absence of the B cell receptor–associated kinase (Bruton's tyrosine kinase) leads to X-linked agammaglobulinemia. This syndrome is characterized by reduced B cell numbers. This leads to high levels of BAFF which alter B cell selection and results in greater survival of autoreactive B cells. Likewise, negative selection of autoreactive T cells in the thymus requires expression of the autoimmune regulator (AIRE) gene that enables the expression of tissue-specific proteins in thymic medullary epithelial cells. Peptides from these proteins are expressed in the context of major histocompatibility complex (MHC) molecules and mediate the central deletion of autoreactive T cells. The absence of AIRE gene expression leads to a failure of negative selection of autoreactive cells, autoantibody production, and severe inflammatory destruction of multiple organs. Individuals deficient in AIRE gene expression develop autoimmune polyendocrinopathy–candidiasis–ectodermal dystrophy (APECED).

Primary alterations in the activity of T and/or B cells, cytokine imbalances, or defective immunoregulatory circuits may also contribute to the emergence of autoimmunity. Diminished production of tumor necrosis factor (TNF) and interleukin (IL) 10 has been reported to be associated with the development of autoimmunity. Overproduction or therapeutic administration of type 1 interferon has also been associated with autoimmunity. Overexpression of costimulatory molecules on T cells similarly can lead to autoantibody production.

Autoimmunity may also result from an abnormality of immunoregulatory mechanisms. Observations made in both human autoimmune disease and animal models suggest that defects in the generation and expression of regulatory T cell (Treg) activity may allow the production of autoimmunity. It has been appreciated that the IPEX (immunodysregulation, polyendocrinopathy, enteropathy X-linked) syndrome results from the failure to express the FOXP3 gene, which encodes a molecule critical in the differentiation of Tregs. Administration of normal Tregs or of factors derived from them can prevent the development of autoimmune disease in rodent models of autoimmunity, and allogeneic stem cell transplantation ameliorates human IPEX. Abnormalities in the function of Tregs have been noted in a number of human autoimmune diseases, including rheumatoid arthritis and SLE, although it remains uncertain whether these functional abnormalities are causative or are secondary to inflammation. One of the mechanisms by which Tregs control immune/inflammatory responses is by the production of the cytokine IL-10. In this regard, children with a deficiency in the expression of IL-10 or the IL-10 receptor develop inflammatory bowel disease that mimics Crohn's disease and that can be cured by allogeneic stem cell transplantation. Finally, recent data indicate that B cells may also exert regulatory function, largely through the production of IL-10. Deficiency of IL-10-producing regulatory B cells can prolong the course of multiple sclerosis in an animal model, and such cells are thought to be functionally diminished in human SLE.

It should be apparent that no single mechanism can explain all the varied manifestations of autoimmunity or autoimmune disease. Furthermore, genetic evaluation has shown that convergence of a number of abnormalities is often required for the induction of an autoimmune disease. Additional factors that appear to be important determinants in the induction of autoimmunity include age, sex (many autoimmune diseases are far more common in women), exposure to infectious agents, and environmental contacts. How all of these disparate factors affect the capacity to develop self-reactivity is currently being investigated intensively.

GENETIC CONSIDERATIONS

Evidence in humans that there are susceptibility genes for autoimmunity comes from family studies and especially from studies of twins. Studies in type 1 diabetes mellitus, rheumatoid arthritis, multiple sclerosis, and SLE have shown that ~15–30% of pairs of monozygotic twins show disease concordance, whereas the figure is <5% for dizygotic twins. The occurrence of different autoimmune diseases within the same family has suggested that certain susceptibility genes may predispose to a variety of autoimmune diseases. Genome-wide association studies have begun to identify polymorphisms in individual genes that are associated with specific autoimmune diseases. More than 100 genetic polymorphisms associated with one or more autoimmune diseases have been identified to date. It is notable that some genes are associated with multiple autoimmune diseases, whereas others are specifically associated with only one autoimmune condition. Moreover, recent genetic evidence suggests that clusters of genetic risk factors can commonly be found in groups of autoimmune diseases. Four general clusters have been identified: one group most frequently associated with Crohn's disease, psoriasis, and multiple sclerosis; a second cluster most strongly associated with celiac disease, rheumatoid arthritis, and SLE; a third cluster most strongly associated with type 1 diabetes, multiple sclerosis, and rheumatoid arthritis; and a fourth cluster most strongly associated with type 1 diabetes, rheumatoid arthritis, celiac disease, Crohn's disease, and SLE. These results imply that autoimmune diseases with widely different clinical presentations and patterns of organ involvement could involve similar immunopathogenic pathways or endophenotypes. For example, the same allele of the gene encoding PTPN22 is associated with multiple autoimmune diseases. Its product is a phosphatase expressed by a variety of hematopoietic cells that downregulates antigen receptor–mediated stimulation of T and B cells. The risk allele is associated with type 1 diabetes mellitus, rheumatoid arthritis, and SLE in some populations. In recent years, genome-wide association studies have demonstrated a variety of other genes that are involved in human autoimmune diseases. Most genes individually confer a relatively low risk for autoimmune diseases and are found in normal individuals. In addition, most polymorphisms associated with autoimmune diseases are in noncoding regions of DNA, implying that expression levels rather that altered function might convey most genetic risk for autoimmune diseases. Abnormalities in epigenetics or the mechanisms controlling and influencing gene expression has also been implicated in contributing to autoimmune diseases. No single gene or epigenetic modification has been identified that is essential for autoimmune diseases. In addition to this evidence from humans, certain inbred mouse strains reproducibly develop specific spontaneous or experimentally induced autoimmune diseases, whereas others do not. These findings have led to an extensive search for genes that determine susceptibility to autoimmune disease and for genes that might be protective.

The strongest consistent association for susceptibility to autoimmune disease is with particular MHC alleles. It has been suggested that the association of MHC genotype with autoimmune disease relates to differences in the ability of different allelic variations of MHC molecules to present autoantigenic peptides to autoreactive T cells. An alternative hypothesis involves the role of MHC alleles in shaping the T cell receptor repertoire during T cell ontogeny in the thymus. In addition, specific MHC gene products may themselves be the source of peptides that can be recognized by T cells. Cross-reactivity between such MHC peptides and peptides derived from proteins produced by common microbes may trigger autoimmunity by molecular mimicry. However, MHC genotype alone does not determine the development

of autoimmunity. Identical twins are far more likely to develop the same autoimmune disease than MHC-identical nontwin siblings; this observation suggests that genetic factors other than the MHC affect disease susceptibility. Studies of the genetics of type 1 diabetes mellitus, SLE, rheumatoid arthritis, and multiple sclerosis in humans and mice have identified several independently segregating disease susceptibility loci in addition to the MHC. Genes that encode molecules of the innate immune response are also involved in autoimmunity. In humans, inherited homozygous deficiency of the early proteins of the classic pathway of complement (C1q, C4, or C2) as well as genes involved in the type 1 interferon pathway are very strongly associated with the development of SLE.

IMMUNOPATHOGENIC MECHANISMS IN AUTOIMMUNE DISEASES

The mechanisms of tissue injury in autoimmune diseases can be divided into antibody-mediated and cell-mediated processes. Representative examples are listed in Table 348-3.

The pathogenicity of autoantibodies can be mediated through several mechanisms, including opsonization of soluble factors or cells, activation of an inflammatory cascade via the complement system, and interference with the physiologic function of soluble molecules or cells.

In autoimmune thrombocytopenic purpura, opsonization of platelets targets them for elimination by phagocytes. Likewise, in autoimmune hemolytic anemia, binding of immunoglobulin to red cell membranes leads to phagocytosis and lysis of the opsonized cell. Goodpasture's syndrome, a disease characterized by lung hemorrhage and severe glomerulonephritis, represents an example of antibody binding leading to local activation of complement and neutrophil accumulation and activation. The autoantibody in this disease binds to the α_3 chain of type IV collagen in the basement membrane. In SLE, activation of the complement cascade at sites of immunoglobulin deposition in renal glomeruli is considered to be a major mechanism of renal damage. Moreover, the DNA- and RNA-containing immune complexes in SLE activate TLR9 and TLR7, respectively, in plasmacytoid dendritic cells and promote the production of type 1 interferon and proinflammatory, cytokines conducive to amplification of the autoimmune response.

Autoantibodies can also interfere with normal physiologic functions of cells or soluble factors. Autoantibodies to hormone receptors can lead to stimulation of cells or to inhibition of cell function through interference with receptor signaling. For example, long-acting thyroid stimulators—autoantibodies that bind to the receptor for thyroid-stimulating hormone (TSH)—are present in Graves' disease and function as agonists, causing the thyroid to respond as if there were an excess of TSH. Alternatively, antibodies to the insulin receptor can cause insulin-resistant diabetes mellitus through receptor blockade. In myasthenia gravis, autoantibodies to the acetylcholine receptor can be detected in 85–90% of patients and are responsible for muscle weakness. The exact location of the antigenic epitope, the valence and affinity of the antibody, and perhaps other characteristics determine whether activation or blockade results from antibody binding.

Antiphospholipid antibodies are associated with thromboembolic events in primary and secondary antiphospholipid syndrome and have also been associated with fetal wastage. The major antibody is directed to the phospholipid–β_2-glycoprotein I complex and appears to exert a procoagulant effect. In pemphigus vulgaris, autoantibodies bind to desmoglein 1 and 3, components of the epidermal cell desmosome, and play a role in the induction of the disease. These antibodies exert their pathologic effect by disrupting cell–cell junctions through stimulation of the production of epithelial proteases, with consequent blister formation. Cytoplasmic antineutrophil cytoplasmic antibody (c-ANCA), found in granulomatosis with polyangiitis, is an antibody to an intracellular antigen, the 29-kDa serine protease (proteinase-3). In vitro experiments have shown that IgG anti-c-ANCA causes cellular activation and degranulation of primed neutrophils.

It is important to note that autoantibodies of a given specificity may cause disease only in genetically susceptible hosts, as has been shown in experimental models of myasthenia gravis, SLE, rheumatic fever, and rheumatoid arthritis. Furthermore, once organ damage is initiated, new inflammatory cascades are initiated that can sustain and amplify the autoimmune process. Finally, some autoantibodies seem to be markers for disease but have, as yet, no known pathogenic potential.

AUTOIMMUNE DISEASES

Manifestations of autoimmunity are found in a large number of pathologic conditions. However, their presence does not necessarily imply that the pathologic process is an autoimmune disease. A number of attempts to establish formal criteria for the classification of diseases as autoimmune have been made, but none is universally accepted. One set of criteria is shown in Table 348-4; however, this scheme should be viewed merely as a guide in consideration of the problem.

To classify a disease as autoimmune, it is necessary to demonstrate that the immune response to a self-antigen causes the observed pathology. Initially, the detection of antibodies to the affected tissue in the serum of patients suffering from various diseases was taken as evidence that these diseases had an autoimmune basis. However, such autoantibodies are also found when tissue damage is caused by trauma or infection and in these cases are secondary to tissue damage. Thus, autoimmunity must be shown to be pathogenic before a disease is categorized as autoimmune.

TABLE 348-3 Mechanisms of Tissue Damage in Autoimmune Disease

EFFECTOR	MECHANISM	TARGET	DISEASE
Autoantibody	Blocking or inactivation	α Chain of the nicotinic acetylcholine receptor	Myasthenia gravis
		Phospholipid–β_2-glycoprotein I complex	Antiphospholipid syndrome
		Insulin receptor	Insulin-resistant diabetes mellitus
		Intrinsic factor	Pernicious anemia
	Stimulation	TSH receptor (LATS)	Graves' disease
		Proteinase-3 (ANCA)	Granulomatosis with polyangiitis
		Epidermal cadherin	Pemphigus vulgaris
		Desmoglein 3	
	Complement activation	α_3 Chain of collagen IV	Goodpasture's syndrome
	Immune complex formation	Double-stranded DNA	Systemic lupus erythematosus
		Immunoglobulin	Rheumatoid arthritis
	Opsonization	Platelet GpIIb:IIIa	Autoimmune thrombocytopenic purpura
		Rh antigens, I antigen	Autoimmune hemolytic anemia
	Antibody-dependent cellular cytotoxicity	Thyroid peroxidase, thyroglobulin	Hashimoto's thyroiditis
T cells	Cytokine production		Rheumatoid arthritis, multiple sclerosis, type 1 diabetes mellitus
	Cellular cytotoxicity		Type 1 diabetes mellitus

Abbreviations: ANCA, antineutrophil cytoplasmic antibody; LATS, long-acting thyroid stimulator; TSH, thyroid-stimulating hormone.

TABLE 348-4 Human Autoimmune Disease: Presumptive Evidence for Immunologic Pathogenesis

Major Criteria
1. Presence of autoantibodies or evidence of cellular reactivity to self
2. Documentation of relevant autoantibody or lymphocytic infiltrate in the pathologic lesion
3. Demonstration that relevant autoantibody or T cells can cause tissue pathology
a. Transplacental transmission
b. Adaptive transfer into animals
c. In vitro impact on cellular function
Supportive Evidence
1. Reasonable animal model
2. Beneficial effect from immunosuppressive agents
3. Association with other evidence of autoimmunity
4. No evidence of infection or other obvious cause

TABLE 348-5 Diseases on the Autoimmune Spectrum

Organ Specific	
Graves' disease	Vitiligo
Hashimoto's thyroiditis	Autoimmune hemolytic anemia
Autoimmune polyglandular syndrome	Autoimmune thrombocytopenic purpura
Type 1 diabetes mellitus	Pernicious anemia
Insulin-resistant diabetes mellitus	Myasthenia gravis
Immune-mediated infertility	Multiple sclerosis
Autoimmune Addison's disease	Guillain-Barré syndrome
Pemphigus vulgaris	Stiff-man syndrome
Pemphigus foliaceus	Acute rheumatic fever
Dermatitis herpetiformis	Sympathetic ophthalmia
Autoimmune alopecia	Goodpasture's syndrome
Primary biliary cirrhosis	
Organ Nonspecific (Systemic)	
Systemic lupus erythematosus	Granulomatosis with polyangiitis
Rheumatoid arthritis	Antiphospholipid syndrome
Systemic necrotizing vasculitis	Sjögren's syndrome

To confirm autoantibody pathogenicity, it may be possible to transfer disease to experimental animals by the administration of autoantibodies from a patient, with the subsequent development of pathology in the recipient similar to that seen in the patient. This scenario has been documented, for example, in Graves' disease. Some autoimmune diseases can be transferred from mother to fetus and are observed in the newborn babies of diseased mothers. The symptoms of the disease in the newborn usually disappear as the levels of maternal antibody decrease. An exception, however, is congenital heart block, in which damage to the developing conducting system of the heart follows in utero transfer of anti-Ro antibody from the mother to the fetus. This antibody transfer can result in a permanent developmental defect in the heart.

In most situations, the critical factors that determine when the development of autoimmunity results in autoimmune disease have not been delineated. The relationship of autoimmunity to the development of autoimmune disease may be associated with the fine specificity of the antibodies or T cells or their specific effector capabilities. In many circumstances, a mechanistic understanding of the pathogenic potential of autoantibodies has not been established. In some autoimmune diseases, biased production of cytokines by helper T (T_H) cells may play a role in pathogenesis. In this regard, T cells can differentiate into specialized effector cells that predominantly produce interferon γ (T_H1), IL-4 (T_H2), or IL-17 (T_H17) or that provide help to B cells (T follicular helper, T_{FH}) **(Chap. 342)**. T_H1 cells facilitate macrophage activation and classic cell–mediated immunity, whereas T_H2 cells are thought to have regulatory functions and are involved in the resolution of normal immune responses as well as in the development of responses to a variety of parasites. T_H17 cells produce a number of inflammatory cytokines, including IL-17 and IL-22, and seem to be prominently involved in host resistance to certain fungal infections. T_{FH} cells help B cells by constitutively producing IL-21. In a number of autoimmune diseases, such as rheumatoid arthritis, multiple sclerosis, type 1 diabetes mellitus, and Crohn's disease, there appears to be biased differentiation of T_H1 and T_H17 cells, with resultant organ damage. Studies suggest an accentuated differentiation of T_H17 cells associated with animal models of inflammatory arthritis, whereas increased differentiation of T_{FH} cells has been associated with animal models of SLE. Importantly, genetically determined or environmentally induced features of the target organ may determine susceptibility of the target organ to autoantibodies or autoreactive T cell-mediated damage.

ORGAN-SPECIFIC VERSUS SYSTEMIC AUTOIMMUNE DISEASES

The spectrum of autoimmune diseases ranges from conditions specifically affecting a single organ to systemic disorders that involve many organs **(Table 348-5)**. Hashimoto's autoimmune thyroiditis is an example of an organ-specific autoimmune disease **(Chap. 375)**. In this disorder, a specific lesion in the thyroid is associated with infiltration of mononuclear cells and damage to follicular cells. Antibody to thyroid constituents can be demonstrated in nearly all cases. Other organ- or tissue-specific autoimmune disorders include pemphigus vulgaris, autoimmune hemolytic anemia, idiopathic thrombocytopenic purpura, Goodpasture's syndrome, myasthenia gravis, and sympathetic ophthalmia. One important feature of some organ-specific autoimmune diseases is the tendency for overlap, such that an individual with one specific syndrome is more likely to develop a second syndrome. For example, there is a high incidence of pernicious anemia in individuals with autoimmune thyroiditis. More striking is the tendency for individuals with an organ-specific autoimmune disease to develop multiple other manifestations of autoimmunity without the development of associated organ pathology. Thus, as many as 50% of individuals with pernicious anemia have non-cross-reacting antibodies to thyroid constituents, whereas patients with myasthenia gravis may develop antinuclear antibodies, antithyroid antibodies, rheumatoid factor, antilymphocyte antibodies, and polyclonal hypergammaglobulinemia. Part of the explanation may relate to the genetic elements shared by individuals with these different diseases.

Systemic autoimmune diseases differ from organ-specific diseases in that pathologic lesions are found in multiple diverse organs and tissues. The hallmark of these conditions is the demonstration of associated relevant autoimmune manifestations that are likely to have an etiologic role in organ pathology. SLE represents the prototype of these disorders because of its abundant autoimmune manifestations. SLE is a disease of protean manifestations that characteristically involves the kidneys, joints, skin, serosal surfaces, blood vessels, and central nervous system **(Chap. 349)**. The disease is associated with a vast array of autoantibodies whose production appears to be a part of a generalized hyperreactivity of the humoral immune system. Other features of SLE include generalized B cell hyperresponsiveness and polyclonal hypergammaglobulinemia. Current evidence suggests that both hypo- and hyperresponsiveness to antigen can lead to survival and activation of autoreactive B cells in SLE. The autoantibodies in SLE are thought to arise as part of an accentuated T cell–dependent B cell response since most pathogenic anti-DNA autoantibodies exhibit evidence of extensive somatic hypermutation.

TREATMENT

Autoimmune Diseases

Treatment of autoimmune diseases can focus on suppressing the induction of autoimmunity, restoring normal regulatory mechanisms, or inhibiting the effector mechanisms. To decrease the number or function of autoreactive cells, immunosuppressive or ablative therapies are most commonly used. In recent years,

cytokine blockade has been demonstrated to be effective in preventing immune activation in some diseases or in inhibiting the extensive inflammatory effector mechanisms characteristic of these diseases. New therapies have also been developed to target lymphoid cells more specifically by blocking a costimulatory signal needed for T or B cell activation, by blocking the migratory capacity of lymphocytes, or by eliminating the effector T cells or B cells. The efficacy of these therapies in some diseases—e.g., SLE (belimumab), rheumatoid arthritis (TNF neutralization, IL-6 receptor blockade, CD28 competition, B cell depletion, IL-1 competition), psoriasis (IL-12/23 depletion, TNF neutralization), and inflammatory bowel disease (TNF neutralization, IL-12/23 neutralization)—has been demonstrated. One major advance in inhibiting effector mechanisms has been the introduction of cytokine blockade that appears to limit organ damage in some diseases, including rheumatoid arthritis, inflammatory bowel disease, psoriasis, and the spondyloarthritides. Small molecules that block cytokine signaling pathways by blocking the Janus kinase (JAK) family of kinases have recently been introduced into the clinic. Biologicals that delete B cells (anti-CD20 antibody) have recently been approved for the treatment of rheumatoid arthritis and have demonstrated efficacy in other autoimmune diseases as well. Their efficacy in diseases characterized by pathogenic effector T cells has highlighted the importance of B cells as antigen presenting cells in autoimmune diseases. Finally, there is renewed interest in cellular therapies in autoimmune diseases, including hematopoietic stem cell reconstitutions and treatment with immunosuppressive mesenchymal stem cells. Therapies that prevent target-organ damage or support target-organ function also remain important in the management of autoimmune disease.

FURTHER READING

Caielli S et al: Oxidized mitochondrial nucleoids release by neutrophils drive type I interferon production in human lupus. J Exp Med 5:697, 2016.

Jackson SW et al: B cells take the front seat; Dysregulated B cell signals orchestrate loss of tolerance and autoantibody production. Curr Opin Immunology 33:70, 2015.

Teruel M, Alacron-Riquelme ME: Genetics of systemic lupus erythematosus and Sjögren's syndrome: An update. Curr Opin Rheumatol 28:506, 2016.

Tsokos GC et al: New insights into the immunopathogenesis of systemic lupus erythematosus. Nat Rev Rheumatol 22:716, 2016.

Ueno H: T follicular helper cells in human autoimmunity. Curr Opin Immunol 43:24, 2016.

Yin Y et al: Normalization of CD4+ T cell metabolism reverses lupus. Science Transl Med 7:274, 2015.

349 Systemic Lupus Erythematosus

Bevra Hannahs Hahn

DEFINITION AND PREVALENCE

Systemic lupus erythematosus (SLE) is an autoimmune disease in which organs and cells undergo damage initially mediated by tissue-binding autoantibodies and immune complexes. In most patients, autoantibodies are present for a few years before the first clinical symptom appears. Ninety percent of patients are women of child-bearing years; people of all genders, ages, and ethnic groups are susceptible. Prevalence of SLE in the United States is 20–150 per 100,000 women depending on race and gender; highest prevalence is in African-American and Afro-Caribbean women, and lowest prevalence is in white men.

PATHOGENESIS AND ETIOLOGY

The proposed pathogenic mechanisms of SLE are illustrated in Fig. 349-1. The abnormal immune responses underlying SLE may be summarized as leading to production of increased quantities and immunogenic forms of nucleic acids, their accompanying proteins, and other self-antigens. The process may begin with autoimmunity-inducing activation of innate immunity, partly through binding of DNA/RNA/proteins by toll-like receptors in those cells. The changes include dendritic cells producing interferon α (IFNα), activated macrophages producing inflammatory cytokines/chemokines such as interleukin (IL12), tumor necrosis factor α (TNFα), and the B cell maturation/survival factor BLys/BAFF, neutrophils releasing DNA/protein-containing nets, and natural killer (NK) cells unable to kill autoreactive T and B cells or to produce the transforming growth factor β (TGFβ) needed for development of regulatory T cells. Upregulation of genes induced by IFNs is a genetic "signature" in peripheral blood cells of 50–80% of SLE patients. The innate immune system interacts with the B and T cells of adaptive immunity, which further drive autoimmune responses. T lymphocytes have altered metabolism (abnormal mitochondrial electron transport, membrane potential, and oxidative stress), increased glucose utilization, increased pyruvate production, activation of mTOR, and increased autophagy. T and B cells are more easily activated and driven into apoptosis than are normal cells, probably due to autoantibodies binding them plus abnormal signaling after engagement of surface molecules resulting in abnormally low production of IL2, which is required for T cell survival. B cells present antigen and secrete IL6 and IL10, further promoting autoreactive B cell survival (which is also favored by estrogen). Lupus phagocytic cells have reduced capacity to clear immune complexes, apoptotic cells, and their DNA/RNA/Ro/La and phospholipid containing surface blebs. The result is persistence of large quantities of autoantigens and resultant large quantities of autoantibodies with increased numbers of activated B cells and plasmablasts/plasma cells, and autoreactive T cells with shifts away from regulatory populations toward increased numbers and functions of Th1, T17, and Tfh cells, all of which promote production of autoantibodies and tissue damage. This damage begins with deposition of autoantibodies and/or immune complexes, followed by destruction mediated by complement activation and release of cytokines/chemokines. Non-immune tissue-fixed cells are then activated to produce more inflammation and damage, such as basal cells of the dermis, synovial fibroblasts, renal mesangial cells, podocytes and tubular epithelium, and endothelial cells throughout the body. Meanwhile, the initial immune attack is attracting into the target tissues additional B and T cells, monocytes/macrophages, dendritic cells, and plasma cells. Inflammation also causes release of vasoactive peptides, oxidative damage, growth factors and fibrosing factors. Sclerosis/fibrosis with irreversible tissue damage can occur in multiple tissues including kidneys, lungs, blood vessels, and skin. Each of these processes depends on the individual's genetic background, environmental influences, and epigenetics. Autoantibodies of SLE are referred to in Fig. 349-1 and described in Table 349-1.

SLE is a multigenic disease. Rare single-gene defects confer high hazard ratios (HRs) for SLE (5–25), including homozygous deficiencies of early components of complement (C1q,r,s; C2; C4) and a mutation in *TREX1* (encoding a DNAase) on the X chromosome. In most genetically susceptible individuals, normal alleles of multiple genes each contribute a small amount to abnormal immune/inflammation/tissue damage responses; if enough predisposing variations are present, disease results. Approximately 60 genes with alleles increasing risk for SLE and/or lupus nephritis (examples listed in Fig. 349-1 which includes most with HR ≥1.5) have been identified in recent genome-wide association studies in different racial groups. Individually, they confer an HR for SLE of 1.5–3 and even in combination account for only 18% of disease susceptibility, suggesting that environmental exposures and epigenetics play major roles. Predisposing, antigen-presenting human leukocyte antigen (HLA) molecules are most commonly found, in multiple ethnic groups (HLA DRB1 *0301 and *1501 and DR3), as well as multiple genes across the major histocompatibility complex (MHC)

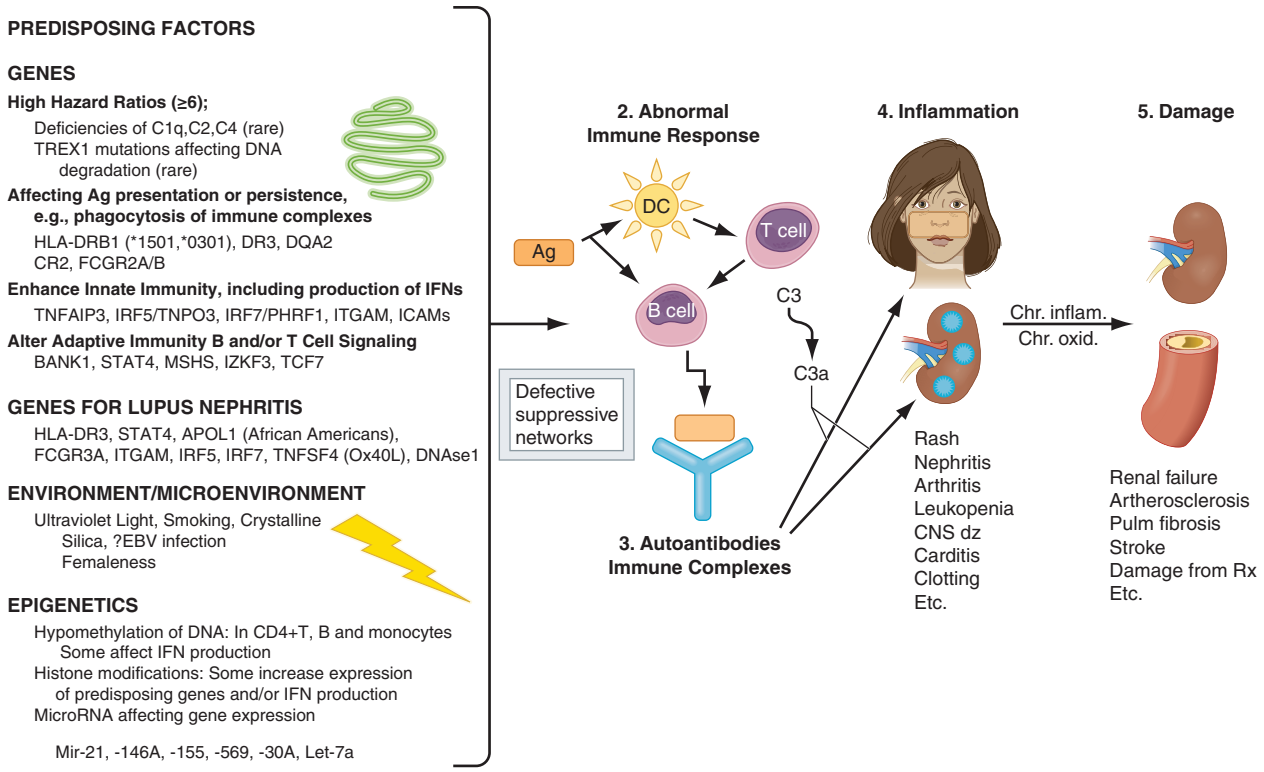


FIGURE 349-1 Pathogenesis of systemic lupus erythematosus (SLE). Pathogenesis is related in large part to production of increased quantities and immunogenic forms of nucleic acids and other self-antigens, which drive autoimmune-inducing activation of innate immunity, autoantibodies, and T cells. Interactions between genes, environment, and epigenetic changes drive increased autophagy, Ag presentation, neutrophil netosis, autoantibody formation with increased plasma cells, and production of pathogenic effector T cells in Th1, Th17, and Tfh subsets, with ineffective regulatory networks. Genes confirmed in more than one genome-wide association analysis in multiple racial groups that increase susceptibility to SLE or lupus nephritis (HR ≥1.5) are listed (reviewed in *Teruel M, Alarcon-Riquelme ME: The genetic basis of systemic lupus erythematosus: What are the risk factors and what have we learned. J Autoimmun 74:161, 2016; and Iwamoto T, Niewold TB: Genetics of human lupus nephritis. Clin Immunol 2016. Epub ahead of print*). Epigenetics are reviewed in *Long H et al: The critical role of epigenetics in systemic lupus erythematosus and autoimmunity, J Autoimmun 2016. Epub ahead of print.* Gene-environment interactions (reviewed in *Barbhaiya M, Costenbader KH: Environmental exposures and the development of systemic lupus erythematosus. Curr Opin Rheumatol 2016. Epub ahead of print*) result in abnormal immune responses that generate pathogenic autoantibodies and immune complexes that deposit in tissue, activate complement, induce cytokine and chemokine release causing inflammation, and over time lead to irreversible organ damage (reviewed in *Anders HJ, Rovin B: A pathophysiology-based approach to the diagnosis and treatment of lupus nephritis. Kidney Int 90:493, 2016;* and *Hahn BH: Pathogenesis of SLE, in Dubois Lupus Erythematosus, 8th ed, DJ Wallace, BH Hahn, (eds). Philadelphia, Elsevier, 2013*). Ag, antigen; C1q, complement system; C3, complement component; CNS, central nervous system; DC, dendritic cell; EBV, Epstein-Barr virus; HLA, human leukocyte antigen; FcR, immunoglobulin Fc-binding receptor; IL, interleukin; MCP, monocyte chemotactic protein; PTPN, phosphotyrosine phosphatase; UV, ultraviolet.

120-gene region. Non-HLA genetic factors are listed in Fig. 349-1 and include polymorphisms that affect innate and adaptive immunity pathways. Note the large number that influences IFN production—the most characteristic gene expression pattern of SLE patients. Other genes affect clearance of apoptotic cells or immune complexes, influence neutrophil adherence (*ITGAM*), and DNA repair (*TREX-1*). Some polymorphisms influence clinical manifestations; such as single nucleotide polymorphisms (SNPs) of *STAT4* that associate with severe disease, anti-DNA, nephritis, and antiphospholipid syndrome (APS), and an allele of *FCGRIIA* encoding a receptor that binds immune complexes poorly and predisposes to nephritis. Some gene effects are in promoter regions (e.g., IL-10), and others are conferred by copy numbers (e.g., C4A, TLR7). In addition, multiple epigenetic changes characterize SLE, including hypomethylation of DNA in CD4+ T cells, B cells, and monocytes, including genes that control production of type 1 interferons, and histone modifications. Some of these changes are mediated by microRNAs associated with SLE including some that control DNA Methyl transferases (DNMTs), such as mIR-146a, that control methylation of DNA in CD4+ T cells and IFN production. Some gene polymorphisms contribute to several autoimmune diseases, such as *STAT4* and *CTLA4*. All of these gene polymorphisms/transcription/epigenetic combinations influence immune responses to the external and internal environment; when such responses are too high and/or too prolonged and/or inadequately regulated, autoimmune disease is favored.

Female sex is permissive for SLE with evidence for hormone effects, genes on the X chromosome, and epigenetic differences between genders playing a role. Females of many mammalian species make higher antibody responses than males. Women exposed to estrogen-containing oral contraceptives or hormone replacement have an increased risk of developing SLE (HR 1.2–2). Estradiol binds to receptors on T and B lymphocytes, increasing activation and survival of those cells, especially autoreactive subsets, thus favoring prolonged immune responses. Genes on the X chromosome that influence SLE, such as *TREX-1*, may play a role in gender predisposition, possibly because some genes on the second X in females are not silent. People with XXY karyotype (Klinefelter's syndrome) have a significantly increased risk for SLE.

Several environmental stimuli may influence SLE (Fig. 349-1). Exposure to ultraviolet light causes flares of SLE in ~70% of patients, possibly by increasing apoptosis in skin cells or by altering DNA and intracellular proteins to make them antigenic. Some infections and lupus-inducing drugs activate autoreactive T and B cells; if such cells are not appropriately regulated, prolonged autoantibody production occurs. Most SLE patients have autoantibodies for 3 years or more before the first symptoms of disease, suggesting that regulation controls the degree of autoimmunity for years before quantities and qualities of autoantibodies, pathogenic B and T cells, and activated tissue-fixed cells such as macrophages cause clinical disease. Epstein-Barr virus (EBV) may be one infectious agent that can trigger SLE in susceptible individuals. Children and adults with SLE are more likely to be infected by EBV than age-, sex-, and ethnicity-matched controls. EBV contains amino acid sequences that mimic sequences on human spliceosomes (RNA/protein antigens) often recognized by autoantibodies in people

TABLE 349-1 Autoantibodies in Systemic Lupus Erythematosus (SLE)

ANTIBODY	PREVALENCE, %	ANTIGEN RECOGNIZED	CLINICAL UTILITY
Antinuclear antibodies	98	Multiple nuclear	Best screening test; repeated negative tests by immunofluorescence make SLE unlikely
Anti-dsDNA	70	DNA (double-stranded)	High titers are SLE-specific and in some patients correlate with disease activity, nephritis, vasculitis. Crithidia immunofluorescence is more specific for SLE than ELISA methods.
Anti-Sm	25	Protein complexed to 6 species of nuclear U1 RNA	Specific for SLE; no definite clinical correlations; most patients also have anti-RNP; more common in blacks and Asians than whites
Anti-RNP	40	Protein complexed to U1 RNA	Not specific for SLE; high titers associated with syndromes that have overlap features of several rheumatic syndromes including SLE; more common in blacks than whites; correlates with high IFN-induced gene signature
Anti-Ro (SS-A)	30	Protein complexed to hY RNA, primarily 60 kDa and 52 kDa	Not specific for SLE; associated with sicca syndrome, predisposes to subacute cutaneous lupus, and to neonatal lupus with congenital heart block; associated with decreased risk for nephritis
Anti-La (SS-B)	10	47-kDa protein complexed to hY RNA	Usually associated with anti-Ro; associated with decreased risk for nephritis
Antihistone	70	Histones associated with DNA (in nucleosome, chromatin)	More frequent in drug-induced lupus than in SLE
Antiphospholipid	50	Phospholipids, β_2 glycoprotein 1 (β_2G1) cofactor, prothrombin	Three tests available—ELISAs for cardiolipin and β_2G1, sensitive prothrombin time (DRVVT) for lupus anticoagulant; predisposes to clotting, fetal loss, thrombocytopenia
Antierythrocyte	60	Erythrocyte membrane	Measured as direct Coombs test; a small proportion develops overt hemolysis
Antiplatelet	30	Surface and altered cytoplasmic antigens on platelets	Associated with thrombocytopenia, but sensitivity and specificity are not good; this is not a useful clinical test
Antineuronal (includes antiglutamate receptor 2)	60	Neuronal and lymphocyte surface antigens	In some series, a positive test in CSF correlates with active CNS lupus
Antiribosomal P	20	Protein in ribosomes	In some series, a positive test in serum correlates with depression or psychosis due to CNS lupus

Abbreviations: CNS, central nervous system; CSF, cerebrospinal fluid; DRVVT, dilute Russell viper venom time; ELISA, enzyme-linked immunosorbent assay.

with SLE. Current tobacco smoking increases risk for SLE (HR 1.5). Prolonged occupational exposure to crystalline silica (e.g., inhalation of soap powder dust or soil in farming activities) increases risk (HR 4.3) in African-American women. Drinking alcohol (2 glasses of wine a week or ½ of an alcoholic drink daily) reduces the risk of SLE. Thus, interplay between genetic susceptibility, environment, gender, race, and abnormal immune responses results in autoimmunity **(Chap. 348)**.

PATHOLOGY

In SLE, biopsies of affected skin show deposition of Ig at the dermal-epidermal junction (DEJ), injury to basal keratinocytes, and inflammation dominated by T lymphocytes in the DEJ and around blood vessels and dermal appendages. Clinically unaffected skin may also show Ig deposition at the DEJ. These patterns are not specific for dermatologic SLE; however, they are highly suggestive.

In renal biopsies, the pattern and severity of injury are important in diagnosis and in selecting the best therapy. Most recent clinical studies of lupus nephritis have used the International Society of Nephrology (ISN) and the Renal Pathology Society (RPS) classification **(Table 349-2)**. In the ISN/RPS classification, the addition of "a" for active and "c" for chronic changes gives physicians information regarding the potential reversibility of disease. The system focuses on glomerular disease, although the presence of tubular interstitial and vascular disease, as well as the chronicity score in both glomeruli and interstitium, are important in predicting clinical outcomes. In general, class III and IV disease, as well as class V accompanied by III or IV disease, should be treated with aggressive immunosuppression if possible, because there is a high risk for end-stage renal disease (ESRD) if patients are untreated or undertreated. In contrast, treatment for lupus nephritis is not recommended in patients with class I or II disease or with extensive irreversible changes. In the recent Systemic Lupus International Collaborating Clinic (SLICC) criteria for classification of SLE, a diagnosis can be established on the basis of renal histology in the presence of lupus autoantibodies, without meeting additional criteria totaling 4 **(Table 349-3)**.

Histologic abnormalities in blood vessels may also determine therapy. Patterns of vasculitis are not specific for SLE but may indicate active disease: leukocytoclastic vasculitis is most common **(Chap. 356)**.

Lymph node biopsies are usually performed to rule out infection or malignancies. In SLE, they show nonspecific diffuse chronic inflammation.

DIAGNOSIS

The diagnosis of SLE is based on characteristic clinical features and autoantibodies. Current criteria for classification are listed in Table 349-3, and an algorithm for diagnosis and initial therapy is shown in **Fig. 349-2**. The criteria are intended for confirming the diagnosis of SLE in patients included in studies; the author uses them in individual patients for estimating the probability that a disease is SLE. Any combination of four or more criteria, with at least one in the clinical and one in the immunologic category, well documented at any time during an individual's history, makes it likely that the patient has SLE. (Specificity and sensitivity are ~93% and ~92%, respectively.) In many patients, criteria accrue over time. Antinuclear antibodies (ANA) are positive in >98% of patients during the course of disease; repeated negative tests by immunofluorescent methods suggest that the diagnosis is not SLE, unless other autoantibodies are present (Fig. 349-2). High-titer IgG antibodies to double-stranded DNA and antibodies to the Sm antigen are both specific for SLE and, therefore, favor the diagnosis in the presence of compatible clinical manifestations. The presence in an individual of multiple autoantibodies without clinical symptoms should not be considered diagnostic for SLE, although such persons are at increased risk.

INTERPRETATION OF CLINICAL MANIFESTATIONS

When a diagnosis of SLE is made, it is important to establish the severity and potential reversibility of the illness and to estimate the possible consequences of various therapeutic interventions. In the following paragraphs, descriptions of some disease manifestations begin with relatively mild problems and progress to those more life-threatening.

TABLE 349-2 Classification of Lupus Nephritis (International Society of Nephrology and Renal Pathology Society)

Class I: Minimal Mesangial Lupus Nephritis

Normal glomeruli by light microscopy, but mesangial immune deposits by immunofluorescence.

Class II: Mesangial Proliferative Lupus Nephritis

Purely mesangial hypercellularity of any degree or mesangial matrix expansion by light microscopy, with mesangial immune deposits. A few isolated subepithelial or subendothelial deposits may be visible by immunofluorescence or electron microscopy, but not by light microscopy.

Class III: Focal Lupus Nephritis

Active or inactive focal, segmental or global endo- or extracapillary glomerulonephritis involving ≤50% of all glomeruli, typically with focal subendothelial immune deposits, with or without mesangial alterations.

- Class III (A): Active lesions—focal proliferative lupus nephritis
- Class III (A/C): Active and chronic lesions—focal proliferative and sclerosing lupus nephritis
- Class III (C): Chronic inactive lesions with glomerular scars—focal sclerosing lupus nephritis

Class IV: Diffuse Lupus Nephritis

Active or inactive diffuse, segmental or global endo- or extracapillary glomerulonephritis involving ≥50% of all glomeruli, typically with diffuse subendothelial immune deposits, with or without mesangial alterations. This class is divided into diffuse segmental (IV-S) lupus nephritis when ≥50% of the involved glomeruli have segmental lesions, and diffuse global (IV-G) lupus nephritis when ≥50% of the involved glomeruli have global lesions. Segmental is defined as a glomerular lesion that involves less than one-half of the glomerular tuft. This class includes cases with diffuse wire loop deposits but with little or no glomerular proliferation.

- Class IV-S (A): Active lesions—diffuse segmental proliferative lupus nephritis
- Class IV-G (A): Active lesions—diffuse global proliferative lupus nephritis
- Class IV-S (A/C): Active and chronic lesions—diffuse segmental proliferative and sclerosing lupus nephritis
- Class IV-G (A/C): Active and chronic lesions—diffuse global proliferative and sclerosing lupus nephritis
- Class IV-S (C): Chronic inactive lesions with scars—diffuse segmental sclerosing lupus nephritis
- Class IV-G (C): Chronic inactive lesions with scars—diffuse global sclerosing lupus nephritis

Class V: Membranous Lupus Nephritis

Global or segmental subepithelial immune deposits or their morphologic sequelae by light microscopy and by immunofluorescence or electron microscopy, with or without mesangial alterations. Class V lupus nephritis may occur in combination with class III or IV, in which case both will be diagnosed. Class V lupus nephritis may show advanced sclerosis.

Class VI: Advanced Sclerotic Lupus Nephritis

≥90% of glomeruli globally sclerosed without residual activity.

Note: Indicate and grade (mild, moderate, severe) tubular atrophy, interstitial inflammation and fibrosis, and severity of arteriosclerosis or other vascular lesions.

Source: JJ Weening et al: Kidney Int 65:521, 2004. Reprinted by permission from Macmillan Publishers Ltd., Copyright 2004.

TABLE 349-3 Systemic Lupus International Collaborating Clinic Criteria for Classification of Systemic Lupus Erythematosus

CLINICAL MANIFESTATIONS	IMMUNOLOGIC MANIFESTATIONS
Skin	ANA > reference negative value
Acute, subacute cutaneous LE (photosensitive, malar, maculopapular, bullous)	Anti-dsDNA >reference, if by ELISA 2x reference
Chronic cutaneous LE (discoid lupus, panniculitis, lichen planus-like, hypertrophic verrucous, chillblains)	Anti-Sm
Oral or nasal ulcers	Antiphospholipid (any of lupus anticoagulant, false-positive RPR, anti-cardiolipin, anti-β glycoprotein I
Nonscarring Alopecia	Low serum complement (C3, C4 or CH50)
Synovitis involving ≥2 joints	Positive direct Coombs test in absence of hemolytic anemia
Serositis (pleurisy, pericarditis)	
Renal	
Prot/Cr ≥0.5	
RBC casts	
Biopsy[a]	
Neurologic	
Seizures, psychosis, mononeuritis, myelitis, peripheral or cranial neuropathies, acute confusional state	
Hemolytic anemia	
Leukopenia (<4000/μL) or Lymphopenia (<1000/μL)	
Thrombocytopenia (<100,000/μL)	

[a]Renal biopsy read as systemic lupus qualifies for classification as SLE if any lupus autoantibodies are present, even if total criteria are fewer than 4.

Interpretation: Presence of any four criteria (must have at least 1 in each category) qualifies patient to be classified as having SLE with 93% specificity and 92% sensitivity. American College of Rheumatology is developing new criteria for SLE. For update see website *Rheumatology.org*.

Abbreviations: ANA, antinuclear antibody; Cr, creatinine; LE, lupus erythematosus; Prot, protein.

Source: M Petri et al: Arthritis Rheum 64:2677, 2012. Because these criteria are relatively new, some currently ongoing clinical studies use prior American College of Rheumatology Criteria; see EM Tan et al: Arthritis Rheum 25:1271, 1982; update MC Hochberg: Arthritis Rheum 40:1725, 1997.

OVERVIEW AND SYSTEMIC MANIFESTATIONS

At its onset, SLE may involve one or several organ systems; over time, additional manifestations may occur (Tables 349-3 and 349-4). Most of the autoantibodies characteristic of each person are present at the time clinical manifestations appear (Tables 349-1 and 349-3). Severity of SLE varies from mild and intermittent to severe and fulminant. Approximately 85% of patients have either continuing active disease (on current treatment) or one or more flares of active disease annually. Permanent complete remissions (absence of symptoms with no treatment) are rare; however, low-level disease activity on treatments such as hydroxychloroquine and/or low dose prednisone is achievable in ~35% of patients. Systemic symptoms, particularly fatigue and myalgias/arthralgias, are present most of the time. Severe systemic illness requiring high dose glucocorticoid therapy can occur with fever, prostration, weight loss, and anemia with or without other organ-targeted manifestations.

MUSCULOSKELETAL MANIFESTATIONS

Most people with SLE have intermittent polyarthritis, varying from mild to disabling, characterized by soft tissue swelling and tenderness in joints and/or tendons, most commonly in hands, wrists, and knees. Joint deformities (hands and feet) develop in only 10%. Erosions on joint x-rays are rare but can be identified by ultrasound in 10–50% of patients. Some individuals have rheumatoid-like arthritis with erosions and fulfill criteria for both RA and SLE ("rhupus"). Joint pain is the most common reason that patients increase their dose of glucocorticoids. If pain persists in a single joint, such as knee, shoulder, or hip, a diagnosis of ischemic necrosis of bone (INB) should be considered, particularly if there are no other manifestations of active SLE, because INB prevalence is increased in SLE, especially in patients treated with systemic glucocorticoids. Myositis with clinical muscle weakness, elevated creatine kinase levels, positive magnetic resonance imaging (MRI) scan, and muscle necrosis and inflammation on biopsy can occur, although most patients have myalgias without frank myositis. Glucocorticoid therapies (commonly) and antimalarial therapies (rarely) can cause muscle weakness; these adverse effects must be distinguished from active inflammatory disease.

CUTANEOUS MANIFESTATIONS

Lupus dermatitis can be classified as acute, subacute, or chronic, and there are many different types of lesions encompassed within these groups. Discoid lupus erythematosus (DLE) is the most common chronic dermatitis in lupus; lesions are roughly circular with slightly raised, scaly hyperpigmented erythematous rims and depigmented, atrophic centers in which all dermal appendages are permanently destroyed. Lesions can be disfiguring, particularly on the face and

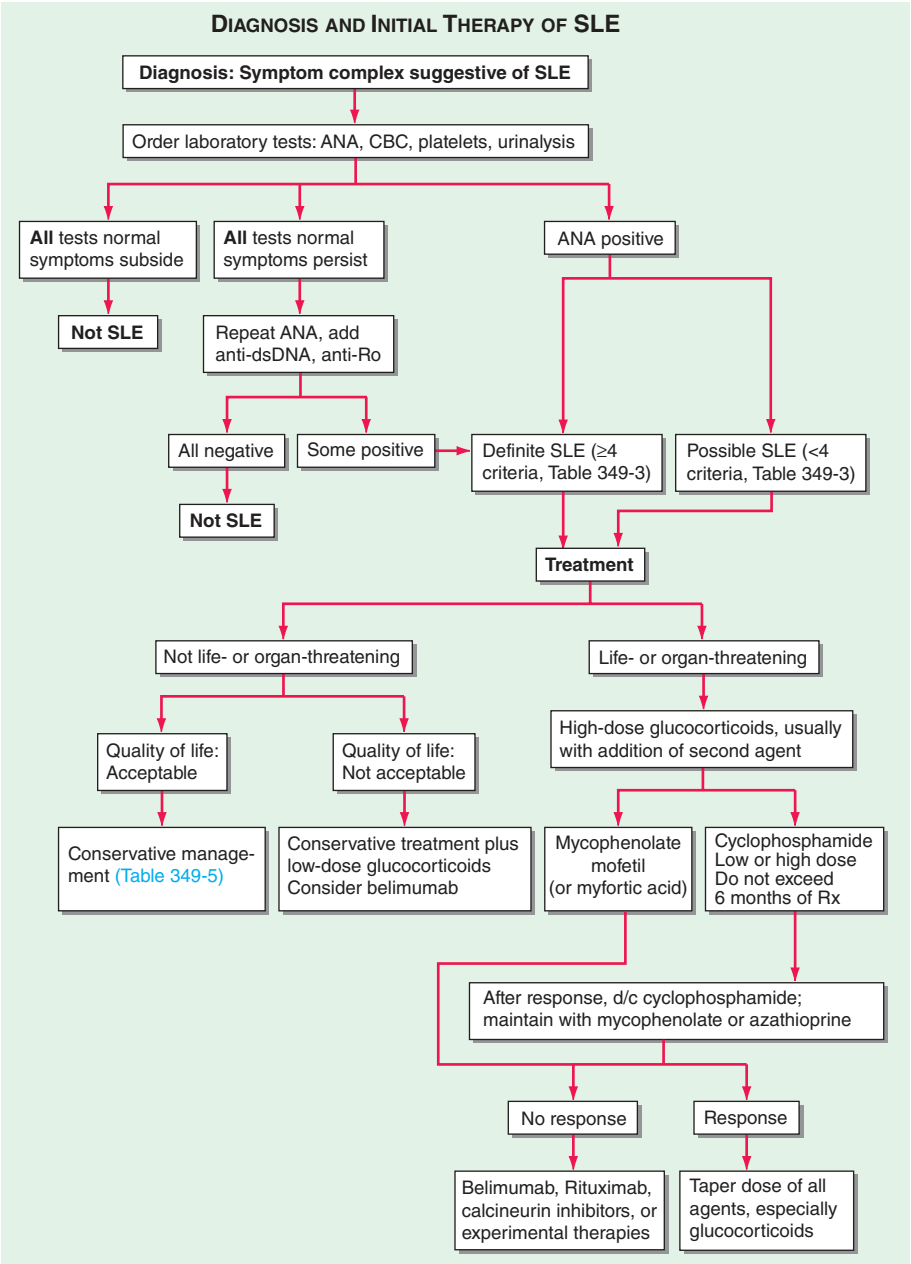


FIGURE 349-2 Algorithm for diagnosis and initial therapy of systemic lupus erythematosus (SLE). For guidelines on management of lupus and lupus nephritis, see Hahn BH et al: Arthritis Care Res (Hoboken) 64:797, 2012; Bertsias GK et al: Ann Rheum Dis 71:1771, 2012; Anders HJ, Rovin B: Kidney Int 2016 Epub ahead of print. For details on mycophenolate and cyclophosphamide induction and maintenance therapies, see Henderson L et al: Cochrane Database Syst Rev 12:CD002922, 2012; Ginzler EM et al: Arthritis Rheum 62:211, 2010; Houssiau FA et al: Ann Rheum Dis 69:61, 2010; and Dooley MA et al: N Engl J Med 365:1886, 2011. For belimumab in treatment, see Bruce IN et al: Long term organ damage accumulation and safety in patients with SLE treated with belimumab plus standard of care. Lupus 25:699, 2016; Hahn BH: N Eng J Med 368:1528, 2013. For rituximab, see Lightstone L: Lupus 22:390, 2013 and Rovin BH et al: Arthritis Rheum 64:1215, 2012. For tacrolimus, see Liu Z et al: Multitarget therapy for induction treatment of lupus nephritis: A randomized trial. Ann Intern Med 162:18, 2015. ANA, antinuclear antibodies; CBC, complete blood count.

scalp. Treatment consists primarily of topical or locally injected glucocorticoids and systemic antimalarials. Only 5% of people with DLE have SLE (although half have positive ANA); however, among individuals with SLE, as many as 20% have DLE. The most common acute SLE rash is a photosensitive, slightly raised erythema, occasionally scaly, on the face (particularly the cheeks and nose—the "butterfly" rash), ears, chin, V region of the neck and chest, upper back, and extensor surfaces of the arms. Worsening of this rash often accompanies flare of systemic disease. Subacute cutaneous lupus erythematosus (SCLE) consists of scaly red patches similar to psoriasis, or circular flat red-rimmed lesions. Patients with these manifestations are exquisitely photosensitive; most have antibodies to Ro (SS-A). Other SLE rashes include recurring urticaria, lichen planus-like dermatitis, bullae, and panniculitis ("lupus profundus"). Rashes can be minor or severe; they may be the major disease manifestation. Small ulcerations on the oral or nasal mucosa are common in SLE; the lesions resemble aphthous ulcers and may or may not be painful.

RENAL MANIFESTATIONS

Nephritis is usually the most serious manifestation of SLE, particularly because nephritis and infection are the leading causes of mortality in the first decade of disease. Because nephritis is asymptomatic in most lupus patients, urinalysis should be ordered in any person suspected of having SLE. The classification of lupus nephritis is primarily histologic (see "Pathology," above, and Table 349-2). Renal biopsy is recommended for every SLE patient with any clinical evidence of nephritis; results are used to plan current and near-future therapies. Patients with dangerous proliferative forms of glomerular damage (ISN III and IV) usually have microscopic hematuria and proteinuria (>500 mg per 24 h); approximately one-half develop nephrotic syndrome, and most develop hypertension. If diffuse proliferative glomerulonephritis (DPGN) is inadequately treated, virtually all patients develop ESRD within 2 years of diagnosis. Therefore, aggressive immunosuppression is indicated (usually systemic glucocorticoids plus another immunosuppressive drug), unless damage is irreversible (Fig. 349-2, Table 349-5). African Americans are more likely to develop ESRD than are whites, even with the most current therapies. Overall in the United States, ~20% of individuals with lupus DPGN die or develop ESRD within 10 years of diagnosis. Such individuals require aggressive control of SLE and of the complications of renal disease and of therapy. Approximately 20% of SLE patients with proteinuria (usually nephrotic) have membranous glomerular changes without proliferative changes on renal biopsy. Their outcome is better than for those with DPGN, but patients with class V and nephrotic range proteinuria should be treated in the same way as those with classes III or IV proliferative disease. Lupus nephritis tends to be an ongoing disease, with flares requiring re-treatment or increased treatment over many years. For most people with lupus nephritis, accelerated atherosclerosis becomes important after several years of disease; attention must be given to control of systemic inflammation, blood pressure, hyperlipidemia, and hyperglycemia.

NERVOUS SYSTEM MANIFESTATIONS

There are many central nervous system (CNS) and peripheral nervous system manifestations of SLE; in some patients, these are the major cause of morbidity and mortality. It is useful to approach this diagnostically by asking first whether the symptoms result from SLE or another condition (such as infection in immunosuppressed individuals or side effects of therapies). If symptoms are related to SLE, it should be determined whether they are caused by a diffuse process (requiring immunosuppression) or vascular occlusive disease (requiring anticoagulation). The most common manifestation of diffuse CNS lupus is

TABLE 349-4 Clinical Manifestations of SLE and Prevalence over the Entire Course of Disease[a]

MANIFESTATION	PREVALENCE, %
Systemic: Fatigue, malaise, fever, anorexia, weight loss	95
Musculoskeletal	95
Arthralgias/myalgias	95
Nonerosive polyarthritis	60
Hand deformities	10
Myopathy/myositis	25/5
Ischemic necrosis of bone	15
Cutaneous	80
Photosensitivity	70
Malar rash	50
Oral ulcers	40
Alopecia	40
Discoid rash	20
Vasculitis rash	20
Other (e.g., urticaria, subacute cutaneous lupus)	15
Hematologic	85
Anemia (chronic disease)	70
Leukopenia (<4000/μL)	65
Lymphopenia (<1500/μL)	50
Thrombocytopenia (<100,000/μL)	15
Lymphadenopathy	15
Splenomegaly	15
Hemolytic anemia	10
Neurologic	60
Cognitive disorder	50
Mood disorder	40
Depression	25
Headache	25
Seizures	20
Mono-, polyneuropathy	15
Stroke, TIA	10
Acute confusional state or movement disorder	2–5
Aseptic meningitis, myelopathy	<1
Cardiopulmonary	60
Pleurisy, pericarditis, effusions	30–50
Myocarditis, endocarditis	10
Lupus pneumonitis	10
Coronary artery disease	10
Interstitial fibrosis	5
Pulmonary hypertension, ARDS, hemorrhage	<5
Shrinking lung syndrome	<5
Renal	30–50
Proteinuria ≥500 mg/24 h, cellular casts	30–60
Nephrotic syndrome	25
End-stage renal disease	5–10
Gastrointestinal	40
Nonspecific (nausea, mild pain, diarrhea)	30
Abnormal liver enzymes	40
Vasculitis	5
Thrombosis	15
Venous	10
Arterial	5
Ocular	15
Sicca syndrome	15
Conjunctivitis, episcleritis	10
Vasculitis	5

[a]Numbers indicate percentage of patients who have the manifestation at some time during the course of illness.

Abbreviations: ARDS, acute respiratory distress syndrome; SLE, Systemic Lupus Erythematosus; TIA, transient ischemic attack.

cognitive dysfunction, including difficulties with memory and reasoning. Headaches are also common. When excruciating, they often indicate SLE flare; when milder, they are difficult to distinguish from migraine or tension headaches. Seizures of any type may be caused by lupus; treatment often requires both antiseizure and immunosuppressive therapies. Psychosis can be the dominant manifestation of SLE; it must be distinguished from glucocorticoid-induced psychosis. The latter usually occurs in the first weeks of glucocorticoid therapy, at daily doses of ≥40 mg of prednisone or equivalent; psychosis resolves over several days after glucocorticoids are decreased or stopped. Myelopathy is not rare and is often disabling; rapid initiation of immunosuppressive therapy starting with high-dose glucocorticoids is standard of care.

VASCULAR OCCLUSIONS INCLUDING STROKE AND MYOCARDIAL INFARCTIONS

The prevalence of transient ischemic attacks, strokes, and myocardial infarctions is increased in patients with SLE. These vascular events are increased in, but not exclusive to, SLE patients with antibodies to phospholipids (antiphospholipid antibodies), which are associated with hypercoagulability and acute thrombotic events **(Chap. 350)**. Ischemia in the brain can be caused by focal occlusion (either noninflammatory or associated with vasculitis) or by embolization from carotid artery plaque or from fibrinous vegetations of Libman-Sacks endocarditis. Appropriate tests for antiphospholipid antibodies (see below) and for sources of emboli should be ordered in such patients to estimate the need for, intensity of, and duration of anti-inflammatory and/or anticoagulant therapies. When it is most likely that a cerebral event results from clotting, long-term anticoagulation is the therapy of choice. Two processes can occur at once—vasculitis plus bland vascular occlusions—in which case it is appropriate to treat with anticoagulation plus immunosuppression.

In SLE, myocardial infarctions are primarily manifestations of accelerated atherosclerosis. The increased risk for vascular events is three- to tenfold overall, and is highest in women aged <49 years. Characteristics associated with increased risk for atherosclerosis include male gender, older age, hypertension, dyslipidemia, diabetes, dysfunctional proinflammatory high-density lipoproteins, repeated high scores for disease activity, high cumulative or daily doses of glucocorticoids, and high serum levels of homocysteine and leptin. Statin therapies reduce levels of low-density lipoproteins (LDL) in SLE patients; significant reduction of cardiac events by statins has been shown in SLE patients with renal transplants and recently in an epidemiologic study of large number of patients in Taiwan.

PULMONARY MANIFESTATIONS

The most common pulmonary manifestation of SLE is pleuritis with or without pleural effusion. This manifestation, when mild, may respond to treatment with nonsteroidal anti-inflammatory drugs (NSAIDs); when more severe, patients require a brief course of glucocorticoid therapy. Pulmonary infiltrates also occur as a manifestation of active SLE and are difficult to distinguish from infection on imaging studies. Life-threatening pulmonary manifestations include interstitial inflammation leading to fibrosis, shrinking lung syndrome, and intraalveolar hemorrhage; all of these probably require early aggressive immunosuppressive therapy as well as supportive care. Pulmonary arterial hypertension occurs in a small proportion of SLE patients and should be treated in the same way as idiopathic pulmonary hypertension.

CARDIAC MANIFESTATIONS

Pericarditis is the most frequent cardiac manifestation; it usually responds to anti-inflammatory therapy and infrequently leads to tamponade. More serious cardiac manifestations are myocarditis and fibrinous endocarditis of Libman-Sacks. The endocardial involvement can lead to valvular insufficiencies, most commonly of the mitral or aortic valves, or to embolic events. It has not been proven that glucocorticoid or other immunosuppressive therapies lead to improvement of lupus myocarditis or endocarditis, but it is usual practice to administer a trial of high-dose steroids along with appropriate supportive

TABLE 349-5 Medications for the Management of SLE

MEDICATION	DOSE RANGE	DRUG INTERACTIONS	SERIOUS OR COMMON ADVERSE EFFECTS
NSAIDs, salicylates (Ecotrin[a] and St. Joseph's aspirin[a] approved by FDA for use in SLE)	Doses toward upper limit of recommended range usually required	A2R/ACE inhibitors, glucocorticoids, fluconazole, methotrexate, thiazides	NSAIDs: Higher incidence of aseptic meningitis, elevated liver enzymes, decreased renal function, vasculitis of skin; entire class, especially COX-2-specific inhibitors, may increase risk for myocardial infarction Salicylates: ototoxicity, tinnitus Both: GI events and symptoms, allergic reactions, dermatitis, dizziness, acute renal failure, edema, hypertension
Topical glucocorticoids	Mid potency for face; mid to high potency for other areas	None known	Atrophy of skin, contact dermatitis, folliculitis, hypopigmentation, infection
Topical sunscreens	SPF 15 at least; 30+ preferred	None known	Contact dermatitis
Hydroxychloroquine[a] (quinacrine can be added or substituted)	200–400 mg qd (100 mg qd); do not exceed 6.5 mg/kg dry weight	None known	Retinal damage, agranulocytosis, aplastic anemia, ataxia, cardiomyopathy, dizziness, myopathy, ototoxicity, peripheral neuropathy, pigmentation of skin, seizures, thrombocytopenia; Quinacrine usually causes diffuse yellow skin coloration
DHEA (dehydroepiandrosterone)	200 mg qd	Unclear	Acne, menstrual irregularities, high serum levels of testosterone
Methotrexate (for dermatitis, arthritis)	10–25 mg once a week, PO or SC, with folic acid; decrease dose if CrCl <60 mL/min	Acitretin, leflunomide, NSAIDs and salicylates, penicillins, probenecid, sulfonamides, trimethoprim	Anemia, bone marrow suppression, leukopenia, thrombocytopenia, hepatotoxicity, nephrotoxicity, infections, neurotoxicity, pulmonary fibrosis, pneumonitis, severe dermatitis, seizures, pseudolymphoma
Glucocorticoids, oral[a] (several specific brands are approved by FDA for use in SLE)	Prednisone, prednisolone: 0.5–1 mg/kg per day for severe SLE 0.07–0.3 mg/kg per day or qod for milder disease	A2R/ACE antagonists, antiarrhythmics class III, cyclosporine, NSAIDs and salicylates, phenothiazines, phenytoins, quinolones, rifampin, risperidone, thiazides, sulfonylureas, warfarin	Infection, VZV infection, hypertension, hyperglycemia, hypokalemia, acne, allergic reactions, anxiety, aseptic necrosis of bone, cushingoid changes, CHF, fragile skin, insomnia, menstrual irregularities, mood swings, osteoporosis, psychosis
Methylprednisolone sodium succinate, IV[a] (FDA approved for lupus nephritis)	For severe disease, 0.5-1 g IV qd × 3 days	As for oral glucocorticoids	As for oral glucocorticoids (if used repeatedly); anaphylaxis
Cyclophosphamide[b] IV	Low dose (for whites of northern European backgrounds): 500 mg every 2 weeks for 6 doses, then begin maintenance with MMF or AZA. High dose: 7–25 mg/kg q month × 6; consider mesna administration with dose	Allopurinol, bone marrow suppressants, colony-stimulating factors, doxorubicin, rituximab, succinylcholine, zidovudine	Infection, VZV infection, bone marrow suppression, leukopenia, anemia, thrombocytopenia, hemorrhagic cystitis (less with IV), carcinoma of the bladder, alopecia, nausea, diarrhea, malaise, malignancy, ovarian and testicular failure. Ovarian failure is probably not a problem with low dose.
Oral	1.5–3 mg/kg per day; decrease dose for CrCl <25 mL/min		
Mycophenolate mofetil (MMF)[b] or mycophenolic acid (MPA)	MMF: 2–3 g/d PO total given bid for induction therapy, 1–2 g/d total given bid for maintenance therapy; max 1 g bid if CrCl <25 mL/min. Begin with low dose and increase every 1–2 weeks to minimize GI side effects. Start treatment at 0.5 g bid. MPA: 360–1080 mg bid; caution if CrCl <25 mL/min	Acyclovir, antacids, azathioprine, bile acid-binding resins, ganciclovir, iron, salts, probenecid, oral contraceptives	Infection, leukopenia, anemia, thrombocytopenia, lymphoma, lymphoproliferative disorders, malignancy, alopecia, cough, diarrhea, fever, GI symptoms, headache, hypertension, hypercholesterolemia, hypokalemia, insomnia, peripheral edema, elevated liver enzymes, tremor, rash. Limited date suggests Asians should begin treatment with doses not exceeding 2 g daily to reduce adverse events.
Azathioprine (AZA)[b]	2–3 mg/kg per day PO for induction; 1–2 mg/kg per day for maintenance; decrease frequency of dose if CrCl <50 mL/min	ACE inhibitors, allopurinol, bone marrow suppressants, interferons, mycophenolate mofetil, rituximab, warfarin, zidovudine	Infection, VZV infection, bone marrow suppression, leukopenia, anemia, thrombocytopenia, pancreatitis, hepatotoxicity, malignancy, alopecia, fever, flulike illness, GI symptoms
Belimumab	10 mg/kg IV wks 0, 2, and 4, then monthly OR subcutaneous 200 mg each week	IVIg	Infusion reactions, allergy, infections. Headache and diffuse body aching.
Rituximab (for patients resistant to above therapies)	375 mg/m^2 q wk × 4 or 1 g q 2 wks × 2	IVIg	Infection (including PML), infusion reactions, headache, arrhythmias, allergic responses
Tacrolimus	Trough blood level should not exceed 5.5 ng/mL to minimize toxicity. Begin dose at 2 mg bid		Infection, nephrotoxicity, neural toxicity

[a]Indicates medication is approved for use in SLE by the U.S. Food and Drug Administration. [b]Indicates the medication has been used with glucocorticoids in the trials showing efficacy.

Abbreviations: A2R, angiotensin II receptor; ACE, angiotensin-converting enzyme; CHF, congestive heart failure; CrCl, creatinine clearance; FDA, U.S. Food and Drug Administration; GI, gastrointestinal; IVIg, intravenous immunoglobulin; NSAIDs, nonsteroidal anti-inflammatory drugs; PML, progressive multifocal leukoencephalopathy; SLE, systemic lupus erythematosus; SPF, sun protection factor; VZV, varicella-zoster virus.

therapy for heart failure, arrhythmia, or embolic events. As discussed above, patients with SLE are at increased risk for myocardial infarction, usually due to accelerated atherosclerosis, which probably results from immune attack, chronic inflammation, and/or chronic oxidative damage to arteries.

HEMATOLOGIC MANIFESTATIONS

The most frequent hematologic manifestation of SLE is anemia, usually normochromic normocytic, reflecting chronic illness. Hemolysis can be rapid in onset and severe, requiring high-dose glucocorticoid therapy, which is effective in most patients. Leukopenia is also common and almost always consists of lymphopenia, not granulocytopenia; lymphopenia rarely predisposes to infections and by itself usually does not require therapy. Thrombocytopenia may be a recurring problem. If platelet counts are >40,000/μL and abnormal bleeding is absent, therapy may not be required. High-dose glucocorticoid therapy (e.g., 1 mg/kg per day of prednisone or equivalent) is usually effective for the first few episodes of severe thrombocytopenia. Recurring or prolonged hemolytic anemia or thrombocytopenia, or disease requiring an unacceptably high dose of daily glucocorticoids, should be treated with additional strategies such as rituximab, platelet growth factors, and/or splenectomy (see "Management of Systemic Lupus Erythematosus" below).

GASTROINTESTINAL MANIFESTATIONS

Nausea, sometimes with vomiting, and diarrhea can be manifestations of an SLE flare, as can diffuse abdominal pain probably caused by autoimmune peritonitis and/or intestinal vasculitis. Increases in serum aspartate aminotransferase (AST) and alanine aminotransferase (ALT) are common when SLE is active. These manifestations usually improve promptly during systemic glucocorticoid therapy. Vasculitis involving the intestine may be life-threatening; perforations, ischemia, bleeding, and sepsis are frequent complications. Aggressive immunosuppressive therapy with high-dose glucocorticoids is recommended for short-term control; evidence of recurrence is an indication for additional therapies.

OCULAR MANIFESTATIONS

Sicca syndrome (Sjögren's syndrome; **Chap. 354**) and nonspecific conjunctivitis are common in SLE and rarely threaten vision. In contrast, retinal vasculitis and optic neuritis are serious manifestations: blindness can develop over days to weeks. Aggressive immunosuppression is recommended, although there are no controlled trials to prove effectiveness. Complications of systemic and intraorbital glucocorticoid therapy include cataracts (common) and glaucoma.

LABORATORY TESTS

Laboratory tests serve (1) to establish or rule out the diagnosis; (2) to follow the course of disease, particularly to suggest that a flare is occurring or organ damage is developing; and (3) to identify adverse effects of therapies.

TESTS FOR AUTOANTIBODIES (TABLES 349-1 AND 349-3)

Diagnostically, the most important autoantibodies to detect are ANA because the test is positive in >95% of patients, usually at the onset of symptoms. A few patients develop ANA within 1 year of symptom onset; repeated testing may thus be useful. ANA tests using immunofluorescent methods are more reliable than enzyme-linked immunosorbent assays (ELISAs) and/or bead assays, which have less specificity. ANA-negative lupus exists but is rare in adults and is usually associated with other autoantibodies (anti-Ro or anti-DNA). High-titer IgG antibodies to double-stranded DNA (dsDNA) (but not to single-stranded DNA) are specific for SLE. ELISA and immunofluorescent reactions of sera with the dsDNA in the flagellate *Crithidia luciliae* have ~60% sensitivity for SLE. Titers of anti-dsDNA vary over time. In some patients, increases in quantities of anti-dsDNA herald a flare, particularly of nephritis or vasculitis, especially when associated with declining levels of C3 or C4 complement. Antibodies to Sm are also specific for SLE and assist in diagnosis; anti-Sm antibodies do not usually correlate with disease activity or clinical manifestations. Antiphospholipid antibodies are not specific for SLE, but their presence fulfills one classification criterion, and they identify patients at increased risk for venous or arterial clotting, thrombocytopenia, and fetal loss. There are three widely accepted tests that measure different antibodies (anticardiolipin, anti-β_2-glycoprotein, and the lupus anticoagulant). ELISA is used for anticardiolipin and anti-β_2-glycoprotein (both internationally standardized with good reproducibility); a sensitive phospholipid-based activated prothrombin time such as the dilute Russell venom viper test is used to identify the lupus anticoagulant. The higher the titers of IgG anticardiolipin (>40 IU is considered high), and the greater the number of different antiphospholipid antibodies that are detected, the greater is the risk for a clinical episode of clotting. Quantities of antiphospholipid antibodies may vary markedly over time; repeated testing is justified if clinical manifestations of the APS appear **(Chap. 350)**. To classify a patient as having APS, with or without SLE, by international criteria requires the presence of one or more clotting episodes and/or repeated fetal losses plus at least two positive tests for antiphospholipid antibodies, at least 12 weeks apart; however, many patients with APS do not meet these stringent criteria, which are intended for inclusion of patients into studies.

An additional autoantibody test with predictive value (not used for diagnosis) detects anti-Ro/SS-A, which indicates increased risk for neonatal lupus, sicca syndrome, and SCLE. Women with child-bearing potential and SLE should be screened for antiphospholipid antibodies and anti-Ro, because both antibodies have the potential to cause fetal harm.

STANDARD TESTS FOR DIAGNOSIS

Screening tests for complete blood count, platelet count, and urinalysis may detect abnormalities that contribute to the diagnosis and influence management decisions.

TESTS FOR FOLLOWING DISEASE COURSE

It is useful to follow tests that indicate the status of organ involvement known to be present during SLE flares. These might include urinalysis for hematuria and proteinuria, hemoglobin levels, platelet counts, and serum levels of creatinine or albumin. There is great interest in identification of additional markers of disease activity. Candidates include levels of anti-DNA and anti-C1q antibodies, several components of complement (C3 is most widely available), activated complement products (an assay is commercially available that measures binding to the C4d receptor on erythrocytes and B cells), IFN-inducible gene expression in peripheral blood cells, serum levels of BLyS (B lymphocyte stimulator, also called BAFF), and urinary levels of TNF-like weak inducer of apoptosis (TWEAK), neutrophil gelatinase-associated lipocalin (NGAL), or monocyte chemotactic protein 1 (MCP-1). None is uniformly agreed upon as a reliable indicator of flare or of response to therapeutic interventions. It is likely that a panel of multiple proteins and nuclear products (and possibly levels of selected miRNAs and methylation profiles of DNA) will be developed to predict both impending flare and response to recently instituted therapies. Increased quantities of plasma cells, and increased expression of their gene signatures in whole blood, are associated with active disease and flares, but measurements are not commercially available. For now, the physician should determine for each patient whether certain available laboratory test changes predict flare (falling complement, rising anti-DNA, increased proteinuria, worsening anemia, etc.). If so, altering therapy in response to these changes may be advisable (30 mg of prednisone daily for 2 weeks has been shown to prevent flares in patients with rising anti-DNA plus falling complement). In addition, given the increased prevalence of atherosclerosis in SLE, it is advisable to follow the recommendations of the National Cholesterol Education Program for testing and treatment, including scoring of SLE as an independent risk factor, similar to diabetes mellitus.

MANAGEMENT OF SYSTEMIC LUPUS ERYTHEMATOSUS

There is no cure for SLE, and complete sustained remissions are rare. There is an international effort to encourage practitioners and patients to aim for low-level disease activity (mild symptoms on the lowest

possible doses of medications) which can be achieved for at least a year in 30–50% of SLE patients. Therefore, the physician should plan to induce remissions of acute flares and then maintain improvements with strategies that suppress symptoms to an acceptable level and prevent organ damage. Therapeutic choices depend on (1) whether disease manifestations are life-threatening or likely to cause organ damage, justifying aggressive therapies; (2) whether manifestations are potentially reversible; and (3) the best approaches to preventing complications of disease and its treatments. Therapies, doses, and adverse effects are listed in Table 349-5.

CONSERVATIVE THERAPIES FOR MANAGEMENT OF NON-LIFE-THREATENING DISEASE

Among patients with fatigue, pain, and autoantibodies indicative of SLE, but without major organ involvement, management can be directed to suppression of symptoms. Analgesics and antimalarials are mainstays. NSAIDs are useful analgesics/anti-inflammatories, particularly for arthritis/arthralgias. However, two major issues indicate caution in using NSAIDs. First, SLE patients compared with the general population are at increased risk for NSAID-induced aseptic meningitis, elevated serum transaminases, hypertension, and renal dysfunction. Second, all NSAIDs, particularly those that inhibit cyclooxygenase-2 specifically, may increase risk for myocardial infarction. Acetaminophen to control pain may be a good strategy, but NSAIDs are more effective in some patients. The relative hazards of NSAIDs compared with low-dose glucocorticoid therapy have not been established. Antimalarials (hydroxychloroquine, chloroquine, and quinacrine) often reduce dermatitis, arthritis, and fatigue. A randomized, placebo-controlled, prospective trial has shown that withdrawal of hydroxychloroquine results in increased numbers of disease flares; hydroxychloroquine also reduces accrual of tissue damage, including renal damage, over time. Some experts recommend a hydroxychloroquine blood level of ≥750 ng/mL to optimize responses in active SLE; after achieving response doses should be reduced. Because of potential retinal toxicity (occurring in 6% of patients after cumulative doses of 1000 g, ~5 years of continuing therapy), patients receiving antimalarials should undergo ophthalmologic examinations annually. A placebo-controlled prospective trial suggests that administration of dehydroepiandrosterone may reduce disease activity. If quality of life is inadequate despite these conservative measures, treatment with low doses of systemic glucocorticoids may be necessary. Belimumab is effective for 50% of patients with fatigue, rash, and/or the arthritis of SLE; it is expensive and should be considered after other approaches fail or are not tolerated. SLE patients most likely to respond to belimumab have robust clinical activity (a Systemic Lupus Erythematosus Disease Activity Index [SLEDAI] score of ≥10), positive anti-DNA, and low serum complement. SLEDAI is a widely used measure of SLE disease activity; scores >3 reflect clinically active disease. Lupus dermatitis should be managed with topical sunscreens, antimalarials, topical glucocorticoids and/or tacrolimus, and if severe or unresponsive, systemic glucocorticoids with or without mycophenolate mofetil, azathioprine, or belimumab.

LIFE-THREATENING SLE: PROLIFERATIVE FORMS OF LUPUS NEPHRITIS

Guidelines for management of lupus nephritis have been published recently by the American College of Rheumatology and the European League Against Rheumatism (encompassed and referenced in Fig. 349-2 and Table 349-5). The mainstay of treatment for any inflammatory life-threatening or organ-threatening manifestations of SLE is systemic glucocorticoids (0.5–1 mg/kg per day PO or 500–1000 mg of methylprednisolone sodium succinate IV daily for 3 days followed by 0.5–1 mg/kg of daily prednisone or equivalent). Evidence that glucocorticoid therapy is life-saving comes from retrospective studies from the predialysis era; survival was significantly better in people with DPGN treated with high-dose daily glucocorticoids (40–60 mg of prednisone daily for 4–6 months) versus lower doses. Currently, high doses are recommended for much shorter periods; recent trials of interventions for severe SLE use 4–6 weeks of 0.5–1 mg/kg per day of prednisone or equivalent. Thereafter, doses are tapered as rapidly as the clinical situation permits, usually to a maintenance dose ranging from 5 to 10 mg of prednisone or equivalent per day. Most patients with an episode of severe SLE require many years of maintenance therapy with low-dose glucocorticoids, which can be increased to prevent or treat disease flares. Frequent attempts to gradually reduce the glucocorticoid requirement are recommended because virtually everyone develops important adverse effects (Table 349-5). High-quality clinical studies regarding initiating therapy for severe, active SLE with IV pulses of high-dose glucocorticoids are not available. Most recent clinical trials in lupus nephritis have initiated therapy with high-dose IV glucocorticoid pulses (500–1000 mg daily for 3–5 days). This approach must be tempered by safety considerations, such as the presence of conditions adversely affected by glucocorticoids (e.g., infection, hyperglycemia, hypertension, osteoporosis). A current clinical trial is evaluating mycophenolate mofetil plus rituximab without maintenance daily glucocorticoids to treat lupus nephritis: if results are positive, the paradigm for short-term and long-term management of SLE is likely to change.

Cytotoxic/immunosuppressive agents added to glucocorticoids are recommended to treat serious SLE. Almost all prospective controlled trials in SLE involving cytotoxic agents have been conducted in combination with glucocorticoids in patients with lupus nephritis. Therefore, the following recommendations apply to treatment of nephritis. Either cyclophosphamide (an alkylating agent) or mycophenolate mofetil (a relatively lymphocyte-specific inhibitor of inosine monophosphatase and therefore of purine synthesis) is an acceptable choice for induction of improvement in severely ill patients; azathioprine (a purine analogue and cycle-specific antimetabolite) may be effective but is associated with more flares. In patients whose renal biopsies show ISN grade III or IV disease, early treatment with combinations of glucocorticoids and cyclophosphamide reduces progression to ESRD and death. Shorter-term studies with glucocorticoids plus mycophenolate mofetil (prospective randomized trials of 6 months, follow-up studies of 5 years) show that this regimen is similar to cyclophosphamide in achieving improvement. Comparisons are complicated by effects of race, since higher proportions of African Americans (and other non-Asian, nonwhite races) respond to mycophenolate than to cyclophosphamide, whereas similar proportions of whites and Asians respond to each drug. Regarding toxicity, diarrhea is more common with mycophenolate mofetil; amenorrhea, leukopenia, and nausea are more common with high dose cyclophosphamide. Importantly, rates of severe infections and death are similar in meta-analyses. Two different regimens of IV cyclophosphamide are available. For white patients with northern European backgrounds, low doses of cyclophosphamide (500 mg every 2 weeks for six total doses, followed by azathioprine or mycophenolate maintenance) are as effective as standard high doses, with less toxicity. Ten-year follow-up has shown no differences between the high-dose and low-dose groups (death or ESRD in 9–20% of patients in each group). It is not clear whether the data apply to U.S. populations, especially African Americans and Latinas. High-dose cyclophosphamide (500–1000 mg/m^2 body surface area given monthly IV for 6 months, followed by azathioprine or mycophenolate maintenance) is an acceptable approach for patients with severe nephritis (e.g., multiple cellular crescents and/or fibrinoid necrosis on renal biopsy, or rapidly progressive glomerulonephritis). Cyclophosphamide and mycophenolate responses begin 3–16 weeks after treatment is initiated, whereas glucocorticoid responses may begin within 24 h.

For maintenance therapy, mycophenolate and azathioprine probably are similar in efficacy and toxicity; both are safer than cyclophosphamide. In a recently published multicenter study, mycophenolate was superior to azathioprine in maintaining renal function and survival in patients who responded to induction therapy with either cyclophosphamide or mycophenolate. The incidence of ovarian failure, a common effect of high-dose cyclophosphamide therapy (but probably not of low-dose therapy), can be reduced by treatment with a gonadotropin-releasing hormone agonist (e.g., leuprolide 3.75 mg intramuscularly) prior to each monthly cyclophosphamide dose. Patients with high serum creatinine levels (e.g., ≥265 μmol/L [≥3.0 mg/dL]) many months in duration and high chronicity scores on renal biopsy are not

likely to respond to any of these therapies. In general, it may be better to induce improvement in an African-American or Hispanic patient with proliferative glomerulonephritis with mycophenolate mofetil (2–3 g daily) rather than cyclophosphamide, with the option to switch if no evidence of response is detectable after 3–6 months of treatment. For whites and Asians, induction with either mycophenolate mofetil or cyclophosphamide is acceptable. Cyclophosphamide may be discontinued when it is clear that a patient is improving. The number of SLE flares is reduced by maintenance therapy with mycophenolate mofetil (1.5–2 g daily) or azathioprine (1–2.5 mg/kg per day). Both cyclophosphamide and mycophenolate mofetil are potentially teratogenic; patients should be off either medication for at least 3 months before attempting to conceive. Azathioprine can be used if necessary to control active SLE in patients who are pregnant. If azathioprine is used either for induction or maintenance therapy, patients may be prescreened for homozygous deficiency of the TMPT enzyme (which is required to metabolize the 6-mercaptopurine product of azathioprine) because they are at higher risk for bone marrow suppression.

Good improvement occurs in ~80% of lupus nephritis patients receiving either cyclophosphamide or mycophenolate at 1–2 years of follow-up. However, in some studies, at least 50% of these individuals have flares of nephritis over the next 5 years, and re-treatment is required; such individuals are more likely to progress to ESRD. Long-term outcome of lupus nephritis to most interventions is better in whites than in African Americans. Methotrexate (a folinic acid antagonist) may have a role in the treatment of arthritis and dermatitis but probably not in nephritis or other life-threatening disease. Small controlled trials (in Asia) of leflunomide, a relatively lymphocyte-specific pyrimidine antagonist licensed for use in rheumatoid arthritis, have suggested it can suppress disease activity in some SLE patients. Cyclosporine and tacrolimus, which inhibit calcium flux and therefore production of IL-2 and T lymphocyte functions, have not been studied in prospective controlled trials in SLE in the United States; several studies in Asia have shown they are effective in lupus nephritis. A recent trial in China showed that a combination of low dose mycophenolate mofetil (one gram daily) plus tacrolimus (4 mg daily) plus prednisone (pulse followed by 0.6 mg/kg/d) had a better response rate than high dose intravenous cyclophosphamide. Because calcineurin blockers have potential nephrotoxicity but little bone marrow toxicity, the author uses them for periods of a few months in patients with steroid-resistant cytopenias of SLE, in steroid-resistant patients who have developed bone marrow suppression from standard cytotoxic agents, or in patients with active SLE in spite of treatment with mycophenolate or cyclosphosphamide.

Most patients with SLE of any type should be treated with hydroxychloroquine since it prevents damage in skin and kidney and reduces overall damage scores. Patients with proteinuria > 500 mg daily should receive ACE inhibitors or ARBs, as they reduce the chance for ESRD.

Use of biologics directed against B cells for active SLE is under intense study. Use of anti-CD20 (rituximab), particularly in patients with SLE who are resistant to the more standard combination therapies discussed above, is controversial. Several open trials have shown efficacy in a majority of such patients, both for nephritis and for extrarenal lupus. However, recent prospective placebo-controlled randomized trials, one in renal and one in nonrenal SLE, did not show a difference between anti-CD20 and placebo when added to standard combination therapies. Belimumab, which is approved by the FDA for use in SLE without active renal disease (indication is serologically positive SLE which has failed standard treatments), is in clinical trials for active lupus nephritis. Drugs that kill plasma cells, used in multiple myeloma, are being studied in SLE, as are molecules and antibodies that prevent activation of B cells and/or T cells, such as Jak/Stat inhibitors.

SPECIAL CONDITIONS IN SLE THAT MAY REQUIRE ADDITIONAL OR DIFFERENT THERAPIES

Crescentic Lupus Nephritis The presence of cellular or fibrotic crescents in glomeruli with proliferative glomerulonephritis indicates a worse prognosis than in patients without this feature. There are no large prospective multinational controlled trials showing efficacy of cyclophosphamide, mycophenolate, cyclosporine, or tacrolimus in such cases. Most authorities recommend high-dose cyclophosphamide as the induction therapy of choice; there is some evidence that high dose mycophenolate mofetil is equally effective.

Membranous Lupus Nephritis Most SLE patients with membranous (INS-V) nephritis also have proliferative changes and should be treated for proliferative disease. However, some have pure membranous changes. Treatment for this group is less well defined. Some authorities do not recommend immunosuppression unless proteinuria is in the nephrotic range (although treatment with angiotensin-converting enzyme inhibitors or angiotensin II receptor blockers is recommended). In those patients, recent prospective controlled trials suggest that alternate-day glucocorticoids plus cyclophosphamide or mycophenolate mofetil or cyclosporine are all effective in the majority of patients in reducing proteinuria. It is more controversial whether they preserve renal function over the long term.

Pregnancy and Lupus Fertility rates for men and women with SLE are probably normal. However, rate of fetal loss is increased (approximately two- to threefold) in women with SLE. Fetal demise is higher in mothers with high disease activity, antiphospholipid antibodies (especially the lupus anticoagulant), hypertension, and/or active nephritis. Suppression of disease activity can be achieved by administration of systemic glucocorticoids. A placental enzyme, 11-β-dehydrogenase 2, deactivates glucocorticoids; it is more effective in deactivating prednisone and prednisolone than the fluorinated glucocorticoids dexamethasone and betamethasone. Glucocorticoids are listed by the FDA as pregnancy category A (no evidence of teratogenicity in human studies); cyclosporine, tacrolimus, and rituximab are listed as category C (may be teratogenic in animals but no good evidence in humans); azathioprine, hydroxychloroquine, mycophenolate mofetil, and cyclophosphamide are category D (there is evidence of teratogenicity in humans, but benefits might outweigh risks in certain situations); and methotrexate is category X (risks outweigh benefits). Therefore, active SLE in pregnant women should be controlled with hydroxychloroquine and, if necessary, prednisone/prednisolone at the lowest effective doses for the shortest time required. Azathioprine may be added if these treatments do not suppress disease activity. Adverse effects of prenatal glucocorticoid exposure (primarily betamethasone, which is not recommended) on offspring may include low birth weight, developmental abnormalities in the CNS, and predilection toward adult metabolic syndrome. It is likely that each of these glucocorticoids and immunosuppressive medications gets into breast milk, at least in low levels; patients should consider not breastfeeding if they need therapy for SLE. In SLE patients with antiphospholipid antibodies and prior fetal losses, treatment with heparin (usually low-molecular-weight) plus low-dose aspirin has been shown in prospective controlled trials to increase significantly the proportion of live births. Aspirin alone may be used, although most consider it less effective than heparin-plus-aspirin. Warfarin is teratogenic. Studies with oral thrombin and Factor Xa inhibitors are in progress for APSs; however, their role in preventing fetal loss is not established. An additional potential problem for the fetus is the presence of antibodies to Ro, sometimes associated with neonatal lupus consisting of rash and congenital heart block with or without cardiomyopathy. The cardiac manifestations can be life-threatening; therefore the presence of anti-Ro requires vigilant monitoring of fetal heart rates with prompt intervention (delivery if possible) if distress occurs. Recent evidence shows that hydroxychloroquine treatment of an anti-Ro-positive mother whose infant develops congenital heart block significantly reduces the chance that subsequent fetuses will develop heart block. There is some evidence that dexamethasone treatment of a mother in whom first- or second-degree heart block is detected in utero sometimes prevents progression of heart block. Women with SLE usually tolerate pregnancy without disease flares. However, a small proportion develops severe flares requiring aggressive glucocorticoid therapy or early delivery.

Lupus and Antiphospholipid Syndrome Patients with SLE who have venous or arterial clotting and/or repeated fetal losses and

at least two positive tests for antiphospholipid antibodies have APS and should be managed with long-term anticoagulation (Chap. 350). With warfarin, a target international normalized ratio (INR) of 2.0–2.5 is recommended for patients with one episode of venous clotting; an INR of 3.0–3.5 is recommended for patients with recurring clots or arterial clotting, particularly in the CNS. Recommendations are based on both retrospective and prospective studies of posttreatment clotting events and adverse effects from anticoagulation. Thrombin and Factor Xa inhibitors are under study.

Microvascular Thrombotic Crisis (Thrombotic Thrombocytopenic Purpura, Hemolytic-Uremic Syndrome) This syndrome of hemolysis, thrombocytopenia, and microvascular thrombosis in kidneys, brain, and other tissues carries a high mortality rate and occurs most commonly in young individuals with lupus nephritis. The most useful laboratory tests are identification of schistocytes on peripheral blood smears, elevated serum levels of lactate dehydrogenase, and antibodies to ADAMS13. Plasma exchange or extensive plasmapheresis is usually life-saving; most authorities recommend concomitant glucocorticoid therapy; there is no evidence that cytotoxic drugs are effective.

Lupus Dermatitis Patients with any form of lupus dermatitis should minimize exposure to ultraviolet light, using appropriate clothing and sunscreens with a sun protection factor of at least 30. Topical glucocorticoids and antimalarials (such as hydroxychloroquine) are effective in reducing lesion severity in most patients and are relatively safe. Systemic treatment with retinoic acid is a useful strategy in patients with inadequate improvement on topical glucocorticoids and antimalarials; adverse effects are potentially severe (particularly fetal abnormalities), and there are stringent reporting requirements for its use in the United States. Extensive, pruritic, bullous, or ulcerating dermatitides usually improve promptly after institution of systemic glucocorticoids; tapering may be accompanied by flare of lesions, thus necessitating use of a second medication such as hydroxychloroquine, retinoids, or belimumab. Cytotoxic medications such as methotrexate, azathioprine, or mycophenolate mofetil may also be effective. In therapy-resistant lupus dermatitis there are reports of success with topical tacrolimus (caution must be exerted because of the possible increased risk for malignancies) or with systemic dapsone or thalidomide (the extreme danger of fetal deformities from thalidomide requires permission from and supervision by the supplier; peripheral neuropathy is also common).

PREVENTIVE THERAPIES

Prevention of complications of SLE and its therapy include providing appropriate vaccinations (the administration of influenza and pneumococcal vaccines has been studied in patients with SLE; flare rates are similar to those receiving placebo) and suppressing recurrent urinary tract infections. In patients receiving glucocorticoids, the higher the daily dose the lower the immune response to vaccination; however, the great majority of patients achieve protective levels. Vaccination with attenuated live viruses is generally discouraged in patients who are immunosuppressed; however, a recent study of vaccination of a small number of SLE patient with zostavax showed safety and efficacy. Strategies to prevent osteoporosis should be initiated in most patients likely to require long-term glucocorticoid therapy and/or with other predisposing factors. Postmenopausal women can be partially protected from steroid-induced osteoporosis with calcium supplementation, Vitamin D, and either bisphosphonates or denosumab. Safety of long-term use of these strategies in premenopausal women is not well established. Control of hypertension and appropriate prevention strategies for atherosclerosis, including monitoring and treatment of dyslipidemias, management of hyperglycemia, and management of obesity, are recommended. There is increasing evidence that statin therapies can reduce deaths from cardiac events in SLE patients. Finally, the physician must keep in mind that some cancers are increased in SLE patients including non-Hodgkin lymphomas and cancers of thyroid, lung, liver, and vulvar/vaginal tissues.

EXPERIMENTAL THERAPIES

Studies of highly targeted experimental therapies for SLE are in progress. They include (1) inhibition of IFN-α, which was promising in phase II clinical trials, (2) inhibition of IL12 and IL23 signaling; (3) inhibition of IL17; (4) inhibition of IL-6; (5) elimination of plasma cells; (6) inhibition of B/T cell second signal coactivation with CTLA-Ig or anti-CD40L; (7) inhibition of innate immune activation via TLR7 or TLR7 and 9; (8) induction of regulatory T cells with peptides from immunoglobulins or autoantigens or with low doses if IL2; and (9) inhibition of lymphocyte activation by blockade of Jak/Stat. A few studies have used vigorous untargeted immunosuppression with high-dose cyclophosphamide plus anti-T cell strategies, with rescue by transplantation of autologous hematopoietic stem cells for the treatment of severe and refractory SLE. One U.S. report showed an estimated mortality rate over 5 years of 15% and sustained remission in 50%. It is hoped that in the next edition of this text, we will be able to recommend more effective and less toxic approaches to treatment of SLE based on some of these strategies.

PATIENT OUTCOMES, PROGNOSIS, AND SURVIVAL

Survival in patients with SLE in the United States, Canada, Europe, and China is ~95% at 5 years, 90% at 10 years, and 78% at 20 years. In the United States, African Americans and Hispanic Americans with a mestizo heritage have a worse prognosis than whites, whereas Africans in Africa and Hispanic Americans with a Puerto Rican origin do not. The relative importance of gene mixtures and environmental differences accounting for ethnic differences is not known. Poor prognosis (~50% mortality in 10 years) in most series is associated with (at the time of diagnosis) high serum creatinine levels (>124 μmol/L [>1.4 mg/dL]), hypertension, nephrotic syndrome (24-h urine protein excretion >2.6 g), anemia (hemoglobin <124 g/L [<12.4 g/dL]), hypoalbuminemia, hypocomplementemia, antiphospholipid antibodies, male sex, ethnicity (African American, Hispanic with mestizo heritage), and low socioeconomic status. Data regarding outcomes in SLE patients with renal transplants show mixed results: some series show a twofold increase in graft rejection compared to patients with other causes of ESRD, whereas others show no differences. Overall patient survival is comparable (85% at 2 years). Lupus nephritis occurs in ~5% of transplanted kidneys. Disability in patients with SLE is common due primarily to chronic fatigue, arthritis, and pain, as well as renal disease. As many as 30–50% of patients may achieve low disease activity (defined as mild activity on hydroxychloroquine with or without low dose glucocorticoids); fewer than 10% experience remissions (defined as no disease activity on no medications). Both of these conditions may persist for a few years, but are usually not permanent, as flares of SLE occur. The leading causes of death in the first decade of disease are systemic disease activity, renal failure, and infections; subsequently, thromboembolic events become increasingly frequent causes of mortality.

DRUG-INDUCED LUPUS

This is a syndrome of positive ANA associated with symptoms such as fever, malaise, arthritis or intense arthralgias/myalgias, serositis, and/or rash. The syndrome appears during therapy with certain medications and biologic agents, is predominant in whites, has less female predilection than SLE, rarely involves kidneys or brain, is rarely associated with anti-dsDNA, is commonly associated with antibodies to histones, and usually resolves over several weeks after discontinuation of the offending medication. The list of substances that can induce lupus-like disease is long. Among the most frequent are the antiarrhythmics procainamide, disopyramide, and propafenone; the anti-hypertensive hydralazine; several angiotensin-converting enzyme inhibitors and beta blockers; the antithyroid propylthiouracil; the antipsychotics chlorpromazine and lithium; the anticonvulsants carbamazepine and phenytoin; the antibiotics isoniazid, minocycline, and nitrofurantoin (Macrodantin); the antirheumatic sulfasalazine; the diuretic hydrochlorothiazide; the antihyperlipidemics lovastatin and simvastatin. Biologics that can cause drug-induced lupus (DIL) include inhibitors

of IFNs and TNF. In DIL, ANA usually appears before symptoms; however, many of the medications mentioned above induce ANA in patients who never develop symptoms of drug-induced lupus. It is appropriate to test for ANA at the first hint of relevant symptoms and to use test results to help decide whether to withdraw the suspect agent.

FURTHER READING

Anders HJ, Rovin B: A pathophysiology-based approach to the diagnosis and treatment of lupus nephritis. Kidney Int 90:493, 2016.

Hahn BH et al: American College of Rheumatology guidelines for screening, treatment, and management of lupus nephritis. Arthritis Care Res (Hoboken) 64:797, 2012.

Hahn BH: Pathogenesis of SLE, in *Dubois Lupus Erythematosus*, 9th ed, DJ Wallace, BH Hahn, (eds). Philadelphia, Elsevier, 2018.

Narain S, Furie R: Update on clinical trials in systemic lupus erythematosus. Curr Opin Rheumatol 28:477, 2016.

Teruel M, Alarcon-Riquelme ME: The genetic basis of systemic lupus erythematosus: What are the risk factors and what have we learned. J Autoimmun 74:161, 2016.

350 Antiphospholipid Syndrome

Haralampos M. Moutsopoulos

DEFINITIONS

Antiphospholipid syndrome (APS) is an autoantibody-mediated acquired thrombophilia characterized by recurrent arterial or venous thrombosis and/or pregnancy morbidity. It affects primarily females. APS may occur alone (primary) or in association with other autoimmune diseases, mainly systemic lupus erythematosus (SLE) (secondary). Catastrophic APS (CAPS) is a life-threatening rapidly progressive thromboembolic disease involving simultaneously three or more organs.

The major autoantibodies detected in patients' sera are directed against phospholipids and/or phospholipid (PL)-binding plasma proteins such as prothrombin and β2 glycoprotein I (β2GPI). PLs are components of the cytoplasmic membrane of all living cells. The antibodies are directed against negatively charged PLs including among others cardiolipin, phosphocholine, and phosphatidylserine. The plasma protein β2GPI is a 43-kDa plasma apolipoprotein, which consists of 326 amino acids arranged in five domains (I through V). Domain V forms a positively charged patch, suitable to interact with negatively charged PLs. In plasma, β2GPI has a circular conformation with domain V binding to and concealing the B cell epitopes lying on domain I. Another group of antibodies termed *lupus anticoagulant* (LA) prolongs clotting times in vitro, which are not corrected by adding normal plasma **(Table 350-1)**. Patients with APS often possess antibodies recognizing *Treponema pallidum* PL/cholesterol complexes, detected by Venereal Disease Research Laboratory (VDRL) tests and characterized as biologic false-positive serologic tests for syphilis (BFP-STS).

EPIDEMIOLOGY

The incidence of APS is estimated to be around 5 cases per 100,000 persons per year. Anti-PL antibodies occur in 1–5% of the general population. Their prevalence increases with age; however, it is questionable whether they are able to induce thrombotic events in elderly individuals. Moreover, one-third of patients with SLE and other autoimmune diseases **(Chap. 349)** possess these antibodies, with only 5–10% of them developing APS.

PATHOGENESIS

The initiating event for the induction of antibodies to PL-binding proteins seems to be infections, oxidative stress, and major physical stresses such as surgery or trauma. All these factors appear to induce increased apoptosis of the vessel endothelial cells and subsequent exposure of PLs. The latter, bound with serum proteins such as β2GPI or prothrombin, lead to neoantigen formation, which in turn triggers the induction of anti-PLs. The binding of anti-PLs to the disrupted endothelial cells leads to initiation of intravascular coagulation and thrombus formation. Complement activation has been also proposed as a mechanism of APS-related fetal injury.

TABLE 350-1 Classification and Nomenclature of Antiphospholipid Antibodies

NAME	ASSAY FOR THEIR DETECTION	COMMENTS
Antibodies against cardiolipin (aCL)	Enzyme-linked immunosorbent assay (ELISA) using as antigen cardiolipin (CL), a negatively charged phospholipid	aCL from patients with APS recognize β2GPI existing in the human serum as well as in bovine serum, which is used to block the nonspecific binding sites on the ELISA plate. CL simply stabilizes β2GPI at high concentration on the polystyrene surface.
Antibodies against β2GPI (anti-β2GPI)	ELISA using as antigen affinity purified or recombinant β2GPI in the absence of PL	Antibodies recognize β2GPI bound in the absence of CL to an oxidized polystyrene surface, where oxygen atoms in the moieties C–O or C=O were introduced by γ-irradiation.
Lupus anticoagulant (LA)	Activated partial thromboplastin time (aPTT) Kaolin clotting time (KCT) Dilute Russel viper venom test (DRVVT)	Antibodies recognize β2GPI or prothrombin (PT) and elongate aPTT, implying that they interfere with the generation of thrombin by prothrombin. Prolongation of the clotting times is an in vitro phenomenon, and LA induces thromboses in vivo.

Abbreviations: APL, antiphospholipid syndrome; β2GPI, β2 glycoprotein I; PL, phospholipid.

CLINICAL MANIFESTATIONS AND LABORATORY FINDINGS

Clinical manifestations represent the consequences of venous or arterial thrombosis and/or pregnancy morbidity **(Table 350-2)**. Venous thrombosis, superficial or deep, occurs primarily in the lower extremities often leading to pulmonary emboli. Thrombosis of the pulmonary arteries leads to pulmonary hypertension and thrombosis of the inferior vena cava to Budd-Chiari syndrome. Cerebral venous thrombosis, presents with signs and symptoms of intracranial hypertension and retinal vein thrombosis. Arterial thrombosis affects more commonly the arteries of the brain and is manifested as migraines, cognitive dysfunction, transient ischemic attacks, stroke, and retinal artery occlusion. Arterial thrombosis of the extremities presents with ischemic leg ulcers, digital gangrene, avascular bone necrosis, while thrombosis of other arteries leads to myocardial infarction, renal artery stenosis, glomerular lesions, and infarcts of spleen, pancreas, and adrenals.

Livedo reticularis consists of a mottled reticular vascular pattern that appears as a lace-like, purplish discoloration of the skin. It is probably caused by swelling of the venules due to obstruction of capillaries by thrombi. This clinical manifestation usually occurs together with vascular lesions in the central nervous system and with aseptic bone necrosis. Libman-Sacks endocarditis consists of very small vegetations, histologically characterized by organized platelet-fibrin microthrombi surrounded by growing fibroblasts and macrophages. Glomerular involvement is manifested with hypertension, mildly elevated serum creatinine levels, as well as mild proteinuria/hematuria. Histologically, in an acute phase, thrombotic microangiopathy is present in the glomerular capillaries. In a chronic phase, fibrous intima hyperplasia, fibrous and/or fibrocellular arteriolar occlusions, and focal cortical atrophy are present (Table 350-2). Pregnancy morbidity manifests with increased risk of recurrent miscarriages, intrauterine growth

TABLE 350-2 Clinical Features of Antiphospholipid Syndrome

MANIFESTATION	%
Venous Thrombosis and Related Consequences	
Deep vein thrombosis	39
Livedo reticularis	24
Pulmonary embolism	14
Superficial thrombophlebitis	12
Thrombosis in various other sites	11
Arterial Thrombosis and Related Consequences	
Stroke	20
Cardiac valve thickening/dysfunction and/or Libman-Sacks vegetations	14
Transient ischemic attack	11
Myocardial ischemia (infarction or angina) and coronary bypass graft thrombosis	10
Leg ulcers and/or digital gangrene	9
Arterial thrombosis in the extremities	7
Retinal artery thrombosis/amaurosis fugax	7
Ischemia of visceral organs or avascular necrosis of bone	6
Multi-infarct dementia	3
Neurologic Manifestations of Uncertain Etiology	
Migraine	20
Epilepsy	7
Chorea	1
Cerebellar ataxia	1
Transverse myelopathy	0.5
Renal Manifestations Due to Various Reasons (Renal Artery/Renal Vein/Glomerular Thrombosis, Fibrous Intima Hyperplasia)	**3**
Musculoskeletal Manifestations	
Arthralgias	39
Arthritis	27
Obstetric Manifestations (Referred to the Number of Pregnancies)	
Preeclampsia	10
Eclampsia	4
Fetal Manifestations (Referred to the Number of Pregnancies)	
Early fetal loss (<10 weeks)	35
Late fetal loss (≥10 weeks)	17
Premature birth among the live births	11
Hematologic Manifestations	
Thrombocytopenia	30
Autoimmune hemolytic anemia	10

Source: Adapted from R Cervera et al: Arthritis Rheum 46:1019, 2002.

retardation, preeclampsia, eclampsia, and preterm birth. The major causes of these complications are infarctions of the placenta.

Premature atherosclerosis has been also recognized as a feature of APS. Musculoskeletal manifestations include in addition to bone necrosis, arthralgia/arthritis, bone marrow necrosis, muscle infarction, non-traumatic fractures, and osteoporosis. Coombs-positive hemolytic anemia and thrombocytopenia are laboratory findings associated with APS.

DIAGNOSIS AND DIFFERENTIAL DIAGNOSIS

The diagnosis of APS should be seriously considered in cases of thrombosis, cerebral vascular accidents in individuals <55 years of age, or pregnancy morbidity in the presence of livedo reticularis or thrombocytopenia. In these cases, aPL antibodies should be measured. The presence of at least one clinical and one laboratory criterion is compatible with the diagnosis, in the absence of other thrombophilia causes. Clinical criteria include: (1) vascular thrombosis defined as one or more clinical episodes of arterial, venous, or small vessel thrombosis in any tissue or organ; and (2) pregnancy morbidity, defined as (a) one or more unexplained deaths of a morphologically normal fetus at or beyond the tenth week of gestation; (b) one or more premature births of a morphologically normal neonate before the thirty-fourth week of gestation because of eclampsia, severe preeclampsia, or placental insufficiency; or (c) three or more unexplained consecutive spontaneous abortions before the tenth week of gestation. Laboratory criteria include (1) LA, (2) anticardiolipin (aCL), and/or (3) anti-β2GPI antibodies, at intermediate or high titers on two occasions, 12 weeks apart.

Differential diagnosis is based on the exclusion of other inherited or acquired causes of thrombophilia **(Chap. 112)**, Coombs-positive hemolytic anemia **(Chap. 96)**, and thrombocytopenia **(Chap. 111)**. Livedo reticularis with or without a painful ulceration on the lower extremities may be also a manifestation of disorders affecting (1) the vascular wall, such as atherosclerosis, polyarteritis nodosa, SLE, cryoglobulinemia, and lymphomas; or (2) the vascular lumen, such as myeloproliferative disorders, hypercholesterolemia, or other causes of thrombophilia.

TREATMENT

Antiphospholipid Syndrome

After the first thrombotic event, APS patients should be placed on warfarin for life, aiming to achieve an international normalized ratio (INR) ranging from 2.5 to 3.5, alone or in combination with 80 mg of aspirin daily. Pregnancy morbidity is prevented by administering low-molecular-weight heparin with aspirin 80 mg daily. IV immunoglobulin (IVIg) 400 mg/kg every day for 5 days may also prevent abortions, whereas glucocorticoids are ineffective. Patients with aPL in the absence of any clinical event who are simultaneously positive for aCL, anti-β2GPI, and LA or have SLE are at risk of developing thrombotic events which can be prevented by taking aspirin 80 mg and hydroxychloroquine 200 mg daily.

Some patients with APS and patients with CAPS have recurrent thrombotic events despite appropriate anticoagulation. In these cases, IVIg 400 mg/kg every day for 5 days may be of benefit. Patients with CAPS, who are treated in the intensive care unit, are unable to receive warfarin; in this situation, therapeutic doses of low-molecular-weight heparin should be administered. In cases of heparin-induced thrombocytopenia and thrombosis syndrome, inhibitors of phospholipid-bound activated factor X (FXa), such as fondaparinux 7.5 mg SC daily or rivaroxaban 10 mg PO daily, are effective. The above drugs are administered by fixed doses and do not require close monitoring; their safety during the first trimester of pregnancy has not been clearly established.

ACKNOWLEDGMENT

I would like to thank Dr. P. G. Vlachoyiannopoulos for his contribution to the previous edition of the chapter.

FURTHER READING

ANDRADE D, TEKTONIDOU M: Emerging therapies in antiphospholipid syndrome. Curr Rheumatol Rep 18:22, 2016.

MERASHLI M et al: Antiphospholipid syndrome: An update. Eur J Clin Invest 45:653, 2015.

351 Rheumatoid Arthritis

Ankoor Shah, E. William St. Clair

INTRODUCTION

Rheumatoid arthritis (RA) is a chronic inflammatory disease of unknown etiology characterized by a symmetric polyarthritis, the most common form of chronic inflammatory arthritis. Since persistently active RA often results in articular cartilage and bone destruction and functional disability, it is vital to diagnose and treat this disease early

and aggressively before damage ensues. RA, a systemic disease, may also lead to a variety of extraarticular manifestations, including fatigue, subcutaneous nodules, lung involvement, pericarditis, peripheral neuropathy, vasculitis, and hematologic abnormalities, which must be managed accordingly.

Insights gained by a wealth of basic and clinical research over the past two decades have revolutionized the contemporary paradigms for the diagnosis and management of RA. Serum antibodies to cyclic citrullinated peptides (anti-CCPs) are routinely included with rheumatoid factor in the diagnostic evaluation of patients with suspected RA, and serve as biomarkers of prognostic significance. Advances in imaging modalities have assisted clinical decision-making by improving the detection of joint inflammation and damage. The science of RA has taken major leaps forward by illuminating new disease-related genes and environmental interactions and elucidating in more detail the molecular components and pathways of disease pathogenesis. The relative contribution of these molecular components and pathways has been further brought to light by the observed benefits of targeted biologic and small-molecule therapies. Despite this progress, incomplete understanding of the initiating events of RA and the factors perpetuating the chronic inflammatory response remains a sizable barrier to its cure and prevention.

The last two decades have witnessed a remarkable improvement in the outcomes of RA. The crippling arthritis of year's past is encountered much less frequently today. Much of this progress can be traced to the expanded therapeutic armamentarium and the adoption of early treatment intervention. The shift in treatment strategy dictates a new mindset for primary care practitioners—namely, one that demands early referral of patients with inflammatory arthritis to a rheumatologist for prompt diagnosis and initiation of therapy. Only then will patients achieve their best outcomes.

CLINICAL FEATURES

The incidence of RA increases between 25 and 55 years of age, after which it plateaus until the age of 75 and then decreases. The presenting symptoms of RA typically result from inflammation of the joints, tendons, and bursae. Patients often complain of early morning joint stiffness lasting more than 1 h that eases with physical activity. The earliest involved joints are typically the small joints of the hands and feet. The initial pattern of joint involvement may be monoarticular, oligoarticular (≤4 joints), or polyarticular (>5 joints), usually in a symmetric distribution. Some patients with inflammatory arthritis will present with too few affected joints to be classified as having RA—so-called undifferentiated inflammatory arthritis. Those with an undifferentiated arthritis who are most likely to be diagnosed later with RA have a higher number of tender and swollen joints, test positive for serum rheumatoid factor (RF) or anti-CCP antibodies, and have higher scores for physical disability.

Once the disease process of RA is established, the wrists, metacarpophalangeal (MCP), and proximal interphalangeal (PIP) joints stand out as the most frequently involved joints **(Fig. 351-1)**. Distal interphalangeal (DIP) joint involvement may occur in RA, but it usually is a manifestation of coexistent osteoarthritis. Flexor tendon tenosynovitis is a frequent hallmark of RA and leads to decreased range of motion, reduced grip strength, and "trigger" fingers. Progressive destruction of the joints and soft tissues may lead to chronic, irreversible deformities. Ulnar deviation results from subluxation of the MCP joints, with subluxation, or partial dislocation, of the proximal phalanx to the volar side of the hand. Hyperextension of the PIP joint with flexion of the DIP joint ("swan-neck deformity"), flexion of the PIP joint with hyperextension of the DIP joint ("boutonnière deformity"), and subluxation of the first MCP joint with hyperextension of the first interphalangeal (IP) joint ("Z-line deformity") also may result from damage to the tendons, joint capsule, and other soft tissues in these small joints. Inflammation about the ulnar styloid and tenosynovitis of the extensor carpi ulnaris may cause subluxation of the distal ulna, resulting in a "piano-key movement" of the ulnar styloid. Although metatarsophalangeal (MTP) joint involvement in the feet is an early feature of disease, chronic inflammation of the ankle and midtarsal regions usually comes later and may lead to pes planovalgus ("flat feet"). Large joints, including the knees and shoulders, are often affected in established disease, although these joints may remain asymptomatic for many years after onset.

FIGURE 351-1 Metacarpophalangeal and proximal interphalangeal joint swelling in rheumatoid arthritis. *(© 2018 American College of Rheumatology. Used with permission.)*

Atlantoaxial involvement of the cervical spine is clinically noteworthy because of its potential to cause compressive myelopathy and neurologic dysfunction. Neurologic manifestations are rarely a presenting sign or symptom of atlantoaxial disease, but they may evolve over time with progressive instability of C1 on C2. The prevalence of atlantoaxial subluxation has been declining in recent years, and occurs now in <10% of patients. Unlike the spondyloarthritides **(Chap. 355)**, RA rarely affects the thoracic and lumbar spine. Radiographic abnormalities of the temporomandibular joint occur commonly in patients with RA, but they are generally not associated with significant symptoms or functional impairment.

Extraarticular manifestations may develop during the clinical course of RA in up to 40% of patients, even prior to the onset of arthritis **(Fig. 351-2)**. Patients most likely to develop extraarticular disease have a history of cigarette smoking, have early onset of significant physical disability, and test positive for serum RF or anti-CCP antibodies. Subcutaneous nodules, secondary Sjögren's syndrome, interstitial lung disease (ILD), pulmonary nodules, and anemia are among the most frequently observed extraarticular manifestations. Recent studies have shown a decrease in the incidence and severity of at least some extraarticular manifestations, particularly Felty's syndrome and vasculitis.

The most common systemic and extraarticular features of RA are described in more detail in the sections below.

CONSTITUTIONAL

These signs and symptoms include weight loss, fever, fatigue, malaise, depression, and in the most severe cases, cachexia; they generally reflect a high degree of inflammation and may even precede the onset of joint symptoms. In general, the presence of a fever of >38.3°C (101°F) at any time during the clinical course should raise suspicion of systemic vasculitis (see below) or infection.

NODULES

Subcutaneous nodules have been reported to occur in 30–40% of patients and more commonly in those with the highest levels of disease activity, the disease-related shared epitope (SE) (see below), a positive test for serum RF, and radiographic evidence of joint erosions. However, more recent cohort studies suggest a declining prevalence of

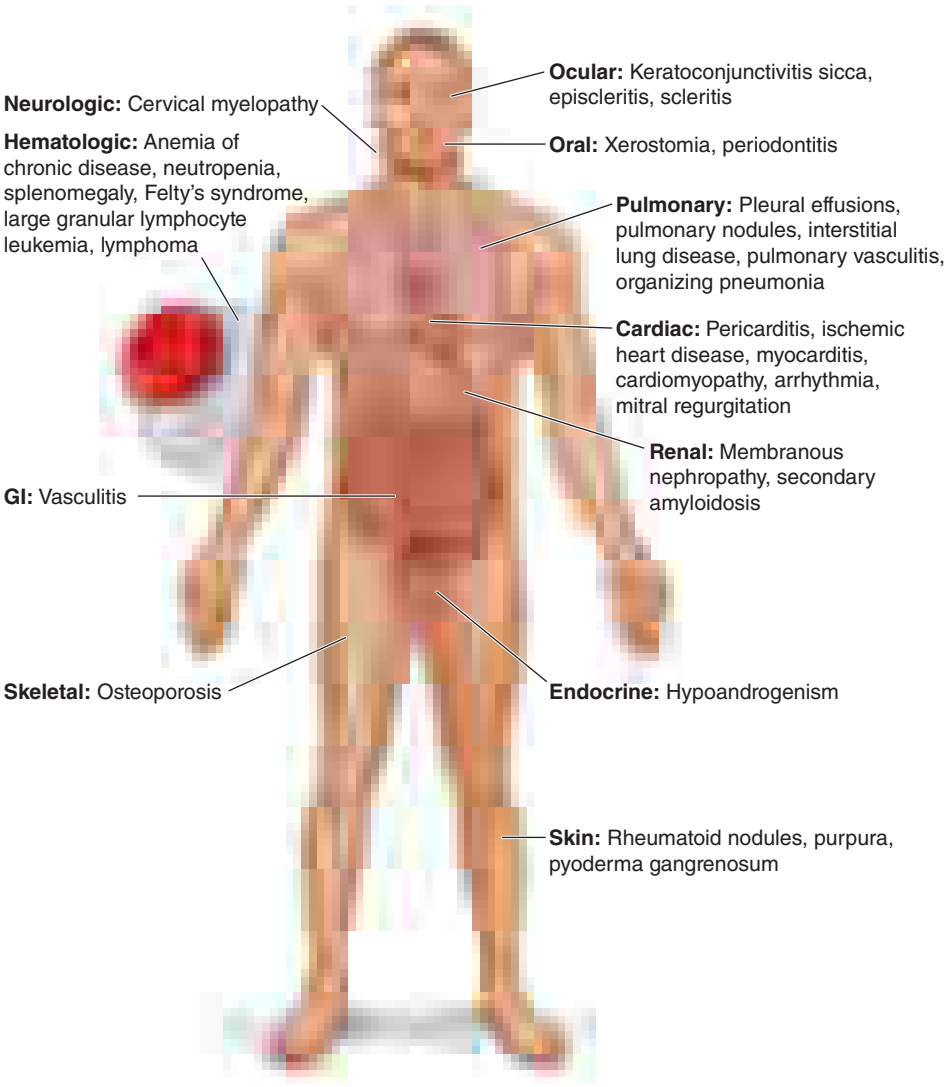


FIGURE 351-2 Extraarticular manifestations of rheumatoid arthritis.

subcutaneous nodules, perhaps, related to early and more aggressive disease-modifying therapy. When palpated, the nodules are generally firm; nontender; and adherent to periosteum, tendons, or bursae; developing in areas of the skeleton subject to repeated trauma or irritation such as the forearm, sacral prominences, and Achilles tendon. They may also occur in the lungs, pleura, pericardium, and peritoneum. Nodules are typically benign, although they can be associated with infection, ulceration, and gangrene.

SJÖGREN'S SYNDROME

Secondary Sjögren's syndrome **(Chap. 354)** is defined by the presence of either keratoconjunctivitis sicca (dry eyes) or xerostomia (dry mouth) in association with another connective tissue disease, such as RA. Approximately 10% of patients with RA have secondary Sjögren's syndrome.

PULMONARY

Pleuritis, the most common pulmonary manifestation of RA, may produce pleuritic chest pain and dyspnea, as well as a pleural friction rub and effusion. Pleural effusions tend to be exudative with increased numbers of monocytes and neutrophils. ILD may also occur in patients with RA and is heralded by symptoms of dry cough and progressive shortness of breath. ILD can be associated with cigarette smoking and is generally found in patients with higher disease activity, although it may be diagnosed in up to 3.5% of patients prior to the onset of joint symptoms. Recent studies have shown the overall prevalence of ILD in RA to be as high as 12%. Diagnosis is readily made by high-resolution chest computed tomography (CT) scan, which shows infiltrative opacification in the periphery of both lungs. Usual interstitial pneumonia (UIP) and non-specific interstitial pneumonia (NSIP) are the main histological and radiologic patterns of ILD. UIP causes progressive scarring of the lungs that produces on chest CT scan honeycomb changes in the periphery and lower portions of the lungs. In contrast, the most common radiographic changes in NSIP are relatively symmetric and bilateral ground glass opacities with associated fine reticulations, with volume loss and traction bronchiectasis. In both cases, pulmonary function testing shows a restrictive pattern (e.g., reduced total lung capacity) with a reduced diffusing capacity for carbon monoxide (DL_{CO}). The presence of ILD confers a poor prognosis, which if present, is associated with a 10% increase in mortality. The prognosis of ILD in RA is not quite as poor as that of idiopathic pulmonary fibrosis (e.g., usual interstitial pneumonitis). ILD secondary to RA responds more favorably than idiopathic ILD to immunosuppressive therapy **(Chap. 287)**. Pulmonary nodules are also common in patients with RA and may be solitary or multiple. Caplan's syndrome is a rare subset of pulmonary nodulosis characterized by the development of nodules and pneumoconiosis following silica exposure. Respiratory bronchiolitis and bronchiectasis are other less common pulmonary disorders associated with RA.

CARDIAC

The most frequent site of cardiac involvement in RA is the pericardium. However, clinical manifestations of pericarditis occur in <10% of patients with RA despite the fact that pericardial involvement is detectable in nearly one-half of cases by echocardiogram or autopsy studies. Cardiomyopathy, another clinically important manifestation of RA, may result from necrotizing or granulomatous myocarditis, coronary artery disease, or diastolic dysfunction. This involvement too may be subclinical and only identified by echocardiography or cardiac magnetic resonance imaging (MRI). Rarely, the heart muscle may contain rheumatoid nodules or be infiltrated with amyloid. Mitral regurgitation is the most common valvular abnormality in RA, occurring at a higher frequency than the general population.

VASCULITIS

Rheumatoid vasculitis **(Chap. 356)** typically occurs in patients with long-standing disease, a positive test for serum RF or anti-CCP antibodies, and hypocomplementemia. The overall incidence has decreased significantly in the last decade to <1% of patients. The cutaneous signs vary and include petechiae, purpura, digital infarcts, gangrene, livedo reticularis, and in severe cases large, painful lower extremity ulcerations. Vasculitic ulcers, which may be difficult to distinguish from those caused by venous insufficiency, may be treated successfully with immunosuppressive agents (requiring cytotoxic treatment in severe cases) as well as skin grafting. Sensorimotor polyneuropathies, such as mononeuritis multiplex, may occur in association with systemic rheumatoid vasculitis.

HEMATOLOGIC

A normochromic, normocytic anemia often develops in patients with RA and is the most common hematologic abnormality. The degree of anemia parallels the degree of inflammation, correlating with the levels

of serum C-reactive protein (CRP) and erythrocyte sedimentation rate (ESR). Platelet counts may also be elevated in RA as an acute-phase reactant. Immune-mediated thrombocytopenia is rare in this disease.

Felty's syndrome is defined by the clinical triad of neutropenia, splenomegaly, and nodular RA and is seen in <1% of patients, although its incidence appears to be declining in the face of more aggressive treatment of the joint disease. It typically occurs in the late stages of severe RA and is more common in whites than other racial groups. T cell large granular lymphocyte leukemia (T-LGL) may have a similar clinical presentation and often occurs in association with RA. T-LGL is characterized by a chronic, indolent clonal growth of LGL cells, leading to neutropenia and splenomegaly. As opposed to Felty's syndrome, T-LGL may develop early in the course of RA. Leukopenia apart from these disorders is uncommon and most often a side effect of drug therapy.

LYMPHOMA

Large cohort studies have shown a two- to fourfold increased risk of lymphoma in RA patients compared with the general population. The most common histopathologic type of lymphoma is a diffuse large B cell lymphoma. The risk of developing lymphoma increases if the patient has high levels of disease activity or Felty's syndrome.

ASSOCIATED CONDITIONS

In addition to extraarticular manifestations, several conditions associated with RA contribute to disease morbidity and mortality rates. They are worthy of mention because they affect chronic disease management.

Cardiovascular Disease The most common cause of death in patients with RA is cardiovascular disease. The incidence of coronary artery disease and carotid atherosclerosis is higher in RA patients than in the general population even when controlling for traditional cardiac risk factors, such as hypertension, obesity, hypercholesterolemia, diabetes, and cigarette smoking. Furthermore, congestive heart failure (including both systolic and diastolic dysfunction) occurs at an approximately twofold higher rate in RA than in the general population. The presence of elevated serum inflammatory markers appears to confer an increased risk of cardiovascular disease in this population.

Osteoporosis Osteoporosis is more common in patients with RA than an age- and sex-matched population, with prevalence rates of 20–30%. The inflammatory milieu of the joint probably spills over into the rest of the body and promotes generalized bone loss by activating osteoclasts. Chronic use of glucocorticoids and disability-related immobility also contributes to osteoporosis. Hip fractures are more likely to occur in patients with RA and are significant predictors of increased disability and mortality rate in this disease.

Hypoandrogenism Men and postmenopausal women with RA have lower mean serum testosterone, luteinizing hormone (LH), and dehydroepiandrosterone (DHEA) levels than control populations. It has thus been hypothesized that hypoandrogenism may play a role in the pathogenesis of RA or arise as a consequence of the chronic inflammatory response. It is also important to realize that patients receiving chronic glucocorticoid therapy may develop hypoandrogenism owing to inhibition of LH and follicle-stimulating hormone (FSH) secretion from the pituitary gland. Because low testosterone levels may lead to osteoporosis, men with hypoandrogenism should be considered for androgen replacement therapy.

EPIDEMIOLOGY

RA affects ~0.5–1% of the adult population worldwide. There is evidence that the overall incidence of RA has been decreasing in recent decades, whereas the prevalence has remained the same because individuals with RA are living longer. The incidence and prevalence of RA varies based on geographic location, both globally and among certain ethnic groups within a country **(Fig. 351-3)**. For example, the Native American Yakima, Pima, and Chippewa tribes of North America have reported prevalence rates in some studies of nearly 7%. In contrast, many population studies from Africa and Asia show lower prevalence rates for RA in the range of 0.2–0.4%.

Like many other autoimmune diseases, RA occurs more commonly in females than in males, with a 2–3:1 ratio. Interestingly, studies of RA from some of the Latin American and African countries show an even greater predominance of disease in females compared to males, with ratios of 6–8:1. Given this preponderance of females, various theories have been proposed to explain the possible role of estrogen in disease pathogenesis. Most of the theories center on the role of estrogens in enhancing the immune response. For example, some experimental studies have shown that estrogen can stimulate production of tumor necrosis factor α (TNF-α), a major cytokine in the pathogenesis of RA.

GENETIC CONSIDERATIONS

It has been recognized for over 30 years that genetic factors contribute to the occurrence of RA as well as to its severity. The likelihood that a first-degree relative of a patient will share the diagnosis of RA is 2–10 times greater than in the general population.

European ancestry:
HLA-DRB1:
- *0401
- *0404
- *0301
- ***0101**

PTPN22: European
STAT4: North American
TNFAIP3: North American
TRAFI/CF: North American
CTLA4: European

Asian ancestry:
HLA-DRB1:
- *0401 (East Asian)
- *0405
- *0901 (Japanese, Malaysian, Korean)

PADI4
CD244

Other:
CD40

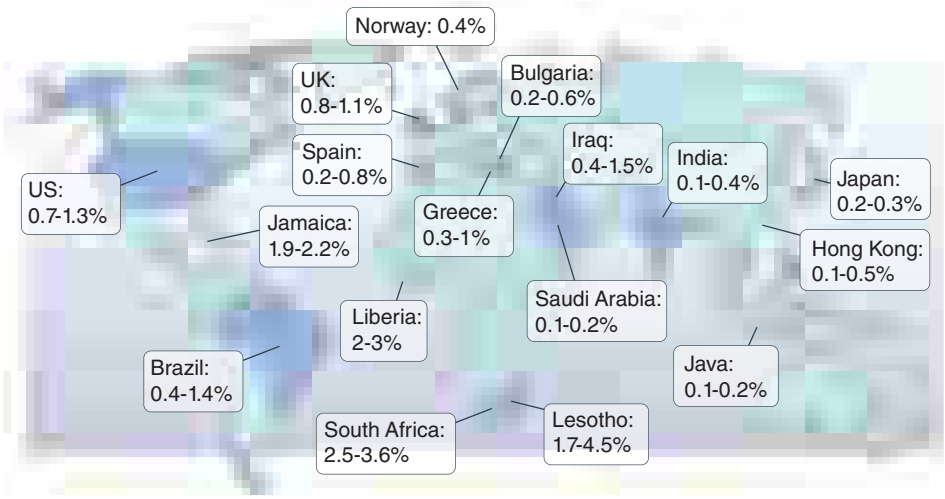


FIGURE 351-3 Global prevalence rates of rheumatoid arthritis (RA) with genetic associations. Listed are the major genetic alleles associated with RA. Although human leukocyte antigen (HLA)-DRB1 mutations are found globally, some alleles have been associated with RA in only certain ethnic groups.

There remains, however, some uncertainty in the extent to which genetics plays a role in the causative mechanisms of RA. Heritability estimates range from 40 to 50% and are approximately the same for autoantibody positive and negative individuals. The estimate of genetic influence may vary across studies due to gene–environment interactions.

The alleles known to confer the greatest risk of RA are located within the major histocompatibility complex (MHC). It has been estimated that about 13% of the genetic risk for RA resides within this locus. Most, but probably not all, of this risk is associated with allelic variation in the HLA-DRB1 gene, which encodes the MHC II β-chain molecule. The disease-associated HLA-DRB1 alleles share an amino acid sequence at positions 70–74 in the third hypervariable regions of the HLA-DR β-chain, termed the *shared epitope*. Carriership of the SE alleles is associated with production of anti-CCP antibodies and worse disease outcomes. Some of these HLA-DRB1 alleles bestow a high risk of disease (*0401), whereas others confer a more moderate risk (*0101, *0404, *1001, and *0901). Additionally, there is regional variation. In Greece, for example, where RA tends to be milder than in western European countries, RA susceptibility has been associated with the *0101 SE allele. By comparison, the *0401 or *0404 alleles are found in ~50–70% of northern Europeans and are the predominant risk alleles in this group. The most common disease susceptibility SE alleles in Asians, namely the Japanese, Koreans, and Chinese, are *0405 and *0901. Lastly, disease susceptibility of Native American populations such as the Pima and Tlingit Indians, where the prevalence of RA can be as high as 7%, is associated with the SE allele *1042. The risk of RA conferred by these SE alleles is less in African and Hispanic Americans than in individuals of European ancestry.

Genome-wide association studies (GWAS) have made possible the identification of several non-MHC-related genes that contribute to RA susceptibility. GWAS are based on the detection of single-nucleotide polymorphisms (SNPs), which allow for examination of the genetic architecture of complex diseases such as RA. There are ~10 million common SNPs within a human genome consisting of 3 billion base pairs. As a rule, GWAS identify only common variants, namely, those with a frequency of >5% in the general population.

Overall, several themes have emerged from GWAS in RA. First, among the more than 100 non-MHC loci identified as risk alleles for RA, they individually have only a modest effect on risk; they also contribute to the risk for developing other autoimmune diseases, such as type 1 diabetes mellitus, systemic lupus erythematosus, and multiple sclerosis. Second, although most of the non-HLA associations are described in patients with anti-CCP antibody-positive disease, there are several risk loci that are unique to anti-CCP antibody-negative disease. Third, risk alleles vary among ethnic groups. And fourth, the risk loci mostly reside in genes encoding proteins involved in the regulation of the immune response. However, the risk alleles identified by GWAS only account at present for ~5% of the genetic risk, suggesting that rare variants or other classes of DNA variants, such as variants in copy number, may be yet found that significantly contribute to the overall risk model.

Recently, imputation of SNP data from a GWAS meta-analysis shows amino acid substitutions in the MHC locus independently associated with the risk for RA are at positions 11, 71, and 74 in HLA-DRβ1, position 9 of HLA-B, and position 9 of HLA-DPβ1. The amino acids at positions 11, 71, and 74 are located in the antigen-binding grove of the HLA-DRβ1 molecule, highlighting positions 71 and 74 that form part of the original SE.

Among the best examples of the non-MHC genes contributing to the risk of RA is the gene encoding protein tyrosine phosphatase non-receptor 22 (*PTPN22*). This gene varies in frequency among patients from different parts of Europe (e.g., 3–10%), but is absent in patients of East Asian ancestry. *PTPN22* encodes lymphoid tyrosine phosphatase, a protein that regulates T and B cell function. Inheritance of the risk allele for *PTPN22* produces a gain-of-function in the protein that is hypothesized to result in the abnormal thymic selection of autoreactive T and B cells and appears to be associated exclusively with anti-CCP-positive disease. The peptidyl arginine deiminase type IV (*PADI4*) gene is another risk allele that encodes an enzyme involved in the conversion of arginine to citrulline and is postulated to play a role in the development of antibodies to citrullinated antigens. A polymorphism in *PADI4* has been associated with RA only in Asian populations. Recently, polymorphisms in apolipoprotein M (APOM) have been demonstrated in an East Asian population to confer an increased risk for RA as well as risk for dyslipidemia, independent of RA disease activity.

Epigenetics is the study of heritable traits that affect gene expression but do not modify DNA sequence. It may provide a link between environmental exposure and predisposition to disease. The best-studied mechanisms include posttranslational histone modifications and DNA methylation. Although studies of epigenetic phenomena are limited, DNA methylation patterns have been shown to differ between RA patients and healthy controls, as well as patients with osteoarthritis. MicroRNAs, which are non-coding RNAs that function as post-transcriptional regulators of gene expression, represent an additional epigenetic mechanism that may potentially influence cellular responses. Many microRNAs have been identified as contributing to the activated phenotype of synovial fibroblasts such as miR146a or miR155.

ENVIRONMENTAL FACTORS

In addition to genetic predisposition, a host of environmental factors have been implicated in the pathogenesis of RA. The most reproducible of these environmental links is cigarette smoking. Numerous cohort and case control studies have demonstrated that smoking confers a relative risk for developing RA of 1.5–3.5. In particular, women who smoke cigarettes have a nearly 2.5 times greater risk of RA, a risk that persists even 15 years after smoking cessation. A twin who smokes will have a significantly higher risk for RA than his or her monozygotic co-twin, theoretically with the same genetic risk, who does not smoke. Interestingly, the risk from smoking is almost exclusively related to RF and anti-CCP antibody-positive disease. However, it has not been shown that smoking cessation, while having many health benefits, improves disease activity.

Researchers began to aggressively seek an infectious etiology for RA after the discovery in 1931 that sera from patients with this disease could agglutinate strains of streptococci. Certain viruses such as Epstein-Barr virus (EBV) have garnered the most interest over the past 30 years given their ubiquity, ability to persist for many years in the host, and frequent association with arthritic complaints. For example, titers of IgG antibodies against EBV antigens in the peripheral blood and saliva are significantly higher in patients with RA than the general population. EBV DNA has also been found in synovial fluid and synovial cells of RA patients. Because the evidence for these links is largely circumstantial, it has not been possible to directly implicate infection as a causative factor in RA.

Recent studies suggest that periodontitis may play a role in the pathogenesis of RA. Multiple studies provide evidence for a link between anti-CCP positive RA and cigarette smoking, periodontal disease, and the oral microbiome, specifically *Porphyromonas gingivalis*. It has been hypothesized that the immune response to *P. gingivalis* may trigger the development of RA and that induction of anti-CCP antibodies results from citrullination of arginine residues in human tissues by the enzyme peptidyl arginine deiminase (PAD). Interestingly, *P. gingivalis* is the only oral bacterial species known to harbor this enzyme. Some studies have shown a relationship between circulating antibodies to *P. gingivalis* and RA, as well as these antibodies and first-degree relatives at risk for this disease.

PATHOLOGY

RA affects the synovial tissue and underlying cartilage and bone. The synovial membrane, which covers most articular surfaces, tendon sheaths, and bursae, normally is a thin layer of connective tissue. In joints, it faces the bone and cartilage, bridging the opposing bony surfaces and inserting at periosteal regions close to the articular cartilage. It consists primarily of two cell types—type A synoviocytes (macrophage-derived) and type B synoviocytes (fibroblast-derived). The synovial fibroblasts are the most abundant and produce the structural components of joints, including collagen, fibronectin, and laminin, as well as other extracellular constituents of the synovial matrix.

The sublining layer consists of blood vessels and a sparse population of mononuclear cells within a loose network of connective tissue. Synovial fluid, an ultrafiltrate of blood, diffuses through the subsynovial lining tissue across the synovial membrane and into the joint cavity. Its main constituents are hyaluronan and lubricin. Hyaluronan is a glycosaminoglycan that contributes to the viscous nature of synovial fluid, which along with lubricin, lubricates the surface of the articular cartilage.

The pathologic hallmarks of RA are synovial inflammation and proliferation, focal bone erosions, and thinning of articular cartilage. Chronic inflammation leads to synovial lining hyperplasia and the formation of pannus, a thickened cellular membrane containing fibroblast-like synoviocytes and granulation-reactive fibrovascular tissue that invades the underlying cartilage and bone. The inflammatory infiltrate is made up of no less than six cell types: T cells, B cells, plasma cells, dendritic cells, mast cells, and, to a lesser extent, granulocytes. The T cells comprise 30–50% of the infiltrate, with the other cells accounting for the remainder. The topographical organization of these cells is complex and may vary among individuals with RA. Most often, the lymphocytes are diffusely organized among the tissue resident cells; however, in some cases, the B cells, T cells, and dendritic cells may form higher levels of organization, such as lymphoid follicles and germinal center–like structures. Growth factors secreted by synovial fibroblasts and macrophages promote the formation of new blood vessels in the synovial sublining that supply the increasing demands for oxygenation and nutrition required by the infiltrating leukocytes and expanding synovial tissue.

The structural damage to the mineralized cartilage and subchondral bone is mediated by the osteoclast. Osteoclasts are multinucleated giant cells that can be identified by their expression of CD68, tartrate-resistant acid phosphatase, cathepsin K, and the calcitonin receptor. They appear at the pannus-bone interface where they eventually form resorption lacunae. These lesions typically localize where the synovial membrane inserts into the periosteal surface at the edges of bones close to the rim of articular cartilage and at the attachment sites of ligaments and tendon sheaths. This process most likely explains why bone erosions usually develop at the radial sites of the MCP joints juxtaposed to the insertion sites of the tendons, collateral ligaments, and synovial membrane. Another form of bone loss is periarticular osteopenia that occurs in joints with active inflammation. It is associated with substantial thinning of the bony trabeculae along the metaphyses of bones, and likely results from inflammation of the bone marrow cavity. These lesions can be visualized on MRI scans, where they appear as signal alterations in the bone marrow adjacent to inflamed joints. Their signal characteristics show they are water-rich with a low fat content and are consistent with highly vascularized inflammatory tissue. These bone marrow lesions are often the forerunner of bone erosions.

The cortical bone layer that separates the bone marrow from the invading pannus is relatively thin and susceptible to penetration by the inflamed synovium. The bone marrow lesions seen on MRI scans are associated with an endosteal bone response characterized by the accumulation of osteoblasts and deposition of osteoid. Thus, in recent years, the concept of joint pathology in RA has been extended to include the bone marrow cavity. Finally, generalized osteoporosis, which results in the thinning of trabecular bone throughout the body, is a third form of bone loss found in patients with RA.

Articular cartilage is an avascular tissue comprised of a specialized matrix of collagens, proteoglycans, and other proteins. It is organized in four distinct regions (superficial, middle, deep, and calcified cartilage zones)—chondrocytes constitute the unique cellular component in these layers. Originally, cartilage was considered to be an inert tissue, but it is now known to be a highly responsive tissue that reacts to inflammatory mediators and mechanical factors, which in turn, alter the balance between cartilage anabolism and catabolism. In RA, the initial areas of cartilage degradation are juxtaposed to the synovial pannus. The cartilage matrix is characterized by a generalized loss of proteoglycan, most evident in the superficial zones adjacent to the synovial fluid. Degradation of cartilage may also take place in the perichondrocytic zone and in regions adjacent to the subchondral bone.

PATHOGENESIS

The pathogenic mechanisms of synovial inflammation are likely to result from a complex interplay of genetic, environmental, and immunologic factors that produces dysregulation of the immune system and a breakdown in self-tolerance **(Fig. 351-4)**. Precisely what triggers these initiating events and what genetic and environmental factors disrupt the immune system remains a mystery. However, a detailed molecular picture is emerging of the mechanisms underlying the chronic inflammatory response and the destruction of the articular cartilage and bone.

In RA, the preclinical stage appears to be characterized by a breakdown in self-tolerance. This idea is supported by the finding that autoantibodies, such as RF and anti-CCP antibodies, may be found in sera from patients many years before onset of clinical disease. However, the antigenic targets of anti-CCP antibodies and RF are not restricted to the joint, and their role in disease pathogenesis remains speculative. Anti-CCP antibodies are directed against deaminated peptides, which result from posttranslational modification by the enzyme PADI4. They recognize citrulline-containing regions of several different matrix proteins, including filaggrin, keratin, fibrinogen, and vimentin, and are present at higher levels in the joint fluid compared to the serum. Other autoantibodies have been found in a minority of patients with RA, but they also occur in the setting of other types of arthritis. They bind to a diverse array of autoantigens, including type II collagen, human cartilage gp-39, aggrecan, calpastatin, immunoglobulin binding protein (BiP), and glucose-6-phosphate isomerase.

In theory, environmental stimulants may synergize with other factors to bring about inflammation in RA. People who smoke display higher citrullination of proteins in bronchoalveolar fluid than those who do not smoke. Thus, it has been speculated that long-term exposure to tobacco smoke might induce citrullination of cellular proteins in the lung and stimulate the expression of a neoepitope capable of inducing self-reactivity, which in turns, leads to formation of immune complexes and joint inflammation. Exposure to silicone dust and mineral oil, which has adjuvant effects, has also been linked to an increased risk for anti-CCP antibody-positive RA. Similarly, periodontal pathogens, such as *P. gingivalis*, may play a pathogenic role and contribute to the citrullination of cellular proteins in the oral cavity. In addition to the possible link between the oral microbiome and RA, investigators are turning their attention toward the intestinal microbiota and whether its altered composition may predispose to disease.

How might microbes or their products be involved in the initiating events of RA? The immune system is alerted to the presence of microbial infections through Toll-like receptors (TLRs). There are 10 TLRs in humans that recognize a variety of microbial products, including bacterial cell-surface lipopolysaccharides and heat-shock proteins (TLR4), lipoproteins (TLR2), double-strand RNA viruses (TLR3), and unmethylated CpG DNA from bacteria (TLR9). TLR2, 3, and 4 are abundantly expressed by synovial fibroblasts in early RA and, when bound by their ligands, upregulate production of proinflammatory cytokines. Although TLR ligands may theoretically amplify inflammatory pathways in RA, their specific role in disease pathogenesis remains uncertain.

The pathogenesis of RA is built upon the concept that self-reactive T cells drive the chronic inflammatory response. In theory, self-reactive T cells might arise in RA from abnormal central (thymic) selection or intrinsic defects lowering the threshold in the periphery for T cell activation. Either mechanism might result in abnormal expansion of the self-reactive T cell repertoire and a breakdown in T cell tolerance. The support for these theories comes mainly from studies of arthritis in mouse models. It has not been shown that patients with RA have abnormal thymic selection of T cells or defective apoptotic pathways regulating cell death. At least some antigen stimulation inside the joint seems likely, owing to the fact that T cells in the synovium express a cell-surface phenotype indicating prior antigen exposure and show evidence of clonal expansion. Of interest, peripheral blood T cells from patients with RA have been shown to display a fingerprint of premature aging that mostly affects inexperienced naïve T cells. In these studies, the most glaring findings have been the loss of telomeric

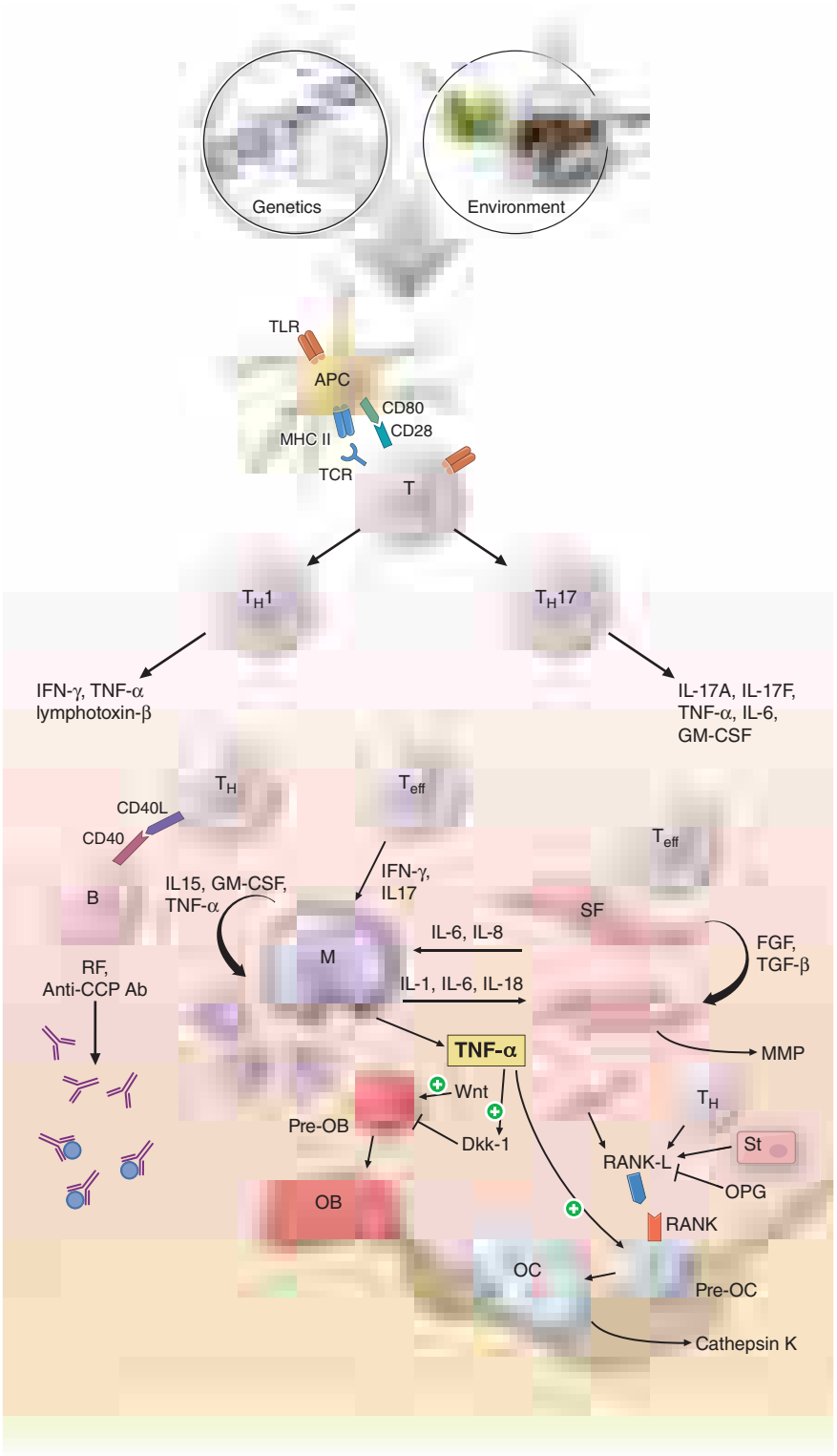


FIGURE 351-4 Pathophysiologic mechanisms of inflammation and joint destruction. Genetic predisposition along with environmental factors may trigger the development of rheumatoid arthritis (RA), with subsequent synovial T cell activation. CD4+ T cells become activated by antigen-presenting cells (APCs) through interactions between the T cell receptor and class II MHC-peptide antigen (signal 1) with co-stimulation through the CD28-CD80/86 pathway, as well as other pathways (signal 2). In theory, ligands binding Toll-like receptors (TLRs) may further stimulate activation of APCs inside the joint. Synovial CD4+ T cells differentiate into T_H1 and T_H17 cells, each with their distinctive cytokine profile. CD4+ T_H cells in turn activate B cells, some of which are destined to differentiate into autoantibody-producing plasma cells. Immune complexes, possibly comprised of rheumatoid factors (RFs) and anti–cyclic citrullinated peptides (CCP) antibodies, may form inside the joint, activating the complement pathway and amplifying inflammation. T effector cells stimulate synovial macrophages (M) and fibroblasts (SF) to secrete proinflammatory mediators, among which is tumor necrosis factor α (TNF-α). TNF-α upregulates adhesion molecules on endothelial cells, promoting leukocyte influx into the joint. It also stimulates the production of other inflammatory mediators, such as interleukin 1 (IL-1), IL-6, and granulocyte-macrophage colony-stimulating factor (GM-CSF). TNF-α has a critically important function in regulating the balance between bone destruction and formation. It upregulates the expression of dickkopf-1 (DKK-1), which can then internalize Wnt receptors on osteoblast precursors. Wnt is a soluble mediator that promotes osteoblastogenesis and bone formation. In RA, bone formation is inhibited through the Wnt pathway, presumably due to the action of elevated levels of DKK-1. In addition to inhibiting bone formation, TNF-α stimulates osteoclastogenesis. However, it is not sufficient by itself to induce the differentiation of osteoclast precursors (Pre-OC) into activated osteoclasts capable of eroding bone. Osteoclast differentiation requires the presence of macrophage colony-stimulating factor (M-CSF) and receptor activator of nuclear factor-κB (RANK) ligand (RANKL), which binds to RANK on the surface of Pre-OC. Inside the joint, RANKL is mainly derived from stromal cells, synovial fibroblasts, and T cells. Osteoprotegerin (OPG) acts as a decoy receptor for RANKL, thereby inhibiting osteoclastogenesis and bone loss. FGF, fibroblast growth factor; IFN, interferon; TGF, transforming growth factor.

sequences and a decrease in the thymic output of new T cells. Although intriguing, it is not clear how generalized T cell abnormalities might provoke a systemic disease with a predominance of synovitis.

There is substantial evidence of a role for CD4+ T cells in the pathogenesis of RA. First, the co-receptor CD4 on the surface of T cells binds to invariant sites on MHC class II molecules, stabilizing the MHC-peptide–T cell receptor complex during T cell activation. Because the SE on MHC class II molecules is a risk factor for RA, it follows that CD4+ T cell activation may play a role in the pathogenesis of this disease. Second, CD4+ memory T cells are enriched in the synovial tissue from patients with RA and can be implicated through "guilt by association." Third, CD4+ T cells have been shown to be important in the initiation of arthritis in animal models. Fourth, some, but not all, T cell–directed therapies have shown clinical efficacy in this disease. Taken together, these lines of evidence suggest that CD4+ T cells play an important role in orchestrating the chronic inflammatory response in RA. However, other cell types, such as CD8+ T cells, natural killer (NK) cells, and B cells are present in synovial tissue and may also influence pathogenic responses.

In the rheumatoid joint, by mechanisms of cell-cell contact and release of soluble mediators, activated T cells stimulate macrophages and fibroblast-like synoviocytes to generate proinflammatory mediators and proteases that drive the synovial inflammatory response and destroy the cartilage and bone. CD4+ T cell activation is dependent on two signals: (1) T cell receptor binding to peptide-MHC on antigen-presenting cells; and (2) CD28 binding to CD80/86 on antigen-presenting cells. CD4+ T cells also provide help to B cells, which in turn, produce antibodies that may promote further inflammation in the joint. The previous T cell–centric model for the pathogenesis of RA was based on a T_H1-driven paradigm, which came from studies indicating that CD4+ T helper (T_H) cells differentiated into T_H1 and T_H2 subsets, each with their distinctive cytokine profiles. T_H1 cells were found to mainly produce interferon γ (IFN-γ), lymphotoxin β, and TNF-α, whereas T_H2 cells predominately secreted interleukin (IL)-4, IL-5, IL-6, IL-10, and IL-13. The recent discovery of another subset of T_H cells, namely the T_H17 lineage, has revolutionized our concepts concerning the pathogenesis of RA. In humans, naïve T cells are induced to differentiate into T_H17 cells by exposure to transforming growth factor β (TGF-β), IL-1, IL-6, and IL-23. Upon activation, T_H17 cells secrete a variety of proinflammatory mediators such as IL-17, IL-21, IL-22, TNF-α, IL-26, IL-6, and granulocyte-macrophage colony-stimulating factor (GM-CSF). Substantial evidence now exists from studies in both animal models and humans that IL-17 plays an important role not only in promoting joint inflammation, but also in destroying cartilage and subchondral bone. Nevertheless, secukinumab, an anti-IL17 receptor antibody, failed to show significant clinical benefit in a phase II trial involving patients with RA, raising new questions about the importance of IL-17 in perpetuating joint inflammation in this disease.

The immune system has evolved mechanisms to counterbalance the potential harmful immune-mediated inflammatory responses provoked by infectious agents and other triggers. Among these negative regulators are regulatory T (T_{reg}) cells, which are produced in the thymus and induced in the periphery to suppress immune-mediated inflammation. They are characterized by the surface expression of CD25 and the expression of the transcription factor forkhead box P3 (FOXP3) and the absence of CD127, the IL-7 receptor. T_{regs} orchestrate dominant tolerance through contact with other immune cells and secretion of inhibitory cytokines, such as TGF-β, IL-10, and IL-35. They are heterogeneous and capable of suppressing distinct classes (T_H1, T_H2, T_H17) of the immune response. In RA, the data that T_{reg} numbers are deficient compared to normal healthy controls are contradictory and inconclusive. Although some experimental evidence suggests that T_{reg} suppressive activity is lost due to dysfunctional expression of cytotoxic T lymphocyte antigen 4 (CTLA-4), the nature of T_{reg} defects in RA and their role in disease mechanisms remains unclear.

Cytokines, chemokines, antibodies, and endogenous danger signals bind to receptors on the surface of immune cells and stimulate a cascade of intracellular signaling events that can amplify the inflammatory response. Signaling molecules and their binding partners in these pathways are the target of small-molecule drugs designed to interfere with signal transduction and in turn, block these reinforcing inflammatory loops. Examples of signaling molecules in these critical inflammatory pathways include Janus kinase (JAK)/signal transducers and activators of transcription (STAT), spleen tyrosine kinase (Syk), mitogen-activated protein kinases (MAPKs), and nuclear factor-κB (NF-κB). These pathways exhibit significant cross-talk and are found in many cell types. Some signal transducers, such as JAK3, are primarily expressed in hematopoietic cells and play an important role in the inflammatory response in RA.

Activated B cells are also important players in the chronic inflammatory response. B cells give rise to plasma cells, which in turn, produce antibodies, including RF and anti-CCP antibodies. RFs may form large immune complexes inside the joint that contribute to the pathogenic process by fixing complement and promoting the release of proinflammatory cytokines and chemokines. In mouse models of arthritis, RF-containing immune complexes and anti-CCP-containing immune complexes synergize with other mechanisms to exacerbate the synovial inflammatory response.

RA is often considered to be a macrophage-driven disease because this cell type is the predominant source of proinflammatory cytokines inside the joint. Key proinflammatory cytokines released by synovial macrophages include TNF-α, IL-1, IL-6, IL-12, IL-15, IL-18, and IL-23. Synovial fibroblasts, the other major cell type in this microenvironment, produce the cytokines IL-1 and IL-6 as well as TNF-α. TNF-α is a pivotal cytokine in the pathobiology of synovial inflammation. It upregulates adhesion molecules on endothelial cells, promoting the influx of leukocytes into the synovial microenvironment; activates synovial fibroblasts; stimulates angiogenesis; promotes pain receptor sensitizing pathways; and drives osteoclastogenesis. Fibroblasts secrete matrix metalloproteinases (MMPs) as well as other proteases that are chiefly responsible for the breakdown of articular cartilage.

Osteoclast activation at the site of the pannus is closely tied to the presence of focal bone erosion. Receptor activator of nuclear factor-κB ligand (RANKL) is expressed by stromal cells, synovial fibroblasts, and T cells. Upon binding to its receptor RANK on osteoclast progenitors, RANKL stimulates osteoclast differentiation and bone resorption. RANKL activity is regulated by osteoprotegerin (OPG), a decoy receptor of RANKL that blocks osteoclast formation. Monocytic cells in the synovium serve as the precursors of osteoclasts and, when exposed to macrophage colony-stimulating factor (M-CSF) and RANKL, fuse to form polykaryons termed *preosteoclasts*. These precursor cells undergo further differentiation into osteoclasts with the characteristic ruffled membrane. Cytokines such as TNF-α, IL-1, IL-6, and IL-17 increase the expression of RANKL in the joint and thus promote osteoclastogenesis. Osteoclasts also secrete cathepsin K, a cysteine protease that degrades the bone matrix by cleaving collagen. Stimulation of osteoclasts also contributes to generalized bone loss and osteoporosis.

Increased bone loss is only part of the story in RA, as decreased bone formation plays a crucial role in bone remodeling at sites of inflammation. Recent evidence shows that inflammation suppresses bone formation. The proinflammatory cytokine TNF-α plays a key role in actively suppressing bone formation by enhancing the expression of dickkopf-1 (DKK-1). DKK-1 is an important inhibitor of the Wnt pathway, which acts to promote osteoblast differentiation and bone formation. The Wnt system is a family of soluble glycoproteins that binds to cell-surface receptors known as frizzled (fz) and low-density lipoprotein (LDL) receptor–related proteins (LRPs) and promotes cell growth. In animal models, increased levels of DKK-1 are associated with decreased bone formation, whereas inhibition of DKK-1 protects against structural damage in the joint. Wnt proteins also induce the formation of OPG and thereby shut down bone resorption, emphasizing their key role in tightly regulating the balance between bone resorption and formation.

DIAGNOSIS

The clinical diagnosis of RA is largely based on signs and symptoms of a chronic inflammatory arthritis, with laboratory and radiographic results providing important corroborating information. In 2010, a collaborative effort between the American College of Rheumatology (ACR) and the European League Against Rheumatism (EULAR)

TABLE 351-1 Classification Criteria for Rheumatoid Arthritis

		SCORE
Joint involvement	1 large joint (shoulder, elbow, hip, knee, ankle)	0
	2–10 large joints	1
	1–3 small joints (MCP, PIP, thumb IP, MTP, wrists)	2
	4–10 small joints	3
	>10 joints (at least 1 small joint)	5
Serology	Negative RF and negative ACPA	0
	Low-positive RF or low-positive anti-CCP antibodies (≤3 times ULN)	2
	High-positive RF or high-positive anti-CCP antibodies (>3 times ULN)	3
Acute-phase reactants	Normal CRP and normal ESR	0
	Abnormal CRP or abnormal ESR	1
Duration of symptoms	<6 weeks	0
	≥6 weeks	1

Note: These criteria are aimed at classification of newly presenting patients who have at least one joint with definite clinical synovitis that is not better explained by another disease. A score of ≥6 fulfills requirements for definite RA.

Abbreviations: ACPA, anti-citrullinated peptide antibodies; CCP, cyclic citrullinated peptides; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; IP, interphalangeal joint; MCP, metacarpophalangeal joint; MTP, metatarsophalangeal joint; PIP, proximal interphalangeal joint; RF, rheumatoid factor; ULN, upper limit of normal.

Source: D Aletaha et al: Arthritis Rheum 62:2569, 2010.

revised the 1987 ACR classification criteria for RA in an effort to improve early diagnosis with the goal of identifying patients who would benefit from early introduction of disease-modifying therapy **(Table 351-1)**. Application of the newly revised criteria yields a score of 0–10, with a score of ≥6 fulfilling the requirements for definite RA. The new classification criteria differ in several ways from the older criteria set. The new criteria include as an item a positive test for serum anti-CCP antibodies (also termed ACPA, anti-citrullinated peptide antibodies), which carries greater specificity for the diagnosis of RA than a positive test for RF. The newer classification criteria also do not take into account whether the patient has rheumatoid nodules or radiographic joint damage because these findings occur rarely in early RA. It is important to emphasize that the new 2010 ACR-EULAR criteria are "classification criteria" as opposed to "diagnostic criteria" and serve to distinguish patients at the onset of disease who have a high likelihood of evolution to chronic disease with persistent synovitis and joint damage. The presence of radiographic joint erosions or subcutaneous nodules may inform the diagnosis in the later stages of the disease.

LABORATORY FEATURES

Patients with systemic inflammatory diseases such as RA will often present with elevated nonspecific inflammatory markers such as an ESR or CRP. Detection of serum RF and anti-CCP antibodies is important in differentiating RA from other polyarticular diseases, although RF lacks diagnostic specificity and may be found in association with other chronic inflammatory diseases in which arthritis figures in the clinical manifestations.

IgM, IgG, and IgA isotypes of RF occur in sera from patients with RA, although the IgM isotype is the one most frequently measured by commercial laboratories. Serum IgM RF has been found in 75–80% of patients with RA; therefore, a negative result does not exclude the presence of this disease. It is also found in other connective tissue diseases, such as primary Sjögren's syndrome, systemic lupus erythematosus, and type II mixed essential cryoglobulinemia, as well as chronic infections such as subacute bacterial endocarditis and hepatitis B and C. Serum RF may also be detected in 1–5% of the healthy population.

The presence of serum anti-CCP antibodies has about the same sensitivity as serum RF for the diagnosis of RA. However, its diagnostic specificity approaches 95%, so a positive test for anti-CCP antibodies in the setting of an early inflammatory arthritis is useful for distinguishing RA from other forms of arthritis. There is some incremental value in testing for the presence of both RF and anti-CCP, as some patients with RA are positive for RF but negative for anti-CCP and vice versa. The presence of RF or anti-CCP antibodies also has prognostic significance, with anti-CCP antibodies showing the most value for predicting worse outcomes.

SYNOVIAL FLUID ANALYSIS

Typically, the cellular composition of synovial fluid from patients with RA reflects an acute inflammatory state. Synovial fluid white blood cell (WBC) counts can vary widely, but generally range between 5000 and 50,000 WBC/μL compared to <2000 WBC/μL for a noninflammatory condition such as osteoarthritis. In contrast to the synovial tissue, the overwhelming cell type in the synovial fluid is the neutrophil. Clinically, the analysis of synovial fluid is most useful for confirming an inflammatory arthritis (as opposed to osteoarthritis), while at the same time excluding infection or a crystal-induced arthritis such as gout or pseudogout **(Chap. 365)**.

JOINT IMAGING

Joint imaging is a valuable tool not only for diagnosing RA, but also for tracking progression of any joint damage. Plain x-ray is the most common imaging modality, but it is limited to visualization of the bony structures and inferences about the state of the articular cartilage based on the amount of joint space narrowing. MRI and ultrasound techniques offer the added value of detecting changes in the soft tissues such as synovitis, tenosynovitis, and effusions, as well as providing greater sensitivity for identifying bony abnormalities. Plain radiographs are usually relied upon in clinical practice for the purpose of diagnosis and monitoring of affected joints. However, in selected cases, MRI and ultrasound can provide additional diagnostic information that may guide clinical decision making. Musculoskeletal ultrasound with power Doppler is increasingly used in rheumatology clinical practice for detecting synovitis and bone erosion.

Plain Radiography Classically in RA, the initial radiographic finding is periarticular osteopenia. Practically speaking, however, this finding is difficult to appreciate on plain films and, in particular, on the newer digitalized x-rays. Other findings on plain radiographs include soft tissue swelling, symmetric joint space loss, and subchondral erosions, most frequently in the wrists and hands (MCPs and PIPs) and the feet (MTPs). In the feet, the lateral aspect of the fifth MTP is often targeted first, but other MTP joints may be involved at the same time. X-ray imaging of advanced RA may reveal signs of severe destruction, including joint subluxation and collapse **(Fig. 351-5)**.

MRI MRI offers the greatest sensitivity for detecting synovitis and joint effusions, as well as early bone and bone marrow changes. These soft tissue abnormalities often occur before osseous changes are noted on x-ray. Presence of bone marrow edema has been recognized to be an early sign of inflammatory joint disease and can predict the subsequent development of erosions on plain radiographs as well as MRI scans.

FIGURE 351-5 X-ray demonstrating progression of erosions on the proximal interphalangeal joint. *(© 2018 American College of Rheumatology. Used with permission.)*

Cost and availability of MRI are the main factors limiting its routine clinical use.

Ultrasound Ultrasound, including power color Doppler, has the ability to detect more erosions than plain radiography, especially in easily accessible joints. It can also reliably detect synovitis, including increased joint vascularity indicative of inflammation. The usefulness of ultrasound is dependent on the experience of the sonographer; however, it does offer the advantages of portability, lack of radiation, and low expense relative to MRI, factors that make it attractive as a clinical tool.

CLINICAL COURSE

The natural history of RA is complex and affected by a number of factors including age of onset, gender, genotype, phenotype (i.e., extraarticular manifestations or variants of RA), and comorbid conditions, which make for a truly heterogeneous disease. There is no simple way to predict the clinical course. It is important to realize that as many as 10% of patients with inflammatory arthritis fulfilling ACR classification criteria for RA will undergo a spontaneous remission within 6 months (particularly seronegative patients). However, the vast majority of patients will exhibit a pattern of persistent and progressive disease activity that waxes and wanes in intensity over time. A minority of patients will show intermittent and recurrent explosive attacks of inflammatory arthritis interspersed with periods of disease quiescence. Finally, an aggressive form of RA may occur in an unfortunate few with inexorable progression of severe erosive joint disease, although this highly destructive course is less common in the modern treatment era.

Disability, as measured by the Health Assessment Questionnaire (HAQ), shows gradual worsening of disability over time in the face of poorly controlled disease activity and disease progression. Disability may result from both a disease activity–related component that is potentially reversible with therapy and a joint damage–related component owing to the cumulative and largely irreversible effects of soft tissue, cartilage and bone breakdown. Early in the course of disease, the extent of joint inflammation is the primary determinant of disability, while in the later stages of disease, the amount of joint damage is the dominant contributing factor. Previous studies have shown that more than one-half of patients with RA are unable to work 10 years after the onset of their disease; however, increased employability and less work absenteeism has been reported recently with the use of newer therapies and earlier treatment intervention.

The overall mortality rate in RA is two times greater than the general population, with ischemic heart disease being the most common cause of death followed by infection. Median life expectancy is shortened by an average of 7 years for men and 3 years for women compared to control populations. Patients at higher risk for shortened survival are those with systemic extraarticular involvement, low functional capacity, low socioeconomic status, low education, and chronic prednisone use.

TREATMENT

Rheumatoid Arthritis

The amount of clinical disease activity in patients with RA reflects the overall burden of inflammation and is the variable most influencing treatment decisions. Joint inflammation is the main driver of joint damage and is the most important cause of functional disability in the early stages of disease. Several composite indices have been developed to assess clinical disease activity. The ACR 20, 50, and 70 improvement criteria (which corresponds to a 20, 50, and 70% improvement, respectively, in joint counts, physician/patient assessment of disease severity, pain scale, serum levels of acute-phase reactants [ESR or CRP], and a functional assessment of disability using a self-administered patient questionnaire) are a composite index with a dichotomous response variable. The ACR improvement criteria are commonly used in clinical trials as an endpoint for comparing the proportion of responders between treatment groups. In contrast, the Disease Activity Score (DAS), Simplified Disease Activity Index (SDAI), the Clinical Disease Activity Index (CDAI), and the Routine Assessment of Patient Index Data 3 (RAPID3) are continuous measures of disease activity. These scales are increasingly used in clinical practice for tracking disease status and, in particular, for documenting treatment response.

Several developments during the past two decades have changed the therapeutic landscape in RA. They include (1) the emergence of methotrexate as the disease-modifying antirheumatic drug (DMARD) of first choice for the treatment of early RA; (2) the development of novel highly efficacious biologicals that can be used alone or in combination with methotrexate; and (3) the proven superiority of combination DMARD regimens over methotrexate alone. The medications used for the treatment of RA may be divided into broad categories: nonsteroidal anti-inflammatory drugs (NSAIDs); glucocorticoids, such as prednisone and methylprednisolone; conventional DMARDs; and biologic DMARDs **(Table 351-2)**. Although disease for some patients with RA is managed adequately with a single DMARD, such as methotrexate, it demands in most cases the use of a combination DMARD regimen that may vary in its components over the treatment course depending on fluctuations in disease activity and emergence of drug-related toxicities and comorbidities.

NSAIDs

NSAIDs were formerly viewed as the core of RA therapy, but they are now considered to be adjunctive agents for management of symptoms uncontrolled by other measures. NSAIDs exhibit both analgesic and anti-inflammatory properties. The anti-inflammatory effects of NSAIDs derive from their ability to nonselectively inhibit cyclooxygenase (COX)-1 and COX-2. Although the results of clinical trials suggest that NSAIDs are roughly equivalent in their efficacy, experience suggests that some individuals may preferentially respond to a particular NSAID. Chronic use should be minimized due to the possibility of side effects, including gastritis and peptic ulcer disease as well as impairment of renal function.

GLUCOCORTICOIDS

Glucocorticoids may serve in several ways to control disease activity in RA. First, they may be administered in low to moderate doses to achieve rapid disease control before the onset of fully effective DMARD therapy, which often takes several weeks or even months. Second, a 1- to 2-week burst of glucocorticoids may be prescribed for the management of acute disease flares, with dose and duration guided by the severity of the exacerbation. Chronic administration of low doses (5–10 mg/d) of prednisone (or its equivalent) may also be warranted to control disease activity in patients with an inadequate response to DMARD therapy. Low-dose prednisone therapy has been shown in prospective studies to retard radiographic progression of joint disease; however, the benefits of this approach must be carefully weighed against the risks. Best practices minimize chronic use of low-dose prednisone therapy owing to the risk of osteoporosis and other long-term complications; however, the use of chronic prednisone therapy is unavoidable in some cases. High-dose glucocorticoids may be necessary for treatment of severe extraarticular manifestations of RA, such as ILD. Finally, if a patient exhibits one or a few actively inflamed joints, the clinician may consider intraarticular injection of an intermediate-acting glucocorticoid such as triamcinolone acetonide. This approach may allow for rapid control of inflammation in a limited number of affected joints. Caution must be exercised to appropriately exclude joint infection, as it often mimics an RA flare.

Osteoporosis ranks as an important long-term complication of chronic prednisone use. Based on a patient's risk factors, including total prednisone dosage, length of treatment, gender, race and bone density, treatment with a bisphosphonate may be appropriate for primary prevention of glucocorticoid-induced osteoporosis. Other agents, including teriparatide and denosomab, have been approved for the treatment of osteoporosis and may be indicated in certain cases. Although prednisone use is known to increase the risk of peptic ulcer disease, especially with concomitant NSAID use, no

TABLE 351-2 DMARDs Used for the Treatment of Rheumatoid Arthritis

DRUG	DOSAGE	SERIOUS TOXICITIES	OTHER COMMON SIDE EFFECTS	INITIAL EVALUATION	MONITORING
Hydroxychloroquine	200–400 mg/d orally (≤5 mg/kg)	Irreversible retinal damage Cardiotoxicity Blood dyscrasia	Nausea Diarrhea Headache Rash	Eye examination if >40 years old or prior ocular disease	Optical coherence tomography and visual field testing every 12 months
Sulfasalazine	Initial: 500 mg orally twice daily Maintenance: 1000–1500 mg twice daily	Granulocytopenia Hemolytic anemia (with G6PD deficiency)	Nausea Diarrhea Headache	CBC, LFTs G6PD level	CBC every 2–4 weeks for first 3 months, then every 3 months
Methotrexate	10–25 mg/week orally or SQ Folic acid 1 mg/d to reduce toxicities	Hepatotoxicity Myelosuppression Infection Interstitial pneumonitis Pregnancy category X	Nausea Diarrhea Stomatitis/mouth ulcers Alopecia Fatigue	CBC, LFTs Viral hepatitis panel[a] Chest x-ray	CBC, creatinine, LFTs every 2–3 months
Leflunomide	10–20 mg/d	Hepatotoxicity Myelosuppression Infection Pregnancy category X	Alopecia Diarrhea	CBC, LFTs Viral hepatitis panel[a]	CBC, creatinine, LFTs every 2–3 months
TNF-α Inhibitors	Infliximab: 3 mg/kg IV at weeks 0, 2, 6, then every 8 weeks. May increase dose up to 10 mg/kg every 4 weeks	↑ Risk bacterial, fungal infections Reactivation of latent TB ↑ Lymphoma risk (controversial) Drug-induced lupus Neurologic deficits	Infusion reaction ↑ LFTs	PPD skin test	LFTs periodically
	Etanercept: 50 mg SQ weekly, or 25 mg SQ biweekly	As above	Injection site reaction	PPD skin test	Monitor for injection site reactions
	Adalimumab: 40 mg SQ every other week	As above	Injection site reaction	PPD skin test	Monitor for injection site reactions
	Golimumab: 50 mg SQ monthly	As above	Injection site reaction	PPD skin test	Monitor for injection site reactions
	Certolizumab: 400 mg SQ weeks 0, 2, 4, then 200 mg every other week	As above	Injection site reaction	PPD skin test	Monitor for injection site reactions
Abatacept	Weight based: <60 kg: 500 mg 60–100 kg: 750 mg >100 kg: 1000 mg IV dose at weeks 0, 2, and 4, and then every 4 weeks OR 125 mg SQ weekly	↑ Risk bacterial, viral infections	Headache Nausea	PPD skin test	Monitor for infusion reactions
Anakinra	100 mg SQ daily	↑ Risk bacterial, viral infections Reactivation of latent TB Neutropenia	Injection site reaction Headache	PPD skin test CBC with differential	CBC every month for 3 months, then every 4 months for 1 year Monitor for injection site reactions
Rituximab	1000 mg IV × 2, days 0 and 14 May repeat course every 24 weeks or more Premedicate with methyl-prednisolone 100 mg to decrease infusion reaction	↑ Risk bacterial, viral infections Infusion reaction Cytopenia Hepatitis B reactivation	Rash Fever	CBC Viral hepatitis panel[a]	CBC at regular intervals
Tocilizumab	4–8 mg/kg 4–8 mg/kg IV monthly OR 162 mg SQ every other week (<100 kg weight) 162 mg SQ every week (≥100 kg weight)	Risk of infection Infusion reaction LFT elevation Dyslipidemia Cytopenias		PPD skin test	CBC and LFTs at regular intervals
Tofacitinib	5 mg orally BID OR 11 mg orally daily	Risk of infection LFT elevation Dyslipidemia Neutropenia	Upper respiratory tract infections Diarrhea Headache Nasopharyngitis	PPD skin test	CBC, LFTs, and lipids at regular intervals

[a]Viral hepatitis panel: hepatitis B surface antigen, hepatitis C viral antibody.

Abbreviations: CBC, complete blood count; DMARDs, disease-modifying antirheumatic drugs; G6PD, glucose-6-phosphate dehydrogenase; IV, intravenous; LFTs, liver function tests; PPD, purified protein derivative; SQ, subcutaneous; TB, tuberculosis.

evidence-based guidelines have been published regarding the use of gastrointestinal ulcer prophylaxis in this situation.

DMARDs

DMARDs are so named because of their ability to slow or prevent structural progression of RA. The conventional DMARDs include hydroxychloroquine, sulfasalazine, methotrexate, and leflunomide; they exhibit a delayed onset of action of ~6–12 weeks. Methotrexate is the DMARD of choice for the treatment of RA and is the anchor drug for most combination therapies. It was approved for the treatment of RA in 1988 and remains the benchmark for the efficacy and safety of new disease-modifying therapies. At the dosages used for the treatment of RA, methotrexate has been shown to stimulate adenosine release from cells, producing an anti-inflammatory effect. The clinical efficacy of leflunomide, an inhibitor of pyrimidine synthesis, appears similar to that of methotrexate; it has been shown in well-designed trials to be effective for the treatment of RA as monotherapy or in combination with methotrexate and other DMARDs.

Although similar to the other DMARDs in its slow onset of action, hydroxychloroquine has not been shown to delay radiographic progression of disease and thus is not considered to be a true DMARD. In clinical practice, hydroxychloroquine is generally used for treatment of early, mild disease or as adjunctive therapy in combination with other DMARDs. Sulfasalazine is used in a similar manner and has been shown in randomized, controlled trials to reduce radiographic progression of disease. Minocycline, gold salts, penicillamine, azathioprine, and cyclosporine have all been used for the treatment of RA with varying degrees of success; however, they are used sparingly now due to their inconsistent clinical efficacy or unfavorable toxicity profile.

BIOLOGICALS

Biologic DMARDs have revolutionized the treatment of RA over the past decade (Table 351-2). They are protein therapeutics designed mostly to target cytokines and cell-surface molecules. The TNF inhibitors were the first biologicals approved for the treatment of RA. Anakinra, an IL-1 receptor antagonist, was approved shortly thereafter; however, its benefits have proved to be relatively modest compared with the other biologicals and therefore this biological is rarely used for the treatment of RA with the availability of other more effective agents. Abatacept, rituximab, and tocilizumab are the newest members of this class.

Anti-TNF Agents The development of TNF inhibitors was originally spurred by the experimental finding that TNF is a critical upstream mediator of joint inflammation. Currently, five agents that inhibit TNF-α are approved for the treatment of RA. There are three different anti-TNF monoclonal antibodies. Infliximab is a chimeric (part mouse and human) monoclonal antibody, whereas adalimumab and golimumab are humanized monoclonal antibodies. Certolizumab pegol is a pegylated Fc-free fragment of a humanized monoclonal antibody with binding specificity for TNF-α. Lastly, etanercept is a soluble fusion protein comprising the TNF receptor 2 in covalent linkage with the Fc portion of IgG1. All of the TNF inhibitors have been shown in randomized controlled clinical trials to reduce the signs and symptoms of RA, slow radiographic progression of joint damage, and improve physical function and quality of life. Anti-TNF drugs are typically used in combination with background methotrexate therapy. This combination regimen, which affords maximal benefit in many cases, is often the next step for treatment of patients with an inadequate response to methotrexate therapy. Etanercept, adalimumab, certolizumab pegol, and golimumab have also been approved for use as monotherapy.

Anti-TNF agents should be avoided in patients with active infection or a history of hypersensitivity to these agents and are contraindicated in patients with chronic hepatitis B infection or class III/IV congestive heart failure. The major concern is the increased risk for infection, including serious bacterial infections, opportunistic fungal infection, and reactivation of latent tuberculosis. For this reason, all patients are screened for latent tuberculosis according to national guidelines prior to starting anti-TNF therapy **(Chap. 173)**. In the United States, patients are skin-tested using an intradermal injection of purified protein derivative (PPD); individuals with skin reactions of >5 mm are presumed to have had previous exposure to tuberculosis and are evaluated for active disease and treated accordingly. Use of an IFN-γ release assay may also be appropriate for screening as some data suggest a lower rate of false-negative and false-positive tests with an IFN-γ release assay compared to skin testing with PPD in patients treated with corticosteroids. While a combination of PPD skin test and IFN-γ release assay may offer the highest sensitivity for screening purposes, no consensus guidelines exist.

Anakinra Anakinra is the recombinant form of the naturally occurring IL-1 receptor antagonist. Although anakinra has seen limited use for the treatment of RA, it has enjoyed a resurgence of late as an effective therapy of some rare inherited syndromes dependent on IL-1 production, including neonatal-onset inflammatory disease, Muckle-Wells syndrome, and familial cold urticaria, as well as systemic juvenile-onset inflammatory arthritis and adult-onset Still's disease. Anakinra should not be combined with an anti-TNF drug due to the high rate of serious infections observed with this regimen in a clinical trial.

Abatacept Abatacept is a soluble fusion protein consisting of the extracellular domain of human CTLA-4 linked to the modified portion of human IgG. It inhibits the co-stimulation of T cells by blocking CD28-CD80/86 interactions and may also inhibit the function of antigen-presenting cells by reverse signaling through CD80 and CD86. Abatacept has been shown in clinical trials to reduce disease activity, slow radiographic progression of damage, and improve functional disability. Many patients receive abatacept in combination with methotrexate or another DMARD such as leflunomide. Abatacept therapy has been associated with an increased risk of infection.

Rituximab Rituximab is a chimeric monoclonal antibody directed against CD20, a cell-surface molecule expressed by most mature B lymphocytes. It works by depleting B cells, which in turn, leads to a reduction in the inflammatory response by unknown mechanisms. These mechanisms may include a reduction in autoantibodies, inhibition of T cell activation, and alteration of cytokine production. Rituximab has been approved for the treatment of refractory RA in combination with methotrexate and has been shown to be more effective for patients with seropositive than seronegative disease. Rituximab therapy has been associated with mild to moderate infusion reactions as well as an increased risk of infection. Notably, there have been rare isolated reports of a potentially lethal brain disorder, progressive multifocal leukoencephalopathy (PML), in association with rituximab therapy, although the absolute risk of this complication appears to be very low in patients with RA. Most of these cases have occurred on a background of previous or current exposure to other potent immunosuppressive drugs.

Tocilizumab Tocilizumab is a humanized monoclonal antibody directed against the membrane and soluble forms of the IL-6 receptor. IL-6 is a proinflammatory cytokine implicated in the pathogenesis of RA, with effects on both joint inflammation and damage. IL-6 binding to its receptor activates intracellular signaling pathways that affect the acute-phase response, cytokine production, and osteoclast activation. Clinical trials attest to the clinical efficacy of tocilizumab therapy for RA, both as monotherapy and in combination with methotrexate and other DMARDs. Tocilizumab has been associated with an increased risk of infection, neutropenia, and thrombocytopenia; the hematologic abnormalities appear to be reversible upon stopping the drug. In addition, this agent has been shown to increase LDL cholesterol. However, it is not known as yet if this effect on lipid levels increases the risk for development of atherosclerotic disease.

SMALL-MOLECULE INHIBITORS

Because some patients do not adequately respond to conventional DMARDs or biologic therapy, other therapeutic targets have been

investigated to fill this gap. Recently, drug development in RA has focused attention on the intracellular signaling pathways that transduce the positive signals of cytokines and other inflammatory mediators that create the positive feedback loops in the immune response. These synthetic DMARDs aim to provide the same efficacy as biological therapies in an oral formulation.

Tofacitinib Tofacitinib is a small-molecule inhibitor that primarily inhibits JAK1 and JAK3, which mediate signaling of the receptors for the common γ-chain-related cytokines IL-2, 4, 7, 9, 15, and 21 as well as IFN-γ and IL-6. These cytokines all play roles in promoting T and B cell activation as well as inflammation. Tofacitinib, an oral agent, has been shown in randomized, placebo-controlled clinical trials to improve the signs and symptoms of RA significantly over placebo. Possible side effects include elevated serum transaminases indicative of liver injury, neutropenia, increased cholesterol levels, and elevation in serum creatinine. Its use is also associated with an increased risk of infections. Tofacitinib can be used as monotherapy or in combination with methotrexate.

TREATMENT OF EXTRAARTICULAR MANIFESTATIONS

In general, treatment of the underlying RA favorably modifies extraarticular manifestations, and it appears that aggressive management of early disease can potentially prevent their occurrence in the first place. RA-ILD, however, can be particularly challenging to treat because some of the DMARDs used for the treatment of RA are associated with pulmonary toxicity, such as methotrexate and leflunomide. High doses of corticosteroids and adjunctive immunosuppressive agents, such as azathioprine, mycophenolate mofetil, and rituximab have been used for treatment of RA-ILD.

APPROACH TO THE PATIENT

Rheumatoid Arthritis

The original treatment pyramid for RA is now considered to be obsolete and has evolved into a new strategy that focuses on several goals: (1) early, aggressive therapy to prevent joint damage and disability; (2) frequent modification of therapy with utilization of combination therapy where appropriate; (3) individualization of therapy in an attempt to maximize response and minimize side effects; and (4) achieving, whenever possible, remission of clinical disease activity. A considerable amount of evidence supports this intensive treatment approach.

As mentioned earlier, methotrexate is the DMARD of first choice for initial treatment of moderate to severe RA. Failure to achieve adequate improvement with methotrexate therapy calls for a change in DMARD therapy, usually a transition to an effective combination regimen. Effective combinations include: methotrexate, sulfasalazine, and hydroxychloroquine (oral triple therapy); methotrexate and leflunomide; and methotrexate plus a biological. The combination of methotrexate and an anti-TNF agent, for example, has been shown in randomized, controlled trials to be superior to methotrexate alone not only for reducing signs and symptoms of disease, but also for retarding the progression of structural joint damage. Predicting which patients are at higher risk for developing radiologic joint damage is imprecise at best, although some factors such as an elevated serum level of acute-phase reactants, high burden of joint inflammation, and the presence of erosive disease are associated with increased likelihood of developing structural injury.

In 2015, the American College of Rheumatology updated and published their guidelines for the treatment of RA. They do make a distinction in the treatment of patients with early (<6 months of disease duration) and established disease and highlight the use of a treat-to-target approach and the need to switch or add therapies for worsening or persistent moderate/high disease activity. For example, in patients with early RA who have persistent moderate/high disease activity on DMARD monotherapy, providers should consider escalation to combination DMARD therapy or switching to an anti-TNF +/– methotrexate or a non-TNF biologic +/– methotrexate. Since a more intensive initial approach (e.g., combination DMARD therapy) has been shown to produce superior long-term outcomes compared with starting methotrexate alone, the usual approach is to begin with methotrexate and rapidly step-up (e.g., after 3–6 months) to combination of DMARD therapy or an anti-TNF or non-TNF biologic agent in the absence of an inadequate therapeutic response.

Some patients may not respond to an anti-TNF drug or may be intolerant of its side effects. Initial responders to an anti-TNF agent that later worsen may benefit from switching to another anti-TNF agent or an alternative biologic with a different mechanism of action. Indeed, some studies suggest that switching to an alternative biologic such as abatacept is more effective than switching to another anti-TNF drug. Unacceptable toxicity from an anti-TNF agent may also call for switching to another biological with a different mechanism of action or a conventional DMARD regimen.

Studies have also shown that oral triple therapy (hydroxychloroquine, methotrexate, and sulfasalazine) may be used effectively for the treatment of early RA. Treatment may be initiated with methotrexate alone and lacking an adequate treatment response followed within 6 months by a step-up to oral triple therapy.

A clinical state defined as low disease activity or remission is the optimal goal of therapy, although most patients never achieve complete remission despite every effort to achieve it. Composite indices, such as the Disease Activity Score-28 (DAS-28), are useful for classifying states of low disease activity and remission; however, they are imperfect tools due to the limitations of the clinical joint examination in which low-grade synovitis may escape detection. Complete remission has been stringently defined as the total absence of all articular and extraarticular inflammation and immunologic activity related to RA. However, evidence for this state can be difficult to demonstrate in clinical practice. In an effort to standardize and simplify the definition of remission for clinical trials, the ACR and EULAR developed two provisional operational definitions of remission in RA **(Table 351-3)**. A patient may be considered in remission if he or she (1) meets all of the clinical and laboratory criteria listed in Table 351-3 or (2) has a composite SDAI score of <3.3. The SDAI is calculated by taking the sum of a tender joint and swollen joint count (using 28 joints), patient global assessment (0–10 scale), physician global assessment (0–10 scale), and CRP (in mg/dL). This definition of remission does not take into account the possibility of subclinical synovitis or that damage alone may produce a tender or swollen joint. Ignoring the semantics of these definitions, the aforementioned remission criteria are nonetheless useful for setting a level of disease control that will likely result in minimal or no progression of structural damage and disability.

PHYSICAL THERAPY AND ASSISTIVE DEVICES

In principle, all patients with RA should receive a prescription for exercise and physical activity. Dynamic strength training, community-based comprehensive physical therapy, and physical-activity coaching (emphasizing 30 min of moderately intensive activity most days a week) have all been shown to improve muscle strength and perceived health status. Foot orthotics for painful valgus deformity decrease foot pain and may reduce disability and functional

TABLE 351-3 ACR/EULAR Provisional Definition of Remission in Rheumatoid Arthritis

At any time point, patient must satisfy all of the following:
Tender joint count ≤1
Swollen joint count ≤1
C-reactive protein ≤1 mg/dL
Patient global assessment ≤1 (on a 0–10 scale)
OR
At any time point, patient must have a Simplified Disease Activity Index score of ≤3.3

Source: Adapted from DT Felson et al: Arthritis Rheum 63:573, 2011.

limitations. Judicious use of wrist splints can also decrease pain; however, their benefits may be offset by decreased dexterity and variably curb grip strength.

SURGERY

Surgical procedures may improve pain and disability in RA with varying degrees of reported long-term success—most notably the hands, wrists, and feet. For large joints, such as the knee, hip, shoulder, or elbow, the preferred option for advanced joint disease may be total joint arthroplasty. A few surgical options exist for dealing with the smaller hand joints. Silicone implants are the most common prosthetic for MCP arthroplasty and are generally implanted in patients with severe decreased arc of motion, marked flexion contractures, MCP joint pain with radiographic abnormalities, and severe ulnar drift. Arthrodesis and total wrist arthroplasty are reserved for patients with severe disease who have substantial pain and functional impairment. These two procedures appear to have equal efficacy in terms of pain control and patient satisfaction. Numerous surgical options exist for correction of hallux valgus in the forefoot, including arthrodesis and arthroplasty, as well as primarily arthrodesis for refractory hindfoot pain.

OTHER MANAGEMENT CONSIDERATIONS

Pregnancy Up to 75% of female RA patients will note overall improvement in symptoms during pregnancy, but often will flare after delivery. Flares during pregnancy are generally treated with low doses of prednisone; hydroxychloroquine and sulfasalazine are probably the safest DMARDs to use during pregnancy. Methotrexate and leflunomide therapy are contraindicated during pregnancy due to their teratogenicity in animals and humans. The experience with biologic agents has been insufficient to make specific recommendations for their use during pregnancy. Ideally, their use should be avoided, but controlling active RA during pregnancy may take precedence in some cases.

Elderly Patients RA presents in up to one-third of patients after the age of 60; however, older individuals may receive less aggressive treatment due to concerns about increased risks of drug toxicity. Studies suggest that conventional DMARDs and biologic agents are equally effective and safe in younger and older patients. Due to comorbidities, many elderly patients have an increased risk of infection. Aging also leads to a gradual decline in renal function that may raise the risk for side effects from NSAIDs and some DMARDS, such as methotrexate. Renal function must be taken into consideration before prescribing methotrexate, which is mostly cleared by the kidneys. To reduce the risks of side effects, methotrexate doses may need to be adjusted downward for the drop in renal function that usually comes with the seventh and eighth decades of life. Methotrexate is usually not prescribed for patients with a serum creatinine >2 mg/dL.

GLOBAL CHALLENGES

Developing countries are finding an increase in the incidence of noncommunicable, chronic diseases such as diabetes, cardiovascular disease, and RA in the face of ongoing poverty, rampant infectious disease, and poor access to modern health care facilities. In these areas, patients tend to have a greater delay in diagnosis and limited access to specialists, and thus greater disease activity and disability at presentation. In addition, infection risk remains a significant issue for the treatment of RA in developing countries because of the immunosuppression associated with the use of glucocorticoids and most DMARDs. For example, in some developing countries, patients undergoing treatment for RA have a substantial increase in the incidence of tuberculosis, which demands the implementation of far more comprehensive screening practices and liberal use of isoniazid prophylaxis than in developed countries. The increased prevalence of hepatitis B and C, as well as human immunodeficiency virus (HIV), in these developing countries also poses challenges. Reactivation of viral hepatitis has been observed in association with some of the DMARDs, such as rituximab. Also, reduced access to antiretroviral therapy may limit the control of HIV infection and therefore the choice of DMARD therapies.

Despite these challenges, one should attempt to initiate early treatment of RA in the developing countries with the resources at hand. Hydroxychloroquine, sulfasalazine and methotrexate are all reasonably accessible throughout the world where they can be used as both monotherapy and in combination with other drugs. The use of biologic agents is increasing in the developed countries as well as in other areas around the world, although their use is limited by high cost; national protocols restrict their use, and concerns remain about the risk for opportunistic infections.

SUMMARY

Improved understanding of the pathogenesis of RA and its treatment has dramatically revolutionized the management of this disease. The outcomes of patients with RA are vastly superior to those of the prebiologic modifier era; more patients than in years past are able to avoid significant disability and continue working, albeit with some job modifications in many cases. The need for early and aggressive treatment of RA as well as frequent follow-up visits for monitoring of drug therapy has implications for our health care system. Primary care physicians and rheumatologists must be prepared to work together as a team to reach the ambitious goals of best practice. In many settings, rheumatologists have reengineered their practice in a way that places high priority on consultations for any new patient with early inflammatory arthritis.

The therapeutic regimens for RA are becoming increasingly complex with the rapidly expanding armamentarium. Patients receiving these therapies must be carefully monitored by both the primary care physician and the rheumatologist to minimize the risk of side effects and identify quickly any complications of chronic immunosuppression. Also, prevention and treatment of RA-associated conditions such as ischemic heart disease and osteoporosis will likely benefit from a team approach owing to the value of multidisciplinary care.

Research will continue to search for new therapies with superior efficacy and safety profiles and investigate treatment strategies that can bring the disease under control more rapidly and nearer to remission. However, prevention and cure of RA will likely require new breakthroughs in our understanding of disease pathogenesis. These insights may come from genetic studies illuminating critical pathways in the mechanisms of joint inflammation. Equally ambitious is the lofty goal of biomarker discovery that will open the door to personalized medicine for the care of patients with RA.

FURTHER READING

Catrina AI et al: Lungs, joints and immunity against citrullinated proteins in rheumatoid arthritis. Nature Rev Rheumatol 10:645, 2014.

Erickson AR et al: Clinical features of rheumatoid arthritis, in *Kelley and Firestein's Textbook of Rheumatology*, 10th ed, Firestein GS et al (eds). Philadelphia, Elsevier, 2017, pp 1167–1186.

McInnes IB, Schett G: The pathogenesis of rheumatoid arthritis. N Engl J Med 365:2205, 2011.

Moreland LW et al: A randomized comparative effectiveness study of oral triple therapy versus methotrexate plus etanercept in early aggressive rheumatoid arthritis: The Treatment of Early Aggressive Rheumatoid Arthritis Trial. Arthritis Rheum 64:2824, 2012.

Singh JA et al: 2015 American College of Rheumatology Guideline for the Treatment of Rheumatoid Arthritis. Arthritis Rheumatol 68:1, 2016.

Yarwood A, Huizinga TW, Worthington J: The genetics of rheumatoid arthritis: Risk and protection in different stages of the evolution of RA. Rheumatology (Oxford) 55:199, 2016.

352 Acute Rheumatic Fever

Jonathan R. Carapetis

Acute rheumatic fever (ARF) is a multisystem disease resulting from an autoimmune reaction to infection with group A streptococcus. Although many parts of the body may be affected, almost all of the manifestations resolve completely. The major exception is cardiac valvular damage (rheumatic heart disease [RHD]), which may persist after the other features have disappeared.

GLOBAL CONSIDERATIONS

ARF and RHD are diseases of poverty. They were common in all countries until the early twentieth century, when their incidence began to decline in industrialized nations. This decline was largely attributable to improved living conditions—particularly less crowded housing and better hygiene—which resulted in reduced transmission of group A streptococci. The introduction of antibiotics and improved systems of medical care had a supplemental effect.

The virtual disappearance of ARF and reduction in the incidence of RHD in industrialized countries during the twentieth century unfortunately was not replicated in developing countries, where these diseases continue unabated. RHD is the most common cause of heart disease in children in developing countries and is a major cause of mortality and morbidity in adults as well. It has been estimated that between 15 and 19 million people worldwide are affected by RHD, with approximately one-quarter of a million deaths occurring each year. Some 95% of ARF cases and RHD deaths now occur in developing countries, with particularly high rates in sub-Saharan Africa, Pacific nations, Australasia, and South and Central Asia. The pathogenetic pathway from exposure to group A streptococcus followed by pharyngeal infection and subsequent development of ARF, ARF recurrences, and development of RHD and its complications is associated with a range of risk factors and, therefore, potential interventions at each point **(Fig. 352-1)**. In affluent countries, many of these risk factors are well controlled, and where needed, interventions are in place. Unfortunately, the greatest burden of disease is found in developing countries, most of which do not have the resources, capacity, and/or interest to tackle this multifaceted disease. In particular, almost none of the developing countries has a coordinated, register-based RHD control program, which is proven to be cost-effective in reducing the burden of RHD. Enhancing awareness of RHD and mobilizing resources for its control in developing countries are issues requiring international attention.

EPIDEMIOLOGY

ARF is mainly a disease of children age 5–14 years. Initial episodes become less common in older adolescents and young adults and are rare in persons aged >30 years. By contrast, recurrent episodes of ARF remain relatively common in adolescents and young adults. This pattern contrasts with the prevalence of RHD, which peaks between 25 and 40 years. There is no clear gender association for ARF, but RHD more commonly affects females, sometimes up to twice as frequently as males.

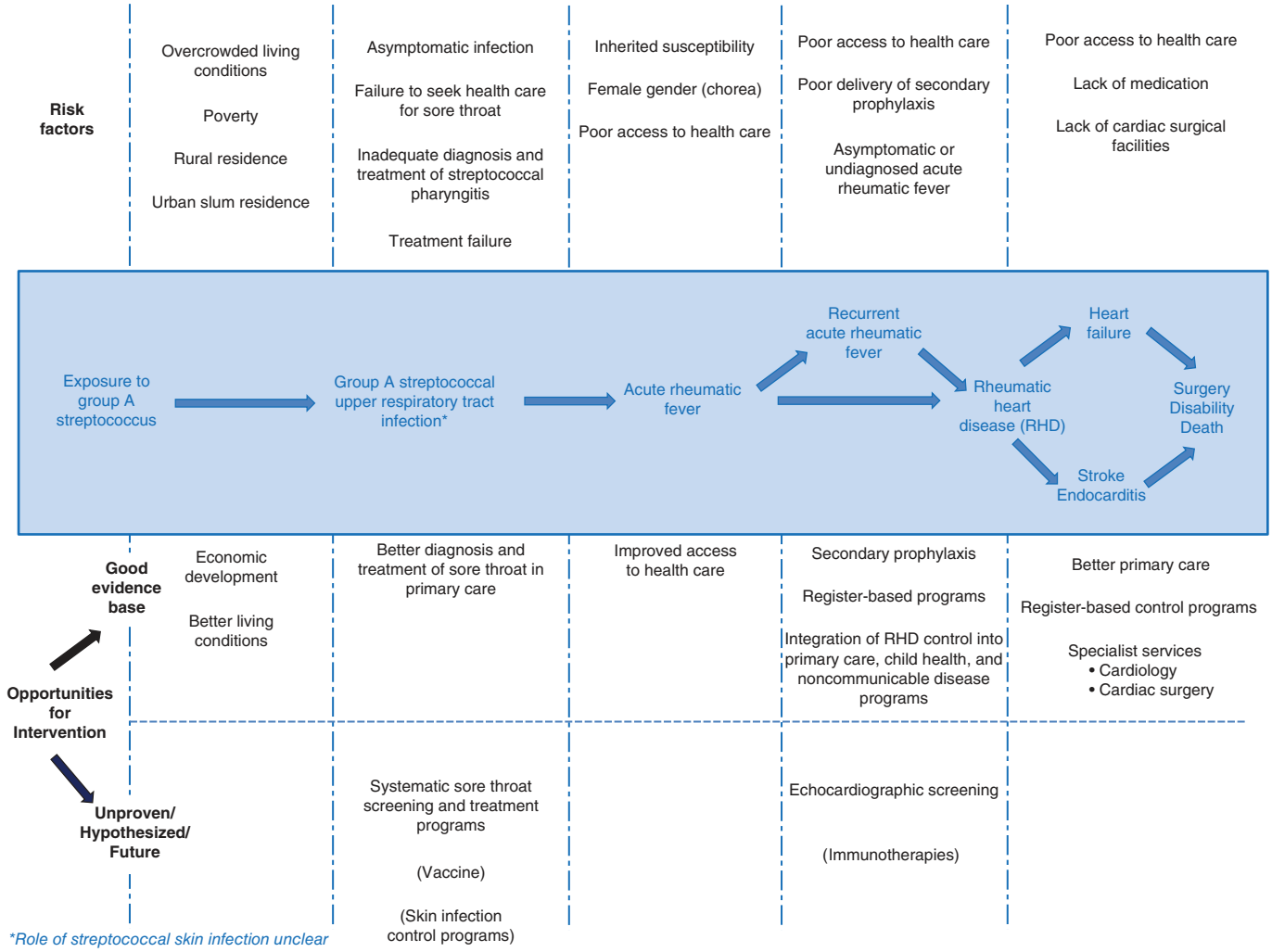


FIGURE 352-1 Pathogenetic pathway for acute rheumatic fever and rheumatic heart disease, with associated risk factors and opportunities for intervention at each step. Interventions in *parentheses* are either unproven or currently unavailable.

PATHOGENESIS

ORGANISM FACTORS

Based on currently available evidence, ARF is exclusively caused by infection of the upper respiratory tract with group A streptococci (Chap. 143). Although classically, certain M-serotypes (particularly types 1, 3, 5, 6, 14, 18, 19, 24, 27, and 29) were associated with ARF, in high-incidence regions, it is now thought that any strain of group A streptococcus has the potential to cause ARF. The potential role of skin infection and of groups C and G streptococci is currently being investigated.

HOST FACTORS

Approximately 3–6% of any population may be susceptible to ARF, and this proportion does not vary dramatically between populations. Findings of familial clustering of cases and concordance in monozygotic twins—particularly for chorea—confirm that susceptibility to ARF is an inherited characteristic, with 44% concordance in monozygotic twins compared to 12% in dizygotic twins, and heritability more recently estimated at 60%. Most evidence for host factors focuses on immunologic determinants. Some human leukocyte antigen (HLA) class II alleles, particularly HLA-DR7 and HLA-DR4, appear to be associated with susceptibility, whereas other class II alleles have been associated with protection (HLA-DR5, HLA-DR6, HLA-DR51, HLA-DR52, and HLA-DQ). Associations have also been described with polymorphisms at the tumor necrosis factor α locus (TNF-α-308 and TNF-α-238), high levels of circulating mannose-binding lectin, and Toll-like receptors.

THE IMMUNE RESPONSE

The most widely accepted theory of rheumatic fever pathogenesis is based on the concept of molecular mimicry, whereby an immune response targeted at streptococcal antigens (mainly thought to be on the M protein and the *N*-acetylglucosamine of group A streptococcal carbohydrate) also recognizes human tissues. In this model, cross-reactive antibodies bind to endothelial cells on the heart valve, leading to activation of the adhesion molecule VCAM-1, with resulting recruitment of activated lymphocytes and lysis of endothelial cells in the presence of complement. The latter leads to release of peptides including laminin, keratin, and tropomyosin, which, in turn, activates cross-reactive T cells that invade the heart, amplifying the damage and causing epitope spreading. An alternative hypothesis proposes that the initial damage is due to streptococcal invasion of epithelial surfaces, with binding of M protein to type IV collagen allowing it to become immunogenic, but not through the mechanism of molecular mimicry.

CLINICAL FEATURES

There is a latent period of ~3 weeks (1–5 weeks) between the precipitating group A streptococcal infection and the appearance of the clinical features of ARF. The exceptions are chorea and indolent carditis, which may follow prolonged latent periods lasting up to 6 months. Although many patients report a prior sore throat, the preceding group A streptococcal infection is commonly subclinical; in these cases, it can only be confirmed using streptococcal antibody testing. The most common clinical features are polyarthritis (present in 60–75% of cases) and carditis (50–60%). The prevalence of chorea in ARF varies substantially between populations, ranging from <2 to 30%. Erythema marginatum and subcutaneous nodules are now rare, being found in <5% of cases.

HEART INVOLVEMENT

Up to 60% of patients with ARF progress to RHD. The endocardium, pericardium, or myocardium may be affected. Valvular damage is the hallmark of rheumatic carditis. The mitral valve is almost always affected, sometimes together with the aortic valve; isolated aortic valve involvement is rare. Damage to the pulmonary or tricuspid valves is usually secondary to increased pulmonary pressures resulting from left-sided valvular disease. Early valvular damage leads to regurgitation. Over ensuing years, usually as a result of recurrent episodes, leaflet thickening, scarring, calcification, and valvular stenosis may develop (Fig. 352-2). See Videos 352-1 and 352-2. Therefore, the characteristic manifestation of carditis in previously unaffected individuals is mitral regurgitation, sometimes accompanied by aortic regurgitation. Myocardial inflammation may affect electrical conduction pathways, leading to P-R interval prolongation (first-degree atrioventricular block or rarely higher level block) and softening of the first heart sound.

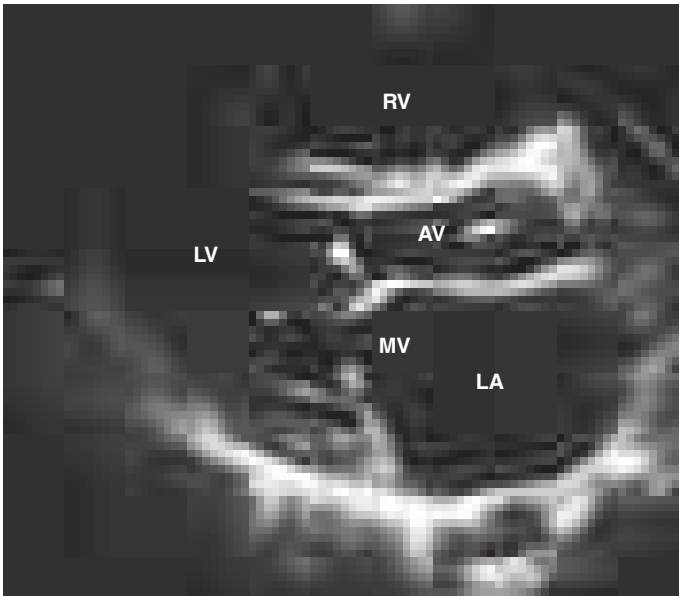


FIGURE 352-2 Transthoracic echocardiographic image from a 5-year-old boy with chronic rheumatic heart disease. This diastolic image demonstrates leaflet thickening, restriction of the anterior mitral valve leaflet tip and doming of the body of the leaflet toward the interventricular septum. This appearance (marked by the *arrowhead*) is commonly described as a "hockey stick" or an "elbow" deformity. AV, aortic valve; LA, left atrium; LV, left ventricle; MV, mitral valve; RV, right ventricle. *(Courtesy of Dr. Bo Remenyi, Department of Paediatric and Congenital Cardiac Services, Starship Children's Hospital, Auckland, New Zealand.)*

People with RHD are often asymptomatic for many years before their valvular disease progresses to cause cardiac failure. Moreover, particularly in resource-poor settings, the diagnosis of ARF is often not made, so children, adolescents, and young adults may have RHD but not know it. These cases can be diagnosed using echocardiography; auscultation is poorly sensitive and specific for RHD diagnosis in asymptomatic patients. Echocardiographic screening of school-aged children in populations with high rates of RHD is becoming more widespread and has been facilitated by improving technologies in portable echocardiography and the availability of consensus guidelines for the diagnosis of RHD on echocardiography (Table 352-1). Although a diagnosis of definite RHD on screening echocardiography should lead to commencement of secondary prophylaxis, the clinical significance of borderline RHD has yet to be determined.

JOINT INVOLVEMENT

The most common form of joint involvement in ARF is arthritis, i.e., objective evidence of inflammation, with hot, swollen, red, and/or tender joints, and involvement of more than one joint (i.e., polyarthritis). Polyarthritis is typically migratory, moving from one joint to another over a period of hours. ARF almost always affects the large joints—most commonly the knees, ankles, hips, and elbows—and is asymmetric. The pain is severe and usually disabling until anti-inflammatory medication is commenced.

Less severe joint involvement is also relatively common and has been recognized as a potential major manifestation in high-risk populations in the most recent revision of the Jones criteria. Arthralgia without objective joint inflammation usually affects large joints in the same migratory pattern as polyarthritis. In some populations, aseptic monoarthritis may be a presenting feature of ARF, which may, in turn, result from early commencement of anti-inflammatory medication before the typical migratory pattern is established.

TABLE 352-1 World Heart Federation Criteria for Echocardiographic Diagnosis of Rheumatic Heart Disease (RHD) in Individuals <20 Years of Age[a]

Definite RHD (either A, B, C, or D)
(A) Pathologic MR and at least two morphologic features of RHD of the mitral valve
(B) MS mean gradient ≥4 mmHg (note: congenital MV anomalies must be excluded)
(C) Pathologic AR and at least two morphologic features of RHD of the AV (note: bicuspid AV and dilated aortic root must be excluded)
(D) Borderline disease of both the MV and AV
Borderline RHD (either A, B, or C)
(A) At least two morphologic features of RHD of the MV without pathologic MR or MS
(B) Pathologic MR
(C) Pathologic AR
Normal Echocardiographic Findings (all of A, B, C, and D)
(A) MR that does not meet all four Doppler criteria (physiologic MR)
(B) AR that does not meet all four Doppler criteria (physiologic AR)
(C) An isolated morphologic feature of RHD of the MV (e.g., valvular thickening), without any associated pathologic stenosis or regurgitation
(D) Morphologic feature of RHD of the AV (e.g., valvular thickening), without any associated pathologic stenosis or regurgitation
Definitions of Pathologic Regurgitation and Morphologic Features of RHD
Pathologic MR: All of the following: seen in two views; in at least one view, jet length 2 cm; peak velocity ≥3 m/s; pansystolic jet in at least one envelope
Pathologic AR: All of the following: seen in two views; in at least one view, jet length ≥1 cm; peak velocity ≥3 m/s; pandiastolic jet in at least one envelope
Morphologic features of RHD in MV: anterior MV leaflet thickening ≥3 mm (age specific); chordal thickening; restricted leaflet motion; excessive leaflet tip motion during systole
Morphologic features of RHD in AV: irregular or focal thickening; coaptation defect; restricted leaflet motion; prolapse

[a]For criteria in individuals >20 years of age, see source document.

Abbreviations: AR, aortic regurgitation; AV, aortic valve; MR, mitral regurgitation; MS, mitral stenosis; MV, mitral valve.

Source: Adapted from B Remenyi et al: World Heart Federation criteria for echocardiographic diagnosis of rheumatic heart disease-an evidence-based guideline. Nat Rev Cardiol 9:297–309, 2012.

The joint manifestations of ARF are highly responsive to salicylates and other nonsteroidal anti-inflammatory drugs (NSAIDs). Indeed, joint involvement that persists for more than 1 or 2 days after starting salicylates is unlikely to be due to ARF.

CHOREA

Sydenham's chorea commonly occurs in the absence of other manifestations, follows a prolonged latent period after group A streptococcal infection, and is found mainly in females. The choreiform movements affect particularly the head (causing characteristic darting movements of the tongue) and the upper limbs **(Chap. 428)**. They may be generalized or restricted to one side of the body (hemi-chorea). In mild cases, chorea may be evident only on careful examination, whereas in the most severe cases, the affected individuals are unable to perform activities of daily living. There is often associated emotional lability or obsessive-compulsive traits, which may last longer than the choreiform movements (which usually resolve within 6 weeks but sometimes may take up to 6 months).

SKIN MANIFESTATIONS

The classic rash of ARF is *erythema marginatum* **(Chap. 16)**, which begins as pink macules that clear centrally, leaving a serpiginous, spreading edge. The rash is evanescent, appearing and disappearing before the examiner's eyes. It occurs usually on the trunk, sometimes on the limbs, but almost never on the face.

Subcutaneous nodules occur as painless, small (0.5–2 cm), mobile lumps beneath the skin overlying bony prominences, particularly of the hands, feet, elbows, occiput, and occasionally the vertebrae. They are a delayed manifestation, appearing 2–3 weeks after the onset of disease, last for just a few days up to 3 weeks, and are commonly associated with carditis.

OTHER FEATURES

Fever occurs in most cases of ARF, although rarely in cases of pure chorea. Although high-grade fever (≥39°C) is the rule, lower grade temperature elevations are not uncommon. Elevated acute-phase reactants are also present in most cases.

EVIDENCE OF A PRECEDING GROUP A STREPTOCOCCAL INFECTION

With the exception of chorea and low-grade carditis, both of which may become manifest many months later, evidence of a preceding group A streptococcal infection is essential in making the diagnosis of ARF. Because most cases do not have a positive throat swab culture or rapid antigen test, serologic evidence is usually needed. The most common serologic tests are the anti-streptolysin O (ASO) and anti-DNase B (ADB) titers. Where possible, age-specific reference ranges should be determined in a local population of healthy people without a recent group A streptococcal infection.

CONFIRMING THE DIAGNOSIS

Because there is no definitive test, the diagnosis of ARF relies on the presence of a combination of typical clinical features together with evidence of the precipitating group A streptococcal infection, and the exclusion of other diagnoses. This uncertainty led Dr. T. Duckett Jones in 1944 to develop a set of criteria (subsequently known as the *Jones criteria*) to aid in the diagnosis. The most recent revision of the Jones criteria **(Table 352-2)** require the clinician to determine if the patient is from a setting or population known to experience low rates of ARF. For this group, there is a set of "low-risk" criteria; for all others, there is a set of more sensitive criteria.

TREATMENT

Acute Rheumatic Fever

Patients with possible ARF should be followed closely to ensure that the diagnosis is confirmed, treatment of heart failure and other symptoms is undertaken, and preventive measures including commencement of secondary prophylaxis, inclusion on an ARF registry, and health education are commenced. Echocardiography should be performed on all possible cases to aid in making the diagnosis and to determine the severity at baseline of any carditis. Other tests that should be performed are listed in **Table 352-3**.

There is no treatment for ARF that has been proven to alter the likelihood of developing, or the severity of, RHD. With the exception of treatment of heart failure, which may be life-saving in cases of severe carditis, the treatment of ARF is symptomatic.

ANTIBIOTICS

All patients with ARF should receive antibiotics sufficient to treat the precipitating group A streptococcal infection **(Chap. 143)**. Penicillin is the drug of choice and can be given orally (as phenoxymethyl penicillin, 500 mg [250 mg for children ≤27 kg] PO twice daily, or amoxicillin, 50 mg/kg [maximum, 1 g] daily, for 10 days) or as a single dose of 1.2 million units (600,000 units for children ≤27 kg) IM benzathine penicillin G.

SALICYLATES AND NSAIDs

These may be used for the treatment of arthritis, arthralgia, and fever, once the diagnosis is confirmed. They are of no proven value in the treatment of carditis or chorea. Aspirin is the drug of choice, delivered at a dose of 50–60 mg/kg per day, up to a maximum of 80–100 mg/kg per day (4–8 g/d in adults) in 4–5 divided doses. At higher doses, the patient should be monitored for symptoms of salicylate toxicity such as nausea, vomiting, or tinnitus; if symptoms appear, lower doses should be used. When the acute symptoms are

TABLE 352-2 Jones Criteria

A. For All Patient Populations with Evidence of Preceding Group A Streptococcal Infection	
Diagnosis: initial ARF	2 major manifestations or 1 major plus 2 minor manifestations
Diagnosis: recurrent ARF	2 major or 1 major and 2 minor or 3 minor
B. Major Criteria	
Low-risk populations[a]	Moderate- and high-risk populations
Carditis[b]	Carditis
• Clinical and/or subclinical	• Clinical and/or subclinical
Arthritis	Arthritis
• Polyarthritis only	• Monoarthritis or polyarthritis • Polyarthralgia[c]
Chorea	Chorea
Erythema marginatum	Erythema marginatum
SC nodules	SC nodules
C. Minor Criteria	
Low-risk populations[a]	Moderate- and high-risk populations
Polyarthralgia	Monoarthralgia
Fever (≥38.5°C)	Fever (≥38°C)
ESR ≥60 mm in the first hour and/or CRP ≥3.0 mg/dL[d]	ESR ≥30 mm/h and/or CRP ≥3.0 mg/dL[d]
Prolonged PR interval, after accounting for age variability (unless carditis is a major criterion)	Prolonged PR interval, after accounting for age variability (unless carditis is a major criterion)

[a]Low-risk populations are those with ARF incidence ≤2 per 100,000 school-age children or all-age rheumatic heart disease prevalence of ≤1 per 1000 population per year. [b]Subclinical carditis indicates echocardiographic valvulitis. (See source document.) [c]Polyarthralgia should only be considered as a major manifestation in moderate- to high-risk populations after exclusion of other causes. As in past versions of the criteria, erythema marginatum and SC nodules are rarely "stand-alone" major criteria. Additionally, joint manifestations can only be considered in either the major or minor categories but not both in the same patient. (See source document for more information.) [d]CRP value must be greater than upper limit of normal for laboratory. Also, because ESR may evolve during the course of ARF, peak ESR values should be used.

Abbreviations: ARF, acute rheumatic fever; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate.

Source: From MH Gewitz et al: Revision of the Jones Criteria for the diagnosis of acute rheumatic fever in the era of Doppler echocardiography: A scientific statement from the American Heart Association. Circulation 131:1806, 2015.

TABLE 352-3 Recommended Tests in Cases of Possible Acute Rheumatic Fever

Recommended for All Cases
White blood cell count
Erythrocyte sedimentation rate
C-reactive protein
Blood cultures if febrile
Electrocardiogram (if prolonged P-R interval or other rhythm abnormality, repeat in 2 weeks and again at 2 months if still abnormal)
Chest x-ray if clinical or echocardiographic evidence of carditis
Echocardiogram (consider repeating after 1 month if negative)
Throat swab (preferably before giving antibiotics)—culture for group A streptococcus
Antistreptococcal serology: both anti-streptolysin O and anti-DNase B titers, if available (repeat 10–14 days later if first test not confirmatory)
Tests for Alternative Diagnoses, Depending on Clinical Features
Repeated blood cultures if possible endocarditis
Joint aspirate (microscopy and culture) for possible septic arthritis
Copper, ceruloplasmin, antinuclear antibody, drug screen for choreiform movements
Serology and autoimmune markers for arboviral, autoimmune, or reactive arthritis

Source: Reprinted with permission from Menzies School of Health Research. Complete information located at *https://www.rhdaustralia.org.au/arf-rhd-guideline*.

substantially resolved, usually within the first 2 weeks, patients on higher doses can have the dose reduced to 50–60 mg/kg per day for a further 2–4 weeks. Fever, joint manifestations, and elevated acute-phase reactants sometimes recur up to 3 weeks after the medication is discontinued. This does not indicate a recurrence and can be managed by recommencing salicylates for a brief period. Naproxen at a dose of 10–20 mg/kg per day is a suitable alternative to aspirin and has the advantage of twice-daily dosing.

CONGESTIVE HEART FAILURE

Glucocorticoids The use of glucocorticoids in ARF remains controversial. Two meta-analyses have failed to demonstrate a benefit of glucocorticoids compared to placebo or salicylates in improving the short- or longer-term outcome of carditis. However, the studies included in these meta-analyses all took place >40 years ago and did not use medications in common usage today. Many clinicians treat cases of severe carditis (causing heart failure) with glucocorticoids in the belief that they may reduce the acute inflammation and result in more rapid resolution of failure. However, the potential benefits of this treatment should be balanced against the possible adverse effects. If used, prednisone or prednisolone is recommended at a dose of 1–2 mg/kg per day (maximum, 80 mg), usually for a few days or up to a maximum of 3 weeks.

MANAGEMENT OF HEART FAILURE

See Chap. 253.

BED REST

Traditional recommendations for long-term bed rest, once the cornerstone of management, are no longer widely practiced. Instead, bed rest should be prescribed as needed while arthritis and arthralgia are present and for patients with heart failure. Once symptoms are well controlled, gradual mobilization can commence as tolerated.

CHOREA

Medications to control the abnormal movements do not alter the duration or outcome of chorea. Milder cases can usually be managed by providing a calm environment. In patients with severe chorea, carbamazepine or sodium valproate is preferred to haloperidol. A response may not be seen for 1–2 weeks, and medication should be continued for 1–2 weeks after symptoms subside. There is recent evidence that corticosteroids are effective and lead to more rapid symptom reduction in chorea. They should be considered in severe or refractory cases. Prednisone or prednisolone may be commenced at 0.5 mg/kg daily, with weaning as early as possible, preferably after 1 week if symptoms are reduced, although slower weaning or temporary dose escalation may be required if symptoms worsen.

INTRAVENOUS IMMUNOGLOBULIN (IVIG)

Small studies have suggested that IVIg may lead to more rapid resolution of chorea but have shown no benefit on the short- or long-term outcome of carditis in ARF without chorea. In the absence of better data, IVIg is *not* recommended except in cases of severe chorea refractory to other treatments.

PROGNOSIS

Untreated, ARF lasts on average 12 weeks. With treatment, patients are usually discharged from hospital within 1–2 weeks. Inflammatory markers should be monitored every 1–2 weeks until they have normalized (usually within 4–6 weeks), and an echocardiogram should be performed after 1 month to determine if there has been progression of carditis. Cases with more severe carditis need close clinical and echocardiographic monitoring in the longer term.

Once the acute episode has resolved, the priority in management is to ensure long-term clinical follow-up and adherence to a regimen of secondary prophylaxis. Patients should be entered onto the local ARF registry (if present) and contact made with primary care practitioners to ensure a plan for follow-up and administration of secondary prophylaxis before the patient is discharged. Patients and their families

should also be educated about their disease, emphasizing the importance of adherence to secondary prophylaxis.

PREVENTION

PRIMARY PREVENTION

Ideally, primary prevention would entail elimination of the major risk factors for streptococcal infection, particularly overcrowded housing. This is difficult to achieve in most places where ARF is common.

Therefore, the mainstay of primary prevention for ARF remains primary prophylaxis (i.e., the timely and complete treatment of group A streptococcal sore throat with antibiotics). If commenced within 9 days of sore throat onset, a course of penicillin (as outlined above for treatment of ARF) will prevent almost all cases of ARF that would otherwise have developed. In settings where ARF and RHD are common but microbiologic diagnosis of group A streptococcal pharyngitis is not available, such as in resource-poor countries, primary care guidelines often recommend that all patients with sore throat be treated with penicillin or, alternatively, that a clinical algorithm be used to identify patients with a higher likelihood of group A streptococcal pharyngitis. Although imperfect, such approaches recognize the importance of ARF prevention at the expense of overtreating many cases of sore throat that are not caused by group A streptococcus.

SECONDARY PREVENTION

The mainstay of controlling ARF and RHD is secondary prevention. Because patients with ARF are at dramatically higher risk than the general population of developing a further episode of ARF after a group A streptococcal infection, they should receive long-term penicillin prophylaxis to prevent recurrences. The best antibiotic for secondary prophylaxis is benzathine penicillin G (1.2 million units, or 600,000 units if ≤27 kg) delivered every 4 weeks. It can be given every 3 weeks, or even every 2 weeks, to persons considered to be at particularly high risk, although in settings where good compliance with an every-4-week dosing schedule can be achieved, more frequent dosing is rarely needed. Oral penicillin V (250 mg) can be given twice daily instead but is less effective than benzathine penicillin G. Penicillin-allergic patients can receive erythromycin (250 mg) twice daily.

The duration of secondary prophylaxis is determined by many factors, in particular the duration since the last episode of ARF (recurrences become less likely with increasing time), age (recurrences are less likely with increasing age), and the severity of RHD (if severe, it may be prudent to avoid even a very small risk of recurrence because of the potentially serious consequences) (Table 352-4). Secondary prophylaxis is best delivered as part of a coordinated RHD control program, based around a registry of patients. Registries improve the ability to follow patients and identify those who default from prophylaxis and to institute strategies to improve adherence.

TABLE 352-4 American Heart Association Recommendations for Duration of Secondary Prophylaxis[a]

CATEGORY OF PATIENT	DURATION OF PROPHYLAXIS
Rheumatic fever without carditis	For 5 years after the last attack or 21 years of age (whichever is longer)
Rheumatic fever with carditis but no residual valvular disease	For 10 years after the last attack, or 21 years of age (whichever is longer)
Rheumatic fever with persistent valvular disease, evident clinically or on echocardiography	For 10 years after the last attack, or 40 years of age (whichever is longer); sometimes lifelong prophylaxis

[a]These are only recommendations and must be modified by individual circumstances as warranted. Note that some organizations recommend a minimum of 10 years of prophylaxis after the most recent episode, or until 21 years of age (whichever is longer), regardless of the presence of carditis with the initial episode.

Source: Adapted from AHA Scientific Statement Prevention of Rheumatic Fever and Diagnosis and Treatment of Acute Streptococcal Pharyngitis. Circulation 119:1541, 2009.

FURTHER READING

Carapetis JR et al: Acute rheumatic fever and rheumatic heart disease. Nat Rev Dis Primers 14:15084, 2016.

Gewitz MH et al: Revision of the Jones Criteria for the diagnosis of acute rheumatic fever in the era of Doppler echocardiography: A scientific statement from the American Heart Association. Circulation 131:1806, 2015.

RHDAustralia (ARF/RHD Writing Group): The Australian guideline for prevention, diagnosis and management of acute rheumatic fever and rheumatic heart disease (2nd edition). Darwin: Menzies School of Health Research; 2012. Available at *https://www.rhdaustralia.org.au/arf-rhd-guideline.*

Steer AC et al: Status of research and development of vaccines for Streptococcus pyogenes. Vaccine 34:2953, 2016.

Zühlke L et al: Clinical outcomes in 3343 children and adults with rheumatic heart disease from 14 low- and middle-income countries: Two-year follow-up of the Global Rheumatic Heart Disease Registry (the REMEDY Study). Circulation 134:1456, 2016.

VIDEO 352-1A Transthoracic echocardiographic images of a 9-year-old girl with first episode of acute rheumatic fever. Images demonstrate the typical echocardiographic findings of acute rheumatic carditis. The valve leaflets are relatively thin and highly mobile. The failure of coaptation of the mitral valve leaflets is the result of chordal elongation and annular dilatation. The mitral valve regurgitation is moderate with a typical posterolaterally directed regurgitant jet of rheumatic carditis. ***A.*** Acute rheumatic carditis (apical four-chamber view echocardiogram).

VIDEO 352-1B Transthoracic echocardiographic images of a 9-year-old girl with first episode of acute rheumatic fever. Images demonstrate the typical echocardiographic findings of acute rheumatic carditis. The valve leaflets are relatively thin and highly mobile. The failure of coaptation of the mitral valve leaflets is the result of chordal elongation and annular dilatation. The mitral valve regurgitation is moderate with a typical posterolaterally directed regurgitant jet of rheumatic carditis. ***B.*** Acute rheumatic carditis (apical four-chamber view color Doppler echocardiogram).

VIDEO 352-1C Transthoracic echocardiographic images of a 9-year-old girl with first episode of acute rheumatic fever. Images demonstrate the typical echocardiographic findings of acute rheumatic carditis. The valve leaflets are relatively thin and highly mobile. The failure of coaptation of the mitral valve leaflets is the result of chordal elongation and annular dilatation. The mitral valve regurgitation is moderate with a typical posterolaterally directed regurgitant jet of rheumatic carditis. ***C.*** Acute rheumatic carditis (parasternal long-axis view echocardiogram).

VIDEO 352-1D Transthoracic echocardiographic images of a 9-year-old girl with first episode of acute rheumatic fever. Images demonstrate the typical echocardiographic findings of acute rheumatic carditis. The valve leaflets are relatively thin and highly mobile. The failure of coaptation of the mitral valve leaflets is the result of chordal elongation and annular dilatation. The mitral valve regurgitation is moderate with a typical posterolaterally directed regurgitant jet of rheumatic carditis. ***D.*** Acute rheumatic carditis (parasternal long-axis view color Doppler echocardiogram).

VIDEO 352-2A Transthoracic echocardiographic images are from a 5-year-old boy with chronic rheumatic heart disease with severe mitral valve regurgitation and moderate mitral valve stenosis. Images demonstrate the typical echocardiographic findings in advanced chronic rheumatic heart disease. Both the anterior and posterior mitral valve leaflets are markedly thickened. During diastole, the motion of the anterior mitral valve leaflet tip is restricted with doming of the body of the leaflet toward the interventricular septum. This appearance is commonly described as a "hockey stick" or an "elbow" deformity. ***A.*** Chronic rheumatic heart disease (parasternal long-axis view).

VIDEO 352-2B Transthoracic echocardiographic images are from a 5-year-old boy with chronic rheumatic heart disease with severe mitral valve regurgitation and moderate mitral valve stenosis. Images demonstrate the typical echocardiographic findings in advanced chronic rheumatic heart disease. Both the anterior and posterior mitral valve leaflets are markedly thickened. During diastole, the motion of the anterior mitral valve leaflet tip is restricted with doming of the body of the leaflet toward the interventricular septum. This appearance is commonly described as a "hockey stick" or an "elbow" deformity. ***B.*** Chronic rheumatic heart disease (apical two-chamber view echocardiogram).

353 Systemic Sclerosis (Scleroderma) and Related Disorders

John Varga

DEFINITION AND CLASSIFICATION

Systemic sclerosis (SSc) is a complex and clinically heterogeneous orphan disease with protean clinical manifestations, a chronic and frequently progressive course, and significant disability, disfigurement and mortality. Virtually every organ can be affected **(Fig. 353-1)**.

There is marked variability among SSc patients in patterns of skin involvement, organ complications, rates of disease progression, response to treatment, and survival. The early stages of SSc are associated with prominent inflammatory features; however, over time, structural alterations in multiple vascular beds and progressive visceral organ dysfunction due to fibrosis and atrophy come to dominate the clinical picture. Classification criteria for diagnosis of SSc are shown in **Table 353-1**.

Although thick and indurated skin (*scleroderma*) is the distinguishing hallmark of SSc, skin changes also occur in localized forms of scleroderma, along with multiple metabolic, inherited and autoimmune disorders **(Table 353-2)**. Patients with SSc can be broadly segregated into two major subsets defined by the pattern of skin involvement, clinical and laboratory features, and natural history **(Table 353-3)**. Diffuse cutaneous SSc (dcSSc) is typically associated with extensive skin induration starting in the fingers (sclerodactyly) and ascending from distal to proximal limbs and the trunk. In these patients, interstitial lung disease (ILD) and acute renal involvement develop relatively early. In contrast, in patients with limited cutaneous SSc (lcSSc), Raynaud's phenomenon generally precedes other disease manifestations, sometimes by years. In these patients, skin involvement remains confined to the fingers, distal limbs, and face, while the trunk is spared. The constellation of calcinosis cutis, Raynaud's phenomenon, esophageal dysmotility, sclerodactyly, and telangiectasia, was historically termed the *CREST syndrome*. In lcSSc, visceral organ involvement tends to show insidious progression, and digital ischemic ulcers, pulmonary arterial hypertension (PAH), hypothyroidism, and primary biliary cirrhosis may occur as late complications. In some patients, Raynaud's phenomenon and characteristic clinical and laboratory features of SSc occur in the absence of detectable skin thickening. This syndrome has been termed *SSc sine scleroderma*.

INCIDENCE AND PREVALENCE

SSc is an acquired sporadic disease with a worldwide distribution and affecting all races. In the United States, the incidence is 9–46 cases per million per year. There are an estimated 100,000 U.S. cases, although this number may be significantly higher if patients who do not meet classification criteria are also included. There are large regional variations in incidence rates, potentially reflecting differences in case definition, environmental exposures or susceptibility genes in populations with different ancestries. Prevalence rates in England, Europe, and Japan appear to be lower than in North America and Australia. Age, sex, and ethnicity influence disease susceptibility, and blacks have higher age-specific incidence rates. In common with other connective tissue diseases, SSc shows a strong female predominance (4.6:1), which is most pronounced in the childbearing years and declines after menopause. An additional risk factor is having an affected first-degree family member, which increases disease risk 13-fold. Although SSc can present at any age, the peak age of onset in women with both lcSSc and dcSSc is 65–74 years, although in blacks, disease onset occurs at an earlier age. Furthermore, blacks with SSc are more likely to have dcSSc, ILD, and a worse prognosis.

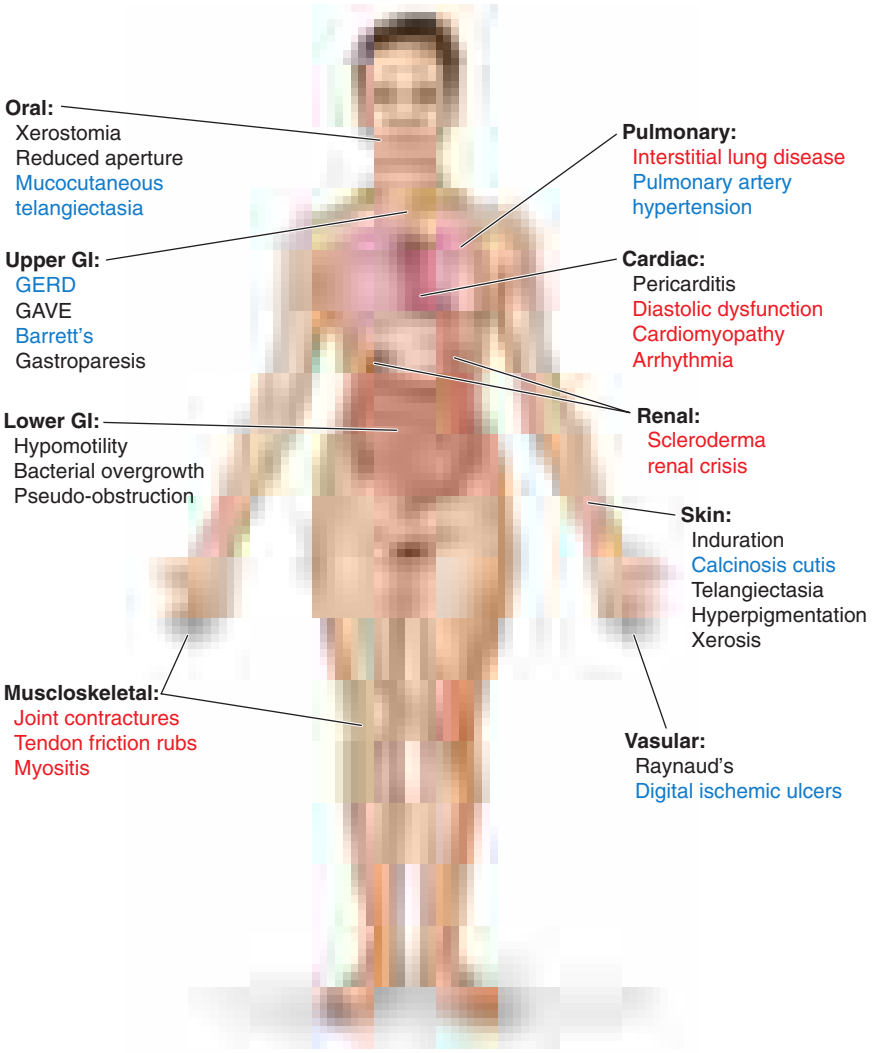


FIGURE 353-1 Multi-organ involvement in systemic sclerosis. Prominent complications more common in diffuse cutaneous SSc are shown in red; more common in limited cutaneous SSc in blue; and common in both forms of SSc shown in black.

GENETIC CONTRIBUTION TO DISEASE PATHOGENESIS

SSc is a polygenic disease. In general, the genetic associations of SSc identified to date make only a small contribution to disease susceptibility. Disease concordance rates are low (4.7%) in monozygotic twins, although concordance for antinuclear antibody (ANA) positivity is significantly higher. On the other hand, evidence for genetic contribution to disease susceptibility is provided by the observation that 1.6% of SSc patients have a first-degree relative with SSc, a prevalence rate markedly increased compared to the general population. The risk of Raynaud's phenomenon, ILD, and other autoimmune diseases, including systemic lupus erythematosus (SLE) **(Chap. 349)**, rheumatoid arthritis **(Chap. 351)**, and autoimmune thyroiditis **(Chap. 375)**, is also increased in first-degree relatives. Current approaches to uncover genetic factors in SSc include DNA sequencing and single nucleotide polymorphism (SNP) analysis of candidate genes, and SNP analysis of the entire genome in a hypothesis-free manner. Genome-wide association studies (GWASs) involve large multi-center and multi-national

TABLE 353-1 Classification Criteria for Diagnosis of Systemic Sclerosis

ITEM	SUB-ITEM	WEIGHT/SCORE
Skin thickening (bilateral)—fingers extending proximal to MCP joints		9
Skin thickening of fingers only	Puffy fingers Sclerodactyly (skin thickened distal to MCP joints)	2 4
Fingertip lesions	Digital tip ulcer or pitting scar	2 3
Mucocutaneous telangiectasia		2
Abnormal nails capillary pattern		2
Lung involvement	PAH Interstitial lung disease	2 2
Raynaud's phenomenon		3
SSc-specific autoantibodies	ACA Scl-70 RNA polymerase III	3

Abbreviations: ACA, anterior cerebral artery; MCP, metacarpophalangeal joint; PAH, pulmonary arterial hypertension.

TABLE 353-3 Subsets of Systemic Sclerosis (SSc): Features of Limited Cutaneous SSc versus Diffuse Cutaneous Disease

CHARACTERISTIC FEATURE	LIMITED CUTANEOUS SSc	DIFFUSE CUTANEOUS SSc
Skin involvement	Indolent onset. Limited to fingers, distal to elbows, face; slow progression	Rapid onset. Diffuse: fingers, extremities, face, trunk; rapid progression
Raynaud's phenomenon	Antedates skin involvement, sometimes by years; may be associated with critical ischemia in the digits	Onset coincident with skin involvement; critical ischemia less common
Musculoskeletal	Mild arthralgia	Severe arthralgia, carpal tunnel syndrome, tendon friction rubs; small and large joint contractures
Interstitial lung disease	Slowly progressive, generally mild	Frequent, early onset and progression, can be severe
Pulmonary arterial hypertension	Frequent, late, may occur as an isolated complication	Often occurs in association with interstitial lung disease
Scleroderma renal crisis	Very rare	Occurs in 15%; onset may be fulminant; generally early (<4 years from disease onset)
Calcinosis cutis	Frequent, prominent	Less common, mild
Characteristic autoantibodies	Anti-centromere	Anti–topoisomerase I (Scl-70), anti-RNA polymerase III

cohorts. A majority of the robustly validated susceptibility loci for SSc are genes involved in innate and adaptive immune responses, highlighting the importance of autoimmunity as the initial trigger for the disease. Genetic studies have shown associations with common (small effect size) variants related to B and T lymphocyte activation and signaling (*BANK1, BLK, CD247, STAT4, IL2RA, CCR6, IDO1, TNFSF4/OX40L, PTPN22*, and *TNIP1*). In addition, candidate gene studies and GWASs identified a strong association with human leukocyte antigen (HLA)-Class II haplotypes on chromosome 6, including *HLA-DRB1*11:04, DQA1*05:01*, and *DQB1*03:01*, and the non-HLA genes histocompatibility complex (MHC) genes *NOTCH4* and *PSORSC1*. Other genetic variants associated with SSc are involved in innate immunity and the interferon pathways (*IRF5, IRF7, STAT4, TNFAIP3/A20, GSDMA, PRDM1 (BLIMP1), TNFAIP3*, and *TLR2*). Additional associations with *IL12RB2, IL-21*, the apoptosis-related genes *DNASE1L3* and *SOX5*, and the fibrosis-related genes *CSK, CAV1, PPARG*, and *GRB10* have been reported. In addition to disease susceptibility, some of these genetic loci are associated with particular disease manifestations or serologic subsets, including ILD (*CTGF, CD226*), PAH (*TNIP1*), and scleroderma renal crisis (*HLA-DRB1**). While the functional consequences of these gene variants and their potential roles in pathogenesis are currently not well understood, it seems likely that in combination they cause a state of altered immune regulation, leading to increased susceptibility to autoimmunity and persistent inflammation. Of note, many of the genetic variants associated with SSc are also implicated in other autoimmune disorders, including SLE, rheumatoid arthritis, and psoriasis, suggesting common pathogenic pathways shared among these phenotypically dissimilar conditions. The genetic associations identified to date only explain a fraction of the heritability of SSc, and GWASs, and whole exome sequencing to identify additional genetic susceptibility factors in SSc, particularly rare (and potentially causal) variants, are currently ongoing.

TABLE 353-2 Conditions Associated with Skin Induration

- Systemic sclerosis (SSc)
  - Limited cutaneous SSc
  - Diffuse cutaneous SSc
- Localized scleroderma
  - Guttate (plaque) morphea, bullous morphea
  - Linear scleroderma, coup de sabre, hemifacial atrophy
- Pansclerotic morphea
- Overlap syndromes
  - Mixed connective tissue disease
  - SSc/polymyositis
- Diabetic scleredema and scleredema of Buschke
- Scleromyxedema (papular mucinosis)
- Chronic graft-versus-host disease
- Diffuse fasciitis with eosinophilia (Shulman's disease, eosinophilic fasciitis)
- Stiff skin syndrome
- Pachydermatoperiostosis (Primary hypertrophic osteoarthropathy)
- Chemically induced and drug-associated scleroderma-like conditions
  - Vinyl chloride–induced disease
  - Eosinophilia-myalgia syndrome (associated with L-tryptophan contaminant exposure)
  - Nephrogenic systemic fibrosis (associated with gadolinium exposure)
- Paraneoplastic syndrome

ENVIRONMENTAL AND OCCUPATIONAL RISK FACTORS

Given the relatively modest genetic contribution to disease susceptibility in SSc, environmental factors, such as infectious agents, intestinal microbiota, and occupational, dietary, lifestyle, and drug exposures, are likely to play a major role. Some evidence suggests potential roles for parvovirus B19, Epstein-Barr virus (EBV), cytomegalovirus (CMV), and *Rhodotorula glutinis* and other microorganisms. An epidemic of a novel syndrome with features suggestive of SSc occurred in Spain in the 1980s. The outbreak, termed *toxic oil syndrome*, was linked to use of contaminated rapeseed oil for cooking. Another epidemic outbreak, termed *eosinophilia-myalgia syndrome* (EMS), was linked to consumption of L-tryptophan-containing dietary supplements. Exposure to gadolinium contrast material in patients with renal compromise undergoing magnetic resonance scanning has been associated with nephrogenic systemic fibrosis. While each of these novel toxic-epidemic syndromes was characterized by chronic indurative skin changes and variable visceral organ involvement, the constellation of associated clinical, pathologic, and laboratory features distinguishes them from SSc. Occupational exposures tentatively linked with SSc include particulate silica (quartz), polyvinyl chloride, epoxy resins, welding fumes, and organic solvents and aromatic hydrocarbons including pain thinners, toluene, xylene, and trichloroethylene. These exposures might elicit

stable and heritable epigenetic changes such as DNA methylation and histone modification underlying pathogenic alterations in gene expression. Drugs implicated in SSc-like illnesses include bleomycin, pentazocine, and cocaine, and appetite suppressants linked with PAH. Radiation therapy for cancer has been linked with *de novo* onset of SSc as well as with exacerbation of pre-existing SSc. In contrast to rheumatoid arthritis, cigarette smoking does not increase the risk of SSc. Although case reports and series of SSc in women with silicone breast implants had raised concern regarding a possible causal role of silicone in SSc, large-scale epidemiologic investigations found no evidence of increased prevalence of SSc.

PATHOGENESIS

Three cardinal pathomechanistic processes underlie the protean clinical manifestations of SSc: (1) diffuse microangiopathy, (2) inflammation and autoimmunity, and (3) visceral and vascular fibrosis in multiple organs **(Fig. 353-2)**. While all three processes are concurrently operative in SSc patients, their activity, relative severity, and contribution to the overall clinical picture vary among individual patients and over time. In general, autoimmunity and altered vascular reactivity occur early, while fibrosis and atrophy occur later in the disease. Complex and dynamic interplay among these processes initiates and sustains the fibrotic process and tissue damage.

ANIMAL MODELS OF DISEASE

No single animal model of SSc fully reproduces the three cardinal processes that underlie pathogenesis, but some recapitulate selected aspects of the human disease. Tight-skin mice (Tsk1/+) spontaneously develop skin fibrosis due to a mutation in the fibrillin-1 gene. Mutant fibrillin-1 protein disrupts extracellular matrix assembly and causes aberrant activation of transforming growth factor β (TGF-β). Fibrillin-1 mutations in humans are associated with Marfan's disease and stiff skin syndrome, but have not been reported in SSc. Skin and lung fibrosis accompanied by variable vasculopathy and autoimmunity can be elicited in mice by injection of bleomycin or Angiotensin II, or by transplantation of HLA-mismatched bone marrow or spleen cells. Targeted genetic modifications in mice give rise to new disease models for investigating the pathogenetic roles of individual molecules, pathways, and cell types. For example, mice lacking IRF5, Smad3, uPAR, or peroxisome proliferator-activated receptor (PPAR)-γ, or constitutively overexpressing β-catenin, Wnt10b, sirtuin 3, Fra-2, TGFB1, PDGFRα, or adiponectin are either resistant or hypersensitive to experimental scleroderma, or spontaneously develop fibrosis. These disease models can contribute to understanding specific aspects of SSc pathogenesis, and to discovery and validation of novel targets for therapy.

MICROANGIOPATHY

Vascular injury is an early and possibly primary pathogenic event in SSc that leads to protean clinical manifestations of small vessel vasculopathy **(Fig. 353-3)**.

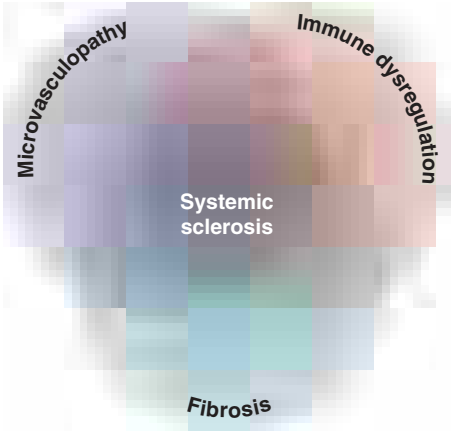


FIGURE 353-2 The characteristic constellation of vasculopathy, autoimmunity/inflammation and fibrosis underlies the protean clinical manifestations of systemic sclerosis.

Prominent microangiopathy in multiple vascular beds has important clinical sequelae including mucocutaneous telangiectasiae, Raynaud's phenomenon, ischemic digital ulcers, scleroderma renal crisis, myocardial involvement, and PAH. Raynaud's phenomenon is characterized by altered blood-flow response to cold challenge in small digital arteries. This initially reversible functional abnormality is associated with autonomic and peripheral nervous system alterations, including impaired production of the neuropeptide calcitonin gene-related peptide from sensory afferent nerves and heightened sensitivity of α_2-adrenergic receptors on vascular smooth-muscle cells. Isolated (primary) Raynaud's disease is common, generally benign and non-progressive. In contrast, secondary Raynaud's phenomenon in SSc is often progressive and complicated by irreversible structural changes, culminating in ischemic digital ulcers, necrosis, and amputation.

Viruses, cytotoxic factors, and chemokines thrombogenic microparticles, alternate complement pathway activation and autoantibodies targeting endothelial cells, phospholipids, and β2 glycoprotein I (β2GPI) are implicated as potential triggers of endothelial cell injury. Endothelial damage results in dysregulated production of vasodilatory (nitric oxide and prostacyclin) and vasoconstricting (endothelin-1) substances, as well as upregulation of intercellular adhesion molecule 1 (ICAM-1) and other surface adhesion molecules. Microvessels show enhanced permeability and transendothelial leukocyte diapedesis, abnormal activation of coagulation cascades, elevated thrombin production, and impaired fibrinolysis. Spontaneous platelet aggregation causes release of serotonin, platelet-derived growth factor (PDGF), and platelet alpha granules including thromboxane, a potent vasoconstrictor. Smooth-muscle cell–like myointimal cells in the media proliferate, the basement membrane is thickened and reduplicated, and perivascular adventitial fibrosis develops. The vasculopathic process affects capillaries, as well as arterioles, and less commonly even large vessels in many organs, resulting in reduced blood flow and tissue ischemia. Progressive luminal occlusion due to intimal and medial hypertrophy, combined with persistent endothelial cell damage and adventitial fibrosis, establish a vicious cycle that culminates in the striking absence of small blood vessels (rarefaction) in late-stage disease. Recurrent ischemia-reperfusion generates reactive oxygen species (ROS) that further damages the endothelium through peroxidation of membrane lipids. Paradoxically, the process of revascularization that normally reestablishes blood flow to ischemic tissue is defective in SSc despite elevated levels of other angiogenic factors. Moreover, bone marrow–derived circulating endothelial progenitor cells are reduced in number and impaired in function. Widespread capillary loss, obliterative vasculopathy of small and medium-sized arteries, and impaired ability to repair and replace damaged vessels are hallmarks of SSc.

INFLAMMATION AND AUTOIMMUNITY

Cellular Immunity The following observations provide support for the inflammatory/autoimmune nature of SSc: near-universal presence of circulating autoantibodies with defined specificities; familial clustering of SSc with other autoimmune diseases; detection of activated immune cells, including T cells with oligoclonal antigen receptors and T follicular helper-like cells, in target organs; prominent type I interferon (IFN) signatures, characterized by elevated expression of IFN-regulated genes, in a variety of cell types; elevated circulating levels and spontaneous secretion from mononuclear cells of cytokines and chemokines such as interleukin-6 (IL-6); tumor necrosis factor, IL-4, IL-10, IL-17, IL-33, CCL2, and CXCL4; genetic association of SSc with variants of MHC and other genes functionally implicated in the immune response; and the rapid clinical response, fibrosis resolution, and vascular regeneration observed in some SSc patients treated with immunomodulatory or immunoablative therapies. Genetic studies reveal strong associations with MHC locus alleles, as well as non-HLA-linked genes encoding mediators of both adaptive and innate immune responses (*CD247*, *STAT4*, *IRF5*, *CD226*, *TNFAIP3/A20*, and *TNFSF4*).

Circulating monocytes from SSc patients overexpress IFN-regulated genes such as Siglec-1, have reduced levels of caveolin-1, and exhibit an inherently profibrotic phenotype. In early (edematous) stage SSc,

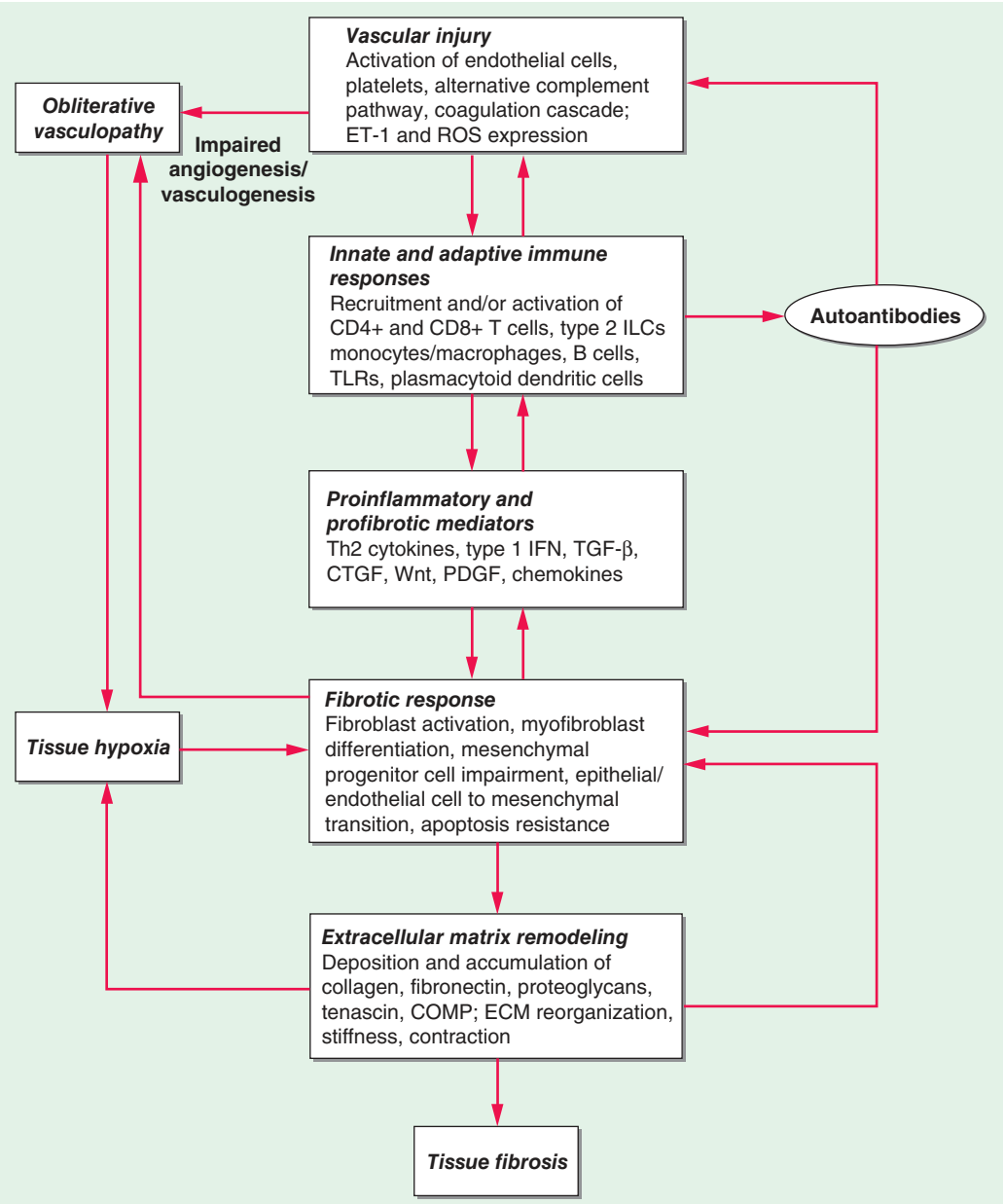


FIGURE 353-3 Initial vascular injury in a genetically susceptible individual triggers functional and structural vascular alterations, inflammation and autoimmunity, culminating in fibrosis. Inflammatory and immune responses initiate and sustain fibroblast activation and differentiation, resulting in pathologic fibrogenesis and irreversible tissue damage. Vascular damage results in tissue ischemia that further contributes to progressive fibrosis and atrophy. COMP, cartilage oligomeric matrix protein; CTGF, connective tissue growth factor; PDGF, platelet-derived growth factor; ROS, reactive oxygen species; TGF-β, transforming growth factor β; TLR, toll-like receptor.

mononuclear cell infiltrates comprised of activated T cells, monocytes/macrophages, and dendritic cells can be seen in skin, lungs, and other affected organs prior to appearance of fibrosis or vascular damage. Dendritic cells can be found in close proximity to activated fibroblasts and myofibroblasts and express toll-like receptors (TLR) and secrete IFN, IL-10, thymic stromal lymphopoietin (TSLP), and CXCL4, shaping the adaptive immune response and contributing to loss of immune tolerance. Tissue-infiltrating T cells express CD45 and HLA-DR activation markers and display restricted T cell receptor signatures indicative of oligoclonal expansion in response to recognition of as-yet unknown antigen. Of note, in patients diagnosed with SSc in close temporal association with cancer who are RNA polymerase III antibody-positive, the tumor may show mutations in RNApol3 autoantigen, which results in the generation of mutant-specific T cell immunity and cross-reactive antibodies. These findings support the premise that an abnormal antigen might act as initial trigger for the autoimmune response in SSc.

Circulating T cells in SSc express chemokine receptors and α_1 integrin, accounting for their enhanced binding to endothelium and to fibroblasts, while endothelial cells express ICAM-1 and other adhesion molecules that facilitate leukocyte diapedesis. Activated T cells show a T_H2-polarized immune response driven by dendritic cells. The T_H2 cytokines IL-4, IL-13, IL-33, and TSLP induce fibroblast activation, whereas the T_H1 cytokine interferon γ (IFN-γ) blocks cytokine-mediated fibroblast activation and exhibits anti-fibrotic properties. Evidence for altered Th17 and regulatory T cell (Treg) numbers and function in SSc has been reported. Type 2 innate lymphoid cells (iLCs), a recently discovered lymphoid cell population implicated in type 2 immunity and tissue remodeling, are also elevated in SSc skin biopsies. Alternately activated M2 macrophages, which produce TGF-β and promote angiogenesis and tissue remodeling, are increased in the skin in SSc. Although the frequency of regulatory T cells that enforce immune tolerance is elevated in the circulation and tissues, their immunosuppressive function appears to be defective. Some evidence implicates altered B cell homeostasis and function in SSc. Circulating B cells show elevated CD19 and co-stimulatory molecules CD80 and CD86, suggesting B cell chronic activation. Serum levels of a proliferation-inducing ligand (APRIL) and

B cell activating factor (BAFF), members of the TNF superfamily with potent effects on B cell activation, are elevated in SSc, and associate with extent of skin and lung involvement. B cells secrete IL-6, TGF-β, and other profibrotic cytokines implicated in pathogenesis. Thus, B cell hyperactivity in SSc might directly contribute to the inflammatory and fibrotic processes, as well as generation of autoantibodies. Microarray analysis identifies a distinct subset of SSc skin biopsies with elevated expression of inflammation-related genes. Evidence of innate immune and TLR signaling, reflecting activation by type 1 IFN from plasmacytoid dendritic cells, is prominent in peripheral blood cells and target organs.

Humoral Autoimmunity Circulating ANAs can be detected by indirect immunofluorescence in virtually all patients with SSc, even in early stages of disease. In addition, several SSc-specific autoantibodies with distinct patterns of immunofluorescence show strong associations with unique disease endophenotypes **(Table 353-4)**. These antibodies are directed mostly against intracellular proteins associated with transcription, DNA repair, and RNA processing. Owing to their high specificity, mutual exclusivity and association with unique disease manifestations, SSc-associated autoantibodies have substantial utility in clinical practice as diagnostic and prognostic markers, while their role in monitoring disease activity remains uncertain. Moreover, antibodies directed against fibrillin-1, matrix metalloproteinases, cell surface markers Angiotensin II receptor, endothelin-1 receptor, muscarinic 3 receptor, or the PDGF receptor, have been described in patients with SSc, although their clinical relevance is not yet established. These antibodies manifest functional receptor agonist activity and might have direct pathogenic roles.

A variety of mechanisms have been proposed to account for the generation of SSc-associated autoantibodies. Proteolytic cleavage, increased expression or altered subcellular localization of normal proteins, or their alterations due to mutation in the case of certain tumors, could lead to immune recognition as neoepitopes, resulting in the breaking of immune tolerance.

TABLE 353-4 Major Systemic Sclerosis-Specific Autoantibodies and Principal Associated Features

TARGET ANTIGEN	SSc SUBSET	PROMINENT CHARACTERISTIC CLINICAL ASSOCIATION
DNA Topoisomerase I (Scl-70) Speckled pattern	dcSSc	Tendon friction rubs, digital ischemic ulcers, scleroderma, extensive skin involvement, early ILD, cardiac involvement, scleroderma renal crisis
Centromere proteins Discreet speckled (centromere) pattern	lcSSc	Digital ischemic ulcers, calcinosis cutis, isolated PAH; renal crisis rare
RNA polymerase III Speckled pattern	dcSSc	Rapidly progressive skin, tendon friction rubs, joint contractures, GAVE, renal crisis, contemporaneous cancers; digital ulcers rare
U3-RNP (fibrillarin) Nucleolar pattern	dc/lcSSc	PAH, ILD, scleroderma renal crisis, GI tract involvement, myositis
Th/T_0 Nucleolar pattern	lcSSc	ILD, PAH
PM/Scl Nucleolar pattern	lcSSc	Calcinosis cutis, ILD, myositis overlap
Ku Speckled pattern	Overlap	SLE, myositis overlap
U1-RNP Speckled pattern	MCTD	PAH, inflammatory arthritis, myositis overlap
U11/U12 RNP Speckled pattern	dc/lcSSc	ILD

Abbreviations: dcSSc, diffuse cutaneous SSc; GAVE, gastric antral vascular ectasia; ILD, interstitial lung disease; lcSSc, limited cutaneous SSc; MCTD, mixed connective tissue disease; PAH, pulmonary arterial hypertension; SLE, systemic lupus erythematosus.

FIBROSIS

Fibrosis affecting multiple organs is a distinguishing feature of SSc. The process is characterized by replacement of normal tissue architecture with dense, rigid, avascular, and relatively acellular connective tissue. Fibrosis in SSc follows, and is a consequence of, inflammation, autoimmunity, and microvascular damage (Fig. 353-3). Fibroblasts are mesenchymal cells primarily responsible for the functional and structural integrity of connective tissue. Upon their activation by extracellular cues, fibroblasts proliferate, migrate, secrete collagens and other matrix molecules, growth factors, chemokines, and cytokines, and transdifferentiate into contractile myofibroblasts. Under normal conditions, these self-limited responses accomplish physiologic repair and regeneration of tissue. In contrast, when these responses become sustained and amplified, pathologic fibrosis results. Stimulatory signaling by endogenous TGF-β and paracrine fibrotic mediators including IL-6, IL-13, Wnt ligands, connective tissue growth factor (CTGF), PDGF, lysophosphatidic acid, endothelin-1, hypoxia, ROS, thrombin, and mechanical forces are responsible for sustained fibroblast activation underlying non-resolving fibrosis in SSc. Buildup of damage-associated endogenous ligands for TLR4 (EDA-fibronectin, high mobility group B1 [HMGB1] and Tenascin-C) and for TLR9 (mitochondrial DNA) within the fibrotic microenvironment further contributes to non-resolving fibrosis.

In addition to tissue-resident fibroblasts and transformed myofibroblasts, bone marrow–derived circulating mesenchymal progenitor cells also contribute to fibrosis. The factors that regulate the differentiation of mesenchymal progenitor cells and their trafficking from the circulation into lesional tissue are unknown. Epithelial and endothelial cells, mesenchymal progenitor cells, preadipocytes and tissue fibroblasts have all been proposed as sources of myofibroblasts in fibrosis. Although myofibroblasts are transiently found in normal wound healing, their persistence in fibrotic tissue, possibly due to resistance to apoptosis, contributes to scar formation.

Explanted SSc fibroblasts display an abnormally activated phenotype ex vivo, with variably increased rates of collagen production, spontaneous ROS generation, prominent stress fibers, and constitutive expression of alpha smooth-muscle actin. Persistence of the "scleroderma phenotype" during serial ex vivo passage of SSc fibroblasts may reflect autocrine TGF-β stimulatory loops, deregulated microRNA expressions, or stable acquired epigenetic modifications in these cells.

PATHOLOGY

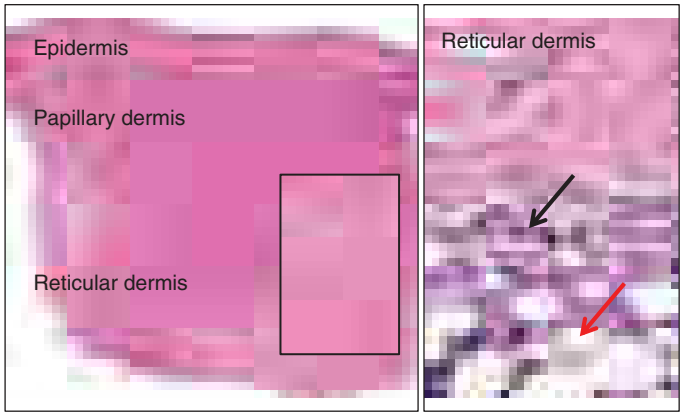
While pathological findings in SSc vary across anatomic sites, the distinguishing hallmark of SSc irrespective of the organ system is the triad of widespread capillary loss and obliterative microangiopathy, combined with fibrosis in the skin and internal organs. In early-stage disease, perivascular inflammatory cell infiltrates composed of T and B lymphocytes, activated monocytes and macrophages and mast cells may be detected in multiple organs. A non-inflammatory obliterative microangiopathy is a prominent late finding in the heart, lungs, kidneys, and gastrointestinal tract. Fibrosis is found in the skin, lungs, cardiovascular and gastrointestinal systems, tendon sheaths, perifascicular tissue surrounding skeletal muscle, and some endocrine organs. Excessive accumulation of collagens, proteoglycans, COMP and other structural matrix macromolecules progressively disrupts normal architecture, resulting in impaired function and failure of affected organs.

SKIN

The dermis is thickened, and accumulation of broad bundles of homogenized collagen oriented parallel to the epithelium is seen **(Fig. 353-4*A*)**. Adnexal glands are atrophic, and loss of periadnexal and intradermal white adipose tissue and its replacement with collagen can be striking. While perivascular mononuclear cell infiltrates may be seen early, established skin fibrosis generally shows absence of inflammation. These findings are histologically indistinguishable from those in localized scleroderma.

LUNGS

Autopsy studies in SSc universally show evidence of lung involvement. Most common is a nonspecific interstitial pneumonia (NSIP)



A

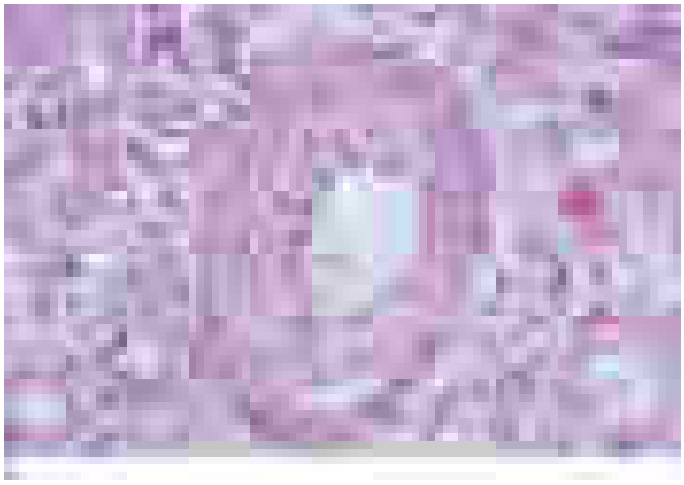


FIGURE 353-4 Pathologic findings in systemic sclerosis (SSc). ***A. Left panel:*** The skin is thickened due to expansion of the dermis. Inset, higher magnification showing thick hyalinized collagen bundles replacing skin appendages. ***Right panel:*** Mononuclear inflammatory cells within the intradermal adipose tissue. Black arrow, collagen; red arrow, dermal adipocytes. ***B.*** Early SSc-ILD. Diffuse fibrosis of the alveolar septae and a chronic inflammatory cell infiltrate. Trichrome stain. ***C.*** Pulmonary arterial obliterative vasculopathy. Striking intimal hyperplasia and luminal narrowing of small pulmonary artery, with little inflammation and minimal interstitial fibrosis, in a patient with SSc-PAH.

pattern characterized by variable interstitial fibrosis and mild chronic inflammation. Patchy infiltration of the alveolar walls with T lymphocytes, macrophages, and eosinophils may occur in early disease. With progression, interstitial fibrosis and vascular damage dominate, often coexisting within the same biopsy. The usual interstitial pneumonia (UIP) pattern of spatial/temporal heterogeneity of inflammation, fibrosis and fibrotic foci seen in idiopathic pulmonary fibrosis is less common in SSc (Fig. 353-4*B*). Fibrosis of the alveolar septae results in obliteration of the airspaces and loss of pulmonary blood vessels. This process impairs gas exchange and contributes to pulmonary hypertension. Intimal thickening of the pulmonary arteries, best seen with elastin stain, underlies SSc-associated PAH (Fig. 353-4*C*) and, at autopsy, is often associated with multiple pulmonary emboli and myocardial fibrosis. Patients may also show fibrosis and intimal proliferation in preseptal venules and veins in the lung, accounting for veno-occlusive disease. Lymphocytic bronchiolitis involving the submucosa of the terminal bronchioles may also be seen.

GASTROINTESTINAL TRACT

Pathologic changes can be found at any level from the mouth to the rectum. Atrophy and fibrosis of the muscularis propria and characteristic vascular lesions are prominent in the lower esophagus, while striated muscle in the upper third of the esophagus is generally spared. Collagenous replacement of the normal intestinal tract architecture results in impaired smooth muscle contractility and diminished peristaltic activity, with dysmotility, bacterial overgrowth, small-bowel obstruction, and perforation. Chronic gastroesophageal reflux is associated with esophageal inflammation, mucosal ulceration, and stricture formation and may lead to Barrett's metaplasia with attendant risk of adenocarcinoma. Esophageal dilation and reflux are associated with ILD due to chronic microaspiration.

KIDNEYS

In the kidneys, vascular lesions affecting the interlobular and arcuate arteries predominate. Chronic renal ischemia is associated with shrunken glomeruli. Patients with scleroderma renal crisis show acute fibrinoid necrosis of afferent arterioles, followed by intimal proliferation (onion-skin pattern), and ischemic collapse of glomeruli. These changes are reminiscent of thrombotic microangiopathies such as atypical hemolytic-uremic syndrome (see Chap. 304), and are accompanied by complement deposition, thrombosis, thrombocytopenia due to platelet consumption, and intravascular hemolysis. Extensive vascular thrombosis, glomerular collapse and sclerosis, and peritubular capillary deposits in renal biopsy are associated with irreversible renal failure.

HEART

Subclinical cardiac pathology is common, with prominent involvement of the myocardium and pericardium. The characteristic arteriolar lesions are concentric intimal hypertrophy and luminal narrowing, accompanied by patchy contraction band necrosis, loss of cardiac myocytes, and myocardial fibrosis due to microvascular involvement and ischemia-reperfusion injury. Fibrosis of the conduction system is common, especially at the sinoatrial node. The frequency of epicardial atherosclerotic coronary artery disease may be increased compared to the general population, similar to other systemic inflammatory diseases. Pericardial involvement with chronic inflammatory infiltrates and fibrinous exudates is common and may be associated with pericardial effusions.

PATHOLOGY IN OTHER ORGANS

Synovitis may be found in early SSc; with disease progression, the synovium becomes fibrotic, and in contrast to rheumatoid disease, pannus formation or bone resorption are uncommon. Fibrosis of tendon sheaths and fascia, sometimes accompanied by calcifications, produces palpable and sometimes audible tendon friction rubs. Inflammation and, in later stages, atrophy and fibrosis of skeletal muscles are common findings, and are similar to those in polymyositis. Fibrosis of the thyroid gland and of the minor salivary glands may be seen. Placentas from SSc pregnancies show decidual vasculopathy, which is associated with poor perinatal outcomes and fetal death.

CLINICAL FEATURES

OVERVIEW

SSc can affect virtually any organ (Fig. 353-1 and Table 353-5). Although a dichotomous approach stratifying SSc into diffuse and limited cutaneous subsets (Table 353-2) is useful, disease expression is far

TABLE 353-5 Frequency of Clinical Organ Involvement in Limited Cutaneous and Diffuse Cutaneous Systemic Sclerosis (SSc)

FEATURES	LIMITED CUTANEOUS SSc (%)	DIFFUSE CUTANEOUS SSc (%)
Skin involvement	90[a]	100
Raynaud's phenomenon	99	98
Ischemic digital ulcers	50	25
Esophageal involvement	90	80
Interstitial lung disease	35	65
Pulmonary arterial hypertension	15	15
Myopathy	11	23
Clinical cardiac involvement	9	12
Scleroderma renal crisis	2	15
Calcinosis cutis	—	—

[a]Approximately 10% of patients have SSc *sine* scleroderma.

more complex, and multiple distinct endophenotypes with unique patterns of manifestations can be recognized within each subset. Unique endophenotypes associate with autoantibodies with distinct and mutually exclusive specificities (Table 353-4). Patients with SSc "overlap" have typical features coexisting with clinical and laboratory evidence of another autoimmune disease, most commonly polymyositis, Sjögren's syndrome, polyarthritis, autoimmune liver disease, or SLE.

INITIAL CLINICAL PRESENTATION

Characteristic initial presentation is quite different in patients with the diffuse (dcSSc) versus limited (lcSSc) cutaneous forms of the disease. In dcSSc, the interval between Raynaud's phenomenon and onset of other disease manifestations is brief (weeks to months). Soft tissue swelling, puffy fingers, and intense pruritus are signs of the early inflammatory "edematous" phase. The fingers, distal limbs, and face are usually affected first. Diffuse hyperpigmentation of the skin, carpal tunnel syndrome arthralgias, muscle weakness, fatigue, and decreased joint mobility are common. During the ensuing weeks to months, the inflammatory edematous phase evolves into the "fibrotic" phase, with skin induration associated with hair loss, reduced production of skin oils, and decline in sweating capacity. Progressive flexion contractures of the fingers ensue. The wrists, elbows, shoulders, hip girdles, knees, and ankles become stiff due to fibrosis of the supporting joint structures. While advancing skin involvement is the most visible manifestation of early dcSSc, important and clinically silent internal organ involvement commonly occurs during this stage. The initial 4 years from disease onset is the period of most rapidly evolving pulmonary and renal damage. If organ failure does not occur during this phase of dcSSc, the systemic process may stabilize.

Compared to dcSSc, the course of lcSSc tends to be more indolent. The interval between onset of Raynaud's phenomenon and disease manifestations such as GERD, cutaneous telangiectasia, or soft tissue calcifications can be as long as years. Scleroderma renal crisis, significant ILD, and tendon friction rubs occur rarely in lcSSc, while PAH, and overlap with keratoconjunctivitis sicca, polyarthritis, cutaneous vasculitis, and biliary cirrhosis can develop many years after disease onset.

ORGAN INVOLVEMENT

RAYNAUD'S PHENOMENON

Raynaud's phenomenon, the most frequent extracutaneous complication of SSc, is characterized by episodes of reversible vasoconstriction in the fingers and toes, sometimes also affecting the tip of the nose and earlobes. Attacks, triggered by a decrease in temperature, as well as emotional stress and vibration, typically start with pallor, followed by cyanosis of variable duration. Hyperemia ensues spontaneously or with rewarming of the digit. The progression of the three color phases reflects the underlying vasoconstriction, ischemia, and reperfusion. Up to 5% of the general population has Raynaud's phenomenon. In the absence of signs or symptoms of an underlying condition, Raynaud's

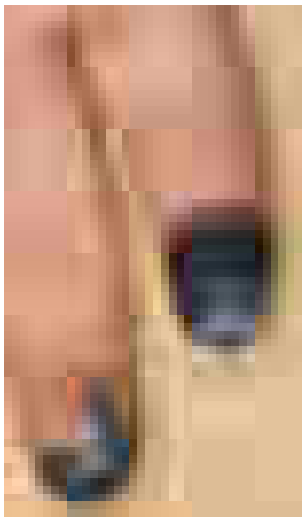


FIGURE 353-5 Digital necrosis. Sharply demarcated necrosis of the fingertip secondary to ischemia in a patient with limited cutaneous systemic sclerosis (SSc) associated with severe Raynaud's phenomenon.

phenomenon is classified as primary (Raynaud's disease), which represents an exaggerated physiologic response to cold. Secondary Raynaud's phenomenon occurs in SSc and other connective tissue diseases, hematologic and endocrine conditions, and occupational disorders, and can complicate treatment with beta blockers and anticancer drugs such as cisplatin and bleomycin. Distinguishing primary Raynaud's disease from secondary Raynaud's phenomenon can present a diagnostic challenge. Raynaud's disease is supported by the following: absence of an underlying cause, a family history of Raynaud's phenomenon, absence of digital tissue necrosis or ulceration, and a negative ANA test. Secondary Raynaud's phenomenon tends to occur at an older age (>30 years), is more severe (episodes more frequent, prolonged, and painful), and is associated with ischemic digital ulcers and loss of digits **(Fig. 353-5)**.

Nailfold capillaroscopy using a low-power stereoscopic microscope or opthalmoscope permits visualization of nailbed cutaneous capillaries under immersion oil **(Fig. 353-6)**. Raynaud's disease is associated with evenly spaced parallel vascular loops, whereas in secondary

FIGURE 353-6 SSc-associated nailfold capillary alterations. Normal nailfold pattern in healthy subjects. Note regularly-arrayed and uniform-size "hairpin" microvessels; "early pattern" showing dilations of microvessels and symmetrically increased microvessels (giant capillaries) representing the first morphological sign of systemic sclerosis; "active pattern" with giant capillaries, collapse with microhemorrages and loss of capillaries; "late pattern" showing massive loss of capillaries, fibrosis (white/yellow background) and neoangogenesis with secondary dilations (nailfold videocapillaroscope VIDEOCAP; magnification 220×). *(Courtesy of Professor Maurizio Cutolo, University of Genoa.)*

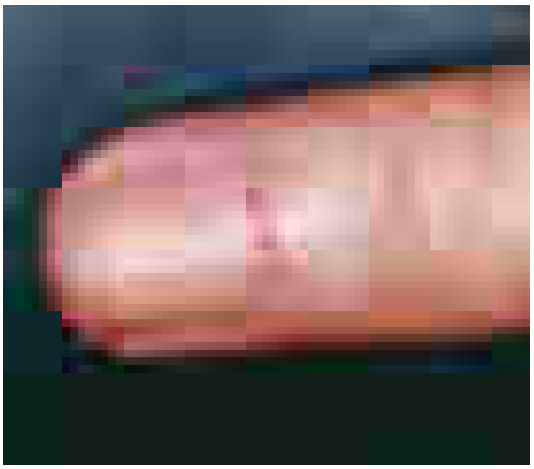
Raynaud's phenomenon, nailfold capillaries are distorted with widened and irregular loops, dilated lumen, microhemorrhages, and areas of vascular "dropout." Thus, nailfold capillaroscopy can be helpful in differentiating primary from secondary Raynaud's phenomenon and in establishing the early diagnosis of SSc.

SKIN FEATURES

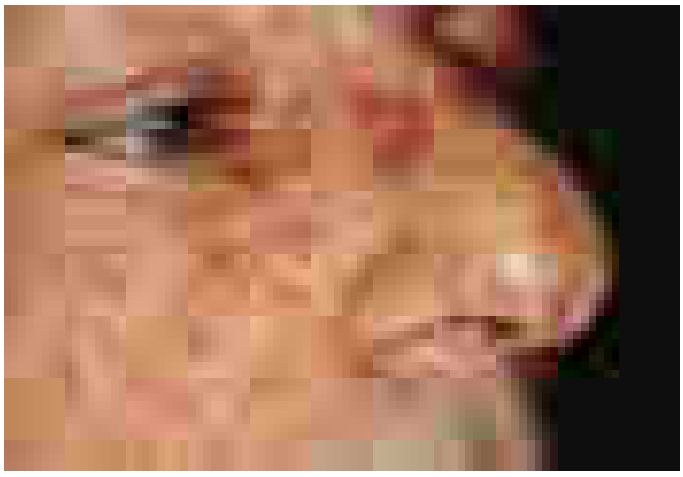
Bilateral symmetrical skin thickening is the hallmark of SSc that distinguishes it from other connective tissue diseases. Skin involvement starts in the fingers and characteristically advances from distal to proximal extremities in an ascending fashion. Some patients note diffuse tanning in the absence of sun exposure as a very early manifestation. In dark-skinned individuals, vitiligo-like hypopigmentation may occur. Because pigment loss spares the perifollicular areas, the skin may have a "salt-and-pepper" appearance, most prominently on the scalp, upper back, and chest. Dermal sclerosis obliterating hair follicles, sweat glands, and eccrine and sebaceous glands cause hair loss, decreased sweating, and dry and itchy skin on the extremities. Transverse creases on the dorsum of the fingers disappear **(Fig. 353-7)**. Fixed flexion contractures of the fingers cause reduced hand mobility and lead to muscle atrophy. Skin and subjacent tendon fibrosis accounts for fixed contractures of the wrists, elbows, and knees. Thick ridges at the neck due to firm adherence of skin to the underlying platysma muscle interfere with neck extension.

In established SSc, the face assumes a characteristic "mauskopf" appearance with taut and shiny skin, loss of wrinkles, and occasionally an expressionless facies due to reduced mobility of the eyelids, cheeks, and mouth. Thinning of the lips with accentuation of the central incisor teeth and prominent perioral radial furrowing (rhytides) complete the picture. Reduced oral aperture (microstomia) interferes with eating and oral hygiene. The nose assumes a pinched, beak-like appearance. In late-stage disease, the skin becomes thin and atrophic, and is firmly bound to the subcutaneous fat (tethering). Dilated skin capillaries 2–20 mm in diameter (telangiectasiae), reminiscent of hereditary hemorrhagic telangiectasia, are frequently seen on the face, hands, lips, and oral mucosa **(Fig. 353-8)**. The number of telangiectasias correlates with the severity of microvascular disease, including PAH. Breakdown of atrophic skin leads to chronic ulcerations at the extensor surfaces of the proximal interphalangeal joints, the volar pads of the fingertips, and bony prominences such as elbows and malleoli. Ulcers are often painful, heal slowly, and become secondarily infected, resulting in osteomyelitis. Healing of ischemic fingertip ulcerations leaves characteristic fixed digital "pits." Loss of soft tissue at the fingertips due to ischemia may be associated with striking resorption of the terminal phalanges (acro-osteolysis) **(Fig. 353-9)**.

Dystrophic calcifications in the skin, subcutaneous, and soft tissues (calcinosis cutis) in the presence of normal serum calcium and phosphate levels occur in up to 40% of patients, most commonly in those with long-standing anti-centromere antibody-positive lcSSc. Calcific deposits, composed of calcium hydroxyapatite crystals, vary in size from tiny punctate lesions to large conglomerate masses can be readily visualized on plain radiographs, or dual-energy CT. These deposits occur when calcium precipitates in tissue damaged by inflammation,



A



B

FIGURE 353-8 Cutaneous vascular changes. ***A.*** Vascular changes at the nailfold in lcSSc. ***B.*** Telangiectasia on the face.

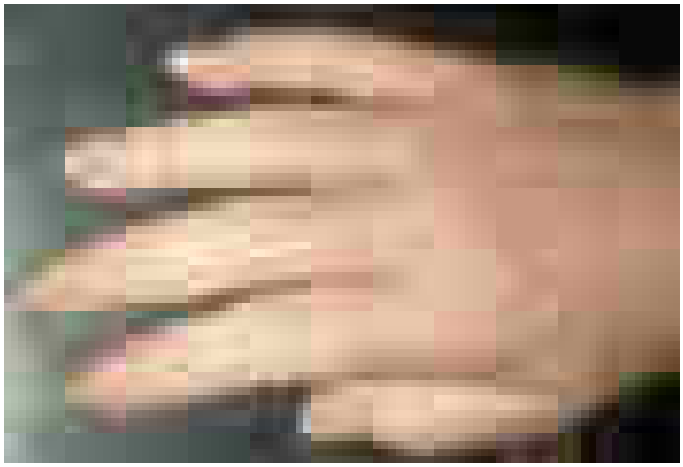


FIGURE 353-7 Sclerodactyly. Note skin induration on the fingers, and fixed flexion contractures of proximal interphalangeal joints, in a patient with limited cutaneous systemic sclerosis (lcSSc).

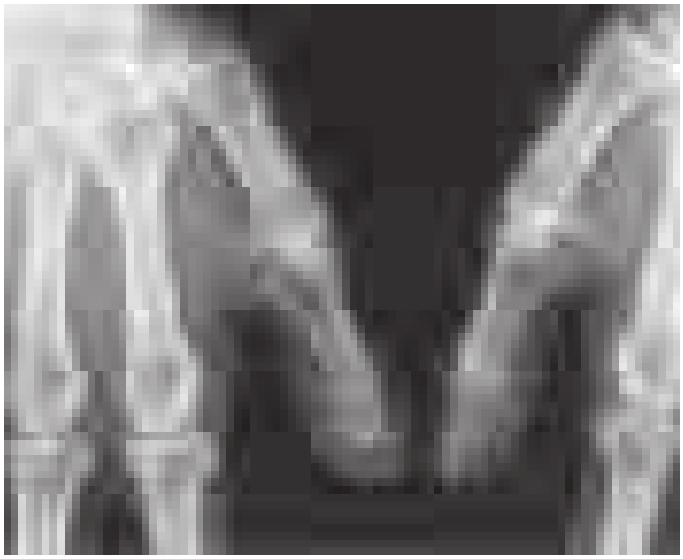


FIGURE 353-9 Acro-osteolysis. Note dissolution of distal terminal phalanges, commonly associated with ischemia, in a patient with long-standing limited cutaneous systemic sclerosis (lcSSc) and Raynaud's phenomenon.

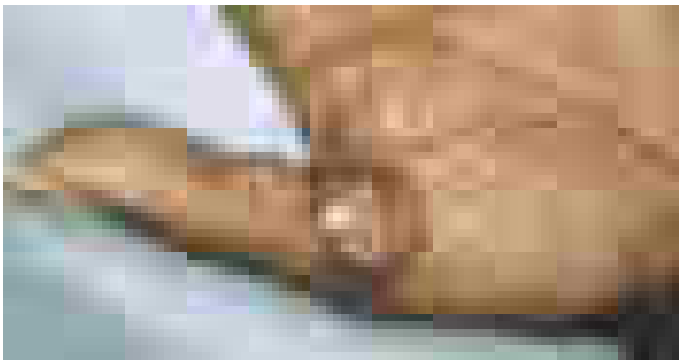


FIGURE 353-10 Calcinosis cutis. Note soft tissue calcific deposit breaking through the skin in a patient with limited cutaneous systemic sclerosis (lcSSc).

hypoxia, or local trauma. Common locations include the finger pads, palms, extensor surfaces of the forearms, and the olecranon and prepatellar bursae (Fig. 353-10). They can cause pain and nerve compression, ulcerate through the overlying skin with drainage of chalky white material, and secondary infections. Paraspinal sheet calcifications may cause neurologic complications.

PULMONARY FEATURES

The two principal forms of lung involvement in SSc, ILD, and pulmonary vascular disease are frequent and account for a majority of SSc-related deaths. Survival is particularly poor in SSc patients with concurrent presence of these two processes. Less common pulmonary complications of SSc include aspiration pneumonitis complicating chronic gastroesophageal reflux, pulmonary hemorrhage due to endobronchial telangiectasia, obliterative bronchiolitis, pleural reactions, restrictive physiology due to chest wall fibrosis, spontaneous pneumothorax, and drug-induced lung toxicity. The incidence of lung cancer is increased in SSc.

Interstitial Lung Disease While evidence of ILD can be found in up to 65% of SSc patients by high-resolution computed tomography (HRCT), clinically significant ILD develops in 16–43%; the frequency varies depending on the detection method used. Risk factors for ILD include male sex, African-American race, diffuse skin involvement, severe gastroesophageal reflux, and the presence of topoisomerase I autoantibodies; in contrast, anti-centromere antibody-positive patients have a reduced risk of ILD. Additional risk factors include low forced vital capacity (FVC) or single-breath diffusing capacity of the lung for carbon monoxide (DLco) at initial presentation. Esophageal dilation with chronic acid reflux in SSc cause micro-aspiration, a risk factor for the development and progression of ILD. The most rapid progression in ILD generally occurs early in the disease course (within the first 3 years), when the FVC can decline by 30% per year.

Pulmonary involvement can remain asymptomatic until it is advanced. The most common presenting respiratory symptoms—exertional dyspnea, fatigue, and reduced exercise tolerance—are subtle and slowly progressive. A chronic dry cough may be present. Physical examination may reveal fine inspiratory "Velcro" crackles at the lung bases. Pulmonary function testing (PFT) is relatively sensitive for detecting early pulmonary involvement, and typically shows a restrictive ventilatory defect (FV<70% predicted and/or FEV1/FVC ratio >0.8), reduced total lung capacity (TLC) and diffusing capacity (DLco). A reduction in DLco that is significantly out of proportion to the reduction in lung volumes should raise suspicion for pulmonary vascular disease, but may also be due to anemia. Oxygen desaturation with exercise is common.

Chest radiography can be used as an initial screening tool to rule out infection and other causes of pulmonary involvement; however, compared to HRCT, it is relatively insensitive for detection of early ILD. It may demonstrate lower lobe subpleural reticular linear opacities and ground-glass opacifications, even in asymptomatic patients with normal PFTs (Fig. 353-11). Additional HRCT findings include mediastinal lymphadenopathy, pulmonary nodules, traction bronchiectasis, and

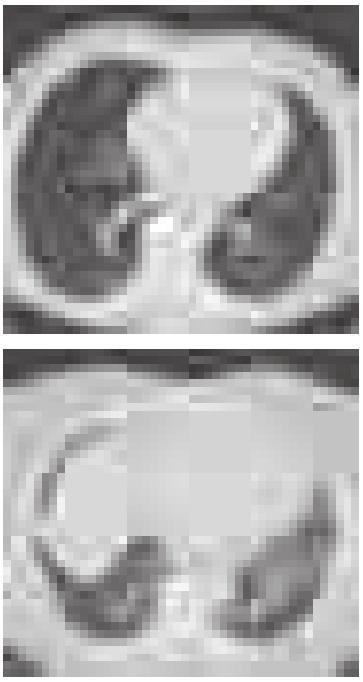


FIGURE 353-11 Chest CT in systemic sclerosis. ***Top panel:*** Early interstitial lung disease with subpleural reticulations and ground glass opacities in the lower lobes. Patient in supine position. ***Bottom panel:*** Extensive lung fibrosis with coarse reticular honeycombing, and traction bronchiectasis. Note dilated esophagus. *(Courtesy of Rishi Agrawal, Northwestern University.)*

uncommonly, honeycomb changes. The extent of interstitial changes on chest HRCT is a predictor of ILD progression and mortality. Bronchoalveolar lavage (BAL) can demonstrate inflammatory cells in the lower respiratory tract, and may be useful for ruling out tuberculosis and other infections. However, BAL does not appear to be useful for SSc diagnosis or for identifying reversible alveolitis, and is used primarily for research. Lung biopsy is indicated only in patients with atypical findings on chest radiographs. The histologic pattern on lung biopsy may predict the risk of progression of ILD, with NSIP, carrying a better prognosis than UIP.

Pulmonary Arterial Hypertension PAH resulting from vascular remodeling of small (<500 μm) pulmonary arteries develops in 8–12% of patients with SSc, and occurs as an isolated abnormality or in association with ILD. PAH is defined hemodynamically as a mean pulmonary artery pressure ≥25 mmHg with a pulmonary capillary wedge pressure ≤15 mmHg and pulmonary vascular resistance >3 Wood units. The natural history of SSc-associated PAH is variable, but often follows a downhill course with onset of right heart failure. The 3-year survival of SSc patients with untreated PAH is <50%. Risk factors include lcSSc, high numbers of cutaneous telangiectasia, older age at disease onset, and the presence of antibodies to centromere, U1-RNP, U3-RNP (fibrillarin), and B23. Mutations in the BMPR2 gene associated with idiopathic PAH are not found in patients with SSc-PAH.

Although patients with PAH are often asymptomatic in early stages, they may present with nonspecific symptoms of exertional dyspnea and reduced exercise capacity. With progression, angina, near-syncope, and symptoms and signs of right-sided heart failure appear. Physical examination may show tachypnea, a loud pulmonic component of the S_2 heart sound, pulmonic/tricuspid regurgitation murmur, palpable right ventricular heave, elevated jugular venous pressure, and dependent edema. Doppler echocardiography provides a noninvasive screening method for estimating the pulmonary arterial pressure. In light of the poor prognosis of untreated PAH and better therapeutic response in patients with early diagnosis, all SSc patients should be screened for PAH at initial evaluation, followed by annual evaluation. Estimated pulmonary artery systolic pressure >40 mmHg at rest or tricuspid regurgitation jet velocities >3 m/sec suggest PAH. PFT may show a reduced DLco in isolation or out of proportion with

the severity of restriction. Because echocardiography can over- or underestimate pulmonary artery pressures, cardiac catheterization is the gold standard required to confirm the diagnosis of suspected PAH, to assess its severity, including the degree of right heart dysfunction, to rule out veno-occlusive disease and other cardiac (post-capillary) causes of pulmonary hypertension, and to provide prognostic parameters. Yearly echocardiographic screening for PAH is recommended in most patients; an isolated decline in DLco may also be indicative of developing PAH. Distinguishing PAH from pulmonary hypertension secondary to pulmonary fibrosis and hypoxia in SSc can be difficult. Serum levels of N-terminal pro-brain natriuretic peptide (NT proBNP) correlate with the presence and severity of PAH in SSc, as well as survival. While NT proBNP measurements can be useful in screening for PAH and in monitoring the response to treatment, elevated levels are not specific for PAH and also occur in other forms of right and left heart disease. Despite more favorable hemodynamics, the prognosis of SSc-associated PAH is worse, and treatment response poorer, than that of idiopathic PAH, most likely due to frequent concurrence of ILD and cardiac complications in these patients.

GASTROINTESTINAL INVOLVEMENT

Involvement of the gastrointestinal tract, which can affect any level, occurs in up to 90% of SSc patients with both lcSSc and dcSSc disease (Table 353-6). The pathologic findings of fibrosis, smooth muscle atrophy, and obliterative small-vessel vasculopathy are similar throughout the length of the gastrointestinal tract, and contribute to reduced quality of life, malnutrition, and increased mortality.

Upper Gastrointestinal Tract Involvement. Decreased oral aperture interferes with regular dental hygiene. Teeth are loosened due to loss of periodontal ligament attaching teeth to the alveolar bone. Additional oropharyngeal manifestations due to a combination of xerostomia, shortened frenulum, and resorption of the mandibular condyles are frequent and cause much distress. Most patients have symptoms of gastroesophageal reflux disease (GERD): heartburn, regurgitation, and dysphagia. A combination of reduced lower esophageal sphincter pressure resulting in reflux, impaired esophageal clearance of refluxed gastric contents due to diminished motility, and delayed gastric emptying accounts for GERD. Calcium channel antagonists and phosphodiesterase inhibitors used to treat Raynaud's phenomenon can further aggravate reflux. Esophageal manometry shows abnormal motility in most patients, even in the absence of symptoms. Extra-esophageal manifestations of GERD include hoarseness, chronic cough, and microaspiration, which can result in infections and may aggravate underlying ILD. Chest CT characteristically shows a dilated patulous esophagus with intraluminal air. Endoscopy may be necessary to rule out opportunistic infections with *Candida*, herpes virus, and CMV. Severe erosive esophagitis may be found on endoscopy in patients with minimal symptoms. Esophageal strictures and Barrett's esophagus may complicate chronic GERD. Because Barrett's metaplasia is associated with increased risk of adenocarcinoma, SSc patients with Barrett's require regular surveillance endoscopy with biopsy.

Gastroparesis with early satiety, abdominal distention, and aggravated reflux symptoms are common. Barium contrast studies are neither sensitive nor specific for evaluation of gastric involvement in SSc. Gastric antral vascular ectasia (GAVE) in the antrum may occur. These subepithelial lesions, reflecting the diffuse small-vessel vasculopathy of SSc, are described as "watermelon stomach" due to their endoscopic appearance. Patients with GAVE can have recurrent episodes of gastrointestinal bleeding, resulting in chronic unexplained anemia.

Lower Gastrointestinal Tract and Anorectal Involvement Weight loss and malnutrition due to impaired intestinal motility, malabsorption, and chronic diarrhea secondary to bacterial overgrowth are common. Fat and protein malabsorption and vitamin B_{12} and vitamin D deficiencies ensue, and may be further exacerbated by pancreatic insufficiency. Disturbed intestinal motor function can also lead to intestinal pseudo-obstruction, with symptoms that are indistinguishable from those of delayed gastric emptying. Patients present with recurrent episodes of acute abdominal pain, nausea, and vomiting, and radiographic studies show acute intestinal obstruction. A major diagnostic challenge is differentiating pseudo-obstruction, which responds to supportive care and intravenous nutritional supplementation, from mechanical obstruction. Colonic involvement may result in severe constipation, occasionally complicated by sigmoid volvulus. Fecal incontinence, gastrointestinal bleeding from telangiectasia, and rectal prolapse, can occur. In late-stage SSc, wide-mouth sacculations or diverticula occur in the colon, occasionally causing perforation and bleeding. An occasional radiologic finding is pneumatosis cystoides intestinalis due to air trapping in the bowel wall that may rarely rupture and cause benign pneumoperitoneum. Although the liver is rarely affected, primary biliary cirrhosis may coexist with SSc.

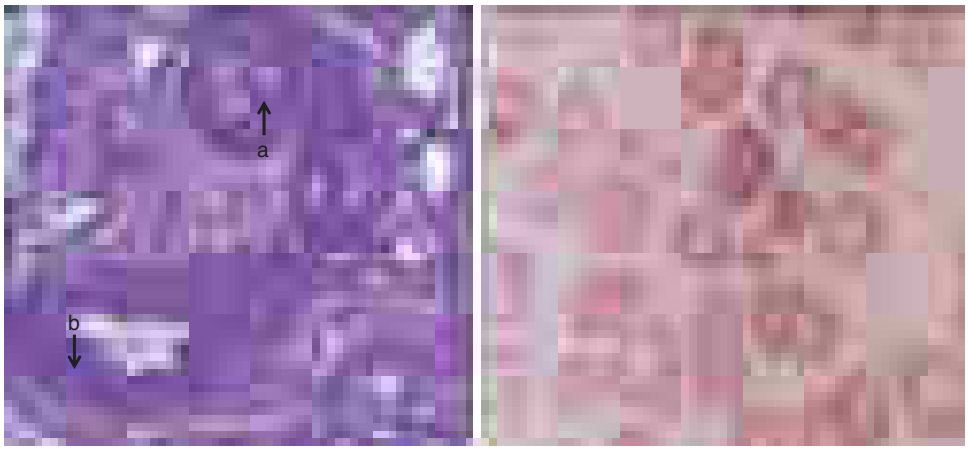
TABLE 353-6 Prominent Gastrointestinal Manifestations of SSc and Their Management

SITE	PRINCIPAL MANIFESTATION	MANAGEMENT
Oropharynx	Diminished oral aperture Dry mouth Periodontitis, gingivitis swallowing	Periodontal care Artificial saliva Swallowing therapy
Esophagus	Reflux Dysphagia Strictures Barret's metaplasia	Lifestyle modifications Prokinetic drugs proton pump inhibitors Endoscopic procedures
Stomach	Gastroparesis Gastric antral vascular ectasia (GAVE, watermelon stomach)	Prokinetic agents Endoscopic laser cryotherapy
Small and large intestines	Bacterial overgrowth Diarrhea/constipation Pseudo-obstruction Pneumatosis intestinalis Malabsorption Colonic pseudodiverticula	Laxatives Prokinetic agents Rotating antibiotics Octreotide Parenteral nutritional support
Anorectum	Sphincter incompetence	Biofeedback, sacral nerve stimulation, surgery

RENAL INVOLVEMENT: SCLERODERMA RENAL CRISIS

Scleroderma renal crisis presents with accelerated hypertension accompanied by acute kidney injury and progressive failure. This acute life-threatening complication of SSc occurs in 10–15% of patients, generally within 4 years of disease onset. Rarely, scleroderma renal crisis can be the initial presenting manifestation of SSc. Prior to the advent of angiotensin-converting enzyme (ACE) inhibitors, short-term survival in scleroderma renal crisis was <10%. The pathogenesis involves obliterative vasculopathy and luminal narrowing of the renal arcuate and interlobular arteries, with consequent intravascular hemolysis, along with evidence of activation of the complement pathways (Fig. 353-12). Progressive reduction in renal blood flow, aggravated by vasospasm, leads to juxtaglomerular renin secretion and activation of Angiotensin II, with further renal vasoconstriction resulting in a vicious cycle that culminates in accelerated hypertension. Risk factors for scleroderma renal crisis include African-American race, male sex, and diffuse or progressive skin involvement. Up to 50% of patients with scleroderma renal crisis have anti-RNA polymerase III antibodies, whereas patients with anti-centromere antibodies appear to be protected from this complication. Palpable tendon friction rubs, pericardial effusion, new unexplained anemia, and thrombocytopenia may be harbingers of impending scleroderma renal crisis. High-risk patients with early SSc should monitor their blood pressure daily. Because glucocorticoid use is associated with scleroderma renal crisis, prednisone in high-risk SSc patients should be taken only when absolutely required and at low doses (<10 mg/d).

Patients characteristically present with accelerated hypertension (generally >150/90 mmHg) and progressive oliguric renal insufficiency. However, ~10% of patients with scleroderma renal crisis present



A *B*

FIGURE 353-12 Renal changes in scleroderma renal crisis. ***A.*** Renal biopsy demonstrating intimal proliferation and myxoid changes in medium-sized renal arteries (*arrows*). ***B.*** Fragmentation of red blood cells due to intravascular hemolysis in scleroderma renal crisis. (*Courtesy of Drs. Edward Stern and Christopher Denton, Royal Free Hospital, London, UK.*)

with normal blood pressure. Normotensive renal crisis is generally associated with a poor outcome. Headache, blurred vision, congestive heart failure, and pulmonary edema may accompany elevation of blood pressure. Urinalysis typically shows mild proteinuria, granular casts, and microscopic hematuria; moderate thrombocytopenia and microangiopathic hemolysis with fragmented red blood cells can be seen. Progressive oliguric renal failure over several days generally follows. Scleroderma renal crisis is occasionally misdiagnosed as thrombotic thrombocytopenic purpura (TTP) or other forms of thrombotic microangiopathy. In such cases, renal biopsy and measuring vWF-cleaving protease activity may be of some benefit. Oliguria or a creatinine >3 mg/dL at presentation predicts poor outcome (permanent hemodialysis and mortality), as do biopsy findings of vascular thrombosis and glomerular ischemic collapse. Rarely, crescentic glomerulonephritis occurs in the setting of SSc and may be associated with myeloperoxidase-specific antineutrophil cytoplasmic antibodies. Membranous glomerulonephritis may occur in patients treated with D-penicillamine. Asymptomatic renal function impairment occurs in up to half of SSc patients. Such subclinical renal involvement is associated with other vascular manifestations of SSc and rarely progresses.

CARDIAC INVOLVEMENT

Although it is often silent, variable cardiac involvement in SSc is detected in 10–50% of patients screened with sensitive diagnostic tools. Clinical cardiac involvement, more frequent in dcSSc than in lcSSc, may be primary or secondary to PAH, ILD, or renal involvement, and is associated with poor outcomes. The endocardium, myocardium, and pericardium may each be affected separately or together. Pericardial involvement is manifested as pericarditis, pericardial effusions, constrictive pericarditis, and rarely, cardiac tamponade. Conduction system fibrosis occurs commonly and may be silent or manifested by heart block. Arrythmias including premature ventricular contractions, atrial fibrillation, and supraventricular tachycardia are common. Microvascular involvement, recurrent vasospasm, and ischemia-reperfusion injury contribute to patchy myocardial fibrosis, resulting in asymptomatic systolic or diastolic left ventricular dysfunction that may progress to overt heart failure. Acute or subacute myocarditis leading to left ventricular dysfunction may occur, and diagnosis requires cardiac magnetic resonance imaging (MRI) or endomyocardial biopsy. While conventional echocardiography has low sensitivity for detecting preclinical heart involvement in SSc, newer modalities such as tissue Doppler echocardiography (TDE), cMRI, and nuclear imaging (single photon emission CT [SPECT]) reveal a high prevalence of abnormal myocardial function or perfusion. The serum levels of N-terminal pro-BNP, a ventricular hormone elevated in SSc-PAH, may also have utility as markers of primary cardiac involvement.

Musculoskeletal Complications

Musculoskeletal complications are very common in SSc. Carpal tunnel syndrome may be a presenting disease manifestation. Generalized arthralgia and stiffness are prominent in early disease. Mobility of both small and large joints is progressively impaired, and fixed contractures develop at the proximal interphalangeal joints and wrists. Large joint contractures, seen in patients with dcSSc, are frequently accompanied by tendon friction rubs characterized by coarse leathery crepitation heard or palpated upon passive joint movement, that are due to extensive fibrosis and adhesion of the tendon sheaths and fascial planes at the affected joint. Tendon friction rubs are associated with increased risk for renal and cardiac complications and reduced survival. Synovitis detected by ultrasound or MRI is common; occasional SSc patients develop erosive polyarthritis in the hands, and some have a seropositive rheumatoid arthritis overlap. Muscle weakness is common and multifactorial: deconditioning, disuse atrophy, malnutrition, inflammation, and fibrosis may all contribute. A chronic non-inflammatory myopathy characterized by atrophy and fibrosis with mildly elevated muscle enzymes can be seen in late-stage SSc. Bone resorption in the terminal phalanges causes loss of the distal tufts (acro-osteolysis) (Fig. 353-9). Resorption of the mandibular condyles can lead to bite difficulties. Osteolysis can also affect the ribs and distal clavicles.

LESS RECOGNIZED DISEASE MANIFESTATIONS

Dry eyes and dry mouth (sicca complex) are common in SSc. Biopsy of the minor salivary glands shows fibrosis rather than focal lymphocytic infiltration characteristic of primary Sjögren's syndrome (Chap. 354). Hypothyroidism resulting from Graves' or Hashimoto's disease is common, particularly in lcSSc, and may be under-recognized. Whereas the central nervous system is generally spared, unilateral or bilateral sensory trigeminal neuropathy can occur. Erectile dysfunction is a frequent, and occasionally initial, disease manifestation. Inability to attain or maintain penile erection is due to vascular insufficiency and fibrosis of corporeal smooth muscle. Sexual performance is also adversely affected in women. While fertility is not impaired in SSc, pregnancy is associated with higher risk of adverse fetal outcomes. Furthermore, cardiopulmonary involvement may worsen during pregnancy, and new onset of scleroderma renal crisis has been described.

Cancer Epidemiologic studies indicate an increased cancer risk in SSc. Lung cancer and esophageal adenocarcinoma typically occur in the setting of long-standing ILD or GERD and may be caused by chronic inflammation and repair. In contrast, breast, lung, and ovarian carcinomas and lymphomas tend to occur in close temporal association with the onset of SSc, particularly in patients who have autoantibodies to RNA polymerase III. In this scenario, SSc may represent a paraneoplastic syndrome triggered by the anti-tumor immune response.

LABORATORY EVALUATION AND BIOMARKERS

Mild microcytic anemia is frequent and may indicate gastrointestinal bleeding caused by GAVE or chronic esophagitis. Macrocytic anemia may be caused by folate and vitamin B_{12} deficiency due to small-bowel bacterial overgrowth and malabsorption or by drugs such as methotrexate. Microangiopathic hemolytic anemia caused by mechanical fragmentation of red blood cells during their passage through microvessels coated with fibrin or platelet thrombi is a hallmark of scleroderma renal crisis. The erythrocyte sedimentation rate (ESR) is generally normal; an elevation may signal coexisting myositis or malignancy.

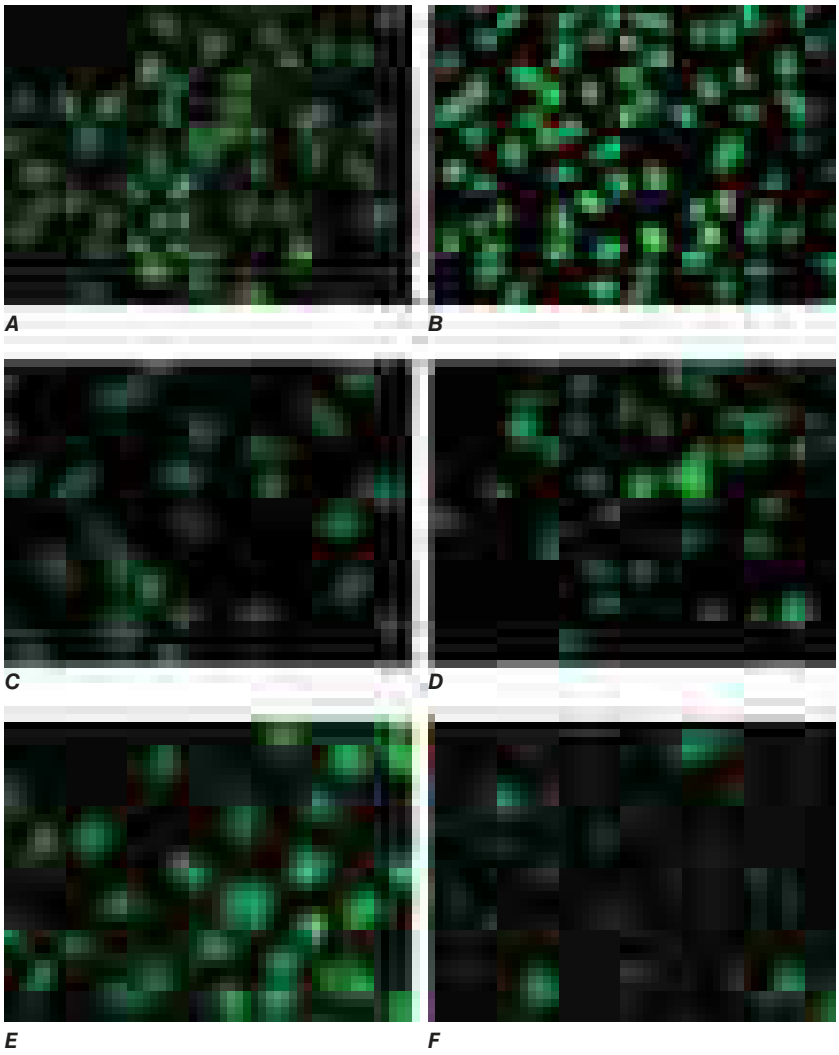


FIGURE 353-13 SSc-associated autoantibodies: immunofluorescence patterns. Indirect immunofluorescence on HEp-2 substrate shows distinct patterns: ***A.*** anti-centromere; ***B.*** anti-Scl-70/topoisomerase I; ***C.*** anti-PM/Scl; ***D.*** anti-Th/To; ***E.*** anti-RNA polymerase III; ***F.*** anti-fibrillarin/U3RNP antibodies. Except for anti-centromere (discrete dots in metaphase nucleus), variations of nucleolar staining are clues to autoantibody specificity. However, immunoassays employing purified autoantigens are recommended to confirm specificity of these autoantibodies. *(Courtesy of Marvin Fritzler and Susan Copple, Inova Diagnostics Inc., San Diego, California.)*

Antinuclear autoantibodies are detected in almost all patients with SSc. Anti-topoisomerase I (Scl-70) and anti-centromere antibodies are mutually exclusive and each is highly specific for SSc. Topoisomerase I antibodies are associated with increased risk of ILD and poor outcomes. Anti-centromere antibodies are associated with PAH, but only infrequently with significant cardiac, pulmonary, or renal involvement. Nucleolar immunofluorescence pattern may indicate antibodies to U3-RNP (fibrillarin), Th/To, or PM/Scl, whereas speckled immunofluorescence indicates antibodies to RNA polymerase III **(Fig. 353-13)**.

DIAGNOSIS, STAGING, AND MONITORING

The diagnosis of SSc is made primarily on clinical grounds and is generally straightforward in patients with established disease. The presence of skin induration with a characteristic symmetric distribution pattern associated with typical visceral organ manifestations establishes the diagnosis with a high degree of certainty. In lcSSc, a history of Raynaud's phenomenon and GERD symptoms, coupled with sclerodactyly and nailfold capillary changes, often in combination with cutaneous telangiectasia and calcinosis cutis, help to establish the diagnosis. Primary Raynaud's disease is a benign condition that must be differentiated from early or limited SSc. Nailfold microscopy is particularly helpful in this situation, because in contrast to SSc, nailfold capillaries are normal. Diagnosing SSc at an early stage may be a challenge. In dcSSc, initial symptoms are often nonspecific, Raynaud's phenomenon may be absent, and physical examination may only show upper extremity edema and puffy fingers. Patients with early SSc might be diagnosed as arthritis, SLE, myositis, or, most commonly, undifferentiated connective tissue disease. Within weeks to months, Raynaud's phenomenon and advancing skin induration appear. SSc-specific autoantibodies provide a high degree of diagnostic certainty. Raynaud's phenomenon with fingertip ulcerations or other evidence of digital ischemia, coupled with telangiectasia, distal esophageal dysmotility, unexplained ILD or PAH, or accelerated hypertension with renal failure in the absence of clinically evident skin induration, suggests the diagnosis of SSc *sine* scleroderma.

APPROACH TO THE PATIENT

Management of Systemic Sclerosis

OVERVIEW: GENERAL PRINCIPLES

To date, with the possible exception of hematopoietic stem cell therapy (HSCT), no therapy has been shown to significantly alter the natural history of SSc. In contrast, multiple interventions are highly effective in alleviating the symptoms, slowing the progression of the cumulative organ damage, and reducing disability. A significant reduction in disease-related mortality has been noted during the past 25 years. In light of the marked heterogeneity in disease manifestations, and natural history, the management of SSc mandates a "personalized medicine" approach that is specifically tailored to each individual patient's unique needs.

TABLE 353-7 Key Principles in Management

- Establish early and accurate diagnosis.
- Detect and evaluate internal organ involvement.
- Define clinical disease stage and activity.
- Tailor individualized therapy to each patient's unique needs.
- Assess treatment response, and adjust therapy as needed; monitor for disease activity, progression and new complications.

The following general principles should guide management **(Table 353-7)**: prompt and accurate diagnosis; classification and risk stratification based on clinical and laboratory evaluation, including prognostic and predictive biomarkers; early recognition of organ-based complications and assessment of their extent, severity, and likelihood of deterioration; regular monitoring for disease progression, new complications, and response to therapy; adjusting therapy; and patient education. In order to minimize irreversible organ damage, management should be proactive, with regular screening and initiation of appropriate, intervention at the earliest possible opportunity. In light of the complex and multisystemic nature of the SSc, a team-oriented management approach integrating appropriate specialists should be pursued. Generally, a combination of drugs that impact different aspects of the disease is used. Patients should be encouraged to become familiar with potential complications and understand therapeutic options, including interventional trials, and natural history, and empowered to partner with their treating physicians. This requires a long-term relationship between patient and physician, with ongoing counseling, encouragement, and two-way dialogue.

DISEASE-MODIFYING THERAPY: IMMUNOSUPPRESSIVE AGENTS

Immuno-suppressive agents used in other autoimmune diseases have generally shown modest or no benefit in SSc. Glucocorticoids alleviate stiffness and aching in early inflammatory-stage dcSSc, but do not influence the progression of skin or internal organ involvement. Since their use is associated with an increased risk of scleroderma renal crisis, glucocorticoids should be given only when absolutely necessary, at the lowest dose possible, and for brief periods only.

Cyclophosphamide has been extensively studied in light of its efficacy in the treatment of vasculitis **(Chap. 356)**, SLE **(Chap. 349)**, and other autoimmune diseases **(Chap. 348)**. Both oral and intravenous cyclophosphamide have been shown to reduce the progression of SSc-associated ILD, with stabilization and, rarely, modest improvement of pulmonary function, HRCT findings, respiratory symptoms, and skin induration. The benefits of cyclophosphamide need to be balanced against its potential toxicity, including bone marrow suppression, opportunistic infections, hemorrhagic cystitis and bladder cancer, premature ovarian failure, and late secondary malignancies.

Methotrexate had modest effect on SSc skin involvement in small studies. Mycophenolate mofetil was evaluated in both open label and randomized control trials. Both skin induration and ILD improved in patients treated with MMF, and the drug was well tolerated. Tocilizumab, a monoclonal antibody directed against the IL-6 receptor that blocks IL-6 signaling, also showed benefit in randomized SSc trials. Open-label studies and small trials provide support for the use of rituximab, a monoclonal antibody directed against the mature B cell marker CD20, along with extracorporeal photopheresis and IV immunoglobulin. Randomized trials in SSc evaluating the efficacy of abatacept, a fusion protein that inhibits T cell co-stimulation and function, are on-going. The use of cyclosporine, azathioprine, plaquenil, thalidomide, and rapamycin is currently not well supported by the literature. Intensive immune ablation using high-dose chemotherapy, (myeloablation) alone, or combined with total body irradiation, followed by autologous stem cell reconstitution has been evaluated in patients with severe early-stage SSc. In selected patients this intensive intervention was associated with durable remission and improved long-term survival in multiple small randomized clinical trials. Since this regimen has been associated with significant morbidity and even treatment-related mortality, its use currently should be restricted to SSc patients with severe, or treatment-refractory, disease.

Antifibrotic Therapy Because tissue fibrosis underlies organ damage in SSc, drugs that interfere with the fibrotic process represent a rational therapeutic approach. In older retrospective studies, D-penicillamine was shown to stabilize skin induration, prevent new internal organ involvement, and improve survival. However, a randomized-controlled clinical trial in early active SSc found no difference in the extent of skin involvement between patients treated with standard-dose (750 mg/d) or very low-dose (125 mg every other day) D-penicillamine. Recent clinical trials show benefit of pirfenidone and of nintedanib in patients with idiopathic pulmonary fibrosis, with significant slowing of the loss of lung function. Whether these anti-fibrotic drugs have comparable efficacy and tolerability in patients with SSc-associated ILD and other fibrotic manifestations of the disease is under investigation.

Vascular Therapy The goal of Raynaud's therapy is to control episodes, prevent and enhance the healing of ischemic complications, and slow the progression of obliterative vasculopathy. Patients should dress warmly, minimize cold exposure, and avoid drugs that precipitate or exacerbate vasospastic episodes. Extended-release dihydropyridine calcium channel blockers such as amlodipine and diltiazem ameliorate Raynaud's phenomenon, but their use is often limited by side effects (palpitations, dependent edema, worsening gastroesophageal reflux). While ACE inhibitors do not reduce the frequency or severity of episodes, Angiotensin II receptor blockers such as losartan are effective and well tolerated. Patients with Raynaud's phenomenon unresponsive to these therapies may require the addition of α_1-adrenergic receptor blockers (e.g., prazosin), 5-phosphodiesterase inhibitors (e.g., sildenafil), topical nitroglycerine, and intermittent IV infusions of prostaglandins. Low-dose aspirin and dipyridamole prevent platelet aggregation and may have a role as adjunctive agents. In patients with ischemic digital tip ulcerations, the endothelin-1 receptor antagonist bosentan reduces the risk of new ulcers. Digital sympathectomy and intradigital injections of botulinum type A (Botox) may be considered in patients with severe on-going ischemia. Empirical long-term therapy with statins and antioxidants may retard the progression of vascular damage and obliteration. There is limited evidence-based information for the treatment of cardiac complications of SSc, which should be guided by specialists experienced in their diagnosis and management. While selective beta blockers such as metoprolol can precipitate vasospasm, non-dihydropyridine calcium channel blockers can be used for rate control in atrial arrhythmias, and non-selective alpha/beta blockers such as carvedilol for improving myocardial perfusion and left ventricular systolic function.

TREATMENT

TREATMENT OF SSc-ASSOCIATED ILD

ILD is a leading cause of death in patients with SSc. However, as SSc-associated ILD is not necessarily progressive, it is important to identify patients who are at high risk for disease progression in the absence of treatment. The extent of ILD on HRCT and the FVC at initial evaluation, and decline in PFTs during the preceding 12-month period, are helpful in identifying these patients. Patients at high risk for ILD should be monitored by performing PFTs every 6 months; serial HRCT imaging is not recommended. Cyclophosphamide, given IV or orally for 6 to 12 months, and mycophenolate mofetil slow the decline in lung function and improve respiratory symptoms; however, cyclophosphamide is associated with more frequent side effects. The safety and efficacy of anti-fibrotic drugs recently approved for idiopathic pulmonary fibrosis in the treatment of SSc-associated ILD are currently under investigation. In certain patients who show continued progression of ILD despite

medical therapy, lung transplantation might be considered as a life-prolonging procedure, although significant GERD is a concern in SSc. Recurrence of SSc-ILD in transplanted lung allografts has not been reported.

TREATMENT OF GASTROINTESTINAL COMPLICATIONS

Because oral problems including decreased oral aperture, decreased saliva production, gum recession, periodontal disease, and teeth loss are common, regular dental care is recommended. Gastroesophageal reflux is very common and may occur in the absence of symptoms. Patients should be instructed to elevate the head of the bed, eat frequent small meals, and avoid alcohol, caffeine, and known reflux exacerbants, or meals before bedtime. Proton pump inhibitors reduce acid reflux and in patients with SSc may need to be given in relatively high doses. Prokinetic agents such as metoclopramide, erythromycin (a motilin agonist), and domperidone may occasionally be helpful, but are frequently associated with side effects. Botulinum toxin injection sometimes ameliorates impaired gastric emptying. Anti-reflux procedures such as Nissen fundoplication can result in secondary achalasia and generally should be avoided. Episodic bleeding from GAVE (watermelon stomach) may be amenable to treatment with endoscopic ablation using laser or argon plasma photocoagulation, although bleeding frequently recurs. Some patients may require enteral feeding and/or decompression via percutaneous gastrostomy or jejunostomy. Small bowel bacterial overgrowth secondary to dysmotility causes abdominal bloating and diarrhea, and may lead to malabsorption and severe malnutrition. Treatment with short courses of rotating broad-spectrum antibiotics such as metronidazole, erythromycin, and rifaximin can eradicate bacterial overgrowth. Small bowel hypomotility may respond to octreotide; however, pseudo-obstruction is difficult to treat. Fecal incontinence, a frequent and under-reported complication, may respond to anti-diarrheal medication, biofeedback therapy, sphincter augmentation, and sacral neuromodulation. Potential malnutrition should be routinely assessed.

TREATMENT OF PAH

In SSc, PAH carries an extremely poor prognosis and accounts for 30% of deaths. Because PAH is asymptomatic until advanced, patients with SSc should be screened at initial evaluation, and regularly thereafter. Treatment is generally started with an oral endothelin-1 receptor antagonist such as bosentan or a phosphodiesterase 5 inhibitor such as sildenafil. Recently, the soluble guanylate cyclase stimulator riociguat, which acts by increasing the production of nitric oxide, and the selective IP prostacyclin receptor agonist selexipag, were shown to improve PAH symptoms and survival. Patients may also require diuretics and digoxin. If hypoxemia is documented, supplemental oxygen should be prescribed in order to avoid secondary pulmonary vasoconstriction. Prostacyclin analogues such as epoprostenol or treprostinil can be given by continuous IV or SC infusion, or via intermittent nebulized inhalations. Combination therapy with different classes of agents acting additively or synergistically is often necessary. Lung transplantation remains an option for selected SSc patients with PAH who fail medical therapy, and 2-year survival rates (64%) are comparable to those of idiopathic ILD or PAH.

MANAGEMENT OF RENAL CRISIS

Scleroderma renal crisis is a medical emergency. Since the outcome is largely determined by the extent of renal damage at the time that aggressive therapy is initiated, prompt recognition of impending or early scleroderma renal crisis is essential, and efforts should be made to avoid its occurrence. High-risk SSc patients with early disease, extensive and progressive skin involvement, tendon friction rubs, and anti-RNA polymerase III antibodies should be instructed to monitor their blood pressure daily and report significant alterations immediately. Potentially nephrotoxic drugs should be avoided, and glucocorticoids should be used only when absolutely necessary and at low doses. Patients presenting with scleroderma renal crisis should be immediately hospitalized. Once other causes of renal disease are excluded, treatment should be started promptly with titration of short-acting ACE inhibitors, with the goal of achieving rapid normalization of the blood pressure. In patients with persistent hypertension, addition of angiotensin II receptor blockers, calcium channel blockers, endothelin-1 receptor blockers, prostacyclins, and direct renin inhibitors should be considered. Up to two-thirds of patients with scleroderma renal crisis will require dialysis. Substantial renal recovery can occur, and dialysis can be discontinued in 30–50% of the patients. Kidney transplantation is appropriate for patients unable to discontinue dialysis after 2 years. Survival of transplanted SSc patients is comparable to that of other diseases, and recurrence of renal crisis is rare.

SKIN CARE

Because skin involvement in SSc is never life-threatening and it stabilizes and may even regress spontaneously, disease management should not be dictated by its cutaneous manifestations. The inflammatory symptoms of early skin involvement can be controlled with antihistamines and short-term use of low-dose glucocorticoids (<5 mg/d of prednisone). Cyclophosphamide and methotrexate have modest effects on skin induration. Because the skin is dry, the use of hydrophilic ointments and bath oils is encouraged, and regular skin massage is helpful. Telangiectasia, which presents a cosmetic problem, especially on the face, can be treated with pulsed dye laser. Ischemic digital ulcerations should be protected by occlusive dressing to promote healing and prevent infection. Infected skin ulcers are treated with topical antibiotics and surgical debridement. While no therapy has been shown to be effective in preventing soft tissue calcific deposits or promoting their dissolution, reports support the use of diltiazem, minocycline, bisphosphonates, and topical or IV sodium thiosulfate (STS). Other therapies that have been used for calcinosis include carbon dioxide laser, extracorporeal shock-wave lithotripsy, and surgical high-speed microdrilling.

TREATMENT OF MUSCULOSKELETAL COMPLICATIONS

Arthralgia and joint stiffness are very common and distressing manifestations in early-stage disease. Short courses of nonsteroidal anti-inflammatory agents, methotrexate, and cautious use of low-dose glucocorticoids alleviate symptoms. Physical and occupational therapy can be effective for preventing loss of musculoskeletal function and joint contractures, and should be initiated early.

COURSE

The natural history of SSc is highly variable and difficult to predict, especially in early stages of the disease. Patients with dcSSc tend to have a more rapidly progressive course and worse prognosis than those with lcSSc. Inflammatory symptoms of early dcSSc, such as fatigue, edema, joint pain and pruritus subside, and skin thickening reach a plateau at 2–4 years after disease onset. It is during the early edematous/inflammatory stage that life-threatening visceral organ involvement may develop. While existing visceral organ involvement, such as ILD, may progress even after skin involvement peaks, new organ involvement is rare. Scleroderma renal crisis generally occurs within the first 4 years of disease. In late-stage disease (>6 years), the skin is usually soft and atrophic. Skin regression characteristically occurs in an order that is the reverse of initial involvement, with softening on the trunks followed by proximal and finally distal extremities; however, sclerodactyly and fixed finger contractures generally persist. Relapse or recurrence of skin thickening after peak skin involvement has been reached is uncommon. Patients with lcSSc follow a clinical course that is markedly different than that of dcSSc. Raynaud's phenomenon typically precedes other disease manifestations by years or even decades. Visceral organ complications such as PAH generally develop late and progress slowly.

PROGNOSIS

SSc confers a substantial increase in the risk of premature death. Age- and gender-adjusted mortality rates are fivefold to eightfold higher compared to the general population, and more than half of all patients

with SSc die from their disease. In one population-based study of SSc, the median survival was 11 years. In patients with dcSSc, 5- and 10-year survival rates are 70% and 55%, respectively, whereas in patients with lcSSc, 5- and 10-year survival rates are 90% and 75%, respectively. The prognosis correlates with the extent of skin involvement, which itself is a surrogate for visceral organ involvement. Major causes of death are PAH, pulmonary fibrosis, gastrointestinal involvement, and cardiac disease. Scleroderma renal crisis is associated with a 30% 3-year mortality. Lung cancer and excess cardiovascular deaths also contribute to increased mortality. Markers of poor prognosis include male gender, African-American race, older age at disease onset, extensive skin thickening with truncal involvement, palpable tendon friction rubs, and evidence of significant or progressive visceral organ involvement. Laboratory predictors of increased mortality at initial evaluation include an elevated ESR, anemia, proteinuria, and anti–topoisomerase I antibodies. In one study, SSc patients with extensive skin involvement, vital capacity <55% predicted, significant gastrointestinal involvement (pseudo-obstruction or malabsorption), clinical evidence of cardiac involvement, or scleroderma renal crisis had a 9-year survival of <40%. The severity of PAH predicts mortality, and patients with mean pulmonary arterial pressure ≥45 mmHg had a 33% 3-year survival. The advent of ACE inhibitors in scleroderma renal crisis had a dramatic impact on survival, increasing from <10% at 1 year in the pre–ACE inhibitor era to >70% 3-year survival at the present time. Moreover, 10-year survival in SSc has improved from <60% in the 1970s to >66–78% in the 1990s, a trend that reflects both earlier detection and better management of complications.

LOCALIZED SCLERODERMA

The term *scleroderma* describes a group of localized skin disorders (Table 353-1). These occur more commonly in children than in adults, and in marked contrast to SSc, are generally not complicated by Raynaud's phenomenon or significant internal organ involvement. Morphea presents as solitary or multiple circular patches of thick skin or, rarely, as widespread induration (generalized or pansclerotic morphea); the fingers are generally spared. Linear scleroderma may affect subcutaneous tissues, leading to fibrosis and atrophy of supporting structures, tendons, muscle, and even bone. In children, the growth of affected long bones can be retarded. When linear scleroderma crosses large joints, significant contractures can develop.

MIXED CONNECTIVE TISSUE DISEASE

Patients who have lcSSc coexisting with features of SLE, polymyositis, and rheumatoid arthritis may have mixed connective tissue disease (MCTD). This overlap syndrome is generally associated with the presence of high titers of autoantibodies to U1-RNP. The characteristic initial presentation is Raynaud's phenomenon associated with puffy fingers and myalgia. Over time, sclerodactyly, soft tissue calcinosis, and cutaneous telangiectasia may appear. Skin rash suggestive of SLE (malar erythema, photosensitivity) or dermatomyositis (heliotrope rash on the eyelids, erythematous rash on knuckles) occur. Arthralgia is common, and some patients develop erosive polyarthritis. Pulmonary fibrosis and isolated or secondary PAH may develop. Other manifestations include esophageal dysmotility, pericarditis, Sjögren's syndrome, and renal disease, especially membranous glomerulonephritis. Laboratory evaluation shows elevated ESR and hypergammaglobulinemia. While anti-U1RNP antibodies are detected in high titers, SSc-specific autoantibodies are absent. In contrast to SSc, MCTD often responds to glucocorticoids, and the long-term prognosis is better than that of SSc. Whether MCTD is truly a distinct entity or is a subset of SLE or SSc, remains controversial.

EOSINOPHILIC FASCIITIS (DIFFUSE FASCIITIS WITH EOSINOPHILIA)

Eosinophilic fasciitis is a rare idiopathic disorder of adults associated with abrupt skin induration. The skin characteristically shows a coarse cobblestone "peau d'orange" appearance. In contrast to SSc, Raynaud's phenomenon and SSc-associated internal organ involvement and autoantibodies are absent. Furthermore, skin involvement spares the fingers. Full-thickness biopsy of the lesional skin reveals fibrosis of the subcutaneous fascia, with variable inflammation and eosinophil infiltration. In the acute phase of the illness, peripheral blood eosinophilia may be prominent. MRI appears to be a sensitive tool for the diagnosis of eosinophilic fasciitis. Eosinophilic fasciitis can occur in association with, or preceding, various myelodysplastic syndromes or multiple myeloma. Although glucocorticoids cause prompt resolution of eosinophilia, the skin shows slow and variable improvement. The prognosis of patients with eosinophilic fasciitis who do not develop hematological complications is generally good.

FURTHER READING

Allanore Y et al: Systemic sclerosis. Nat Rev Dis Primers 1:15002, 2015.

Herzog EL et al: Interstitial lung disease associated with systemic sclerosis and idiopathic pulmonary fibrosis: How similar or distinct? Arthritis Rheum 66:1967, 2014.

Joseph CG et al: Association of the autoimmune disease scleroderma with an immunologic response to cancer. Science 343:152, 2014.

Martyanov V, Whitfield ML: Molecular stratification and precision medicine in systemic sclerosis from genomic and proteomic data. Curr Opin Rheumatol 28:83, 2016.

Tashkin DP et al: Mycophenolate mofetil versus oral cyclophosphamide in scleroderma-related interstitial lung disease (SLS II): A randomised controlled, double-blind, parallel group trial. Lancet Respir Med 4:708, 2016.

354 Sjögren's Syndrome

Haralampos M. Moutsopoulos

DEFINITION, INCIDENCE, AND PREVALENCE

Sjögren's syndrome is a chronic, slowly progressing autoimmune disease characterized by lymphocytic infiltration of the exocrine glands resulting in xerostomia and dry eyes (keratoconjunctivitis sicca). The syndrome has unique features since it presents with a wide clinical spectrum from organ-specific autoimmune exocrinopathy to systemic disease. A small but significant number of patients develop malignant lymphoma. The disease can present as an entity alone or in association with other autoimmune diseases **(Table 354-1)**. Finally, the histopathologic lesion in the labial minor salivary glands is easily accessible aiding the diagnosis, prognosis and disease pathogenesis.

Middle-aged women (female-to-male ratio, 9:1) are primarily affected, although Sjögren's syndrome may occur at any age, including childhood. The prevalence of primary Sjögren's syndrome is ~0.5–1%, while 5–20% of patients with other autoimmune diseases suffer from Sjögren's syndrome (secondary).

PATHOGENESIS

Sjögren's syndrome is characterized by both lymphocytic infiltration of the exocrine glands and B lymphocyte hyperreactivity. An oligomonoclonal B cell process, which is characterized by cryoprecipitable

TABLE 354-1 Association of Sjögren's Syndrome with Other Autoimmune Diseases

Rheumatoid arthritis
Systemic lupus erythematosus
Scleroderma
Mixed connective tissue disease
Primary biliary cirrhosis
Autoimmune thyroid disease
Chronic active hepatitis

monoclonal immunoglobulins (IgMκ or IgAκ) with rheumatoid factor activity, is evident in up to 10% of patients.

Sera from patients with Sjögren's syndrome often contain autoantibodies to non-organ-specific antigens such as immunoglobulins (rheumatoid factors) and extractable nuclear and cytoplasmic antigens (Ro/SS-A, La/SS-B). Ro/SS-A autoantigen consists of two polypeptides (52 and 60 kDa, respectively) in conjunction with cytoplasmic RNAs, whereas the 48-kDa La/SS-B protein is bound to RNA III polymerase transcripts. Autoantibodies to Ro/SS-A and La/SS-B antigens are usually present prior to diagnosis and are associated with earlier disease onset, longer disease duration, salivary gland enlargement, extraglandular (systemic) manifestations, and more intense lymphocytic infiltration of minor salivary glands.

The major infiltrating cells in the affected exocrine glands are activated T lymphocytes in mild lesions, whereas B cells prevail in severe lesions. Macrophages and dendritic cells are also found. The number of macrophages positive for interleukin (IL) 18 has been shown to be associated with parotid gland enlargement and low serum levels of the C4 component of complement, both of which are adverse predictors for lymphoma development.

Ductal and acinar epithelial cells appear to play a significant role in the initiation and perpetuation of autoimmune injury. These cells (1) express costimulatory molecules, and inappropriately the intracellular autoantigens Ro/SS-A and La/SS-B on their membranes, acquiring the capacity to provide signals essential for lymphocytic activation; (2) produce proinflammatory cytokines and lymphocyte attracting chemokines necessary for sustaining the autoimmune lesion and allowing the formation of ectopic germinal centers, a finding predicting lymphoma development; and (3) express functional receptors of innate immunity, particularly Toll-like receptors (TLRs) 3, 7, and 9, molecules which may account for the initiation of the autoimmune reactivity.

Both infiltrating T and B cells have a tendency to be resistant to apoptosis. Levels of B cell–activating factor (BAFF) have been found to be elevated in the serum and tissues of Sjögren's syndrome patients, especially those with hypergammaglobulinemia, and probably accounts for the anti-apoptotic effect on B lymphocytes. Glandular epithelial cells seem to have an active role in the production of BAFF, which may be expressed and secreted after stimulation with type I and II interferons. The latter have been detected in ductal epithelial cells and T cells. The triggering factor for epithelial activation appears to be enteroviral infection.

Molecular analysis of human leukocyte antigen (HLA) class II genes has revealed that Sjögren's syndrome, regardless of the patient's ethnic origin, is highly associated with the HLA DQA1*0501 allele. Genome-wide association studies have disclosed an increased prevalence of single-nucleotide polymorphisms in genes of IRF-5 and STAT-4, which participate in the activation of the type I interferon pathway.

CLINICAL MANIFESTATIONS

The majority of patients with Sjögren's syndrome have symptoms related to impaired lacrimal and salivary gland function. The disease evolution is slow and in the majority of patients runs a benign course. Studies have shown that prior to disease onset, patients with Sjögren's syndrome experience major stressful life events with which they cannot cope adequately.

The principal oral symptom of Sjögren's syndrome is dryness (xerostomia). Patients report difficulty in swallowing dry food, a burning mouth sensation, an increase in dental caries, and problems in wearing complete dentures. Physical examination shows a dry, erythematous, sticky oral mucosa. There is atrophy of the filiform papillae on the dorsum of the tongue, and saliva from the major glands is either not expressible or cloudy. Enlargement of the parotid or other major salivary glands occurs in two-thirds of patients with primary Sjögren's syndrome but is uncommon in those in association with rheumatoid arthritis. Diagnostic tests include sialometry and newer imaging techniques, including ultrasound, MRI, and magnetic resonance sialography of the major salivary glands. Biopsy of the labial minor salivary gland permits histopathologic confirmation of focal lymphocytic infiltrates.

Ocular involvement is the other major manifestation of Sjögren's syndrome. Patients usually describe a sandy or gritty feeling under the eyelids. Other ocular symptoms include burning, accumulation of secretions in thick strands at the inner canthi, decreased tearing, redness, itching, eye fatigue, and increased photosensitivity. These symptoms, which define *keratoconjunctivitis sicca*, are attributed to the destruction of corneal and bulbar conjunctival epithelium. Diagnostic evaluation of keratoconjunctivitis sicca includes measurement of tear flow by Schirmer's I test and determination of tear composition, with assessment of tear breakup time or tear lysozyme content. Slit-lamp examination of the cornea and conjunctiva after lissamine green or Rose Bengal staining reveals punctuate corneal ulcerations and attached filaments of corneal epithelium.

Involvement of other exocrine glands, which occurs less frequently, includes a decrease in mucous gland secretions of the upper and lower respiratory tree, resulting in dry nose, throat, and trachea (xerotrachea). In addition, diminished secretion of the exocrine glands of the gastrointestinal tract leads to esophageal mucosal atrophy and atrophic gastritis. Dyspareunia due to dryness of the external genitalia and dry skin also may occur.

Extraglandular (systemic) manifestations are seen in one-third of patients with Sjögren's syndrome **(Table 354-2)** but are very rare in patients whose Sjögren's syndrome is associated with rheumatoid arthritis. They can be categorized as follows: Non-specific, involvement of parenchymal organs by lymphocytes (peri-epithelial), immune complex-mediated pathology, and lymphoma development. In the first category easy fatigability, low-grade fever, Raynaud's phenomenon, myalgias, arthralgias, and arthritis are included. Arthritis in patients with primary Sjögren's syndrome is non-erosive. Involvement of parenchymal organs such as the lungs, kidneys and the liver is due to peri-epithelial accumulation of lymphocytes. On the basis of this observation the term **autoimmune epithelitis** has been coined. Lung involvement is usually manifested with dry cough and rarely with dyspnea. The underlying lung pathology includes peribronchial infiltrates and rarely lymphocyte interstitial pneumonitis. Renal involvement includes interstitial nephritis, clinically manifested by hyposthenuria and renal tubular dysfunction with or without acidosis. Untreated acidosis may lead to nephrocalcinosis. Immune complex-mediated

TABLE 354-2 Prevalence of Extraglandular Manifestations in Primary Sjögren's Syndrome

CLINICAL MANIFESTATION	PERCENT	REMARKS
Non-Specific		
Fatigability/Myalgias	25	Fibromyalgia
Arthralgias/Arthritis	60	Usually non-erosive, leading to Jaccoud's arthropathy
Raynaud's phenomenon	37	In one-third of patients, precedes sicca manifestations
Peri-Epithelial		
Lung involvement	14	Small airway disease/lymphocyte interstitial pneumonitis
Kidney involvement	9	Interstitial kidney disease is usually asymptomatic
Liver involvement	6	Primary biliary cirrhosis stage I
Immune-Complex mediated		
Small vessel vasculitis		Purpura, urticarial lesions
Peripheral neuropathy	2	Polyneuropathy, either sensory or sensorimotor
Glomerulonephritis		Membranoproliferative
Lymphoma		
Lymphoma	6	Glandular MALT[a] lymphoma is most common

[a]Mucosa-associated lymphoid tissue.

disease is expressed with vasculitis affecting primarily small-sized vessels, mainly manifested with purpura and rarely with urticarial rash, skin ulcerations, mononeuritis multiplex, and membranoproliferative glomerulonephritis associated with mixed cryoglobulinemia. Central nervous system involvement is rarely recognized. A few cases of myelitis associated with antibody to aquaporin 4 have been described.

Patients with Sjögren's syndrome associated with rheumatoid arthritis and systemic lupus erythematosus have an increased cardiovascular risk.

Lymphoma in Sjögren's syndrome usually presents later in the disease course. Persistent parotid gland enlargement, purpura, leukopenia, cryoglobulinemia, low serum C4 complement levels, autoantibodies (anti-Ro/SS-A, anti-La/SS-B), and ectopic germinal center formation in minor salivary glands are manifestations predicting the development of lymphoma. Most lymphomas are extranodal, low-grade marginal-zone B cell lymphomas and are usually detected incidentally during evaluation of the labial minor salivary gland biopsy. The affected lymph nodes are usually peripheral. Survival rates are decreased in patients with B symptoms, lymph node mass >7 cm in diameter, and high or intermediate histologic grade.

Routine laboratory tests in Sjögren's syndrome reveal mild normochromic, normocytic anemia. An elevated erythrocyte sedimentation rate is found in ~70% of patients. Certain autoantibodies may determine different disease phenotypes. Patients positive for anticentromere autoantibody present with a clinical picture similar to that of limited scleroderma **(Chap. 353)**. Antimitochondrial antibodies may connote liver involvement in the form of primary biliary cirrhosis **(Chap. 339)**. Autoantibodies to 21-hydroxylase are found in almost 20% of patients in association with a blunted adrenal response.

DIAGNOSIS AND DIFFERENTIAL DIAGNOSIS

Primary Sjögren's syndrome is diagnosed if (1) the patient presents with eye and/or mouth dryness, (2) eye tests disclose keratoconjunctivitis sicca, (3) mouth evaluation reveals dry oral mucosa, and/or (4) the patient's serum reacts with immunoglobulins (rheumatoid factors), Ro/SS-A, and/or La/SS-B autoantigens. Labial biopsy is needed for diagnostic and prognostic purposes as well as to rule out other conditions that may cause dry mouth or eyes or parotid gland enlargement **(Tables 354-3 and 354-4)**. Enlargement of major salivary glands, particularly in patients without autoantibodies, should raise the suspicion of IgG4-related syndrome. Validated methods of disease activity and classification criteria have been established **(Table 354-5)**.

TABLE 354-3 Differential Diagnosis of Sicca Symptoms

XEROSTOMIA	DRY EYE	BILATERAL PAROTID GLAND ENLARGEMENT
Viral infections (HCV, HIV)	Inflammation	Viral infections
Drugs	Stevens-Johnson syndrome	Mumps
Psychotherapeutic	Pemphigoid	Influenza
Parasympatholytic	Chronic conjunctivitis	Epstein-Barr virus
Antihypertensive	Chronic blepharitis	Coxsackievirus A
Psychogenic origin	Sjögren's syndrome	Cytomegalovirus
Irradiation	Toxicity	HIV, HCV
Diabetes mellitus	Burns	Sarcoidosis, Tuberculosis
Trauma	Drugs	IgG4 syndrome
Sjögren's syndrome	Neurologic conditions	Sjögren's syndrome
Amyloidosis	Impaired lacrimal gland function	Metabolic disorders
	Impaired eyelid function	Diabetes mellitus
	Miscellaneous	Hyperlipoproteinemias (types IV and V)
	Trauma	Chronic pancreatitis
	Hypovitaminosis A	Hepatic cirrhosis
	Blink abnormality	Endocrine
	Anesthetic cornea	Acromegaly
	Lid scarring	Gonadal hypofunction
	Epithelial irregularity	

TABLE 354-4 Differential Diagnosis of Sjögren's Syndrome

HIV INFECTION AND SICCA SYNDROME	SJÖGREN'S SYNDROME	SARCOIDOSIS
Predominant in young males	Predominant in middle-aged women	No age or sex preference
Lack of autoantibodies to Ro/SS-A and/or La/SS-B	Presence of autoantibodies	Lack of autoantibodies to Ro/SS-A and/or La/SS-B
Lymphoid infiltrates of salivary glands by CD8+ T lymphocytes	Lymphoid infiltrates of salivary glands by CD4+ T lymphocytes	Granulomas in salivary glands
Association with HLA-DR5	Association with HLA-DR3 and DRw52	Unknown
Positive serologic tests for HIV	Negative serologic tests for HIV	Negative serologic tests for HIV

TREATMENT

Sjögren's Syndrome

Treatment of Sjögren's syndrome aims to relieve symptoms and limit the damage from chronic xerostomia and keratoconjunctivitis sicca through substitution or stimulation of impaired secretions **(Fig. 354-1)**.

To replace deficient tears, several ophthalmic preparations are readily available (hydroxypropyl methylcellulose; polyvinyl alcohol; 0.5% methylcellulose; Hypo Tears). If corneal ulcerations are present, eye patching and boric acid ointments are recommended. Certain drugs that may decrease lacrimal and salivary secretions,

TABLE 354-5 Revised International Classification Criteria for Sjögren's Syndrome[a,b,c]

I. Ocular symptoms: a positive response to at least one of three validated questions.
   1. Have you had daily, persistent, troublesome dry eyes for >3 months?
   2. Do you have a recurrent sensation of sand or gravel in the eyes?
   3. Do you use tear substitutes more than three times a day?

II. Oral symptoms: a positive response to at least one of three validated questions.
   1. Have you had a daily feeling of dry mouth for >3 months?
   2. Have you had recurrent or persistently swollen salivary glands as an adult?
   3. Do you frequently drink liquids to aid in swallowing dry foods?

III. Ocular signs: objective evidence of ocular involvement defined as a positive result to at least one of the following two tests:
   1. Shirmer's I test, performed without anesthesia (≤5 mm in 5 min)
   2. Rose Bengal score or other ocular dye score (≥4 according to van Bijsterveld's scoring system)

IV. Histopathology: In minor salivary glands focal lymphocytic sialoadenitis, with a focus score ≥1.

V. Salivary gland involvement: objective evidence of salivary gland involvement defined by a positive result to at least one of the following diagnostic tests:
   1. Unstimulated whole salivary flow (≤1.5 mL in 15 min)
   2. Parotid sialography
   3. Salivary scintigraphy

VI. Antibodies in the serum to Ro/SS-A or La/SS-B antigens, or both.

[a]Exclusion criteria: past head and neck radiation treatment, hepatitis C infection, AIDS, preexisting lymphoma, sarcoidosis, graft-versus-host disease, use of anticholinergic drugs. [b]Primary Sjögren's syndrome: any four of the six items, as long as item IV (histopathology) or VI (serology) is positive; or any three of the four objective-criteria items (III, IV, V, VI). [c]In patients with a potentially associated disease (e.g., another well-defined connective tissue disease), the presence of item I or item II plus any two from among items III, IV, and V may be considered indicative of secondary Sjögren's syndrome.

Source: From C Vitali et al: Ann Rheum Dis 61:554, 2002. ©2002 with permission from BMJ Publishing Group Ltd.

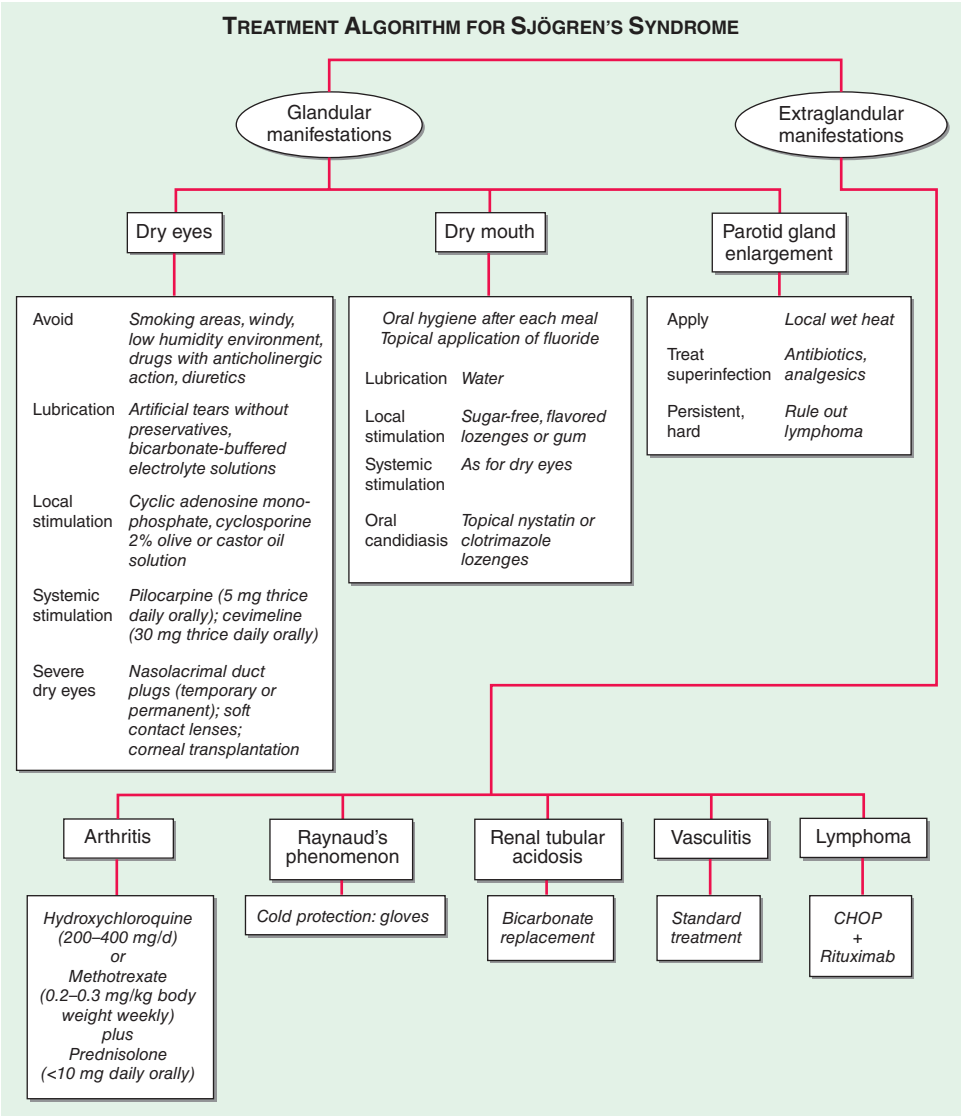


FIGURE 354-1 Treatment algorithm for Sjögren's syndrome. CHOP, cyclophosphamide, adriamycin (hydroxydaunorubicin), vincristine (oncovin), and prednisone.

such as diuretics, antihypertensive drugs, anticholinergics, and antidepressants, should be avoided.

For xerostomia, the best replacement is water. Propionic acid gels may be used to treat vaginal dryness. To stimulate secretions, orally administered pilocarpine (5 mg thrice daily) or cevimeline (30 mg thrice daily) appears to improve sicca manifestations, and both are well tolerated. Hydroxychloroquine (200 mg daily) is helpful for arthralgias and mild arthritis.

Patients with renal tubular acidosis should receive sodium bicarbonate by mouth (0.5–2 mmol/kg in four divided doses). Glucocorticoids and monoclonal antibody to CD20 (Rituximab) appear to be effective in patients with systemic disease, particularly in those with purpura, arthritis, and fatigability. Combination of anti-CD-20 with a classic CHOP regimen (cyclosporine, adriamycin [hydroxydaunorubicin], vincristine [oncovin], and prednisone) leads to increased survival rates among patients with high-grade lymphomas.

ACKNOWLEDGMENT

I would like to thank Dr. Athanasios G. Tzioufas for his contribution in the previous edition of the chapter.

FURTHER READING

MAVRAGANI CP, FRAGOULIS GE, MOUTSOPOULOS HM: Endocrine alterations in primary Sjogren's syndrome: An overview. J Autoimmun 39:354, 2012.

MAVRAGANI CP, MOUTSOPOULOS HM: Sjögren's syndrome. CMAJ 186:579, 2014.

MAVRAGANI CP, NEZOS A, MOUTSOPOULOS HM: New advances in the classification, pathogenesis and treatment of Sjogren's syndrome. Curr Opin Rheumatol 25:623, 2013.

MOUTSOPOULOS HM: Sjögren's syndrome: A forty-year scientific journey. J Autoimmun 51:1, 2014.

355 The Spondyloarthritides

Joel D. Taurog

The spondyloarthritides are a group of overlapping disorders that share certain clinical features, genetic associations, and pathogenic mechanisms. The classic designations include ankylosing spondylitis (AS), reactive arthritis (ReA), psoriatic arthritis (PsA) and spondylitis, enteropathic arthritis and spondylitis, juvenile-onset spondyloarthritis (JSpA), and undifferentiated SpA. More recently, these disorders have been broadly classified as predominantly axial SpA, affecting the spine, pelvis, and thoracic cage, or predominantly peripheral SpA, affecting the extremities.

ANKYLOSING SPONDYLITIS AND AXIAL SPONDYLOARTHRITIS

AS is an inflammatory disorder of unknown cause that primarily affects the axial skeleton; peripheral joints and extraarticular structures are also frequently involved. The disease usually begins in the second or third decade. The term *axial spondyloarthritis* (ax-SpA) is now in common use, supported by criteria formulated in 2009 **(Table 355-1)**. This classification includes definite AS, early stages that will progress to meet classical criteria for AS, and one or more nonprogressing phenotypes. The estimated prevalence of ax-SpA in the US adult population is from 0.9 to 1.4%, similar to that of rheumatoid arthritis (RA).

EPIDEMIOLOGY

AS shows a striking correlation with the histocompatibility antigen HLA-B27 and occurs worldwide roughly in proportion to the prevalence of B27 **(Chap. 343)**. In North American whites, the prevalence of B27 is 7%, whereas it is 75–90% in patients with AS.

In population surveys, AS is present in 1–6% of adults inheriting B27, whereas the prevalence is 10–30% among B27+ adult first-degree relatives of AS probands. Concordance rate in identical twins is about 65%. Susceptibility to AS is determined largely by genetic factors, with B27 estimated to comprise about 20% of the genetic component. Genome-wide single-nucleotide polymorphism (SNP) analysis has identified over 100 additional non-HLA susceptibility alleles.

Patients with ax-SpA that do not have radiologic criteria for AS (see below) are said to have *non-radiographic axial SpA* (nr-ax-SpA). The prevalence of HLA-B27 in these patients is similar to that in AS; however, the proportion of females is much higher (>50% vs ~30%). No information is yet available about other susceptibility loci in nr-ax-SpA.

PATHOLOGY

Sacroiliitis is often an early manifestation of AS and nr-ax-SpA. Knowledge of the pathology comes from both biopsy and autopsy studies that cover a range of disease durations, all from patients with AS. Synovitis and myxoid marrow represent the earliest changes, followed by pannus and subchondral granulation tissue. Marrow edema, enthesitis, and chondroid differentiation are also found. Macrophages, T cells, plasma cells, and osteoclasts are prevalent. If the process continues to progress, eventually the eroded joint margins are gradually replaced by fibrocartilage regeneration and then by ossification.

In the spine, surgically resected or autopsy specimens show inflammatory granulation tissue in the paravertebral connective tissue at the junction of annulus fibrosus and vertebral bone, and in some cases along the entire outer annulus. The outer annular fibers are eroded and eventually replaced by bone, forming the beginning of a syndesmophyte, which then grows by continued endochondral ossification, ultimately bridging the adjacent vertebral bodies. Ascending progression of this process can lead to the "bamboo spine." Other lesions in the spine include diffuse osteoporosis (loss of trabecular bone despite accretion of periosteal bone), erosion of vertebral bodies at the disk margin, and inflammation and destruction of the disk-bone border. Inflammatory arthritis of the apophyseal (facet) joints is common, with synovitis, inflammation at the bony attachment of the joint capsule, and subchondral bone marrow granulation tissue. Erosion of joint cartilage by pannus is often followed by bony ankylosis. This may precede formation of syndesmophytes bridging the adjacent disks. Bone mineral density is diminished in the spine and proximal femur early in the disease course.

Peripheral synovitis in AS shows marked vascularity, evident as tortuous macrovascularity seen during arthroscopy. Lining layer hyperplasia, lymphoid infiltration, and pannus formation are also found. Central cartilaginous erosions caused by proliferation of subchondral granulation tissue are common. The characteristics of peripheral arthritis in AS and other forms of SpA are similar, and distinct from those of RA.

Inflammation in the fibrocartilaginous *enthesis*, the region where a tendon, ligament, or joint capsule attaches to bone, is a characteristic lesion in AS and other SpAs, both at axial and peripheral sites. Enthesitis is associated with prominent edema of the adjacent bone marrow and is often characterized by erosive lesions that eventually undergo ossification.

Subclinical intestinal inflammation has been found in the colon or distal ileum in a majority of patients with SpA. The histology is described below under "Enteropathic Arthritis."

TABLE 355-1 ASAS Criteria for Classification of Axial Spondyloarthritis (to be applied for patients with back pain ≥3 months and age of onset <45 years)[a]

SACROILIITIS ON IMAGING PLUS ≥1 SpA FEATURE	OR	HLA-B27 PLUS ≥2 OTHER SpA FEATURES
Sacroiliitis on imaging • Active (acute) inflammation on MRI highly suggestive of SpA-associated sacroiliitis[b] and/or • Definite radiographic sacroiliitis according to modified New York criteria[c]		SpA features • Inflammatory back pain[d] • Arthritis[e] • Enthesitis (heel)[f] • Anterior uveitis[g] • Dactylitis[e] • Psoriasis[e] • Crohn's disease or ulcerative colitis[e] • Good response to NSAIDs[h] • Family history of SpA[i] • HLA-B27 • Elevated CRP[j]

[a]Sensitivity 83%, specificity 84%. The imaging arm (sacroiliitis) alone has a sensitivity of 66% and a specificity of 97%. [b]Bone marrow edema and/or osteitis on short tau inversion recovery (STIR) or gadolinium-enhanced T1 image. [c]Bilateral grade ≥2 or unilateral grade 3 or 4. [d]See text for criteria. [e]Past or present, diagnosed by a physician. [f]Past or present pain or tenderness on examination at calcaneus insertion of Achilles tendon or plantar fascia. [g]Past or present, confirmed by an ophthalmologist. [h]Substantial relief of back pain at 24–48 h after a full dose of NSAID. [i]First- or second-degree relatives with ankylosing spondylitis (AS), psoriasis, uveitis, reactive arthritis (ReA), or inflammatory bowel disease (IBD). [j]After exclusion of other causes of elevated CRP.

Abbreviations: ASAS, Assessment of Spondyloarthritis International Society; CRP, C-reactive protein; MRI, magnetic resonance imaging; NSAIDs, nonsteroidal anti-inflammatory drugs; SpA, spondyloarthritis.

Source: From M Rudwaleit et al: Ann Rheum Dis 68:777, 2009. Copyright 2009, with permission from BMJ Publishing Group Ltd.

PATHOGENESIS

The pathogenesis of AS is immune-mediated, but there is little direct evidence for antigen-specific autoimmunity, and there is increasing evidence to suggest more of an autoinflammatory pathogenesis. Uncertainty remains regarding the primary site of disease initiation. The dramatic response of the disease to therapeutic blockade of tumor necrosis factor α (TNF-α) or IL-17A indicates that these cytokines play a central role in the immunopathogenesis of AS. Other genes related to TNF pathways show association with AS, including *TNFRSF1A*, *LTBR*, and *TBKBP1*. At least five genes in IL-23/IL-17 pathway show association with AS, including *IL23R*, *PTER4*, *IL12B*, *CARD9*, and *TYK2*. All of these genes are also associated with inflammatory bowel disease (IBD), and three of them are associated with psoriasis. Serum levels of IL-23 and IL-17 are elevated in AS patients. In mice, a population of thymus-dependent CD3+CD4–CD8–T cells expressing γ/δ T cell receptors and IL-23 receptors has been found to reside at entheses, in the aortic root, and near the ciliary body in the eye. These cells express abundant IL-17 and IL-22 upon exposure to systemic IL-23. This finding suggests that site-specific innate immune cells play a critical role in the anatomic specificity of these lesions. In other murine studies, mechanical strain was shown to induce inflammation and new bone formation at enthesial sites.

Mast cells and, to a lesser extent, neutrophils appear to be the major IL-17-producing cells in peripheral arthritis, whereas neutrophils producing IL-17 are prominent in apophyseal joints. High levels of circulating γδ T cells expressing IL-23 receptors and producing IL-17 have been found in AS patients.

Other associated genes encode other cytokines or cytokine receptors (*IL6R*, *IL1R1*, *IL1R2*, *IL7R*, *IL27*), transcription factors involved in the differentiation of immune cells (*RUNX3*, *EOMES*, *BACH2*, *NKX2-3*, *TBX21*), or other molecules involved in activation or regulation of immune or inflammatory responses (*FCGR2A*, *ZMIZ1*, *NOS2*, *ICOSLG*).

The inflamed sacroiliac joint is infiltrated with CD4+ and CD8+ T cells and macrophages and shows high levels of TNF-α, particularly early in the disease. Abundant transforming growth factor β (TGF-β) is found in more advanced lesions. Peripheral synovitis in AS and the other spondyloarthritides is characterized by neutrophils, macrophages expressing CD68 and CD163, CD4+ and CD8+ T cells, and B cells. There is prominent staining for intercellular adhesion molecule 1 (ICAM-1), vascular cell adhesion molecule 1 (VCAM-1), matrix metalloproteinase 3 (MMP-3), and myeloid-related proteins 8 and 14 (MRP-8 and MRP-14). Unlike RA synovium, citrullinated proteins and cartilage gp39 peptide–major histocompatibility complexes (MHCs) are absent.

No specific event or exogenous agent that triggers the onset of disease has been identified, although overlapping features with ReA and IBD and the involvement of the IL-23/IL-17 pathway suggest that enteric bacteria may play a role, and microdamage from mechanical stress at enthesial sites has also been implicated.

It is firmly established that HLA-B27 plays a direct role in AS pathogenesis, but its precise molecular role remains unresolved. Rats transgenic for HLA-B27 develop arthritis and spondylitis, and this is unaffected by the absence of CD8. It thus appears that classical peptide antigen presentation to CD8+ T cells may not be the primary disease mechanism. However, the association of AS with ERAP1, which strongly influences the MHC class I peptide repertoire, suggests that peptide binding to B27 is nonetheless important. The B27 heavy chain has an unusual tendency to misfold, a process that can be proinflammatory. Genetic and functional studies in humans have suggested a role for natural killer (NK) cells in AS, possibly through interaction with B27 heavy chain homodimers. SpA-prone B27 rats show defective dendritic cell function and share with AS patients a characteristic "reverse interferon" gene expression signature in antigen-presenting cells.

New bone formation in AS appears to be largely based on enchondral bone formation and occurs only in the periosteal compartment. It correlates with lack of regulation of the Wnt signaling pathway, which controls the differentiation of mesenchymal cells into osteophytes, by the inhibitors DKK-1 and sclerostin. Indirect evidence and data from animal models also implicate bone morphogenic proteins, hedgehog proteins, and prostaglandin E_2. There is controversy as to whether vertebral new bone formation in AS is a sequela of inflammation or whether it arises independently of inflammation. The second hypothesis is based on the observation that syndesmophyte formation is not suppressed by anti-TNF-α therapy that potently suppresses inflammation. TNF-α is also a known inducer of DKK-1, which inhibits bone formation. Magnetic resonance imaging (MRI) studies suggest that vertebral inflammatory lesions that undergo metaplasia to fat (increased T1-weighted signal) are a preferential site of subsequent syndesmophyte formation despite anti-TNF-α therapy, whereas early acute inflammatory lesions resolve. The rate of syndesmophyte formation appears to decrease after >4 years of anti-TNF-α therapy.

CLINICAL MANIFESTATIONS

In patients eventually diagnosed with AS, symptoms are usually first noticed in late adolescence or early adulthood, at a median age in the mid-twenties. In 5% of patients, symptoms begin after age 40. The initial symptom is usually dull pain, insidious in onset, felt deep in the lower lumbar or gluteal region, accompanied by low-back morning stiffness of up to a few hours' duration that improves with activity and returns following inactivity. Within a few months, the pain has usually become persistent and bilateral. Nocturnal exacerbation of pain often forces the patient to rise and move around.

In some patients, bony tenderness (presumably reflecting enthesitis or osteitis) may accompany back pain or stiffness, whereas in others it may be the predominant complaint. Common sites include the costosternal junctions, spinous processes, iliac crests, greater trochanters, ischial tuberosities, tibial tubercles, and heels. Hip and shoulder ("root" joint) arthritis is considered part of axial disease. Hip arthritis occurs in 25–35% of patients. Shoulder involvement may be at least as common, but is usually less symptomatic. Severe isolated hip arthritis or bony chest pain may be the presenting complaint, and symptomatic hip disease can dominate the clinical picture. Arthritis of peripheral joints other than the hips and shoulders, usually asymmetric, may occur at any point in the disease course. Neck pain and stiffness from involvement of the cervical spine are usually relatively late manifestations, but are occasionally dominant symptoms.

In juvenile onset spondyloarthritis, peripheral arthritis and enthesitis predominate, with axial symptoms supervening in late adolescence.

Initially, axial physical findings mirror the inflammatory process. The most specific findings involve loss of spinal mobility, with limitation of anterior and lateral flexion and extension of the lumbar spine and of chest expansion. Limitation of motion is usually out of proportion to the degree of bony ankylosis and is thought to possibly reflect muscle spasm secondary to pain and inflammation. Pain in the sacroiliac joints may be elicited either with direct pressure or with stress on the joints. In addition, there is commonly tenderness upon palpation of the posterior spinous processes and other sites of symptomatic bony tenderness.

The modified Schober test is a useful measure of lumbar spine flexion. The patient stands erect, with heels together, and marks are made on the spine at the lumbosacral junction (identified by a horizontal line between the posterosuperior iliac spines) and 10 cm above. The patient then bends forward maximally with knees fully extended, and the distance between the two marks is measured. This distance increases by ≥5 cm in the case of normal mobility and by <4 cm in the case of decreased mobility. Chest expansion is measured as the difference between maximal inspiration and maximal forced expiration in the fourth intercostal space in males or just below the breasts in females, with the patient's hands resting on or just behind the head. Normal chest expansion is ≥5 cm. Lateral bending measures the distance the patient's middle finger travels down the leg with maximal lateral bending. Normal is >10 cm.

Limitation or pain with motion of the hips or shoulders is usually present if these joints are involved. It should be emphasized that in early mild, or atypical cases, the symptoms and/or physical findings may be subtle and/or nonspecific.

The course of ax-SpA is extremely variable, ranging from the individual with mild stiffness and normal radiographs to the patient with a totally fused spine and severe bilateral hip arthritis, accompanied by severe peripheral arthritis and extraarticular manifestations. Most of the available data on natural history are from observations of patients with AS, although the prevalence of peripheral arthritis, enthesitis, psoriasis, and IBD appears to be similar in nr-ax-SpA and AS. Pain tends to be persistent early in the disease and intermittent later, with alternating exacerbations and quiescent periods. In a typical severe untreated case with progression of the spondylitis to syndesmophyte formation, the patient's posture undergoes characteristic changes, with obliterated lumbar lordosis, buttock atrophy, and accentuated thoracic kyphosis. There may be a forward stoop of the neck or flexion contractures at the hips, compensated by flexion at the knees. Disease progression can be estimated clinically from loss of height, limitation of chest expansion and spinal flexion, and occiput-to-wall distance. Occasional individuals are encountered with advanced deformities who report having never had significant symptoms.

The factors most predictive of radiographic progression (see below) are the presence of existing syndesmophytes, high inflammatory markers, and smoking. In some but not all studies, onset of AS in adolescence and early hip involvement correlate with a worse prognosis. In women, AS tends to progress less frequently to total spinal ankylosis, although there may be an increased prevalence of isolated cervical ankylosis and peripheral arthritis. Peripheral arthritis (distal to hips and shoulders) occurs in up to 30% of patients in the United States, Canada, and Western Europe, usually as a late manifestation. In Eastern Europe, Latin America, and Asia, the prevalence is over half, with onset more commonly early in the disease course. Pregnancy has no consistent effect on AS, with symptoms improving, remaining the same, or deteriorating in one-third of pregnant patients, respectively.

The most serious complication of the spinal disease is spinal fracture, which can occur with even minor trauma to the rigid, osteoporotic spine. The lower cervical spine is most commonly involved. These fractures are often displaced, causing spinal cord injury. A recent

survey suggested a >10% lifetime risk of fracture. Occasionally, fracture through a diskovertebral junction and adjacent neural arch, termed *pseudoarthrosis*, most common in the thoracolumbar spine, can be an unrecognized source of persistent localized pain and/or neurologic dysfunction. Wedging of thoracic vertebrae is common and correlates with accentuated kyphosis.

The most common extraarticular manifestation is acute anterior uveitis, which occurs in up to 40% of patients and can antedate the spondylitis. Attacks are typically unilateral, causing pain, photophobia, and increased lacrimation. These tend to recur, often in the opposite eye. Cataracts and secondary glaucoma may ensue. Up to 60% of patients with AS have inflammation in the colon or ileum. This is usually asymptomatic, but frank IBD occurs in 5–10% of patients with AS (see "Enteropathic Arthritis," below). About 10% of patients meeting criteria for AS have psoriasis (see "Psoriatic Arthritis," below). Occasional patients are seen with AS in association with skin manifestations seen in SAPHO syndrome (see below), such as acne fulminans or hidradenitis suppurativa. There is an apparently increased risk of ischemic heart disease. Aortic insufficiency occurs in a small percentage of patients. Third-degree heart block may occur alone or together with aortic insufficiency, and association with lesser degrees of heart block has been described. Cauda equina syndrome and upper pulmonary lobe fibrosis are rare late complications. Prostatitis has been reported to have an increased prevalence. Amyloidosis is rare **(Chap. 108)**.

Several validated measures of disease activity and functional outcome are in widespread use in the study and management of AS, particularly the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) and the Ankylosing Spondylitis Disease Activity Score (ASDAS), both measures of disease activity; the Bath Ankylosing Spondylitis Functional Index (BASFI), a measure of limitation in activities of daily living; and several measures of radiographic changes. The Harris hip score, although not specific for AS, can be helpful. Despite persistence of the disease, most patients remain gainfully employed. Some but not all studies of survival in AS have suggested that AS shortens life span, compared with the general population. Mortality attributable to AS is largely the result of spinal trauma, aortic insufficiency, respiratory failure, amyloid nephropathy, or complications of therapy such as upper gastrointestinal hemorrhage. The impact of anti-TNF therapy on outcome and mortality is not yet known, except for significantly improved work productivity.

LABORATORY FINDINGS

No laboratory test is diagnostic of AS. In most ethnic groups, HLA-B27 is present in 75–90% of patients. Erythrocyte sedimentation rate (ESR) and C-reactive protein (CRP) are often, but not always, elevated. Mild anemia may be present. Patients with severe disease may show an elevated alkaline phosphatase level. Elevated serum IgA levels are common. Rheumatoid factor, anti-cyclic citrullinated peptide (CCP), and antinuclear antibodies (ANAs) are largely absent unless caused by a coexistent disease, although ANAs may appear with anti-TNF therapy. Circulating levels of CD8+ T cells tend to be low, and serum matrix metalloproteinase 3 levels correlate with disease activity. Synovial fluid from peripheral joints in AS is nonspecifically inflammatory. Restricted chest wall motion causes decreased vital capacity, but ventilatory function is usually well maintained.

RADIOGRAPHIC FINDINGS

By definition, the diagnosis of AS is associated with advanced radiographically demonstrable sacroiliitis, usually symmetric. The earliest changes by standard radiography are blurring of the cortical margins of the subchondral bone, followed by erosions and sclerosis. Progression of the erosions leads to "pseudowidening" of the joint space; as fibrous and then bony ankylosis supervene, the joints may become obliterated.

In the lumbar spine, progression of the disease can lead to loss of lordosis, and osteitis of the anterior corners of the vertebral bodies with subsequent erosion, leading to "squaring" or even "barreling" of one or more vertebral bodies. Progressive ossification leads to eventual formation of marginal syndesmophytes, visible on plain films as bony bridges connecting successive vertebral bodies anteriorly and laterally.

A recent study showed that only a quarter of patients meeting criteria for nr-ax-SpA developed radiographic sacroiliitis within 15 years. MRI is thus much more useful for the timely diagnosis of ax-SpA. Active sacroiliitis is best visualized by dynamic MRI on tilted coronal slices with fat saturation, either T2-weighed turbo spin-echo sequence or short tau inversion recovery (STIR) with high resolution, or T1-weighted images with contrast enhancement. These techniques identify early intraarticular inflammation, cartilage changes, and underlying bone marrow edema in sacroiliitis **(Fig. 355-1)**. The presence of erosions enhances specificity and is best detected on conventional T1-weighted images. These protocols are also sensitive for evaluation of acute and chronic spinal changes. MRI protocols routinely used to evaluate low back pain have low sensitivity for detecting inflammation and often give false-negative results in ax-SpA. Optimal results require a high index of suspicion, an appropriate protocol, an experienced radiologist, and close communication between radiologist and clinician.

Reduced bone mineral density can be detected by dual-energy x-ray absorptiometry of the femoral neck and the lumbar spine. Use of a lateral projection of the L3 vertebral body can prevent falsely elevated readings related to spinal ossification.

DIAGNOSIS

It is important to recognize ax-SpA before the development of irreversible deformity. This goal is challenging for several reasons: (1) only a minority of back pain patients have ax-SpA; (2) an early diagnosis often relies on clinical grounds and/or an appropriate MRI protocol requiring considerable expertise; (3) young individuals with symptoms of ax-SpA often do not seek medical care; (4) reliance on definite radiographic sacroiliitis causes early or mild cases to be missed. The classification criteria for axial SpA proposed by the Assessment of Spondyloarthritis International Society (ASAS) are shown in Table 355-1. They are applicable to individuals with ≥3 months of back pain with age of onset <45 years. Active inflammation of the sacroiliac joints as determined by dynamic MRI is considered equivalent to definite radiographic sacroiliitis (see below).

ax-SpA must be differentiated from numerous other causes of low-back pain, some substantially more common than ax-SpA. Increased specificity is obtained when the nature and pattern of the pain and the age of the patient are considered. The most typical symptom is inflammatory back pain (IBP), present in 70–80% of patients with ax-SpA, but relatively uncommon otherwise. In chronic (≥3 months) back pain, IBP has the following characteristic features: (1) age of onset <40 years; (2) insidious onset; (3) improvement with exercise; (4) no improvement with rest; and (5) pain at night with improvement upon getting up; (6) morning stiffness >30 min; (7) awakening from back pain during only the second half of the night; and (8) alternating buttock pain. The presence of two or more of these features should arouse suspicion for IBP, and four or more can be considered diagnostic. The most common causes of back pain other than SpA are primarily mechanical or degenerative rather than primarily inflammatory and tend not to show clustering of these features.

Less-common metabolic, infectious, and malignant causes of back pain must also be differentiated from AS, including infectious spondylitis, spondylodiskitis, and sacroiliitis, and primary or metastatic tumor. Ochronosis can produce a phenotype similar to AS. Calcification and ossification of paraspinous ligaments occur in *diffuse idiopathic skeletal hyperostosis* (DISH), which occurs in the middle-aged and elderly and is usually not symptomatic. Ligamentous calcification gives the appearance of "flowing wax" on the anterior bodies of the vertebrae. Intervertebral disk spaces are preserved, and sacroiliac and apophyseal joints appear normal, helping to differentiate DISH from spondylosis and from AS, respectively. Both primary and secondary hyperparathyroidism can cause subchondral bone resorption around the SI joints, with bilateral widened and ill-defined joints on radiographs, but without joint space narrowing.

An algorithm for making or excluding the diagnosis of ax-SpA in patients with chronic back pain starting before age 45 is shown in **Fig. 355-2**.

FIGURE 355-1 Early sacroiliitis in a patient with ankylosing spondylitis, indicated by prominent edema in the juxtaarticular bone marrow (*asterisks*), synovium and joint capsule (*thin arrow*), and interosseous ligaments (*thick arrow*) on a short tau inversion recovery (STIR) magnetic resonance image. *(From M Bollow et al: Zeitschrift für Rheumatologie 58:61, 1999. Reproduced with permission.)*

TREATMENT

Axial Spondyloarthritis

All management of ax-SpA should include an exercise program to maintain posture and range of motion. Nonsteroidal anti-inflammatory drugs (NSAIDs) are the first line of pharmacologic therapy. They reduce pain and tenderness and increase mobility in many patients. Continuous high-dose NSAID therapy may slow radiographic progression, particularly in patients who are at higher risk for progression. However, many patients have continued symptoms despite NSAID therapy and are likely to benefit from anti-TNF-α therapy. Patients with AS treated with infliximab (chimeric human/mouse anti-TNF-α monoclonal antibody), etanercept (soluble p75 TNF-α receptor–IgG fusion protein), adalimumab, or golimumab (human anti-TNF-α monoclonal antibodies, or certolizumab pegol [humanized mouse anti-TNF-α monoclonal antibody]) have shown rapid, profound, and sustained reductions in all clinical and laboratory measures of disease activity. In a good response, there is significant improvement in both objective and subjective indicators of disease activity and function, including morning stiffness, pain, spinal mobility, peripheral joint swelling, CRP, ESR, and bone mineral density. MRI studies indicate substantial resolution of bone marrow edema, enthesitis, and joint effusions in the sacroiliac joints, spine, and peripheral joints. Similar results have been obtained in large randomized controlled trials of all five agents and many open-label studies. About one-half of the patients achieve a ≥50% reduction in the BASDAI. The response tends to persist over time, and partial or full remissions are common. Predictors of the best responses include younger age, shorter disease duration, higher baseline inflammatory markers, and lower baseline functional disability. Nonetheless, some patients with long-standing disease and even spinal ankylosis can obtain significant benefit. Syndesmophyte formation may continue despite the therapy, but this may apply mainly during the early years of therapy. Although less well studied, the response of patients with nr-ax-SpA to anti-TNF therapy is generally similar to that of patients with AS.

Typically, infliximab is given intravenously, 5 mg/kg body weight, and then repeated 2 weeks later, again 6 weeks later, and then at 6- to 8-week intervals. Etanercept is given by subcutaneous injection, 50 mg once weekly. Adalimumab is given by subcutaneous injection, 40 mg biweekly. Golimumab is given by subcutaneous injection, 50 or 100 mg every 4 weeks. Certolizumab pegol is given by subcutaneous injection, 400 mg every 4 weeks. Dosage adjustments can be considered in selected cases.

These potent immunosuppressive agents are relatively safe, but patients are at increased risk for serious infections, including disseminated tuberculosis. Hypersensitivity infusion or injection site reactions are not uncommon. Cases of anti-TNF-induced psoriasis have been increasingly recognized. Rare cases of systemic lupus erythematosus (SLE)–related disease have been reported, as have hematologic disorders such as pancytopenia, demyelinating disorders, exacerbation of congestive heart failure, and severe liver disease. The overall incidence of malignancy does not appear to be increased in AS patients treated with anti-TNF therapy, but isolated cases of hematologic malignancy have occurred shortly after the start of treatment.

Because of the expense, potentially serious side effects, and unknown long-term effects of these agents, their use should be restricted to patients with a definite diagnosis and active disease (BASDAI ≥4 out of 10 and expert opinion) that is inadequately responsive to therapy with at least two different NSAIDs. Before initiation of anti-TNF therapy, all patients should be tested for tuberculin (TB) reactivity, and reactors (≥5 mm on PPD testing or a positive quantiferon test) should be treated with anti-TB agents. Contraindications include active infection or high risk of infection; malignancy or premalignancy; and history of SLE, multiple sclerosis, or related autoimmunity. Pregnancy and breast-feeding are no longer considered contraindications if appropriate precautions are taken. Infants exposed to anti-TNF in utero should not be given live vaccines before age 6 months. Continuation beyond 12 weeks of therapy requires either a 50% reduction in BASDAI or absolute reduction of ≥2 out of 10, and favorable expert opinion. Switching to a second anti-TNF agent may be effective, especially if there was a response to the first that was lost rather than primary failure.

Secukinumab, a human monoclonal antibody to IL-17A, shows dramatic efficacy in AS, similar to that seen with TNF inhibitors, and is effective in some patients who have failed or not tolerated anti-TNF therapy. The recommended dose is 150 mg subcutaneously weekly for 4 weeks, and then at 4 week intervals. Precautions regarding infection are similar to those for anti-TNF agents. An additional concern is potential exacerbation of underlying IBD, whether previously recognized or not, and careful monitoring is advised.

Sulfasalazine, in doses of 2–3 g/d, has modest benefit, primarily for peripheral arthritis. Methotrexate, although widely used, has not been shown to be of benefit in AS, nor has any therapeutic role for gold or oral glucocorticoids been documented. Potential benefit in AS has been reported for thalidomide, 200 mg/d, perhaps acting through inhibition of TNF-α. The oral Jak inhibitor, tofacitinib, showed efficacy in AS, with reduction of inflammation evident on MRI, in a 16 week phase 2 study.

The most common indication for surgery in patients with AS is severe hip joint arthritis, the pain and stiffness of which are usually dramatically relieved by total hip arthroplasty. Rare patients may benefit from surgical correction of extreme flexion deformities of the spine or of atlantoaxial subluxation.

Attacks of uveitis are usually managed effectively with local glucocorticoids and mydriatic agents, although systemic glucocorticoids, immunosuppressive drugs, or anti-TNF therapy may be required. TNF inhibitors reduce the frequency of attacks of uveitis in patients with ax-SpA, and adalimumab has recently been approved by the FDA for treating uveitis. Cases of new or recurrent uveitis after use of a TNF inhibitor have been observed, especially with etanercept. Anti-IL-17 does not appear as effective for uveitis as anti-TNF therapy.

Management of axial osteoporosis is at present similar to that used for primary osteoporosis, since data specific for AS are not available.

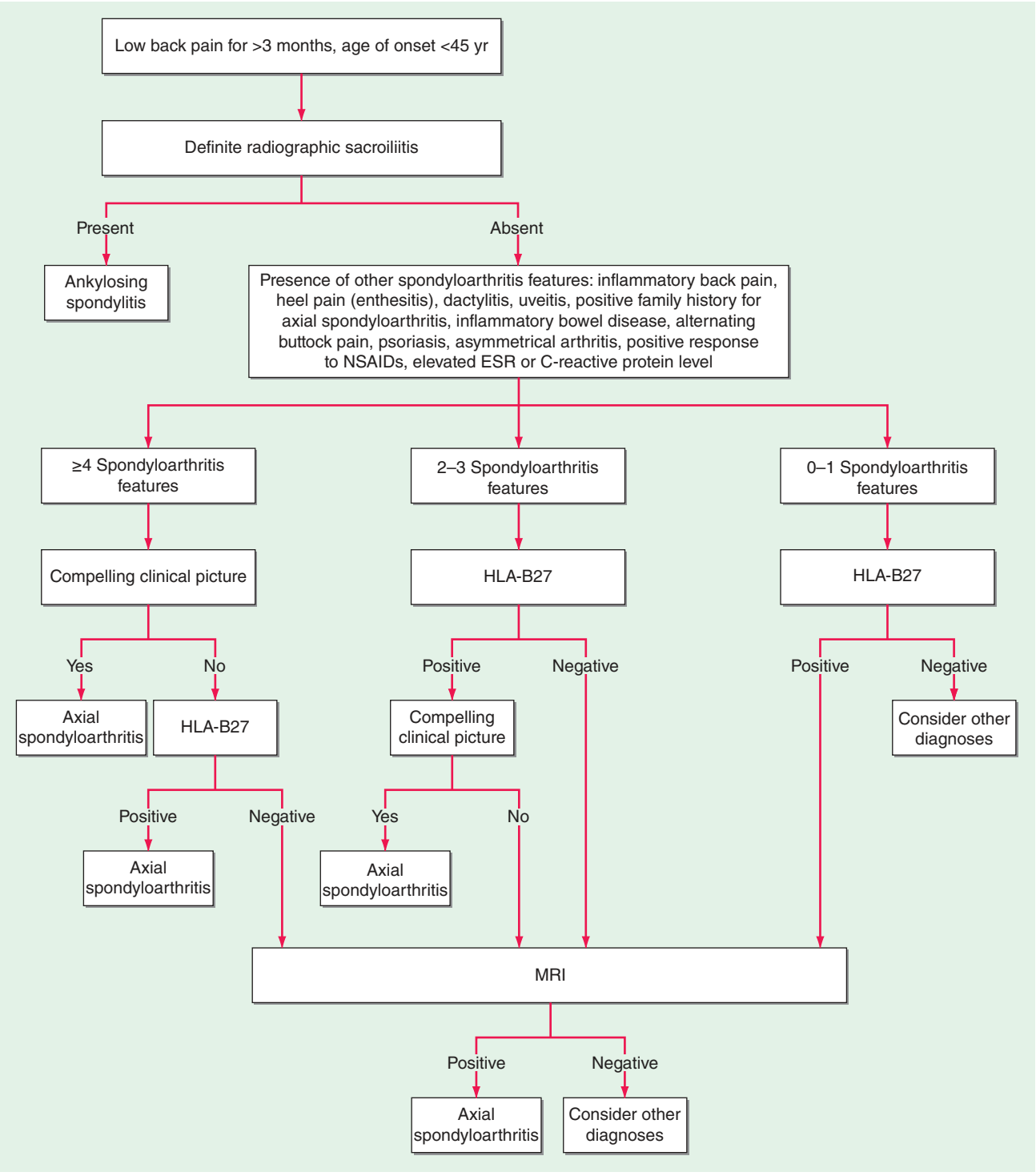


FIGURE 355-2 Algorithm for the diagnosis or exclusion of axial spondyloarthritis. The algorithm is designed for use in patients with at least a 3-month history of unexplained chronic low back pain that started before the age of 45 years. Definite radiographic sacroiliitis is based on the modified New York criteria for ankylosing spondylitis (*van der Linden S et al: Arthritis Rheum 27:361, 1984*). The algorithm is adapted from *van den Berg R et al: Ann Rheum Dis 72:1646, 2013*. The determination of whether or not a clinical picture is compelling is based on the relative weights of the spondyloarthritis features (*Feldtkeller E et al: Rheumatology [Oxford] 52:1648, 2013*) and on clinical judgment. The list of clinical features includes features of both axial and peripheral spondyloarthritis. *(From Taurog JD et al: N Engl J Med 374:2563, 2016.)*

REACTIVE ARTHRITIS

ReA refers to acute nonpurulent arthritis complicating an infection elsewhere in the body. In recent years, the term has been used primarily to refer to SpA following enteric or urogenital infections.

Other forms of reactive and infection-related arthritis not associated with B27 and showing a spectrum of clinical features different from SpA, such as Lyme disease, rheumatic fever, and poststreptococcal reactive arthritis, are discussed in Chaps. 181 and 352.

HISTORIC BACKGROUND

The association of acute arthritis with episodes of diarrhea or urethritis has been recognized for centuries. A large number of cases during World Wars I and II focused attention on the triad of arthritis, urethritis, and conjunctivitis, often with additional mucocutaneous lesions, which became widely known by eponyms that are now of historic interest only.

The identification of bacterial species triggering the clinical syndrome and the finding of an association with HLA-B27 led to the

unifying concept of ReA as a clinical syndrome triggered by specific etiologic agents in a genetically susceptible host. A similar spectrum of clinical manifestations can be triggered by enteric infection with any of several *Shigella*, *Salmonella*, *Yersinia*, and *Campylobacter* species; by genital infection with *Chlamydia trachomatis*; by many other agents as well, apparently in some cases via nasopharyngeal infection with *Chlamydia pneumoniae* or other agents. The "classic triad" represents a small part of the spectrum of the clinical manifestations of ReA and is present only in a small minority of patients. For the purposes of this chapter, the use of the term *ReA* will be restricted to those cases of SpA in which there is at least presumptive evidence for a related symptomatic antecedent infection.

EPIDEMIOLOGY

In early reports, 60–85% of patients who developed ReA triggered by *Shigella*, *Yersinia*, or *Chlamydia* were HLA-B27-positive, but the true figure is probably lower. Other studies demonstrated a lower prevalence of B27 in ReA triggered by *Salmonella*, and little or no association in *Campylobacter*-induced ReA. More recent community-based or common-source epidemic studies showed a prevalence of B27 in ReA <50%. The most common age range is 18–40 years, but ReA can occur rarely in children and occasionally in older adults.

The attack rate of postenteric ReA generally ranges from 1 to about 30% depending on the study and causative organism, whereas the attack rate of postchlamydial ReA is about 4–8%. The gender ratio in ReA following enteric infection is nearly 1:1, whereas venereally acquired ReA occurs mainly in men. The overall prevalence and incidence of ReA are difficult to assess because of the lack of validated diagnostic criteria, variable prevalence and arthritogenic potential of the triggering microbes, and inconstant genetic susceptibility factors in different populations. In Scandinavia, an annual incidence of 10–28:100,000 has been reported. The spondyloarthritides were formerly almost unknown in sub-Saharan Africa. However, ReA and other peripheral SpAs have become common in black Africans in the wake of the AIDS epidemic, without association to B27, which is very rare in these populations. ReA is often the first manifestation of HIV infection and often remits with disease progression. In contrast, Western white patients with HIV and SpA are usually B27-positive, and the arthritis flares as AIDS advances.

PATHOLOGY

Synovial histology is similar to that of other SpAs. Enthesitis shows increased vascularity and macrophage infiltration of fibrocartilage. Microscopic histopathologic evidence of inflammation mimicking IBD has routinely been demonstrated in the colon and ileum of patients with postenteric ReA and less commonly in postvenereal ReA. The skin lesions of keratoderma blennorrhagica, associated mainly with venereally acquired ReA, are histologically indistinguishable from pustular psoriasis.

ETIOLOGY AND PATHOGENESIS

Definite bacterial triggers of ReA include several *Salmonella* spp., *Shigella* spp., *Yersinia enterocolitica*, *Yersinia pseudotuberculosis*, *Campylobacter jejuni*, and *Chlamydia trachomatis*. These triggering microbes are gram-negative bacteria with a lipopolysaccharide (LPS) component to their cell walls. All *Shigella* species have been implicated in cases of ReA, with *S. flexneri* and *S. sonnei* being the most common. After *Salmonella* infection, individuals of Caucasian descent may be more likely than those of Asian descent to develop ReA. Children may be less susceptible to ReA caused by *Salmonella* and *Campylobacter*. *Yersinia* species in Europe and Scandinavia may have greater arthritogenic potential than in other parts of the world, and *C. trachomatis* appears to be a common cause worldwide. The ocular serovars of *C. trachomatis* appear to be particularly, perhaps uniquely, arthritogenic.

There is also evidence implicating *Clostridium difficile*, *Campylobacter coli*, certain toxigenic *Escherichia coli*, and possibly *Ureaplasma urealyticum* and *Mycoplasma genitalium* as potential triggers of ReA. *Chlamydia pneumoniae* is also a trigger of ReA, but far less commonly so than *C. trachomatis*. There have been numerous isolated reports of acute arthritis preceded by many other bacterial, viral, or parasitic infections, and arthritis following intravesicular bacillus Calmette-Guérin (BCG) treatment for bladder cancer is well documented.

It is not known whether there is a common pathogenic mechanism for triggering of ReA that is shared by all of these microorganisms, nor has the mechanism been elucidated in the case of any one of the known triggers. Many of the well established triggers share a capacity to attack mucosal surfaces, to invade host cells, and to survive intracellularly. Antigens from *Chlamydia*, *Yersinia*, *Salmonella*, and *Shigella* have been shown to be present in the synovium and/or synovial fluid leukocytes of patients with ReA for long periods following the acute attack. In ReA triggered by *Y. enterocolitica*, bacterial LPS and heat-shock protein antigens have been found in peripheral blood cells years after the triggering infection. *Yersinia* DNA and *C. trachomatis* DNA and RNA have been detected in synovial tissue from ReA patients, suggesting the presence of viable organisms despite uniform failure to culture the organism from these specimens. In *C. trachomatis*–induced ReA specifically, the bacterial load in synovial tissue of patients with remitting disease is lower than that of active disease, but mRNAs encoding proinflammatory proteins are equal to or higher than those of active disease. The specificity of these findings is unclear, however, since chromosomal bacterial DNA and 16S rRNA from a wide variety of bacteria have also been found in synovium in other rheumatic diseases, albeit less frequently.

Synovial T cells specific for antigens of the inciting organism were reported in the 1980s and early 1990s. More recent work has documented high levels of IL-17 in ReA synovial fluid, but the source has not been identified. HLA-B27 seems to be associated with more severe and chronic ReA, but its pathogenic role remains to be determined. HLA-B27 significantly prolongs the intracellular survival of *Y. enterocolitica* and *Salmonella enteritidis* in human and mouse cell lines. Prolonged intracellular bacterial survival may permit trafficking of infected leukocytes from the site of primary infection to joints, where an innate and/or adaptive immune response to persistent bacterial antigens may then promote arthritis.

CLINICAL FEATURES

The clinical manifestations of ReA range from an isolated, transient monoarthritis or enthesitis to severe multisystem disease. A careful history will often elicit evidence of an antecedent infection 1–4 weeks before onset of symptoms of the reactive disease, particularly in postenteric ReA. However, in a sizable minority, no clinical or laboratory evidence of an antecedent infection can be found, particularly in the case of postchlamydial ReA. In cases of presumed venereally acquired reactive disease, there is often a history of a recent new sexual partner, even without laboratory evidence of infection.

Constitutional symptoms are common, including fatigue, malaise, fever, and weight loss. The musculoskeletal symptoms are usually acute in onset. Arthritis is usually asymmetric and additive, with involvement of new joints occurring over a few days to 1–2 weeks. The joints of the lower extremities, especially the knee, ankle, subtalar, metatarsophalangeal, and toe interphalangeal joints, are most commonly involved, but the wrist and fingers may be involved. The arthritis is usually quite painful, and tense joint effusions are not uncommon, especially in the knee. Dactylitis, or "sausage digit," a diffuse swelling of a solitary finger or toe, is a distinctive feature of ReA and other peripheral spondyloarthritides but can be seen in polyarticular gout and sarcoidosis. Tendinitis and fasciitis are particularly characteristic lesions, producing pain at multiple insertion sites (entheses), especially the Achilles insertion, the plantar fascia, and sites along the axial skeleton. Back and buttock pain are quite common and may be caused by insertional inflammation, muscle spasm, acute sacroiliitis, or, presumably, arthritis in intervertebral joints.

Urogenital lesions may occur throughout the course of the disease. In males, urethritis may be marked or relatively asymptomatic and may be either an accompaniment of the triggering infection or a result of the reactive phase of the disease; interestingly, it occurs in both postvenereal and postenteric ReA. Prostatitis is also common. Similarly, in females, cervicitis or salpingitis may be caused either by the infectious trigger or by the sterile reactive process.

Ocular disease is common, ranging from transient, asymptomatic conjunctivitis to an aggressive anterior uveitis that occasionally proves refractory to treatment and may result in blindness.

Mucocutaneous lesions are frequent. Oral ulcers tend to be superficial, transient, and often asymptomatic. The characteristic skin lesions, *keratoderma blennorrhagica*, consist of vesicles and/or pustules that become hyperkeratotic, ultimately forming a crust before disappearing. They are most common on the palms and soles but may occur elsewhere as well. In patients with HIV infection, these lesions are often severe and extensive, sometimes dominating the clinical picture **(Chap. 197)**. Lesions on the glans penis, termed *circinate balanitis*, consist of vesicles that quickly rupture to form painless superficial erosions, which in circumcised individuals can form crusts similar to those of keratoderma blennorrhagica. Nail changes are common and consist of onycholysis, distal yellowish discoloration, and/or heaped-up hyperkeratosis.

Less-frequent or rare manifestations of ReA include cardiac conduction defects, aortic insufficiency, central or peripheral nervous system lesions, and pleuropulmonary infiltrates.

Arthritis typically persists for 3–5 months, but more chronic courses do occur. Chronic joint symptoms persist in about 15% of patients and in up to 60% of patients in hospital-based series, but these tend to be less severe than in the acute stage. Recurrences of the acute syndrome may occur. Work disability or forced change in occupation is common in those with persistent joint symptoms. Chronic heel pain is often particularly distressing. Low-back pain, sacroiliitis, and frank AS are also common sequelae. In most studies, HLA-B27–positive patients have shown a worse outcome than B27-negative patients. Patients with *Yersinia*- or *Salmonella*-induced arthritis have less chronic disease than those whose initial episode follows epidemic shigellosis.

LABORATORY AND RADIOGRAPHIC FINDINGS

The ESR and acute-phase reactants are usually elevated during the acute phase of the disease, often markedly so. Mild anemia may be present. Synovial fluid is nonspecifically inflammatory. In most ethnic groups, 30–50% of the patients are B27-positive. The triggering infection usually does not persist at the site of primary mucosal infection through the time of onset of the reactive disease, but it may be possible to culture the organism, for example, in the case of *Yersinia*- or *Chlamydia*-induced disease. Serologic evidence of exposure to one of the causative organisms with elevation of antibodies is nonspecific and of questionable utility. Polymerase chain reaction (PCR) for chlamydial DNA in first-voided urine specimens may have high sensitivity in the acute stage but is less useful with chronic disease.

In early or mild disease, radiographic changes may be absent or confined to juxtaarticular osteoporosis. With long-standing disease, radiographic features share those of PsA; marginal erosions and loss of joint space can be seen in affected joints. Periostitis with reactive new bone formation is characteristic, as in all the SpAs. Spurs at the insertion of the plantar fascia are common.

Sacroiliitis and spondylitis may be seen as late sequelae. Sacroiliitis is more commonly asymmetric than in AS, and spondylitis, rather than ascending symmetrically, can begin anywhere along the lumbar spine. The syndesmophytes are described as nonmarginal; they are coarse, asymmetric, and "comma"-shaped, arising from the middle of a vertebral body, a pattern less commonly seen in primary AS. Progression to spinal fusion is uncommon.

DIAGNOSIS

ReA is a clinical diagnosis with no definitively diagnostic laboratory test or radiographic finding. The diagnosis should be entertained in any patient with an acute inflammatory, asymmetric, additive arthritis or tendinitis. The evaluation should include thorough but tactful questioning regarding possible triggering events. On physical examination, attention must be paid to the distribution of the joint and tendon involvement and to possible sites of extraarticular involvement, including the eyes, mucous membranes, skin, nails, and genitalia. Synovial fluid analysis is usually necessary to exclude septic or crystal-induced arthritis. Culture, serology, or molecular methods may help identify a triggering infection, but they cannot be relied upon.

Although typing for B27 has low negative predictive value in ReA, it may have prognostic significance in terms of severity, chronicity, and the propensity for spondylitis and uveitis. Furthermore, if positive, it can be helpful diagnostically in atypical cases. HIV testing is often indicated and may be necessary in selecting therapy.

Both ReA and disseminated gonococcal disease **(Chap. 151)** can be venereally acquired and associated with urethritis. Unlike ReA, gonococcal arthritis and tenosynovitis tend to involve both upper and lower extremities equally, spare the axial skeleton, and be associated with characteristic vesicular skin lesions. A positive gonococcal culture from the urethra or cervix does not exclude a diagnosis of ReA; however, culturing gonococci from blood, skin lesion, or synovium establishes the diagnosis of disseminated gonococcal disease. PCR assay for *Neisseria gonorrhoeae* and *C. trachomatis* may be helpful. Occasionally, only a therapeutic trial of antibiotics can distinguish the two.

ReA shares many features in common with psoriatic arthropathy. However, PsA is usually gradual in onset; the arthritis tends to affect primarily the upper extremities; and there are usually no associated mouth ulcers, urethritis, or bowel symptoms.

TREATMENT

Reactive Arthritis

Most patients with ReA benefit to some degree from high-dose NSAIDs, although acute symptoms are rarely completely ameliorated, and some patients fail to respond at all.

Prompt, appropriate antibiotic treatment of acute chlamydial urethritis or enteric infection may prevent the emergence of ReA, but is not universally successful. Data regarding the potential benefit of antibiotic therapy that is initiated after onset of arthritis are conflicting; however, a recent systematic review and meta-analysis of 10 controlled trials suggested no benefit. The only one of these trials to use combination antibiotics showed that a majority of patients with chronic ReA associated with *C. trachomatis* or *C. pneumoniae* benefited significantly from a 6-month course of rifampin 300 mg daily plus azithromycin 500 mg daily for 5 days, then twice weekly, or 6 months of rifampin 300 mg daily plus doxycycline 100 mg twice daily. This study awaits further confirmation.

Multicenter trials have suggested that sulfasalazine, up to 3 g/d in divided doses, may be beneficial to patients with persistent ReA.[1] Patients with persistent disease may respond to azathioprine, 1–2 mg/kg per day, or to methotrexate, up to 20 mg per week; however, these therapeutic regimens have never formally been studied. Although no controlled trials of anti-TNF-α in ReA have been reported, anecdotal evidence supports the use of these agents in severe chronic cases, although lack of response has also been observed.[1]

Tendinitis and other enthesitic lesions may benefit from intralesional glucocorticoids. Uveitis may require aggressive treatment to prevent serious sequelae (see above). Skin lesions ordinarily require only symptomatic topical treatment. In patients with HIV infection and ReA, many of whom have severe skin lesions, the skin lesions in particular respond to antiretroviral therapy. Cardiac complications are managed conventionally; management of neurologic complications is symptomatic.

Comprehensive management includes counseling of patients in the avoidance of sexually transmitted disease and exposure to enteropathogens, as well as appropriate use of physical therapy, vocational counseling, and continued surveillance for long-term complications such as AS. Patients with a history of ReA are at increased risk for recurrent attacks following repeated exposures.

[1]Azathioprine, methotrexate, sulfasalazine, pamidronate, thalidomide, and anti-TNFα agents have not been approved for this purpose by the FDA at the time of publication.

PSORIATIC ARTHRITIS

Psoriatic arthritis refers to an inflammatory musculoskeletal disease that has both autoimmune and autoinflammatory features characteristically occurring in individuals with psoriasis.

HISTORIC BACKGROUND

The association between arthritis and psoriasis was noted in the nineteenth century. In the 1960s, it became clear that unlike RA, the arthritis associated with psoriasis was usually seronegative, often involved the distal interphalangeal (DIP) joints of the fingers and the spine and sacroiliac joints, had distinctive radiographic features, and showed considerable familial aggregation. In the 1970s, PsA was included in the broader category of the spondyloarthritides because of features similar to those of AS and ReA.

EPIDEMIOLOGY

The prevalence of PsA appears to be increasing in parallel with disease awareness. Recent data suggest that up to 30% of patients with psoriasis develop PsA. The duration and severity of psoriasis increase the likelihood of developing PsA. In white populations, psoriasis is estimated to have a prevalence of 1–3%. Psoriasis and PsA are less common in other races in the absence of HIV infection, and the prevalence of PsA in individuals with psoriasis may be less common. First-degree relatives of PsA patients have an elevated risk for psoriasis, for PsA, and for other forms of SpA. Of patients with psoriasis, up to 30% have an affected first-degree relative. In monozygotic twins, the reported concordance for psoriasis varies from 35 to 72%, and for PsA from 10 to 30%. A variety of HLA associations have been found. HLA-Cw0602 is directly associated with psoriasis, particularly familial juvenile-onset (type I) psoriasis. HLA-B27 is associated with psoriatic spondylitis (see below). HLA-DR7, -DQ3, and -B57 are associated with PsA because of linkage disequilibrium with Cw6. A recent study found additive associations of PsA with haplotypes containing HLA-B08, HLA-Cw0602, HLA-B27, -B38, and -B39. A correlation was also found between different haplotype combinations and entheseal, synovial, or axial predominant phenotypes. Genome-wide analyses have identified associations of PsA with polymorphisms in the IL-23 receptor (*IL23R*), molecules involved in nuclear factor κB gene expression (*TNIP1*) and signaling (*TNFAIP3*), and cytokines *TNF*, *IL12A*, and *IL12B*. A specific *IL23R* SNP is associated with PsA distinct from psoriasis without arthritis.

PATHOLOGY

The inflamed synovium in PsA resembles that of RA, although with somewhat less hyperplasia and cellularity than in RA. As noted with AS above, the synovial vascular pattern in PsA is generally greater and more tortuous than in RA, independent of disease duration. Some studies have indicated a higher tendency to synovial fibrosis in PsA. Unlike RA, PsA shows prominent enthesitis, with histology similar to that of the other spondyloarthritides.

PATHOGENESIS

PsA is almost certainly immune-mediated and presumably shares pathogenic mechanisms with psoriasis. PsA synovium is characterized by lining layer hyperplasia; diffuse infiltration with T cells, B cells, macrophages, and NK receptor–expressing cells, with upregulation of leukocyte homing receptors; and neutrophil proliferation with angiogenesis. Clonally expanded T cell subpopulations are frequent and have been demonstrated both in the synovium and the skin. Plasmacytoid dendritic cells are thought to play a key role in psoriasis, and there is some evidence for their participation in PsA. There is abundant synovial overexpression of proinflammatory cytokines, and synovial tissue staining has identified an overexpression of monocyte-derived cytokines, such as myeloid-related protein (S100A8/A9). Interferon γ, TNF-α, and IL-1β, 2, 6, 8, 10, 12, 13, and 15 are found in PsA synovium or synovial fluid. T_H17-derived cytokines are important in PsA, given the genetic association with genes in the IL-23/17 pathway and the therapeutic response to agents targeting this pathway. T_H17 cells have been identified from the dermal extracts of psoriatic lesions and the synovial fluid of PsA patients. The majority of these CD4+ IL-17+ T cells are of memory phenotype (CD4RO[+]CD45RA[–]CD11a[+]). Consistent with the extensive bone remodeling in PsA, patients with PsA have been found to have a marked increase in osteoclastic precursors in peripheral blood and upregulation of receptor activator of nuclear factor κB ligand (RANKL) in the synovial lining layer. Increased serum levels of TNF-α, RANKL, leptin, and omentin positively correlate with these osteoclastic precursors.

CLINICAL FEATURES

In 60–70% of cases, psoriasis precedes joint disease. In 15–20% of cases, the two manifestations appear within 1 year of each other. In about 15–20% of cases, the arthritis precedes the onset of psoriasis and can present a diagnostic challenge. The frequency in men and women is almost equal, although the frequency of disease patterns differs somewhat in the two sexes. The disease can begin in childhood or late in life but typically begins in the fourth or fifth decade, at an average age of 37 years.

Many classification schemes have been proposed for the broad spectrum of arthropathy in PsA. Wright and Moll described five patterns: (1) arthritis of the DIP joints; (2) asymmetric oligoarthritis; (3) symmetric polyarthritis similar to RA; (4) axial involvement (spine and sacroiliac joints); and (5) arthritis mutilans, a highly destructive form of the disease. These patterns frequently coexist, and the pattern that persists chronically often differs from that of the initial presentation. A simpler scheme in recent use contains three patterns: oligoarthritis, polyarthritis, and axial arthritis.

Nail changes in the fingers or toes occur in most patients with PsA, compared with only a minority of psoriatic patients without arthritis, and pustular psoriasis is said to be associated with more severe arthritis. Dactylitis and enthesitis are common in PsA and help to distinguish it from other joint disorders. Dactylitis occurs in >30%; enthesitis and tenosynovitis are probably present in most patients, although often not appreciated on physical examination. Shortening of digits because of underlying osteolysis is particularly characteristic of PsA **(Fig. 355-3)**, and there is a much greater tendency than in RA for both fibrous and bony ankylosis of small joints. Rapid ankylosis of one or more proximal interphalangeal (PIP) joints early in the course of disease is not uncommon. Joint involvement tends to follow a "ray" distribution, with all of the joints of one finger involved, while sparing adjacent fingers entirely. Back and neck pain and stiffness are also common in PsA.

Arthropathy confined to the DIP joints occurs in about 5% of cases. Accompanying nail changes in the affected digits are almost always present. These joints are also often affected in the other patterns of PsA. Approximately 30% of patients have asymmetric oligoarthritis. This pattern commonly involves a knee or another large joint with a

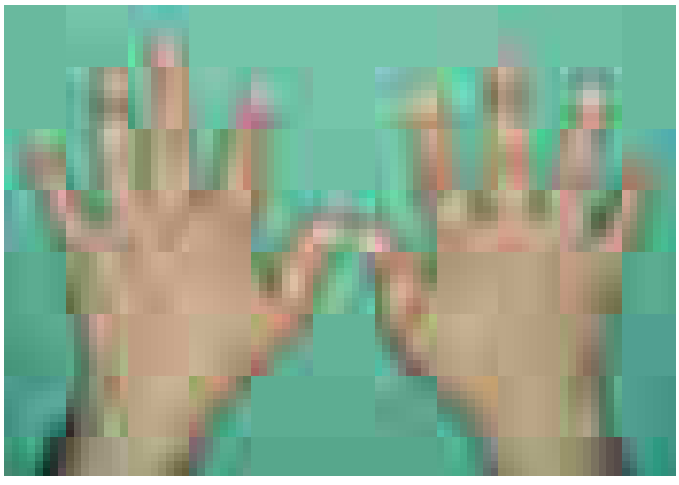


FIGURE 355-3 Characteristic lesions of psoriatic arthritis. Inflammation is prominent in the distal interphalangeal joints (left 5th, 4th, 2nd; right 2nd, 3rd, and 5th) and proximal interphalangeal joints (left 2nd, right 2nd, 4th, and 5th). There is dactylitis in the left 2nd finger and thumb, with pronounced telescoping of the left 2nd finger. Nail dystrophy (hyperkeratosis and onycholysis) affects each of the fingers except the left 3rd finger, the only finger without arthritis. *(Courtesy of Donald Raddatz, MD; with permission.)*

few small joints in the fingers or toes, often with dactylitis. Symmetric polyarthritis occurs in about 40% of PsA patients at presentation. It may be indistinguishable from RA in terms of the joints involved, but other features characteristic of PsA are usually also present. Almost any peripheral joint can be involved. Axial arthropathy without peripheral involvement is found in about 5% of PsA patients. It may be clinically indistinguishable from idiopathic AS, although more neck involvement and less thoracolumbar spinal involvement are characteristic, and nail changes are not found in idiopathic AS. A small percentage of PsA patients have arthritis mutilans, in which there can be widespread shortening of digits ("telescoping"), sometimes coexisting with ankylosis and contractures in other digits.

Six patterns of nail involvement are identified: pitting, horizontal ridging, onycholysis, yellowish discoloration of the nail margins, dystrophic hyperkeratosis, and combinations of these findings. Extraarticular and extradermal manifestations are common. Eye involvement, either conjunctivitis or uveitis, is reported in 7–33% of PsA patients. Unlike the uveitis associated with AS, the uveitis in PsA is more often bilateral, chronic, and/or posterior. Aortic valve insufficiency has been found in <4% of patients, usually after long-standing disease.

Widely varying estimates of clinical outcome have been reported in PsA. At its worst, severe PsA with arthritis mutilans is potentially at least as crippling and ultimately fatal as severe RA. Unlike RA, however, many patients with PsA experience temporary remissions. Overall, erosive disease develops in the majority of patients, progressive disease with deformity and disability is common, and in some large series, mortality was found to be significantly increased compared with the general population. There appears to be a greater incidence of cardiovascular death in psoriatic disease.

The psoriasis and associated arthropathy seen with HIV infection both tend to be severe and can occur in populations with low prevalence of psoriasis. Severe enthesopathy, dactylitis, and rapidly progressive joint destruction are seen, but axial involvement is very rare. This condition is prevented by or responds well to antiretroviral therapy.

LABORATORY AND RADIOGRAPHIC FINDINGS

There are no laboratory tests diagnostic of PsA. ESR and CRP are often elevated. A small percentage of patients may have low titers of rheumatoid factor or ANAs. About 10% of patients have anti-CCP antibodies. Uric acid may be elevated in the presence of extensive psoriasis. HLA-B27 is found in 50–70% of patients with axial disease, but in ≤20% of patients with only peripheral joint involvement.

The peripheral and axial arthropathies in PsA show a number of radiographic features that distinguish them from RA and AS, respectively. Characteristics of peripheral PsA include DIP involvement, including the classic "pencil-in-cup" deformity; marginal erosions with adjacent bony proliferation ("whiskering"); small-joint ankylosis; osteolysis of phalangeal and metacarpal bone, with telescoping of digits; periostitis and proliferative new bone at sites of enthesitis, and a "ray" distribution of lesions. Characteristics of axial PsA that differ from idiopathic AS include asymmetric sacroiliitis; less zygapophyseal joint arthritis; nonmarginal, bulky, "comma"-shaped syndesmophytes that tend to be fewer, less symmetric, and less delicate than the marginal syndesmophytes of AS; fluffy hyperperiostosis on anterior vertebral bodies; severe cervical spine involvement, with a tendency to atlantoaxial subluxation but relative sparing of the thoracolumbar spine; and paravertebral ossification. Ultrasound and MRI both readily demonstrate enthesitis and tendon sheath effusions that can be difficult to assess on physical examination. A recent MRI study of 68 PsA patients found sacroiliitis in 35%, unrelated to B27 but correlated with restricted spinal movement.

DIAGNOSIS

Classification criteria for PsA were published in 2006 (Classification of Psoriatic Arthritis [CASPAR] criteria) **(Table 355-2)**. The sensitivity and specificity of these criteria exceed 90%, and they are useful for early diagnosis in clinical practice. Diagnosis can be challenging when the arthritis precedes psoriasis, the psoriasis is undiagnosed or obscure, or the joint involvement closely resembles another form of arthritis. A high index of suspicion is needed in any patient with an undiagnosed inflammatory arthropathy. The history should include inquiry about psoriasis in the patient and family members. Patients should disrobe for the physical examination, and psoriasiform lesions should be sought in the scalp, ears, umbilicus, and gluteal folds in addition to more accessible sites; the finger and toe nails should also be carefully examined. Axial symptoms or signs, dactylitis, enthesitis, ankylosis, the pattern of joint involvement, and characteristic radiographic changes can be helpful clues. The differential diagnosis includes all other forms of arthritis, which can occur coincidentally in individuals with psoriasis. The differential diagnosis of isolated DIP involvement is short. Osteoarthritis (Heberden's nodes) is usually not inflammatory; gout involving more than one DIP joint often involves other sites and may be accompanied by tophi; the very rare entity multicentric reticulohistiocytosis involves other joints and has characteristic small pearly periungual skin nodules; and the uncommon entity inflammatory osteoarthritis, like the others, lacks the nail changes of PsA. Radiography can be helpful in all of these cases and in distinguishing between psoriatic spondylitis and idiopathic AS. A history of trauma to an affected joint preceding the onset of arthritis may occur more frequently in PsA than in other types of arthritis, perhaps reflecting the Koebner phenomenon in which psoriatic skin lesions arise at sites of skin trauma.

TABLE 355-2 The CASPAR (*C*lassification Criteria for *P*soriatic *A*rthritis) Criteria[a]

To meet the CASPAR criteria, a patient must have inflammatory articular disease (joint, spine, or entheseal) with ≥3 points from any of the following five categories:

1. Evidence of current psoriasis,[b,c] a personal history of psoriasis, or a family history of psoriasis[d]
2. Typical psoriatic nail dystrophy[e] observed on current physical examination
3. A negative test result for rheumatoid factor
4. Either current dactylitis[f] or a history of dactylitis recorded by a rheumatologist
5. Radiographic evidence of juxtaarticular new bone formation[g] in the hand or foot

[a]Specificity of 99% and sensitivity of 91%. [b]Current psoriasis is assigned 2 points; all other features are assigned 1 point. [c]Psoriatic skin or scalp disease present at the time of examination, as judged by a rheumatologist or dermatologist. [d]History of psoriasis in a first- or second-degree relative. [e]Onycholysis, pitting, or hyperkeratosis. [f]Swelling of an entire digit. [g]Ill-defined ossification near joint margins, excluding osteophyte formation.

Source: From W Taylor et al: Arthritis Rheum, 54:2665, 2006.

TREATMENT

Psoriatic Arthritis

Ideally, coordinated therapy is directed at both the skin and joints in PsA. Use of the anti-TNF-α agents has revolutionized the treatment of PsA. Prompt and dramatic resolution of both arthritis and skin lesions has been observed in large, randomized controlled trials of all five agents. Many of the responding patients had long-standing disease that was resistant to all previous therapy, as well as extensive skin disease. The clinical response is often more dramatic than in RA, and delay of disease progression has been demonstrated radiographically. The potential additive effect of methotrexate to anti-TNF-α agents in PsA remains uncertain. As noted above, anti-TNF therapy, paradoxically, has been reported to trigger exacerbation or de novo appearance of psoriasis, typically the palmoplantar pustular variety. In some cases, the therapy can nevertheless be continued.

The anti-IL-17A monoclonal antibody secukinumab is effective in treating both psoriasis and PsA. Ixekizumab, another IL-17 antagonist, is approved for treatment of psoriasis and PsA. Ustekinumab, a monoclonal antibody to the shared IL-23/IL-12p40 subunit, is an efficacious treatment for psoriasis and has some efficacy for PsA. Apremilast, an oral phosphodiesterase-4 inhibitor, is approved for both psoriasis and PsA. Although not quite as effective for PsA as the biologics, apremilast has a more favorable safety profile. It is not indicated in patients with radiographically evident joint damage or axial involvement. The oral Jak inhibitor, tofacitinib, is approved for

treatment of PsA in patients with an inadequate response to another disease-modifying agent. In a clinical trial it efficacy in PsA comparable to adalimumab.

Other treatment for PsA has been based on drugs that have efficacy in RA and/or in psoriasis. Methotrexate in doses of 15–25 mg/week has moderate efficacy for psoriasis, but limited efficacy for PsA. Sulfasalazine (usually given in doses of 2–3 g/d) may not have efficacy for either, and neither regimen halts progression of erosive joint disease. Other agents with efficacy in psoriasis reported to benefit PsA are cyclosporine, retinoic acid derivatives, and psoralens plus ultraviolet A light (PUVA). The pyrimidine synthetase inhibitor leflunomide has been shown to be beneficial in PsA, with modest benefit for psoriasis. There is controversy regarding the efficacy in PsA of gold and antimalarials, which have been widely used in RA.

All of these treatments require careful monitoring. Immunosuppressive therapy may be used cautiously in HIV-associated PsA if the HIV infection is well controlled.

UNDIFFERENTIATED AND JUVENILE-ONSET SPONDYLOARTHRITIS

Many patients present with some features of one or more of the spondyloarthritides discussed above. Until recently, these patients were said to have *undifferentiated spondyloarthritis*, or simply *spondyloarthritis*, as defined by the 1991 European Spondyloarthropathy Study Group criteria. For example, a patient may present with inflammatory synovitis of one knee, Achilles tendinitis, and dactylitis of one digit. Some of these patients may have ReA in which the triggering infection remains clinically silent. In some other cases, the patient subsequently develops IBD or psoriasis. The diagnosis of undifferentiated SpA was also commonly applied to patients with IBP who did meet modified New York criteria for AS. Most of these would now be classified as nr-ax-SpA (Table 355-1).

Comparable to the classification criteria for axial symptoms, the ASAS has formulated criteria for peripheral SpA. This is intended to exclude patients with axial symptoms and thus to divide the universe of patients with SpA into predominantly axial and predominantly peripheral subsets. These criteria are shown in **Table 355-3**.

At most only one-half of the patients with undifferentiated SpA are HLA-B27-positive.

In juvenile-onset SpA, which begins between ages 7 and 16 years, an asymmetric, predominantly lower-extremity oligoarthritis and enthesitis without extraarticular features is the typical mode of presentation. There is male predominance (60–80%), and the prevalence of B27 in this condition, which has been termed the *seronegative enthesopathy and arthropathy* (*SEA*) *syndrome*, is approximately 80%. Despite the absence of axial symptoms, active sacroiliitis by MRI has been found to be common at diagnosis. Many, but not all, of these patients go on to develop AS in late adolescence or adulthood.

Management of undifferentiated SpA is similar to that of the other spondyloarthritides. Anti-TNF-α therapy is indicated in severe, persistent cases not responsive to other treatment.

Current pediatric textbooks and journals should be consulted for information on management of juvenile-onset SpA.

TABLE 355-3 ASAS Criteria for Peripheral Spondyloarthritis[a]

ARTHRITIS[b] *OR* ENTHESITIS *OR* DACTYLITIS
PLUS EITHER
One or more of the following SpA features:
• Psoriasis
• Crohn's disease or ulcerative colitis
• Preceding infection
• Uveitis
• HLA-B27
• Sacroiliitis on imaging (radiographs or MRI)
OR two or more of the following SpA features:
• Arthritis
• Enthesitis
• Dactylitis
• Inflammatory back pain ever
• Family history for SpA

[a]Sensitivity 78%, specificity 82%. [b]Peripheral arthritis, usually predominantly lower limb and/or asymmetric. The various SpA features are as defined in Table 355-1. Preceding infection refers to preceding gastrointestinal or urogenital infection.

Source: M Rudawaleit et al: Ann Rheum Dis 70:25, 2011.

ENTEROPATHIC ARTHRITIS

HISTORIC BACKGROUND

A relationship between arthritis and IBD was observed in the 1930s. The relationship was further defined by epidemiologic studies in the 1950s and 1960s and included in the concept of the spondyloarthritides in the 1970s.

EPIDEMIOLOGY

Both of the common forms of IBD, ulcerative colitis (UC) and Crohn's disease (CD) **(Chap. 319)**, are associated with SpA. UC and CD both have an estimated prevalence of 0.1–0.2%, and the incidence of each is thought to have increased in recent decades. AS, nr-ax-SpA, and peripheral arthritis are all associated with UC and CD. Wide variations have been reported in the estimated frequencies of these associations. In recent series, AS was diagnosed in 1–10%, and peripheral arthritis in 10–50% of patients with IBD. IBP and enthesopathy are common, and many patients have sacroiliitis on imaging studies.

The prevalence of UC or CD in patients with AS is thought to be 5–10%, and a recent meta-analysis found the prevalence in patients with nr-ax-SpA to be 6.4%. However, investigation of unselected SpA patients by ileocolonoscopy has revealed that from one-third to two-thirds of patients with AS have subclinical intestinal inflammation that is evident either macroscopically or histologically. These lesions have also been found in patients with undifferentiated SpA or ReA (both enterically and urogenitally acquired).

Both UC and CD have a tendency to familial aggregation, more so for CD. HLA associations have been weak and inconsistent. HLA-B27 is found in up to 70% of patients with IBD and AS, but in ≤15% of patients with IBD and peripheral arthritis or IBD alone. Three alleles of the *NOD2/CARD15* gene on chromosome 16 have been found in approximately one-half of patients with CD. These alleles are not associated with SpA per se. However, they are found significantly more often in (1) CD patients with sacroiliitis than in those without sacroiliitis, and (2) SpA patients with chronic inflammatory gut lesions than in those with normal gut histology. These associations are independent of HLA-B27. In addition to *NOD2*, over 200 other genes have been found to be associated with CD, UC, or both. Many of the SNPs associated with AS are also associated with IBD.

PATHOLOGY

Available data for IBD-associated peripheral arthritis suggest a synovial histology similar to other spondyloarthritides. Association with arthropathy does not affect the gut histology of UC or CD **(Chap. 319)**. The subclinical inflammatory lesions in the colon and distal ileum associated with SpA have been classified as either acute or chronic. The former resemble acute bacterial enteritis, with largely intact architecture and neutrophilic infiltration in the lamina propria. The latter resemble the lesions of CD, with distortion of villi and crypts, aphthoid ulceration, and mononuclear cell infiltration in the lamina propria.

PATHOGENESIS

Both IBD and SpA are immune-mediated, but the specific pathogenic mechanisms are poorly understood, and the connection between the two is obscure. The shared genetics evidently reflects shared pathogenic mechanisms. A number of rodent models showing various immune perturbations manifest both IBD and arthritis. Resident innate immune cells and intestinal dysbiosis have been implicated in both conditions. Several lines of evidence indicate trafficking of leukocytes between the gut and the joint. Mucosal leukocytes from IBD patients have been shown to bind avidly to synovial vasculature through several different adhesion molecules. Macrophages expressing CD163 are prominent in the inflammatory lesions of both gut and synovium in the spondyloarthritides.

CLINICAL FEATURES

AS associated with IBD is clinically indistinguishable from idiopathic AS. It runs a course independent of the bowel disease, and in some patients, it precedes the onset of IBD, sometimes by many years. Peripheral arthritis may also begin before onset of overt bowel disease. The spectrum of peripheral arthritis includes acute self-limited attacks of oligoarthritis that often coincide with relapses of IBD, and more chronic and symmetric polyarticular arthritis that runs a course independent of IBD activity. The patterns of joint involvement are similar in UC and CD. In general, erosions and deformities are infrequent in IBD-associated peripheral arthritis. Isolated destructive hip arthritis is a rare complication of CD, apparently distinct from osteonecrosis and septic arthritis. Dactylitis and enthesopathy are occasionally found. In addition to the ~20% of IBD patients with SpA, a comparable percentage have arthralgias or fibromyalgia symptoms.

Other extraintestinal manifestations of IBD are seen in addition to arthropathy, including uveitis, pyoderma gangrenosum, erythema nodosum, and finger clubbing, all somewhat more commonly in CD than UC. The uveitis shares the features described above for PsA-associated uveitis.

LABORATORY AND RADIOGRAPHIC FINDINGS

Laboratory findings reflect the inflammatory and metabolic manifestations of IBD. Joint fluid is usually at least mildly inflammatory. Of patients with AS and IBD, 30–70% carry the HLA-B27 gene, compared with 75–90% of patients with AS alone and 50–70% of those with AS and psoriasis. Hence, definite or probable AS in a B27-negative individual in the absence of psoriasis should prompt a search for occult IBD. Radiographic changes in the axial skeleton are the same as in uncomplicated AS. Erosions are uncommon in peripheral arthritis but may occur, particularly in the metatarsophalangeal joints.

DIAGNOSIS

Diarrhea and arthritis are both common conditions that can coexist for a variety of reasons. When etiopathogenically related, ReA and IBD-associated arthritis are the most common causes. Rare causes include celiac disease, blind loop syndromes, and Whipple's disease. In most cases, diagnosis depends on investigation of the bowel disease.

TREATMENT

Enteropathic Arthritis

Treatment of CD has been improved by therapy with anti-TNF agents. Infliximab, adalimumab, and certolizumab pegol are effective for induction and maintenance of clinical remission in CD, and infliximab has been shown to be effective in fistulizing CD. IBD-associated arthritis also responds to these agents. Other treatment for IBD, including sulfasalazine and related drugs, systemic glucocorticoids, and immunosuppressive drugs, is also usually of benefit for associated peripheral arthritis. NSAIDs are generally helpful and well tolerated, but they can precipitate flares of IBD. As noted above for psoriasis, rare cases of IBD, either CD or UC, have apparently been precipitated by anti-TNF therapy, usually etanercept, given for any of several rheumatic diseases.

SAPHO SYNDROME

The syndrome of synovitis, acne, pustulosis, hyperostosis, and osteitis (SAPHO) is characterized by a variety of skin and musculoskeletal manifestations. Dermatologic manifestations include palmoplantar pustulosis, acne conglobata, acne fulminans, and hidradenitis suppurativa. The main musculoskeletal findings are sternoclavicular and spinal hyperostosis, chronic recurrent foci of sterile osteomyelitis, and axial or peripheral arthritis. Cases with one or a few manifestations are probably the rule. The ESR and/or CRP are usually mildly to moderately elevated, occasionally dramatically. In some cases, bacteria, most often *Propionibacterium acnes*, have been cultured from bone biopsy specimens and occasionally other sites. IBD was coexistent in 8% of patients in one large series. B27 is not associated. Either bone scan or computed tomography scan is helpful diagnostically. An MRI report described characteristic vertebral body corner cortical erosions in 12 of 12 patients. High-dose NSAIDs may provide relief from bone pain. A number of uncontrolled series and case reports describe successful therapy with pamidronate or other bisphosphonates. Response to anti-TNF-α therapy has also been observed, although in a few cases this has been associated with a flare of skin manifestations. Successful prolonged antibiotic therapy has also been reported. Recent reports suggest a possible autoinflammatory pathogenesis and successful treatment with the IL-1 receptor antagonist anakinra.

FURTHER READING

Bowness P: HLA-B27. Annu Rev Immunol 33:29, 2015.

Carter JD, Hudson AP: Recent advances and future directions in understanding and treating Chlamydia-induced reactive arthritis. Expert Rev Clin Immunol. 13:197, 2017.

de Winter JJ et al: Prevalence of peripheral and extra-articular disease in ankylosing spondylitis versus non-radiographic axial spondyloarthritis: A meta-analysis. Arthritis Res Ther 18:196, 2016.

Ellinghaus D et al: Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. Nat Genet 48:510, 2016.

Gladman DD: Editorial: What is peripheral spondyloarthritis? Arthritis Rheumatol 67:865, 2015.

Gmuca S, Weiss PF: Juvenile spondyloarthritis. Curr Opin Rheumatol 27:364, 2015.

Maksymowych WP: The role of MRI in the evaluation of spondyloarthritis: A clinician's guide. Clin Rheumatol 35:1447, 2016.

Reinhardt A et al: Interleukin-23-dependent gamma/delta T cells produce interleukin-17 and accumulate in the enthesis, aortic valve, and ciliary body in mice. Arthritis Rheumatol 68:2476, 2016.

Ritchlin CT et al: Psoriatic arthritis. N Engl J Med 376:957, 2017.

Sieper J, et al: Secukinumab efficacy in anti-TNF-naive and anti-TNF-experienced subjects with active ankylosing spondylitis: Results from the MEASURE 2 Study. Ann Rheum Dis 76:571, 2017.

Sieper J, Poddubnyy D: Axial spondyloarthritis. Lancet 390:73, 2017.

Taurog JD et al: Ankylosing spondylitis and axial spondyloarthritis. New Engl J Med 374:2563, 2016.

Ward MM et al: American College of Rheumatology/Spondylitis Association of America/Spondyloarthritis Research and Treatment Network 2015 recommendations for the treatment of ankylosing spondylitis and nonradiographic axial spondyloarthritis. Arthritis Rheumatol 68:282, 2016.

356 The Vasculitis Syndromes

Carol A. Langford, Anthony S. Fauci

DEFINITION

Vasculitis is a clinicopathologic process characterized by inflammation of and damage to blood vessels. The vessel lumen is usually compromised, and this is associated with ischemia of the tissues supplied by the involved vessel. A broad and heterogeneous group of syndromes may result from this process, since any type, size, and location of blood vessel may be involved. Vasculitis and its consequences may be the primary or sole manifestation of a disease; alternatively, vasculitis may be a secondary component of another disease. Vasculitis may be confined to a single organ, such as the skin, or it may simultaneously involve several organ systems.

CLASSIFICATION

A major feature of the vasculitic syndromes as a group is the fact that there is a great deal of heterogeneity at the same time as there is considerable overlap among them. **Table 356-1** lists the major vasculitis syndromes. The distinguishing and overlapping features of these syndromes are discussed below.

TABLE 356-1 Vasculitis Syndromes

PRIMARY VASCULITIS SYNDROMES	SECONDARY VASCULITIS SYNDROMES
Granulomatosis with polyangiitis (Wegener's)	Vasculitis associated with probable etiology
Microscopic polyangiitis	Drug-induced vasculitis
Eosinophilic granulomatosis with polyangiitis (Churg-Strauss)	Hepatitis C virus–associated cryoglobulinemic vasculitis
IgA vasculitis (Henoch-Schönlein)	Hepatitis B virus–associated vasculitis
Cryoglobulinemic vasculitis	Cancer-associated vasculitis
Polyarteritis nodosa	Vasculitis associated with systemic disease
Kawasaki disease	Lupus vasculitis
Giant cell arteritis	Rheumatoid vasculitis
Takayasu arteritis	Sarcoid vasculitis
Behçet's disease	
Cogan's syndrome	
Single-organ vasculitis	
Cutaneous leukocytoclastic angiitis	
Cutaneous arteritis	
Primary central nervous system vasculitis	
Isolated aortitis	

Source: Adapted from JC Jennette et al: Arthritis Rheum 65:1, 2013.

PATHOPHYSIOLOGY AND PATHOGENESIS

Generally, most of the vasculitic syndromes are assumed to be mediated at least in part by immunopathogenic mechanisms that occur in response to certain antigenic stimuli. However, evidence supporting this hypothesis is for the most part indirect and may reflect epiphenomena as opposed to true causality. Furthermore, it is unknown why some individuals might develop vasculitis in response to certain antigenic stimuli, whereas others do not. It is likely that a number of factors are involved in the ultimate expression of a vasculitic syndrome. These include the genetic predisposition, environmental exposures, and the regulatory mechanisms associated with immune response to certain antigens. Although immune complex formation, antineutrophil cytoplasmic antibodies (ANCAs), and pathogenic T lymphocyte responses **(Table 356-2)** have been among the prominent hypothesized mechanisms, it is likely that the pathogenesis of individual forms of vasculitis is complex and varied.

PATHOGENIC IMMUNE-COMPLEX FORMATION

Deposition of immune complexes was the first and most widely accepted pathogenic mechanism of vasculitis. However, the causal role of immune complexes has not been clearly established in most of the vasculitic syndromes. Circulating immune complexes need not result in deposition of the complexes in blood vessels with ensuing vasculitis, and many patients with active vasculitis do not have demonstrable circulating or deposited immune complexes. The actual antigen contained in the immune complex has only rarely been identified in vasculitic syndromes. In this regard, hepatitis B antigen has been identified in both the circulating and deposited immune complexes in a subset of patients who have features of a systemic vasculitis, most notably in polyarteritis nodosa (see "Polyarteritis Nodosa"). Cryoglobulinemic vasculitis is strongly associated with hepatitis C virus infection; hepatitis C virions and hepatitis C virus antigen-antibody complexes have been identified in the cryoprecipitates of these patients (see "Cryoglobulinemic Vasculitis").

The mechanisms of tissue damage in immune complex–mediated vasculitis resemble those described for serum sickness. In this model, antigen-antibody complexes are formed in antigen excess and are deposited in vessel walls whose permeability has been increased by vasoactive amines such as histamine, bradykinin, and leukotrienes released from platelets or from mast cells as a result of IgE-triggered mechanisms. The deposition of complexes results in activation of complement components, particularly C5a, which is strongly chemotactic for neutrophils. These cells then infiltrate the vessel wall, phagocytose the immune complexes, and release their intracytoplasmic enzymes, which damage the vessel wall. As the process becomes subacute or chronic, mononuclear cells infiltrate the vessel wall. The common denominator of the resulting syndrome is compromise of the vessel lumen with ischemic changes in the tissues supplied by the involved vessel. Several variables may explain why only certain types of immune complexes cause vasculitis and why only certain vessels are affected in individual patients. These include the ability of the reticuloendothelial system to clear circulating complexes from the blood, the size and physicochemical properties of immune complexes, the relative degree of turbulence of blood flow, the intravascular hydrostatic pressure in different vessels, and the preexisting integrity of the vessel endothelium.

TABLE 356-2 Potential Mechanisms of Vessel Damage in Vasculitis Syndromes

Pathogenic immune-complex formation and/or deposition
IgA vasculitis (Henoch-Schönlein)
Lupus vasculitis
Serum sickness and cutaneous vasculitis syndromes
Hepatitis C virus–associated cryoglobulinemic vasculitis
Hepatitis B virus–associated vasculitis
Production of antineutrophilic cytoplasmic antibodies
Granulomatosis with polyangiitis (Wegener's)
Microscopic polyangiitis
Eosinophilic granulomatosis with polyangiitis (Churg-Strauss)
Pathogenic T lymphocyte responses and granuloma formation
Giant cell arteritis
Takayasu arteritis
Granulomatosis with polyangiitis (Wegener's)
Eosinophilic granulomatosis with polyangiitis (Churg-Strauss)

Source: Adapted from MC Sneller, AS Fauci: Med Clin North Am 81:221, 1997.

ANTINEUTROPHIL CYTOPLASMIC ANTIBODIES

Antineutrophil cytoplasmic antibodies (ANCA) are antibodies directed against certain proteins in the cytoplasmic granules of neutrophils and monocytes. These autoantibodies are present in a high percentage of patients with active granulomatosis with polyangiitis (Wegener's) and microscopic polyangiitis, and in a lower percentage of patients with eosinophilic granulomatosis with polyangiitis (Churg-Strauss). Because these diseases share the presence of ANCA and small-vessel vasculitis, some investigators have come to refer to them collectively as "ANCA-associated vasculitis." However, as these diseases possess unique clinical phenotypes in which ANCA may be absent, it remains our opinion that granulomatosis with polyangiitis (Wegener's), microscopic polyangiitis and eosinophilic granulomatosis with polyangiitis (Churg-Strauss) should continue to be viewed as separate entities.

There are two major categories of ANCA based on different targets for the antibodies. The terminology of *cytoplasmic ANCA* (cANCA) refers to the diffuse, granular cytoplasmic staining pattern observed by immunofluorescence microscopy when serum antibodies bind to indicator neutrophils. Proteinase-3, a 29-kDa neutral serine proteinase present in neutrophil azurophilic granules, is the major cANCA antigen. More than 90% of patients with typical active granulomatosis with polyangiitis (Wegener's) have detectable antibodies to proteinase-3 (see below). The terminology of *perinuclear ANCA* (pANCA) refers to the more localized perinuclear or nuclear staining pattern of the indicator neutrophils. The major target for pANCA is the enzyme myeloperoxidase; other targets that can produce a pANCA pattern of staining include elastase, cathepsin G, lactoferrin, lysozyme, and bactericidal/permeability-increasing protein. However, only antibodies to myeloperoxidase have been convincingly associated with vasculitis. Antimyeloperoxidase antibodies have been reported to occur in variable percentages of patients with microscopic polyangiitis, eosinophilic granulomatosis with polyangiitis (Churg-Strauss), isolated necrotizing crescentic glomerulonephritis, and granulomatosis with polyangiitis (Wegener's) (see below). A pANCA pattern of staining that is not due

to antimyeloperoxidase antibodies has been associated with nonvasculitic entities such as rheumatic and nonrheumatic autoimmune diseases, inflammatory bowel disease, certain drugs, and infections such as endocarditis and bacterial airway infections in patients with cystic fibrosis.

It is unclear why patients with these vasculitis syndromes develop antibodies to myeloperoxidase or proteinase-3 or what role these antibodies play in disease pathogenesis. There are a number of in vitro observations that suggest possible mechanisms whereby these antibodies can contribute to the pathogenesis of the vasculitis syndromes. Proteinase-3 and myeloperoxidase reside in the azurophilic granules and lysosomes of resting neutrophils and monocytes, where they are apparently inaccessible to serum antibodies. However, when neutrophils or monocytes are primed by tumor necrosis factor α (TNF-α) or interleukin 1 (IL-1), proteinase-3 and myeloperoxidase translocate to the cell membrane, where they can interact with extracellular ANCA. The neutrophils then degranulate and produce reactive oxygen species that can cause tissue damage. Furthermore, ANCA-activated neutrophils can adhere to and kill endothelial cells in vitro. Activation of neutrophils and monocytes by ANCA also induces the release of proinflammatory cytokines such as IL-1 and IL-8. Adoptive transfer experiments in genetically engineered mice provide further evidence for a direct pathogenic role of ANCA in vivo. In contradiction, however, a number of clinical and laboratory observations argue against a primary pathogenic role for ANCA. Patients may have active granulomatosis with polyangiitis (Wegener's) in the absence of ANCA; the absolute height of the antibody titers does not correlate well with disease activity; and patients with granulomatosis with polyangiitis (Wegener's) in remission may continue to have high antiproteinase-3 (cANCA) titers for years (see below).

PATHOGENIC T LYMPHOCYTE RESPONSES AND GRANULOMA FORMATION

The histopathologic feature of granulomatous vasculitis has provided evidence to support a role of pathogenic T lymphocyte responses and cell-mediated immune injury. Vascular endothelial cells can express human leukocyte antigen (HLA) class II molecules following activation by cytokines such as interferon (IFN) γ. This allows these cells to participate in immunologic reactions such as interaction with CD4+ T lymphocytes in a manner similar to antigen-presenting macrophages. Endothelial cells can secrete IL-1, which may activate T lymphocytes and initiate or propagate in situ immunologic processes within the blood vessel. In addition, IL-1 and TNF-α are potent inducers of endothelial-leukocyte adhesion molecule 1 (ELAM-1) and vascular cell adhesion molecule 1 (VCAM-1), which may enhance the adhesion of leukocytes to endothelial cells in the blood vessel wall.

APPROACH TO THE PATIENT

General Principles of Diagnosis

The diagnosis of vasculitis should be considered in any patient with an unexplained systemic illness. However, there are certain clinical abnormalities that when present alone or in combination should suggest a diagnosis of vasculitis. These include palpable purpura, pulmonary infiltrates and microscopic hematuria, chronic inflammatory sinusitis, mononeuritis multiplex, unexplained ischemic events, and glomerulonephritis with evidence of multisystem disease. A number of nonvasculitic diseases may also produce some or all of these abnormalities. Thus, the first step in the workup of a patient with suspected vasculitis is to exclude other diseases that produce clinical manifestations that can mimic vasculitis **(Table 356-3)**. It is particularly important to exclude infectious diseases with features that overlap those of vasculitis, especially if the patient's clinical condition is deteriorating rapidly and empirical immunosuppressive treatment is being contemplated.

Once diseases that mimic vasculitis have been excluded, the workup should follow a series of progressive steps that establish the diagnosis of vasculitis and determine, where possible, the category of the vasculitis syndrome **(Fig. 356-1)**. This approach is of considerable importance since several of the vasculitis syndromes require aggressive therapy with glucocorticoids and other immunosuppressive agents, whereas other syndromes usually resolve spontaneously and require symptomatic treatment only. The definitive diagnosis of vasculitis is usually made based on biopsy of involved tissue. The yield of "blind" biopsies of organs with no subjective or objective evidence of involvement is very low and should be avoided. When syndromes such as polyarteritis nodosa, Takayasu's arteritis, or primary central nervous system (CNS) vasculitis are suspected, arteriogram of organs with suspected involvement should be performed.

TABLE 356-3 Conditions That Can Mimic Vasculitis

Infectious Diseases
Bacterial endocarditis
Disseminated gonococcal infection
Pulmonary histoplasmosis
Coccidioidomycosis
Syphilis
Lyme disease
Rocky Mountain spotted fever
Whipple's disease
Coagulopathies/Thrombotic Microangiopathies
Antiphospholipid syndrome
Thrombotic thrombocytopenic purpura
Neoplasms
Atrial myxoma
Lymphoma
Carcinomatosis
Drug Toxicity
Cocaine
Levamisole
Amphetamines
Ergot alkaloids
Methysergide
Arsenic
Other
Sarcoidosis
Atheroembolic disease
Antiglomerular basement membrane disease (Goodpasture's syndrome)
Amyloidosis
Migraine
Fibromuscular dysplasia
Heritable disorders of connective tissue
Segmental arterial mediolysis (SAM)
Reversible cerebral vasoconstrictive syndrome

GENERAL PRINCIPLES OF TREATMENT

Once a diagnosis of vasculitis has been established, a decision regarding therapeutic strategy must be made (Fig. 356-1). If an offending antigen that precipitates the vasculitis is recognized, the antigen should be removed where possible. If the vasculitis is associated with an underlying disease such as an infection, neoplasm, or connective tissue disease, the underlying disease should be treated. If the syndrome represents a primary vasculitic disease, treatment should be initiated according to the category of the vasculitis syndrome. Specific therapeutic regimens are discussed below for the individual vasculitis syndromes; however, certain general principles regarding therapy should be considered. Decisions regarding treatment should be based on the use of regimens for which there has been published literature supporting efficacy for that particular vasculitic disease. Since the potential toxic side effects of certain therapeutic regimens may be substantial, the risk-versus-benefit ratio of any therapeutic approach should be weighed carefully. On

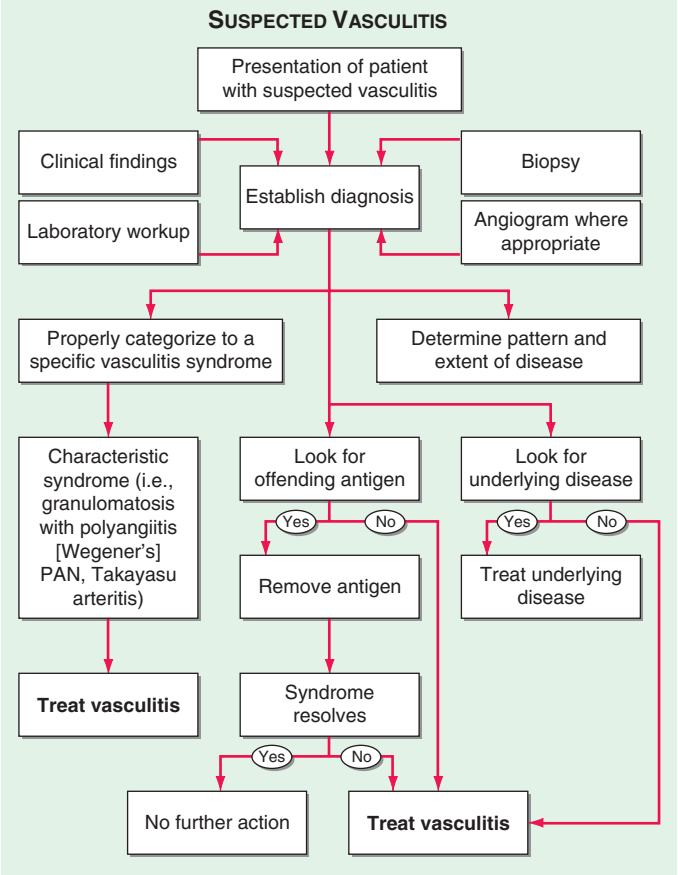


FIGURE 356-1 Algorithm for the approach to a patient with suspected diagnosis of vasculitis. PAN, polyarteritis nodosa.

the one hand, glucocorticoids and/or other immunosuppressive agents should be instituted immediately in diseases where irreversible organ system dysfunction and high morbidity and mortality rates have been clearly established. Granulomatosis with polyangiitis (Wegener's) is the prototype of a severe systemic vasculitis requiring such a therapeutic approach (see below). On the other hand, when feasible, aggressive therapy should be avoided for vasculitic manifestations that rarely result in irreversible organ system dysfunction and that usually do not respond to such therapy. For example, isolated idiopathic cutaneous vasculitis usually resolves with symptomatic treatment, and prolonged courses of glucocorticoids uncommonly result in clinical benefit. Cytotoxic agents have not proved to be beneficial in idiopathic cutaneous vasculitis, and their toxic side effects generally outweigh any potential beneficial effects. Glucocorticoids should be initiated in those systemic vasculitides that cannot be specifically categorized or for which there is no established standard therapy, or other immunosuppressive therapy should be added in these diseases only if an adequate response does not result or if remission can only be achieved and maintained with an unacceptably toxic regimen of glucocorticoids. When remission is achieved, one should continually attempt to taper glucocorticoids and discontinue when possible. When using other immunosuppressive regimens, one should base the choice of agent upon the available therapeutic data supporting efficacy in that disease, the site and severity of organ involvement, and the toxicity profile of the drug.

Physicians should be thoroughly aware of the toxic side effects of therapeutic agents employed that can include both acute and long-term complications (Table 356-4). Morbidity and mortality can occur as a result of treatment, and strategies to monitor for and prevent toxicity represent an essential part of patient care. Glucocorticoids are an important part of treatment for most vasculitides but are associated with substantial toxicities. Monitoring and prevention of glucocorticoid-induced bone loss are important in all patients.

TABLE 356-4 Major Toxic Side Effects of Drugs Used in the Treatment of Systemic Small-Vessel Vasculitis

Glucocorticoids	
Osteoporosis	Growth suppression in children
Cataracts	Hypertension
Glaucoma	Avascular necrosis of bone
Diabetes mellitus	Myopathy
Electrolyte abnormalities	Alterations in mood
Metabolic abnormalities	Psychosis
Suppression of inflammatory and immune responses leading to opportunistic infections	Pseudotumor cerebri
	Peptic ulcer diathesis
Cushingoid features	Pancreatitis
Cyclophosphamide	
Bone marrow suppression	Hypogammaglobulinemia
Cystitis	Pulmonary fibrosis
Bladder carcinoma	Myelodysplasia
Gonadal suppression	Oncogenesis
Gastrointestinal intolerance	Teratogenicity
	Opportunistic infections
Methotrexate	
Gastrointestinal intolerance	Pneumonitis
Stomatitis	Teratogenicity
Bone marrow suppression	Opportunistic infections
Hepatotoxicity (may lead to fibrosis or cirrhosis)	
Azathioprine	
Gastrointestinal intolerance	Opportunistic infections
Bone marrow suppression	Hypersensitivity
Hepatotoxicity	
Rituximab	
Infusion reactions	Opportunistic infections
Progressive multifocal leuko-encephalopathy	Hepatitis B reactivation
	Tumor lysis syndrome
Mucocutaneous reactions	Late-onset neutropenia

With the use of daily cyclophosphamide, strategies are particularly important and are directed toward minimization of bladder toxicity and prevention of leukopenia. Instructing the patient to take cyclophosphamide all at once in the morning with a large amount of fluid throughout the day in order to maintain dilute urine can reduce the risk of bladder injury. Bladder cancer can occur several years after discontinuation of cyclophosphamide therapy; therefore, monitoring for bladder cancer should continue indefinitely in patients who have received cyclophosphamide. Bone marrow suppression is an important toxicity of cyclophosphamide and can be observed during glucocorticoid tapering or over time, even after periods of stable measurements. Monitoring of the complete blood count every 1–2 weeks for as long as the patient receives cyclophosphamide can effectively prevent cytopenias. Maintaining the white blood cell (WBC) count at >3000/μL and the neutrophil count at >1500/μL is essential to reduce the risk of life-threatening infections.

Methotrexate and azathioprine are also associated with bone marrow suppression, and complete blood counts should be obtained every 1–2 weeks for the first 1–2 months after their initiation and once a month thereafter. To lessen toxicity, methotrexate is often given together with folic acid, 1 mg daily, or folinic acid, 5–10 mg once a week 24 h following methotrexate. Prior to initiation of azathioprine, thiopurine methyltransferase (TPMT), an enzyme involved in the metabolism of azathioprine, should be assayed because inadequate levels may result in severe cytopenia.

Rituximab (anti-CD20) can be associated with infusion reactions. In addition to administering this within in skilled infusion center, these reactions can be lessened by the use of pre-medications. There

is a risk of hepatitis B reactivation with rituximab such that all patients should be screened for this infection prior to its use.

Infection represents a significant toxicity for all vasculitis patients treated with immunosuppressive therapy. Infections with *Pneumocystis jirovecii* and certain fungi can be seen even in the face of WBCs that are within normal limits, particularly in patients receiving glucocorticoids. All vasculitis patients who are receiving daily glucocorticoids in combination with another immunosuppressive agent should receive trimethoprim-sulfamethoxazole (TMP-SMX) or another prophylactic therapy to prevent *P. jirovecii* infection.

Finally, it should be emphasized that each patient is unique and requires individual decision-making. The above outline should serve as a framework to guide therapeutic approaches; however, flexibility should be practiced to provide maximal therapeutic efficacy with minimal toxic side effects in each patient.

GRANULOMATOSIS WITH POLYANGIITIS (WEGENER'S)

DEFINITION

Granulomatosis with polyangiitis (Wegener's) is a distinct clinicopathologic entity characterized by granulomatous vasculitis of the upper and lower respiratory tracts together with glomerulonephritis. In addition, variable degrees of disseminated vasculitis involving both small arteries and veins may occur.

INCIDENCE AND PREVALENCE

Granulomatosis with polyangiitis (Wegener's) is an uncommon disease with an estimated prevalence of 3 per 100,000. It is extremely rare in blacks compared with whites; the male-to-female ratio is 1:1. The disease can be seen at any age; ~15% of patients are <19 years of age, but only rarely does the disease occur before adolescence; the mean age of onset is ~40 years.

PATHOLOGY AND PATHOGENESIS

The histopathologic hallmarks of granulomatosis with polyangiitis (Wegener's) are necrotizing vasculitis of small arteries and veins together with granuloma formation, which may be either intravascular or extravascular **(Fig. 356-2)**. Lung involvement typically appears as multiple, bilateral, nodular cavitary infiltrates **(Fig. 356-3)**, which on biopsy almost invariably reveal the typical necrotizing granulomatous vasculitis. Upper airway lesions, particularly those in the sinuses and nasopharynx, typically reveal inflammation, necrosis, and granuloma formation, with or without vasculitis.

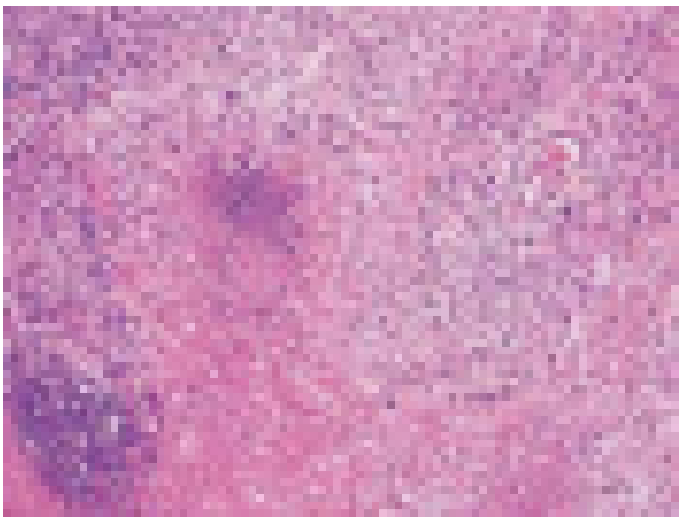


FIGURE 356-2 Lung histology in granulomatosis with polyangiitis (Wegener's). This area of geographic necrosis has a serpiginous border of histiocytes and giant cells surrounding a central necrotic zone. Vasculitis is also present with neutrophils and lymphocytes infiltrating the wall of a small arteriole (*upper right*). *(Courtesy of William D. Travis, MD; with permission.)*

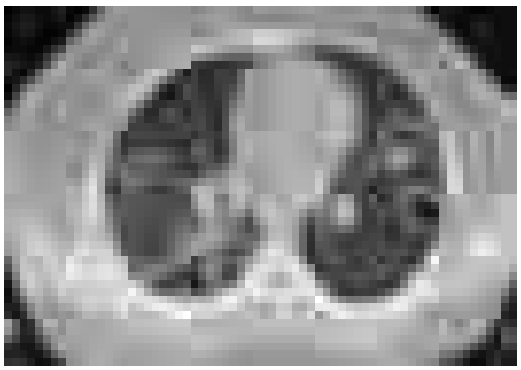


FIGURE 356-3 Computed tomography scan of a patient with granulomatosis with polyangiitis (Wegener's). The patient developed multiple, bilateral, and cavitary infiltrates.

In its earliest form, renal involvement is characterized by a focal and segmental glomerulitis that may evolve into a rapidly progressive crescentic glomerulonephritis. Granuloma formation is only rarely seen on renal biopsy. In contrast to other forms of glomerulonephritis, evidence of immune complex deposition is not found in the renal lesion of granulomatosis with polyangiitis (Wegener's). In addition to the classic triad of disease of the upper and lower respiratory tracts and kidney, virtually any organ can be involved with vasculitis, granuloma, or both.

The immunopathogenesis of this disease is unclear, although the involvement of upper airways and lungs with granulomatous vasculitis suggests an aberrant cell-mediated immune response to an exogenous or even endogenous antigen that enters through or resides in the upper airway. Chronic nasal carriage of *Staphylococcus aureus* has been reported to be associated with a higher relapse rate of granulomatosis with polyangiitis (Wegener's); however, there is no evidence for a role of this organism in the pathogenesis of the disease.

Peripheral blood mononuclear cells obtained from patients with granulomatosis with polyangiitis (Wegener's) manifest increased secretion of IFN-γ but not of IL-4, IL-5, or IL-10 compared to normal controls. In addition, TNF-α production from peripheral blood mononuclear cells and CD4+ T cells is elevated. Furthermore, monocytes from patients with granulomatosis with polyangiitis (Wegener's) produce increased amounts of IL-12. These findings indicate an unbalanced T_H1-type T cell cytokine pattern in this disease that may have pathogenic and perhaps ultimately therapeutic implications.

A high percentage of patients with granulomatosis with polyangiitis (Wegener's) develop ANCA, and these autoantibodies may play a role in the pathogenesis of this disease (see above).

CLINICAL AND LABORATORY MANIFESTATIONS

Involvement of the upper airways occurs in 95% of patients with granulomatosis with polyangiitis (Wegener's). Patients often present with severe upper respiratory tract findings such as paranasal sinus pain and drainage and purulent or bloody nasal discharge, with or without nasal mucosal ulceration **(Table 356-5)**. Nasal septal perforation may follow, leading to saddle nose deformity. Serous otitis media may occur as a result of eustachian tube blockage. Subglottic tracheal stenosis resulting from active disease or scarring occurs in ~16% of patients and may result in severe airway obstruction.

Pulmonary involvement may be manifested as asymptomatic infiltrates or may be clinically expressed as cough, hemoptysis, dyspnea, and chest discomfort. It is present in 85–90% of patients. Endobronchial disease, either in its active form or as a result of fibrous scarring, may lead to obstruction with atelectasis.

Eye involvement (52% of patients) may range from a mild conjunctivitis to dacryocystitis, episcleritis, scleritis, granulomatous sclerouveitis, ciliary vessel vasculitis, and retroorbital mass lesions leading to proptosis.

Skin lesions (46% of patients) appear as papules, vesicles, palpable purpura, ulcers, or subcutaneous nodules; biopsy reveals vasculitis,

TABLE 356-5 Granulomatosis with Polyangiitis (Wegener's): Frequency of Clinical Manifestations in 158 Patients Studied at the National Institutes of Health

MANIFESTATION	PERCENTAGE AT DISEASE ONSET	PERCENTAGE THROUGHOUT COURSE OF DISEASE
Kidney		
Glomerulonephritis	18	77
Ear/Nose/Throat	**73**	**92**
Sinusitis	51	85
Nasal disease	36	68
Otitis media	25	44
Hearing loss	14	42
Subglottic stenosis	1	16
Ear pain	9	14
Oral lesions	3	10
Lung	**45**	**85**
Pulmonary infiltrates	25	66
Pulmonary nodules	24	58
Hemoptysis	12	30
Pleuritis	10	28
Eyes		
Conjunctivitis	5	18
Dacryocystitis	1	18
Scleritis	6	16
Proptosis	2	15
Eye pain	3	11
Visual loss	0	8
Retinal lesions	0	4
Corneal lesions	0	1
Iritis	0	2
Other[a]		
Arthralgias/arthritis	32	67
Fever	23	50
Cough	19	46
Skin abnormalities	13	46
Weight loss (>10% body weight)	15	35
Peripheral neuropathy	1	15
Central nervous system disease	1	8
Pericarditis	2	6
Hyperthyroidism	1	3

[a]Fewer than 1% had parotid, pulmonary artery, breast, or lower genitourinary (urethra, cervix, vagina, testicular) involvement.

Source: GS Hoffman et al: Ann Intern Med 116:488, 1992.

granuloma, or both. Cardiac involvement (8% of patients) manifests as pericarditis, coronary vasculitis, or, rarely, cardiomyopathy. Nervous system manifestations (23% of patients) include cranial neuritis, mononeuritis multiplex, or, rarely, cerebral vasculitis and/or granuloma.

Renal disease (77% of patients) generally dominates the clinical picture and, if left untreated, accounts directly or indirectly for most of the mortality rate in this disease. Although it may smolder in some cases as a mild glomerulitis with proteinuria, hematuria, and red blood cell casts, it is clear that once clinically detectable renal functional impairment occurs, rapidly progressive renal failure usually ensues unless appropriate treatment is instituted.

While the disease is active, most patients have nonspecific symptoms and signs such as malaise, weakness, arthralgias, anorexia, and weight loss. Fever may indicate activity of the underlying disease but more often reflects secondary infection, usually of the upper airway.

Characteristic laboratory findings include a markedly elevated erythrocyte sedimentation rate (ESR), mild anemia and leukocytosis, mild hypergammaglobulinemia (particularly of the IgA class), and mildly elevated rheumatoid factor. Thrombocytosis may be seen as an acute-phase reactant. Approximately 90% of patients with active granulomatosis with polyangiitis (Wegener's) have a positive antiproteinase-3 ANCA. However, in the absence of active disease, the sensitivity drops to ~60–70%. A small percentage of patients with granulomatosis with polyangiitis (Wegener's) may have antimyeloperoxidase rather than antiproteinase-3 antibodies, and up to 20% may lack ANCA.

Patients with granulomatosis with polyangiitis (Wegener's) have been found to have an increased incidence of venous thrombotic events. Although routine anticoagulation for all patients is not recommended, a heightened awareness for any clinical features suggestive of deep venous thrombosis or pulmonary emboli is warranted.

DIAGNOSIS

The diagnosis of granulomatosis with polyangiitis (Wegener's) is made by the demonstration of necrotizing granulomatous vasculitis on tissue biopsy in a patient with compatible clinical features. Pulmonary tissue offers the highest diagnostic yield, almost invariably revealing granulomatous vasculitis. Biopsy of upper airway tissue usually reveals granulomatous inflammation with necrosis but may not show vasculitis. Renal biopsy can confirm the presence of pauci-immune glomerulonephritis.

The specificity of a positive antiproteinase-3 ANCA for granulomatosis with polyangiitis (Wegener's) is very high, especially if active glomerulonephritis is present. However, the presence of ANCA should be adjunctive and, with rare exceptions, should not substitute for a tissue diagnosis. False-positive ANCA titers have been reported in certain infectious and neoplastic diseases.

In its typical presentation, the clinicopathologic complex of granulomatosis with polyangiitis (Wegener's) usually provides ready differentiation from other disorders. However, if all the typical features are not present at once, it needs to be differentiated from the other vasculitides, antiglomerular basement membrane disease (Goodpasture's syndrome) **(Chap. 308)**, relapsing polychondritis **(Chap. 359)**, tumors of the upper airway or lung, and infectious diseases such as histoplasmosis **(Chap. 207)**, mucocutaneous leishmaniasis **(Chap. 221)**, and rhinoscleroma **(Chap. 31)** as well as noninfectious granulomatous diseases.

Of particular note is the differentiation from other *midline destructive diseases*. These diseases lead to extreme tissue destruction and mutilation localized to the midline upper airway structures including the sinuses; erosion through the skin of the face commonly occurs, a feature that is extremely rare in granulomatosis with polyangiitis (Wegener's). Although blood vessels may be involved in the intense inflammatory reaction and necrosis, primary vasculitis is not seen. *Upper airway neoplasms* and specifically *extranodal natural killer (NK)/T cell lymphoma (nasal type)* are important causes of midline destructive disease. These lesions are diagnosed based on histology, which reveals polymorphous atypical lymphoid cells with an NK cell immunophenotype, typically Epstein-Barr virus **(Chap. 104)**. Such cases are treated based on their degree of dissemination, and localized lesions have responded to irradiation. Upper airway lesions should never be irradiated in granulomatosis with polyangiitis (Wegener's). Cocaine-induced tissue injury can be another important mimic of granulomatosis with polyangiitis (Wegener's) in patients who present with isolated midline destructive disease. ANCA that target human neutrophil elastase can be found in patients with cocaine-induced midline destructive lesions and can complicate the differentiation from granulomatosis with polyangiitis (Wegener's). This has been further confounded by the high frequency of levamisole adulteration of cocaine, which can result in cutaneous infarction and serologic changes that may mimic vasculitis. Granulocytopenia is a common finding in levamisole-induced disease that would not be associated with granulomatosis with polyangiitis (Wegener's).

Granulomatosis with polyangiitis (Wegener's) must also be differentiated from *lymphomatoid granulomatosis*, which is an Epstein-Barr virus–positive B cell proliferation that is associated with an exuberant T cell reaction. Lymphomatoid granulomatosis is characterized by lung, skin, CNS, and kidney involvement in which atypical lymphocytoid and plasmacytoid cells infiltrate nonlymphoid tissue in an angioinvasive

manner. In this regard, it clearly differs from granulomatosis with polyangiitis (Wegener's) in that it is not an inflammatory vasculitis in the classic sense but an angiocentric perivascular infiltration of atypical mononuclear cells. Up to 50% of patients may develop a true malignant lymphoma.

TREATMENT

Granulomatosis with Polyangiitis (Wegener's)

Prior to the introduction of effective therapy, granulomatosis with polyangiitis (Wegener's) was universally fatal within a few months of diagnosis. Glucocorticoids alone led to some symptomatic improvement, with little effect on the ultimate course of the disease. The development of treatment with cyclophosphamide dramatically changed patient outcome such that marked improvement was seen in >90% of patients, complete remission in 75% of patients, and 5-year patient survival was seen in >80%.

Despite the ability to successfully induce remission, 50–70% of remissions are later associated with one or more relapses. The determination of relapse should be based on objective evidence of disease activity, taking care to rule out other features that may have a similar appearance such as infection, medication toxicity, or chronic disease sequelae. The ANCA titer can be misleading and should not be used to assess disease activity. Many patients who achieve remission continue to have elevated titers for years. Results from a large prospective study found that increases in ANCA were not associated with relapse and that only 43% relapsed within 1 year of an increase in ANCA levels. Thus, a rise in ANCA by itself is not a harbinger of immediate disease relapse and should not lead to reinstitution or increase in immunosuppressive therapy. Reinduction of remission after relapse is almost always achieved; however, a high percentage of patients ultimately have some degree of damage from irreversible features of their disease, such as varying degrees of renal insufficiency, hearing loss, tracheal stenosis, saddle nose deformity, and chronically impaired sinus function. Patients who developed irreversible renal failure but who achieved subsequent remission have undergone successful renal transplantation.

Treatment of granulomatosis with polyangiitis (Wegener's) is currently viewed as having two phases: *induction*, where active disease is put into remission, followed by *maintenance*. The decision regarding which agents to use for induction and maintenance is based on disease severity together with individual patient factors that include contraindication, relapse history, and comorbidities.

CYCLOPHOSPHAMIDE INDUCTION FOR SEVERE DISEASE

For patients with severe disease, daily cyclophosphamide combined with glucocorticoids has been proven to effectively induce remission and prolong survival. At the initiation of therapy, glucocorticoids are usually given as prednisone, 1 mg/kg/d for the first month, followed by gradual tapering on an alternate-day or daily schedule with discontinuation after ~6–9 months.

Cyclophosphamide is given in doses of 2 mg/kg/d orally, but as it is renally eliminated, dosage reduction should be considered in patients with renal insufficiency. Some reports have indicated therapeutic success with less frequent and severe toxic side effects using IV cyclophosphamide. In a randomized trial, IV cyclophosphamide 15 mg/kg, three infusions given every 2 weeks, then every 3 weeks thereafter, was compared to cyclophosphamide 2 mg/kg daily given for 3 months followed by 1.5 mg/kg daily. Although IV cyclophosphamide was found to have a comparable rate of remission with a lower cumulative cyclophosphamide dose and occurrence of leukopenia, the use of a consolidation phase and an insufficient frequency of blood count monitoring may have negatively influenced the results in those who received daily cyclophosphamide. Of note in this study was that relapse occurred in 19% of those who received IV cyclophosphamide as compared to 9% who received daily oral administration. We continue to strongly favor daily rather than intermittent cyclophosphamide with utilization of blood count monitoring every 1–2 weeks (as discussed above) and limiting the duration of induction exposure to 3–6 months.

In patients with imminently life-threatening disease, such as rapidly progressive glomerulonephritis with a creatinine >4.0 mg/dL or pulmonary hemorrhage requiring mechanical ventilation, a regimen of daily cyclophosphamide and glucocorticoids is favored to induce remission. Adjunctive plasmapheresis has been used in fulminant disease, but as its role remains uncertain, this is currently being investigated in an international trial.

RITUXIMAB INDUCTION FOR SEVERE DISEASE

Rituximab is a chimeric monoclonal antibody directed against CD20 present on normal and malignant B lymphocytes that is U.S. Food and Drug Administration (FDA) approved for the treatment of granulomatosis with polyangiitis (Wegener's) and microscopic polyangiitis. In two randomized trials that enrolled ANCA-positive patients with severe active granulomatosis with polyangiitis (Wegener's) or microscopic polyangiitis, rituximab 375 mg/m^2 once a week for 4 weeks in combination with glucocorticoids was found to be as effective as cyclophosphamide with glucocorticoids for inducing disease remission. In the trial that also enrolled patients with relapsing disease, rituximab was found to be statistically superior to cyclophosphamide. Although rituximab does not have the bladder toxicity or infertility concerns, as can occur with cyclophosphamide, in both of the randomized trials, the rate of adverse events was similar in the rituximab and cyclophosphamide arms. In addition, there are no long-term safety data with rituximab in granulomatosis with polyangiitis (Wegener's) or microscopic polyangiitis.

The decision about whether to utilize cyclophosphamide or rituximab for remission induction must be individually based. Factors to consider include the severity of the disease, whether the patient has newly diagnosed or relapsing disease, medication contraindications, and individual patient factors particularly including fertility concerns.

REMISSION MAINTENANCE

The approach to remission maintenance is influenced by a number of elements that include the medication that is used for remission induction, the presence of prior relapses, the disease characteristics, medication contraindications, and individual patient factors. When cyclophosphamide is given for induction, it should be stopped after 3–6 months and switched to another agent for remission maintenance. The agents with which there has been the greatest published experience are methotrexate and azathioprine and most recently rituximab. Methotrexate is administered orally or subcutaneously starting at a dosage of 0.3 mg/kg as a single weekly dose, not to exceed 15 mg/week. If the treatment is well tolerated after 1–2 weeks, the dosage should be increased by 2.5 mg weekly up to a dosage of 20–25 mg/week and maintained at that level. Azathioprine, 2 mg/kg/d, has also proved effective in maintaining remission following induction with daily cyclophosphamide. In a randomized trial comparing methotrexate to azathioprine for remission maintenance, comparable rates of toxicity and relapse were seen. Therefore, the choice of agent is often based on toxicity profile, because methotrexate cannot be given to patients with renal insufficiency or chronic liver disease, as well as on other individual patient factors. In patients who are unable to receive methotrexate or azathioprine or who have relapsed through such treatment, mycophenolate mofetil, 1000 mg twice a day, may also sustain remission following cyclophosphamide induction. Rituximab 500 mg given intravenously every 6 months recently was compared to azathioprine given after intravenous cyclophosphamide induction in a randomized trial. Overall, a lower rate of relapse was observed with rituximab compared to azathioprine. However, the short duration of the trial and the chronic relapsing nature of these diseases continue to raise many questions as to the long-term role of rituximab for maintenance. Nonetheless, these data demonstrate that rituximab is an effective maintenance option which can be considered within the armamentarium.

For patients who receive rituximab for remission induction, the maintenance approach has not yet clearly been established. The options include to clinically observe the patient and retreat with rituximab should a relapse occur, or to pursue maintenance after rituximab with methotrexate, azathioprine, mycophenolate mofetil, or rituximab. Until further data become available, this decision is determined between the patient and physician.

The optimal duration of continuing maintenance therapy is uncertain. In the absence of toxicity, maintenance therapy is usually given for a minimum of 2 years past remission, after which time consideration can be given for tapering over a 6–12 month period until discontinuation. Patients with significant organ damage or a history of relapse may benefit from longer-term continuation of a maintenance agent.

OTHER BIOLOGIC THERAPIES

Etanercept, a dimeric fusion protein containing the 75-kDa TNF receptor bound to human IgG1, was not found to sustain remission when used adjunctively to standard therapy and should not be used in the treatment of granulomatosis with polyangiitis (Wegener's). Abatacept (CTLA4-Ig) was examined in an open-label pilot study of nonsevere relapsing disease with favorable results, but further investigation is needed before application to clinical practice. Blockade of the activity of complement C5a is also being investigated.

METHOTREXATE INDUCTION FOR NONSEVERE DISEASE

For selected patients whose disease is not immediately life threatening, methotrexate together with glucocorticoids given at the dosages described above may be considered as an alternative for induction therapy, which is then continued for maintenance.

TRIMETHOPRIM-SULFAMETHOXAZOLE

Although certain reports have indicated that trimethoprim-sulfamethoxazole (TMP-SMX) may be of benefit in the treatment of granulomatosis with polyangiitis (Wegener's) isolated to the sinonasal tissues, it should never be used alone to treat active granulomatosis with polyangiitis (Wegener's) outside of the upper airway such as in patients with renal or pulmonary disease. In a study examining the effect of TMP-SMX on relapse, decreased relapses were shown only with regard to upper airway disease, and no differences in major organ relapses were observed.

ORGAN-SPECIFIC TREATMENT

Not all manifestations of granulomatosis with polyangiitis (Wegener's) require or respond to immunosuppressive therapy. In managing non–major organ disease, such as that isolated to the sinus, joints, or skin, the risks of treatment should be carefully weighed against the benefits. Treatment with cyclophosphamide is rare if ever justified for the treatment of isolated sinus disease in granulomatosis with polyangiitis (Wegener's). Differentiation of active disease from damage is also important. Subglottic stenosis is an example of a disease manifestation that can often scar and responds optimally to nonmedical intervention rather than systemic immunosuppressive treatment.

MICROSCOPIC POLYANGIITIS

DEFINITION

The term *microscopic polyarteritis* was introduced into the literature by Davson in 1948 in recognition of the presence of glomerulonephritis in patients with polyarteritis nodosa. In 1992, the Chapel Hill Consensus Conference on the Nomenclature of Systemic Vasculitis adopted the term *microscopic polyangiitis* to connote a necrotizing vasculitis with few or no immune complexes affecting small vessels (capillaries, venules, or arterioles). Glomerulonephritis is very common in microscopic polyangiitis, and pulmonary capillaritis often occurs. The absence of granulomatous inflammation in microscopic polyangiitis is said to differentiate it from granulomatosis with polyangiitis (Wegener's).

INCIDENCE AND PREVALENCE

The incidence of microscopic polyangiitis is estimated to be 3–5/100,000. The mean age of onset is ~57 years, and males are slightly more frequently affected than females.

PATHOLOGY AND PATHOGENESIS

The vasculitis seen in microscopic polyangiitis has a predilection to involve capillaries and venules in addition to small- and medium-sized arteries. Immunohistochemical staining reveals a paucity of immunoglobulin deposition in the vascular lesion of microscopic polyangiitis, suggesting that immune-complex formation does not play a role in the pathogenesis of this syndrome. The renal lesion seen in microscopic polyangiitis is identical to that of granulomatosis with polyangiitis (Wegener's). Like granulomatosis with polyangiitis (Wegener's), microscopic polyangiitis is highly associated with the presence of ANCA, which may play a role in pathogenesis of this syndrome (see above).

CLINICAL AND LABORATORY MANIFESTATIONS

Because of its predilection to involve the small vessels, microscopic polyangiitis and granulomatosis with polyangiitis (Wegener's) share similar clinical features. Disease onset may be gradual, with initial symptoms of fever, weight loss, and musculoskeletal pain; however, it is often acute. Glomerulonephritis occurs in at least 79% of patients and can be rapidly progressive, leading to renal failure. Hemoptysis may be the first symptom of alveolar hemorrhage, which occurs in 12% of patients. Other manifestations include mononeuritis multiplex and gastrointestinal tract and cutaneous vasculitis. Upper airway disease and pulmonary nodules are not typically found in microscopic polyangiitis and, if present, suggest granulomatosis with polyangiitis (Wegener's).

Features of inflammation may be seen, including an elevated ESR, anemia, leukocytosis, and thrombocytosis. ANCA are present in 75% of patients with microscopic polyangiitis, with antimyeloperoxidase antibodies being the predominant ANCA associated with this disease.

DIAGNOSIS

The diagnosis is based on histologic evidence of vasculitis or pauci-immune glomerulonephritis in a patient with compatible clinical features of multisystem disease. Although microscopic polyangiitis is strongly ANCA-associated, no studies have as yet established the sensitivity and specificity of ANCA in this disease.

TREATMENT

Microscopic Polyangiitis

The 5-year survival rate for patients with treated microscopic polyangiitis is 74%, with disease-related mortality occurring from alveolar hemorrhage or gastrointestinal, cardiac, or renal disease. Studies on treatment have come from trials that have included patients with granulomatosis with polyangiitis (Wegener's) or microscopic polyangiitis. Currently, the treatment approach for microscopic polyangiitis is the same as is used for granulomatosis with polyangiitis (Wegener's) (see "Granulomatosis with Polyangiitis [Wegener's]" for a detailed description of this therapeutic regimen), and patients with immediately life-threatening disease should be treated with the combination of prednisone and daily cyclophosphamide or rituximab. Disease relapse has been observed in at least 34% of patients. Treatment for such relapses would be similar to that used at the time of initial presentation and based on site and severity of disease.

EOSINOPHILIC GRANULOMATOSIS WITH POLYANGIITIS (CHURG-STRAUSS)

DEFINITION

Eosinophilic granulomatosis with polyangiitis (Churg-Strauss) was described in 1951 by Churg and Strauss and is characterized by asthma, peripheral and tissue eosinophilia, extravascular granuloma formation, and vasculitis of multiple organ systems.

INCIDENCE AND PREVALENCE

Eosinophilic granulomatosis with polyangiitis (Churg-Strauss) is an uncommon disease with an estimated annual incidence of 1–3 per million. The disease can occur at any age with the possible exception of infants. The mean age of onset is 48 years, with a female-to-male ratio of 1.2:1.

PATHOLOGY AND PATHOGENESIS

The necrotizing vasculitis of eosinophilic granulomatosis with polyangiitis (Churg-Strauss) involves small- and medium-sized muscular arteries, capillaries, veins, and venules. A characteristic histopathologic feature of eosinophilic granulomatosis with polyangiitis (Churg-Strauss) is granulomatous reactions that may be present in the tissues or even within the walls of the vessels themselves. These are usually associated with infiltration of the tissues with eosinophils. This process can occur in any organ in the body; lung involvement is predominant, with skin, cardiovascular system, kidney, peripheral nervous system, and gastrointestinal tract also commonly involved. Although the precise pathogenesis of this disease is uncertain, its strong association with asthma and its clinicopathologic manifestations, including eosinophilia, granuloma, and vasculitis, point to aberrant immunologic phenomena.

CLINICAL AND LABORATORY MANIFESTATIONS

Patients with eosinophilic granulomatosis with polyangiitis (Churg-Strauss) often exhibit nonspecific manifestations such as fever, malaise, anorexia, and weight loss, which are characteristic of a multisystem disease. The pulmonary findings in eosinophilic granulomatosis with polyangiitis (Churg-Strauss) clearly dominate the clinical picture with severe asthmatic attacks and the presence of pulmonary infiltrates. Mononeuritis multiplex is the second most common manifestation and occurs in up to 72% of patients. Allergic rhinitis and sinusitis develop in up to 61% of patients and are often observed early in the course of disease. Clinically recognizable heart disease occurs in ~14% of patients and is an important cause of mortality. Skin lesions occur in ~51% of patients and include purpura in addition to cutaneous and subcutaneous nodules. The renal disease in eosinophilic granulomatosis with polyangiitis (Churg-Strauss) is less common and generally less severe than that of granulomatosis with polyangiitis and microscopic polyangiitis.

The characteristic laboratory finding in virtually all patients with eosinophilic granulomatosis with polyangiitis (Churg-Strauss) is a striking eosinophilia, which reaches levels >1000 cells/μL in >80% of patients. Evidence of inflammation as evidenced by elevated ESR, fibrinogen, or α_2-globulins can be found in 81% of patients. The other laboratory findings reflect the organ systems involved. Approximately 48% of patients with eosinophilic granulomatosis with polyangiitis (Churg-Strauss) have circulating ANCA that is usually antimyeloperoxidase.

DIAGNOSIS

Although the diagnosis of eosinophilic granulomatosis with polyangiitis (Churg-Strauss) is optimally made by biopsy in a patient with the characteristic clinical manifestations (see above), histologic confirmation can be challenging because the pathognomonic features often do not occur simultaneously. In order to be diagnosed with eosinophilic granulomatosis with polyangiitis (Churg-Strauss), a patient should have evidence of asthma, peripheral blood eosinophilia, and clinical features consistent with vasculitis.

TREATMENT

Eosinophilic Granulomatosis with Polyangiitis (Churg-Strauss)

The prognosis of untreated eosinophilic granulomatosis with polyangiitis (Churg-Strauss) is poor, with a reported 5-year survival of 25%. With treatment, prognosis is favorable, with one study finding a 78-month actuarial survival rate of 72%. Myocardial involvement is the most frequent cause of death and is responsible for 39% of patient mortality. Echocardiography should be performed in all newly diagnosed patients because this may influence therapeutic decisions.

Glucocorticoids alone appear to be effective in many patients. Dosage tapering is often limited by asthma, and many patients require low-dose prednisone for persistent asthma many years after clinical recovery from vasculitis. In patients who present with fulminant multisystem disease, particularly cardiac involvement, the treatment of choice is a combined regimen of daily cyclophosphamide and prednisone followed by azathioprine or methotrexate (see "Granulomatosis with Polyangiitis [Wegener's]" for a detailed description of this therapeutic regimen).

Mepolizumab (anti-IL-5 antibody) was studied in a randomized trial and found to be more effective than placebo in eosinophilic granulomatosis with polyangiitis (Churg-Strauss). Patients with life-threatening eosinophilic granulomatosis with polyangiitis (Churg-Strauss) were excluded from the mepolizumab trial and should continue to be treated with cyclophosphamide and glucocorticoids. Mepoluzimab is FDA approved for the treatment of severe eosinophilic asthma and may particularly have a role in the setting of relapsing or resistant asthma in eosinophilic granulomatosis with polyangiitis (Churg-Strauss). Rituximab has been examined only in small retrospective series, primarily in patients who have active disease despite conventional agents or who are intolerant of these medications.

POLYARTERITIS NODOSA

DEFINITION

Polyarteritis nodosa was described in 1866 by Kussmaul and Maier. It is a multisystem, necrotizing vasculitis of small- and medium-sized muscular arteries in which involvement of the renal and visceral arteries is characteristic. Polyarteritis nodosa does not involve pulmonary arteries, although bronchial vessels may be involved; granulomas, significant eosinophilia, and an allergic diathesis are not observed.

INCIDENCE AND PREVALENCE

It is difficult to establish an accurate incidence of polyarteritis nodosa because previous reports have included polyarteritis nodosa and microscopic polyangiitis as well as other related vasculitides. Polyarteritis nodosa, as currently defined, is felt to be a very uncommon disease.

PATHOLOGY AND PATHOGENESIS

The vascular lesion in polyarteritis nodosa is a necrotizing inflammation of small- and medium-sized muscular arteries. The lesions are segmental and tend to involve bifurcations and branchings of arteries. They may spread circumferentially to involve adjacent veins. However, involvement of venules is not seen in polyarteritis nodosa and, if present, suggests microscopic polyangiitis (see below). In the acute stages of disease, polymorphonuclear neutrophils infiltrate all layers of the vessel wall and perivascular areas, which results in intimal proliferation and degeneration of the vessel wall. Mononuclear cells infiltrate the area as the lesions progress to the subacute and chronic stages. Fibrinoid necrosis of the vessels ensues with compromise of the lumen, thrombosis, infarction of the tissues supplied by the involved vessel, and, in some cases, hemorrhage. As the lesions heal, there is collagen deposition, which may lead to further occlusion of the vessel lumen. Aneurysmal dilations up to 1 cm in size along the involved arteries are characteristic of polyarteritis nodosa. Granulomas and substantial eosinophilia with eosinophilic tissue infiltrations are not characteristically found and suggest eosinophilic granulomatosis with polyangiitis (Churg-Strauss) (see above).

Multiple organ systems are involved, and the clinicopathologic findings reflect the degree and location of vessel involvement and the resulting ischemic changes. As mentioned above, pulmonary arteries are not involved in polyarteritis nodosa, and bronchial artery involvement is uncommon. The pathology in the kidney in classic polyarteritis nodosa is that of arteritis without glomerulonephritis. In patients with

significant hypertension, typical pathologic features of glomerulosclerosis may be seen. In addition, pathologic sequelae of hypertension may be found elsewhere in the body.

The presence of a polyarteritis nodosa–like vasculitis in patients with hepatitis B together with the isolation of circulating immune complexes composed of hepatitis B antigen and immunoglobulin and the demonstration by immunofluorescence of hepatitis B antigen, IgM, and complement in the blood vessel walls strongly suggest the role of immunologic phenomena in the pathogenesis of this disease. A polyarteritis nodosa–like vasculitis has also been reported in patients with hepatitis C. Hairy cell leukemia can be associated with polyarteritis nodosa; the pathogenic mechanisms of this association are unclear.

CLINICAL AND LABORATORY MANIFESTATIONS

Nonspecific signs and symptoms are the hallmarks of polyarteritis nodosa. Fever, weight loss, and malaise are present in over one-half of cases. Patients usually present with vague symptoms such as weakness, malaise, headache, abdominal pain, and myalgias that can rapidly progress to a fulminant illness. Specific complaints related to the vascular involvement within a particular organ system may also dominate the presenting clinical picture as well as the entire course of the illness (Table 356-6). In polyarteritis nodosa, renal involvement most commonly manifests as hypertension, renal insufficiency, or hemorrhage due to microaneurysms.

There are no diagnostic serologic tests for polyarteritis nodosa. In >75% of patients, the leukocyte count is elevated with a predominance of neutrophils. Eosinophilia is seen only rarely and, when present at high levels, suggests the diagnosis of eosinophilic granulomatosis with polyangiitis (Churg-Strauss). The anemia of chronic disease may be seen, and an elevated ESR is almost always present. Other common laboratory findings reflect the particular organ involved. Hypergammaglobulinemia may be present, and all patients should be screened for hepatitis B and C. Antibodies against myeloperoxidase or proteinase-3 (ANCA) are rarely found in patients with polyarteritis nodosa.

DIAGNOSIS

The diagnosis of polyarteritis nodosa is based on the demonstration of characteristic findings of vasculitis on biopsy material of involved organs. In the absence of easily accessible tissue for biopsy, the arteriographic demonstration of involved vessels, particularly in the form of aneurysms of small- and medium-sized arteries in the renal, hepatic, and visceral vasculature, is sufficient to make the diagnosis. This should consist of a catheter-directed dye arteriogram because magnetic resonance and computed tomography arteriograms do not have sufficient resolution at the current time to visualize the vessels affected in polyarteritis nodosa. Aneurysms of vessels are not pathognomonic of polyarteritis nodosa; furthermore, aneurysms need not always be present, and arteriographic findings may be limited to stenotic segments and obliteration of vessels. Biopsy of symptomatic organs such as nodular skin lesions, painful testes, and nerve/muscle provides the highest diagnostic yields.

TABLE 356-6 Clinical Manifestations Related to Organ System Involvement in Polyarteritis Nodosa

ORGAN SYSTEM	PERCENT INCIDENCE	CLINICAL MANIFESTATIONS
Renal	60	Renal failure, hypertension
Musculoskeletal	64	Arthritis, arthralgia, myalgia
Peripheral nervous system	51	Peripheral neuropathy, mononeuritis multiplex
Gastrointestinal tract	44	Abdominal pain, nausea and vomiting, bleeding, bowel infarction and perforation, cholecystitis, hepatic infarction, pancreatic infarction
Skin	43	Rash, purpura, nodules, cutaneous infarcts, livedo reticularis, Raynaud's phenomenon
Cardiac	36	Congestive heart failure, myocardial infarction, pericarditis
Genitourinary	25	Testicular, ovarian, or epididymal pain
Central nervous system	23	Cerebral vascular accident, altered mental status, seizure

Source: From TR Cupps, AS Fauci: *The Vasculitides*. Philadelphia, Saunders, 1981.

TREATMENT

Polyarteritis Nodosa

The prognosis of untreated polyarteritis nodosa is extremely poor, with a reported 5-year survival rate between 10 and 20%. Death usually results from gastrointestinal complications, particularly bowel infarcts and perforation, and cardiovascular causes. Intractable hypertension often compounds dysfunction in other organ systems, such as the kidneys, heart, and CNS, leading to additional late morbidity and mortality in polyarteritis nodosa. With the introduction of treatment, survival rate has increased substantially. Favorable therapeutic results have been reported in polyarteritis nodosa with the combination of prednisone and cyclophosphamide (see "Granulomatosis with Polyangiitis [Wegener's]" for a detailed description of this therapeutic regimen). In less severe cases of polyarteritis nodosa, glucocorticoids alone have resulted in disease remission. In patients with hepatitis B who have a polyarteritis nodosa–like vasculitis, antiviral therapy represents an important part of therapy and has been used in combination with glucocorticoids and plasma exchange. Careful attention to the treatment of hypertension can lessen the acute and late morbidity and mortality rates associated with renal, cardiac, and CNS complications of polyarteritis nodosa. Following successful treatment, relapse of polyarteritis nodosa has been estimated to occur in 10–20% of patients.

GIANT CELL ARTERITIS AND POLYMYALGIA RHEUMATICA

DEFINITION

Giant cell arteritis, historically referred to as *temporal arteritis*, is an inflammation of medium- and large-sized arteries. It characteristically involves one or more branches of the carotid artery, particularly the temporal artery. However, it is a systemic disease that can involve arteries in multiple locations, particularly the aorta and its main branches.

Giant cell arteritis is closely associated with *polymyalgia rheumatica*, which is characterized by stiffness, aching, and pain in the muscles of the neck, shoulders, lower back, hips, and thighs. Most commonly, polymyalgia rheumatica occurs in isolation, but it may be seen in 40–50% of patients with giant cell arteritis. In addition, ~10–20% of patients who initially present with features of isolated polymyalgia rheumatica later go on to develop giant cell arteritis. This strong clinical association together with data from pathophysiologic studies has increasingly supported that giant cell arteritis and polymyalgia rheumatica represent differing clinical spectrums of a single disease process.

INCIDENCE AND PREVALENCE

Giant cell arteritis occurs almost exclusively in individuals aged >50 years. It is more common in women than in men and is rare in blacks. The incidence of giant cell arteritis varies widely in different studies and in different geographic regions. A high incidence has been found in Scandinavia and in regions of the United States with large Scandinavian populations, compared to a lower incidence in southern Europe. The annual incidence rates in individuals aged ≥50 years range from 6.9 to 32.8 per 100,000 population. Familial aggregation has been reported, as has an association with HLA-DR4. In addition, genetic linkage studies have demonstrated an association of giant cell arteritis with alleles at the HLA-DRB1 locus, particularly HLA-DRB1*04 variants. In Olmsted County, Minnesota, the annual incidence of polymyalgia rheumatica in individuals aged ≥50 years is 58.7 per 100,000 population.

PATHOLOGY AND PATHOGENESIS

Although the temporal artery is most frequently involved in giant cell arteritis, patients often have a systemic vasculitis of multiple medium- and large-sized arteries, which may go undetected. Histopathologically, the disease is a panarteritis with inflammatory mononuclear cell infiltrates within the vessel wall with frequent giant cell formation. There is proliferation of the intima and fragmentation of the internal elastic lamina. Pathophysiologic findings in organs result from the ischemia related to the involved vessels.

Experimental data support that giant cell arteritis is an antigen-driven disease in which activated T lymphocytes, macrophages, and dendritic cells play a critical role in the disease pathogenesis. Sequence analysis of the T cell receptor of tissue-infiltrating T cells in lesions of giant cell arteritis indicates restricted clonal expansion, suggesting the presence of an antigen residing in the arterial wall. Giant cell arteritis is believed to be initiated in the adventitia where CD4+ T cells enter through the vasa vasorum, become activated, and orchestrate macrophage differentiation. T cells recruited to vasculitic lesions in patients with giant cell arteritis produce predominantly IL-2 and IFN-γ, and the latter has been suggested to be involved in the progression to overt arteritis. Laboratory-based data demonstrate that at least two separate lineages of CD4 T cells—IFN-γ-producing T_H1 cells and IL-17-producing T_H17 cells—participate in vascular inflammation and may have differing levels of responsiveness to glucocorticoids.

CLINICAL AND LABORATORY MANIFESTATIONS

Giant cell arteritis is most commonly characterized clinically by the complex of fever, anemia, high ESR, and headaches in a patient aged >50 years. Other phenotypic manifestations include features of systemic inflammation, including malaise, fatigue, anorexia, weight loss, sweats, arthralgias, polymyalgia rheumatica, or large-vessel disease.

In patients with involvement of the cranial arteries, headache is the predominant symptom and may be associated with a tender, thickened, or nodular artery, which may pulsate early in the disease but may become occluded later. Scalp pain and claudication of the jaw and tongue may occur. A well-recognized and dreaded complication of giant cell arteritis, particularly in untreated patients, is ischemic optic neuropathy, which may lead to serious visual symptoms, even sudden blindness in some patients. However, most patients have complaints relating to the head or eyes before visual loss. Attention to such symptoms with institution of appropriate therapy (see below) will usually avoid this complication. Other cranial ischemic complications include strokes and scalp or tongue infarction.

Up to one-third of patients can have large-vessel disease that can be the primary presentation of giant cell arteritis or can emerge at a later point in patients who have had previous cranial arteritis features or polymyalgia rheumatica. Manifestations of large-vessel disease can include subclavian artery stenosis that can present as arm claudication or aortic aneurysms involving the thoracic and to a lesser degree the abdominal aorta, which carry risks of rupture or dissection.

Characteristic laboratory findings in addition to the elevated ESR include a normochromic or slightly hypochromic anemia. Liver function abnormalities are common, particularly increased alkaline phosphatase levels. Increased levels of IgG and complement have been reported. Levels of enzymes indicative of muscle damage such as serum creatine kinase are not elevated.

DIAGNOSIS

The diagnosis of giant cell arteritis and its associated clinicopathologic syndrome can often be suggested clinically by the demonstration of the complex of fever, anemia, and high ESR with or without symptoms of polymyalgia rheumatica in a patient >50 years. The diagnosis can be confirmed by biopsy of the temporal artery but may not be positive in all patients due to patchy histologic findings. Since involvement of the vessel may be segmental, positive yield is increased by obtaining a biopsy segment of 3–5 cm together with serial sectioning of biopsy specimens. Ultrasonography of the temporal artery has been reported to be helpful in diagnosis and has been increasingly used by some physicians. Therapy should not be delayed pending the performance of diagnostic studies. In this regard, it has been reported that temporal artery biopsies may show vasculitis even after ~14 days of glucocorticoid therapy. A dramatic clinical response to a trial of glucocorticoid therapy can further support the diagnosis.

Large-vessel disease may be suggested by symptoms and findings on physical examination such as diminished pulses or bruits. It is confirmed by vascular imaging, most commonly through magnetic resonance or computed tomography.

Isolated polymyalgia rheumatica is a clinical diagnosis made by the presence of typical symptoms of stiffness, aching, and pain in the muscles of the hip and shoulder girdle, an increased ESR, the absence of clinical features suggestive of giant cell arteritis, and a prompt therapeutic response to low-dose prednisone.

TREATMENT

Giant Cell Arteritis and Polymyalgia Rheumatica

Acute disease-related mortality directly from giant cell arteritis is uncommon, with fatalities occurring from cerebrovascular events or myocardial infarction. However, patients are at risk of late mortality from aortic aneurysm rupture or dissection as patients with giant cell arteritis are 18 times more likely to develop thoracic aortic aneurysms than the general population.

The goals of treatment in giant cell arteritis are to reduce symptoms and, most importantly, to prevent visual loss. The treatment approach for cranial and large-vessel disease in giant cell arteritis is currently the same. Giant cell arteritis and its associated symptoms are exquisitely sensitive to glucocorticoid therapy. Treatment should begin with prednisone, 40–60 mg/d for ~1 month, followed by a gradual tapering. When ocular signs and symptoms occur, consideration should be given for the use of methylprednisolone 1000 mg daily for 3 days to protect remaining vision. Although the optimal duration of glucocorticoid therapy has not been established, most series have found that patients require treatment for ≥2 years. Symptom recurrence during prednisone tapering develops in 60–85% of patients with giant cell arteritis, requiring a dosage increase. The ESR can serve as a useful indicator of inflammatory disease activity in monitoring and tapering therapy and can be used to judge the pace of the tapering schedule. However, minor increases in the ESR can occur as glucocorticoids are being tapered and do not necessarily reflect an exacerbation of arteritis, particularly if the patient remains symptom-free. Under these circumstances, the tapering should continue with caution. Glucocorticoid toxicity occurs in 35–65% of patients and represents an important cause of patient morbidity. Aspirin 81 mg daily has been found to reduce the occurrence of cranial ischemic complications in giant cell arteritis and should be given in addition to glucocorticoids in patients who do not have contraindications. The use of weekly methotrexate as a glucocorticoid-sparing agent has been examined in two randomized placebo-controlled trials that reached conflicting conclusions. Infliximab, a monoclonal antibody to TNF, was studied in a randomized trial and was not found to provide benefit.

Tocilizumab (anti-IL-6 receptor) was found to be more effective than prednisone alone in a recent large-scale randomized trial of giant cell arteritis and was FDA approved for this indication. It is used adjunctively to glucocorticoids and its optimal role in patient management will continue to be defined over time. The side effect profile of tocilizumab which includes leukopenia, thrombocytopenia, transaminnase elevation, and hyperlipidemia must be weighed. Because of the risk of gastrointestinal perforation, patients with prior diverticulitis were excluded from the giant cell arteritis trial. By nature of its mechanism, tocilizumab impacts laboratory parameters of the acute phase response which will eliminate the ability to utilize these in disease activity assessment.

Abatacept (CTLA4-Ig) was examined in a small randomized trial in giant cell arteritis and demonstrated greater efficacy than glucocorticoids alone.

Patients with isolated polymyalgia rheumatica respond promptly to prednisone, which can be started at a lower dose of 10–20 mg/d. Similar to giant cell arteritis, the ESR can serve as a useful indicator in monitoring and prednisone reduction. Recurrent polymyalgia symptoms develop in the majority of patients during prednisone tapering. One study of weekly methotrexate found that the use of this drug reduced the prednisone dose on average by only 1 mg and did not decrease prednisone-related side effects. A randomized trial in polymyalgia rheumatica did not find infliximab to lessen relapse or glucocorticoid requirements.

TAKAYASU ARTERITIS

DEFINITION

Takayasu arteritis is an inflammatory and stenotic disease of medium- and large-sized arteries characterized by a strong predilection for the aortic arch and its branches.

INCIDENCE AND PREVALENCE

Takayasu arteritis is an uncommon disease with an estimated annual incidence rate of 1.2–2.6 cases per million. It is most prevalent in adolescent girls and young women. Although it is more common in Asia, it is neither racially nor geographically restricted.

PATHOLOGY AND PATHOGENESIS

The disease involves medium- and large-sized arteries, with a strong predilection for the aortic arch and its branches; the pulmonary artery may also be involved. The most commonly affected arteries seen by arteriography are listed in Table 356-7. The involvement of the major branches of the aorta is much more marked at their origin than distally. The disease is a panarteritis with inflammatory mononuclear cell infiltrates and occasionally giant cells. There are marked intimal proliferation and fibrosis, scarring and vascularization of the media, and disruption and degeneration of the elastic lamina. Narrowing of the lumen occurs with or without thrombosis. The vasa vasorum are frequently involved. Pathologic changes in various organs reflect the compromise of blood flow through the involved vessels.

Immunopathogenic mechanisms, the precise nature of which is uncertain, are suspected in this disease. As with several of the vasculitis syndromes, circulating immune complexes have been demonstrated, but their pathogenic significance is unclear.

TABLE 356-7 Frequency of Arteriographic Abnormalities and Potential Clinical Manifestations of Arterial Involvement in Takayasu Arteritis

ARTERY	PERCENTAGE OF ARTERIOGRAPHIC ABNORMALITIES	POTENTIAL CLINICAL MANIFESTATIONS
Subclavian	93	Arm claudication, Raynaud's phenomenon
Common carotid	58	Visual changes, syncope, transient ischemic attacks, stroke
Abdominal aorta[a]	47	Abdominal pain, nausea, vomiting
Renal	38	Hypertension, renal failure
Aortic arch or root	35	Aortic insufficiency, congestive heart failure
Vertebral	35	Visual changes, dizziness
Coeliac axis[a]	18	Abdominal pain, nausea, vomiting
Superior mesenteric[a]	18	Abdominal pain, nausea, vomiting
Iliac	17	Leg claudication
Pulmonary	10–40	Atypical chest pain, dyspnea
Coronary	<10	Chest pain, myocardial infarction

[a]Arteriographic lesions at these locations are usually asymptomatic but may potentially cause these symptoms.

Source: G Kerr et al: Ann Intern Med 120:919, 1994.

CLINICAL AND LABORATORY MANIFESTATIONS

Takayasu arteritis is a systemic disease with generalized as well as vascular symptoms. The generalized symptoms include malaise, fever, night sweats, arthralgias, anorexia, and weight loss, which may occur months before vessel involvement is apparent. These symptoms may merge into those related to vascular compromise and organ ischemia. Pulses are commonly absent in the involved vessels, particularly the subclavian artery. The frequency of arteriographic abnormalities and the potentially associated clinical manifestations are listed in Table 356-7. Hypertension occurs in 32–93% of patients and contributes to renal, cardiac, and cerebral injury.

Characteristic laboratory findings include an elevated ESR, mild anemia, and elevated immunoglobulin levels.

DIAGNOSIS

The diagnosis of Takayasu arteritis should be suspected strongly in a young woman who develops a decrease or absence of peripheral pulses, discrepancies in blood pressure, and arterial bruits. The diagnosis is confirmed by the characteristic pattern on arteriography, which includes irregular vessel walls, stenosis, poststenotic dilation, aneurysm formation, occlusion, and evidence of increased collateral circulation. Complete aortic arteriography by catheter-directed dye arteriography or magnetic resonance arteriography should be obtained to fully delineate the distribution and degree of arterial disease. Histopathologic demonstration of vessel wall inflammation that is predominantly lymphocytic with granuloma formation and giant cells involving the media and adventitia adds confirmatory data; however, tissue is rarely readily available for examination. IgG4-related disease is a potential cause of aortitis and periaortitis that is histologically differentiated from Takayasu arteritis by a dense lymphoplasmacytic infiltrate rich in IgG4-positive plasma cells, a storiform pattern of fibrosis, and obliterative phlebitis.

TREATMENT

Takayasu Arteritis

The long-term outcome of patients with Takayasu arteritis has varied widely between studies. Although two North American reports found overall survival to be ≥94%, the 5-year mortality rate from other studies has ranged from 0 to 35%. Disease-related mortality most often occurs from congestive heart failure, cerebrovascular events, myocardial infarction, aneurysm rupture, or renal failure. Even in the absence of life-threatening disease, Takayasu arteritis can be associated with significant morbidity. The course of the disease is variable, and although spontaneous remissions may occur, Takayasu arteritis is most often chronic and relapsing. Although glucocorticoid therapy in doses of 40–60 mg prednisone per day alleviates symptoms, there are no convincing studies that indicate that it increases survival. The combination of glucocorticoid therapy for acute signs and symptoms and an aggressive surgical and/or arterioplastic approach to stenosed vessels has markedly improved outcome and decreased morbidity by lessening the risk of stroke, correcting hypertension due to renal artery stenosis, and improving blood flow to ischemic viscera and limbs. Unless it is urgently required, surgical correction of stenosed arteries should be undertaken only when the vascular inflammatory process is well controlled with medical therapy. In individuals who are refractory to or unable to taper glucocorticoids, methotrexate in doses up to 25 mg per week has yielded encouraging results. Preliminary results with anti-TNF therapies and tocilizumab have been encouraging, but will require further study through randomized trials to determine efficacy.

Abatacept was recently examined in the first randomized trial to be conducted in Takayasu arteritis but did not demonstrate efficacy beyond glucocorticoids alone. There have been a number

of published retrospective studies with the use of tocilizumab in Takayasu arteritis that have supported benefit but further investigation is necessary to fully determine efficacy.

IgA VASCULITIS (HENOCH-SCHÖNLEIN)

DEFINITION

IgA vasculitis (Henoch-Schönlein) is a small-vessel vasculitis characterized by palpable purpura (most commonly distributed over the buttocks and lower extremities), arthralgias, gastrointestinal signs and symptoms, and glomerulonephritis.

INCIDENCE AND PREVALENCE

IgA vasculitis (Henoch-Schönlein) is usually seen in children; most patients range in age from 4 to 7 years; however, the disease may also be seen in infants and adults. It is not a rare disease; in one series it accounted for between 5 and 24 admissions per year at a pediatric hospital. The male-to-female ratio is 1.5:1. A seasonal variation with a peak incidence in spring has been noted.

PATHOLOGY AND PATHOGENESIS

The presumptive pathogenic mechanism for IgA (Henoch-Schönlein) vasculitis is immune-complex deposition. A number of inciting antigens have been suggested including upper respiratory tract infections, various drugs, foods, insect bites, and immunizations. IgA is the antibody class most often seen in the immune complexes and has been demonstrated in the renal biopsies of these patients.

CLINICAL AND LABORATORY MANIFESTATIONS

In pediatric patients, palpable purpura is seen in virtually all patients; most patients develop polyarthralgias in the absence of frank arthritis. Gastrointestinal involvement, which is seen in almost 70% of pediatric patients, is characterized by colicky abdominal pain usually associated with nausea, vomiting, diarrhea, or constipation, and is frequently accompanied by the passage of blood and mucus per rectum; bowel intussusception may occur. Renal involvement occurs in 10–50% of patients and is usually characterized by mild glomerulonephritis leading to proteinuria and microscopic hematuria, with red blood cell casts in the majority of patients; it usually resolves spontaneously without therapy. Rarely, a progressive glomerulonephritis will develop. In adults, presenting symptoms are most frequently related to the skin and joints, while initial complaints related to the gut are less common. Although certain studies have found that renal disease is more frequent and more severe in adults, this has not been a consistent finding. However, the course of renal disease in adults may be more insidious and thus requires close follow-up. Myocardial involvement can occur in adults but is rare in children.

Laboratory studies generally show a mild leukocytosis, a normal platelet count, and occasionally eosinophilia. Serum complement components are normal, and IgA levels are elevated in about one-half of patients.

DIAGNOSIS

The diagnosis of IgA vasculitis (Henoch-Schönlein) is based on clinical signs and symptoms. Skin biopsy specimen can be useful in confirming leukocytoclastic vasculitis with IgA and C3 deposition by immunofluorescence. Renal biopsy is rarely needed for diagnosis but may provide prognostic information in some patients.

TREATMENT

IgA Vasculitis (Henoch-Schönlein)

The prognosis of IgA vasculitis (Henoch-Schönlein) is excellent. Mortality is exceedingly rare, and 1–5% of children progress to end-stage renal disease. Most patients recover completely, and some do not require therapy. Treatment is similar for adults and children. When glucocorticoid therapy is required, prednisone, in doses of 1 mg/kg/d and tapered according to clinical response, has been shown to be useful in decreasing tissue edema, arthralgias, and abdominal discomfort; however, it has not proved beneficial in the treatment of skin or renal disease and does not appear to shorten the duration of active disease or lessen the chance of recurrence. Patients with rapidly progressive glomerulonephritis have been anecdotally reported to benefit from intensive plasma exchange combined with cytotoxic drugs. Disease recurrences have been reported in 10–40% of patients.

CRYOGLOBULINEMIC VASCULITIS

DEFINITION

Cryoglobulins are cold-precipitable monoclonal or polyclonal immunoglobulins. Cryoglobulinemia may be associated with a systemic vasculitis characterized by palpable purpura, arthralgias, weakness, neuropathy, and glomerulonephritis. Although this can be observed in association with a variety of underlying disorders including multiple myeloma, lymphoproliferative disorders, connective tissue diseases, infection, and liver disease, in many instances it appears to be idiopathic. Because of the apparent absence of an underlying disease and the presence of cryoprecipitate containing oligoclonal/polyclonal immunoglobulins, this entity was referred to as *essential mixed cryoglobulinemia*. Since the discovery of hepatitis C, it has been established that the vast majority of patients who were considered to have essential mixed cryoglobulinemia have cryoglobulinemic vasculitis related to hepatitis C infection.

INCIDENCE AND PREVALENCE

The incidence of cryoglobulinemic vasculitis has not been established. It has been estimated, however, that 5% of patients with chronic hepatitis C will develop cryoglobulinemic vasculitis.

PATHOLOGY AND PATHOGENESIS

Skin biopsies in cryoglobulinemic vasculitis reveal an inflammatory infiltrate surrounding and involving blood vessel walls, with fibrinoid necrosis, endothelial cell hyperplasia, and hemorrhage. Deposition of immunoglobulin and complement is common. Abnormalities of uninvolved skin including basement membrane alterations and deposits in vessel walls may be found. Membranoproliferative glomerulonephritis is responsible for 80% of all renal lesions in cryoglobulinemic vasculitis.

The association between hepatitis C and cryoglobulinemic vasculitis has been supported by the high frequency of documented hepatitis C infection, the presence of hepatitis C RNA and anti–hepatitis C antibodies in serum cryoprecipitates, evidence of hepatitis C antigens in vasculitic skin lesions, and the effectiveness of antiviral therapy (see below). Current evidence suggests that in the majority of cases, cryoglobulinemic vasculitis occurs when an aberrant immune response to hepatitis C infection leads to the formation of immune complexes consisting of hepatitis C antigens, polyclonal hepatitis C–specific IgG, and monoclonal IgM rheumatoid factor. The deposition of these immune complexes in blood vessel walls triggers an inflammatory cascade that results in cryoglobulinemic vasculitis.

CLINICAL AND LABORATORY MANIFESTATIONS

The most common clinical manifestations of cryoglobulinemic vasculitis are cutaneous vasculitis, arthritis, peripheral neuropathy, and glomerulonephritis. Renal disease develops in 10–30% of patients. Life-threatening rapidly progressive glomerulonephritis or vasculitis of the CNS, gastrointestinal tract, or heart occurs infrequently.

The presence of circulating cryoprecipitates is the fundamental finding in cryoglobulinemic vasculitis. Rheumatoid factor is almost always found and may be a useful clue to the disease when cryoglobulins are not detected. Hypocomplementemia occurs in 90% of patients. An elevated ESR and anemia occur frequently. Evidence for hepatitis C infection must be sought in all patients by testing for hepatitis C antibodies and hepatitis C RNA.

TREATMENT

Cryoglobulinemic Vasculitis

Acute mortality directly from cryoglobulinemic vasculitis is uncommon, but the presence of glomerulonephritis is a poor prognostic sign for overall outcome. In such patients, 15% progress to end-stage renal

disease, with 40% later experiencing fatal cardiovascular disease, infection, or liver failure. As indicated above, the majority of cases are associated with hepatitis C infection. In such patients, treatment with antiviral therapy (Chap. 332) is first-line therapy for hepatitis C–associated cryoglobulinemic vasculitis, particularly given the efficacy of current hepatitis C therapies. Clinical improvement with antiviral therapy is dependent on the virologic response. Patients who clear hepatitis C from the blood have objective improvement in their vasculitis along with significant reductions in levels of circulating cryoglobulins, IgM, and rheumatoid factor. While transient improvement can be observed with glucocorticoids, a complete response is seen in only 7% of patients. Plasmapheresis and cytotoxic agents have been used in anecdotal reports. These observations have not been confirmed, and such therapies carry significant risks. Randomized trials with rituximab (anti-CD20) in hepatitis C–associated cryoglobulinemic vasculitis have provided evidence of benefit such that this agent should be considered in patients with active vasculitis either in combination with antiviral therapy or alone in patients who have relapsed through, are intolerant to, or have contraindications to antiviral agents.

SINGLE-ORGAN VASCULITIS

The potential for vasculitis to affect single organs has become increasingly recognized. This has been defined as vasculitis in arteries or veins of any size in a single organ that has no features that indicate that it is a limited expression of a systemic vasculitis. Examples include isolated aortitis, testicular vasculitis, vasculitis of the breast, isolated cutaneous vasculitis, and primary CNS vasculitis. In some instances, this may be discovered at the time of surgery such as orchiectomy for a testicular mass where there is concern for neoplasm that is found instead to be vasculitis. Some patients originally diagnosed with single-organ vasculitis may later develop additional manifestations of a more systemic disease. In instances where there is no evidence of systemic vasculitis and the affected organ has been removed in its entirety, the patient may be followed closely without immunosuppressive therapy. In other instances, such as primary CNS vasculitis or some patients with isolated cutaneous vasculitis, medical intervention is warranted.

IDIOPATHIC CUTANEOUS VASCULITIS

DEFINITION

The term *cutaneous vasculitis* is defined broadly as inflammation of the blood vessels of the dermis. Because of its heterogeneity, cutaneous vasculitis has been described by a variety of terms including *hypersensitivity vasculitis* and *cutaneous leukocytoclastic angiitis*. However, cutaneous vasculitis is not one specific disease but a manifestation that can be seen in a variety of settings. In >70% of cases, cutaneous vasculitis occurs either as part of a primary systemic vasculitis or as a secondary vasculitis related to an inciting agent or an underlying disease (see "Secondary Vasculitis," below). In the remaining 30% of cases, cutaneous vasculitis occurs idiopathically.

INCIDENCE AND PREVALENCE

Cutaneous vasculitis represents the most commonly encountered vasculitis in clinical practice. The exact incidence of idiopathic cutaneous vasculitis has not been determined due to the predilection for cutaneous vasculitis to be associated with an underlying process and the variability of its clinical course.

PATHOLOGY AND PATHOGENESIS

The typical histopathologic feature of cutaneous vasculitis is the presence of vasculitis of small vessels. Postcapillary venules are the most commonly involved vessels; capillaries and arterioles may be involved less frequently. This vasculitis is characterized by a *leukocytoclasis*, a term that refers to the nuclear debris remaining from the neutrophils that have infiltrated in and around the vessels during the acute stages. In the subacute or chronic stages, mononuclear cells predominate; in certain subgroups, eosinophilic infiltration is seen. Erythrocytes often extravasate from the involved vessels, leading to palpable purpura. *Cutaneous arteritis* can also occur, which involves slightly larger-sized vessels within the dermis.

CLINICAL AND LABORATORY MANIFESTATIONS

The hallmark of idiopathic cutaneous vasculitis is the predominance of skin involvement. Skin lesions may appear typically as palpable purpura; however, other cutaneous manifestations of the vasculitis may occur, including macules, papules, vesicles, bullae, subcutaneous nodules, ulcers, and recurrent or chronic urticaria. The skin lesions may be pruritic or even quite painful, with a burning or stinging sensation. Lesions most commonly occur in the lower extremities in ambulatory patients or in the sacral area in bedridden patients due to the effects of hydrostatic forces on the postcapillary venules. Edema may accompany certain lesions, and hyperpigmentation often occurs in areas of recurrent or chronic lesions.

There are no specific laboratory tests diagnostic of idiopathic cutaneous vasculitis. A mild leukocytosis with or without eosinophilia is characteristic, as is an elevated ESR. Laboratory studies should be aimed toward ruling out features to suggest an underlying disease or a systemic vasculitis.

DIAGNOSIS

The diagnosis of cutaneous vasculitis is made by the demonstration of vasculitis on biopsy. An important diagnostic principle in patients with cutaneous vasculitis is to search for an etiology of the vasculitis—be it an exogenous agent, such as a drug or an infection, or an endogenous condition, such as an underlying disease (Fig. 356-1). In addition, a careful physical and laboratory examination should be performed to rule out the possibility of systemic vasculitis. This should start with the least invasive diagnostic approach and proceed to the more invasive only if clinically indicated.

TREATMENT

Idiopathic Cutaneous Vasculitis

When an antigenic stimulus is recognized as the precipitating factor in the cutaneous vasculitis, it should be removed; if this is a microbe, appropriate antimicrobial therapy should be instituted. If the vasculitis is associated with another underlying disease, treatment of the latter often results in resolution of the former. In situations where disease is apparently self-limited, no therapy, except possibly symptomatic therapy, is indicated. When cutaneous vasculitis persists and when there is no evidence of an inciting agent, an associated disease, or an underlying systemic vasculitis, the decision to treat should be based on weighing the balance between the degree of symptoms and the risk of treatment. Some cases of idiopathic cutaneous vasculitis resolve spontaneously, whereas others remit and relapse. In patients with persistent vasculitis, a variety of therapeutic regimens have been tried with variable results. In general, the treatment of idiopathic cutaneous vasculitis has not been satisfactory. Fortunately, since the disease is generally limited to the skin, this lack of consistent response to therapy usually does not lead to a life-threatening situation. Agents with which there have been anecdotal reports of success include dapsone, colchicine, hydroxychloroquine, and nonsteroidal anti-inflammatory agents. Glucocorticoids are often used in the treatment of idiopathic cutaneous vasculitis. Therapy is usually instituted as prednisone, 1 mg/kg/d, with rapid tapering where possible, either directly to discontinuation or by conversion to an alternate-day regimen followed by ultimate discontinuation. In cases that prove refractory to glucocorticoids, a trial of another immunosuppressive agent may be indicated. Patients with chronic vasculitis isolated to cutaneous venules rarely respond dramatically to any therapeutic regimen, and cytotoxic agents should be used only as a last resort in these patients. Methotrexate and azathioprine have been used in such situations in anecdotal reports. Although

cyclophosphamide is the most effective therapy for the systemic vasculitides, it should almost never be used for idiopathic cutaneous vasculitis because of the potential toxicity.

PRIMARY CENTRAL NERVOUS SYSTEM VASCULITIS

Primary central nervous system (CNS) vasculitis is an uncommon clinicopathologic entity characterized by vasculitis restricted to the vessels of the CNS without other apparent systemic vasculitis. The inflammatory process is usually composed of mononuclear cell infiltrates with or without granuloma formation.

Patients may present with headaches, altered mental function, and focal neurologic defects. Systemic symptoms are generally absent. Devastating neurologic abnormalities may occur depending on the extent of vessel involvement. The diagnosis can be suggested by abnormal magnetic resonance imaging of the brain, an abnormal lumbar puncture, and/or demonstration of characteristic vessel abnormalities on arteriography **(Fig. 356-4)**, but it is confirmed by biopsy of the brain parenchyma and leptomeninges. In the absence of a brain biopsy, care should be taken not to misinterpret as true primary vasculitis arteriographic abnormalities that might actually be related to another cause. An important entity in the differential diagnosis is reversible cerebral vasoconstrictive syndrome, which typically presents with "thunderclap" headache and is associated with arteriographic abnormalities that mimic primary CNS vasculitis that are reversible. Other diagnostic considerations include infection, atherosclerosis, emboli, connective tissue disease, sarcoidosis, malignancy, and drug-associated causes. The prognosis of granulomatous primary CNS vasculitis is poor; however, some reports indicate that glucocorticoid therapy, alone or together with cyclophosphamide administered as described above, has induced clinical remissions.

BEHÇET'S DISEASE

Behçet's disease is a clinicopathologic entity characterized by recurrent episodes of oral and genital ulcers, iritis, and cutaneous lesions. The underlying pathologic process is a leukocytoclastic venulitis, although vessels of any size and in any organ can be involved. **This disorder is described in detail in Chap. 357.**

COGAN'S SYNDROME

Cogan's syndrome is characterized by interstitial keratitis together with vestibuloauditory symptoms. It may be associated with a systemic vasculitis, particularly aortitis with involvement of the aortic valve. Glucocorticoids are the mainstay of treatment. Initiation of treatment as early as possible after the onset of hearing loss improves the likelihood of a favorable outcome.

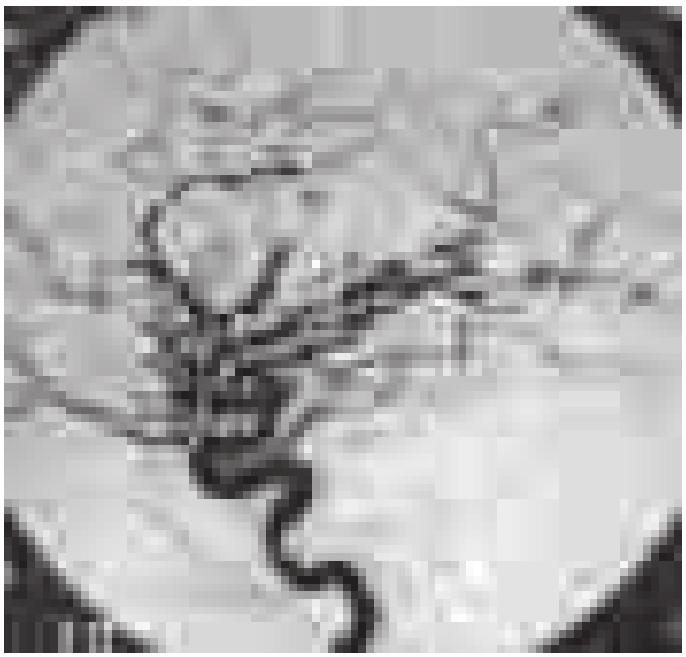


FIGURE 356-4 Cerebral arteriogram from a 32-year-old man with primary central nervous system vasculitis. Dramatic beading (*arrow*) typical of vasculitis is seen.

KAWASAKI'S DISEASE

Kawasaki's disease is an acute, febrile, multisystem disease of children. Some 80% of cases occur prior to the age of 5, with the peak incidence occurring at ≤2 years. It is characterized by nonsuppurative cervical adenitis and changes in the skin and mucous membranes such as edema; congested conjunctivae; erythema of the oral cavity, lips, and palms; and desquamation of the skin of the fingertips. Although the disease is generally benign and self-limited, it is associated with coronary artery aneurysms in ~25% of cases, with an overall case fatality rate of 0.5–2.8%. These complications usually occur between the third and fourth weeks of illness during the convalescent stage. Vasculitis of the coronary arteries is seen in almost all the fatal cases that have been autopsied. There is typical intimal proliferation and infiltration of the vessel wall with mononuclear cells. Beadlike aneurysms and thromboses may be seen along the artery. Other manifestations include pericarditis, myocarditis, myocardial ischemia and infarction, and cardiomegaly.

Apart from the up to 2.8% of patients who develop fatal complications, the prognosis of this disease for uneventful recovery is excellent. High-dose IV γ-globulin (2 g/kg as a single infusion over 10 h) together with aspirin (100 mg/kg/d for 14 days followed by 3–5 mg/kg/d for several weeks) have been shown to be effective in reducing the prevalence of coronary artery abnormalities when administered early in the course of the disease. Surgery may be necessary for Kawasaki disease patients who have giant coronary artery aneurysms or other coronary complications. Surgical treatment most commonly includes thromboendarterectomy, thrombus clearing, aneurysmal reconstruction, and coronary artery bypass grafting.

POLYANGIITIS OVERLAP SYNDROMES

Some patients with systemic vasculitis manifest clinicopathologic characteristics that do not fit precisely into any specific disease but have overlapping features of different vasculitides. Active systemic vasculitis in such settings has the same potential for causing irreversible organ system damage as when it occurs in one of the defined syndromes listed in Table 356-1. The diagnostic and therapeutic considerations as well as the prognosis for these patients depend on the sites and severity of active vasculitis. Patients with vasculitis that could potentially cause irreversible damage to a major organ system should be treated as described under "Granulomatosis with Polyangiitis (Wegener's)."

SECONDARY VASCULITIS

DRUG-INDUCED VASCULITIS

Vasculitis associated with drug reactions usually presents as palpable purpura that may be generalized or limited to the lower extremities or other dependent areas; however, urticarial lesions, ulcers, and hemorrhagic blisters may also occur **(Chap. 56)**. Signs and symptoms may be limited to the skin, although systemic manifestations such as fever, malaise, and polyarthralgias may occur. Although the skin is the predominant organ involved, systemic vasculitis may result from drug reactions. Drugs that have been implicated in vasculitis include allopurinol, thiazides, gold, sulfonamides, phenytoin, and penicillin **(Chap. 56)**.

An increasing number of drugs have been reported to cause vasculitis associated with antimyeloperoxidase ANCA. Of these, the best evidence of causality exists for hydralazine and propylthiouracil. The clinical manifestations in ANCA-positive drug-induced vasculitis can range from cutaneous lesions to glomerulonephritis and pulmonary hemorrhage. Outside of drug discontinuation, treatment should be based on the severity of the vasculitis. Patients with immediately life-threatening small-vessel vasculitis should initially be treated with glucocorticoids and cyclophosphamide as described for granulomatosis with polyangiitis (Wegener's). Following clinical improvement, consideration may be given for tapering such agents along a more rapid schedule.

SERUM SICKNESS AND SERUM SICKNESS–LIKE REACTIONS

These reactions are characterized by the occurrence of fever, urticaria, polyarthralgias, and lymphadenopathy 7–10 days after primary exposure and 2–4 days after secondary exposure to a heterologous protein (classic serum sickness) or a nonprotein drug such as penicillin or sulfa (serum sickness–like reaction). Most of the manifestations are not due to a vasculitis; however, occasional patients will have typical cutaneous venulitis that may progress rarely to a systemic vasculitis.

VASCULITIS ASSOCIATED WITH OTHER UNDERLYING DISEASES

Certain *infections* may directly trigger an inflammatory vasculitic process. For example, rickettsias can invade and proliferate in the endothelial cells of small blood vessels causing a vasculitis **(Chap. 182)**. In addition, the inflammatory response around blood vessels associated with certain systemic fungal diseases such as histoplasmosis **(Chap. 207)** may mimic a primary vasculitic process. A leukocytoclastic vasculitis predominantly involving the skin with occasional involvement of other organ systems may be a minor component of many other infections. These include *subacute bacterial endocarditis*, *Epstein-Barr virus infection*, *HIV infection*, and a number of other infections.

Vasculitis can be associated with certain *malignancies*, particularly lymphoid or reticuloendothelial neoplasms. Leukocytoclastic venulitis confined to the skin is the most common finding; however, widespread systemic vasculitis may occur. Of particular note is the association of *hairy cell leukemia* **(Chap. 106)** with polyarteritis nodosa.

A number of *connective tissue diseases* have vasculitis as a secondary manifestation of the underlying primary process. Foremost among these are *systemic lupus erythematosus* **(Chap. 349)**, *rheumatoid arthritis* **(Chap. 351)**, *inflammatory myositis* **(Chap. 358)**, *relapsing polychondritis* **(Chap. 359)**, and *Sjögren's syndrome* **(Chap. 354)**. The most common form of vasculitis in these conditions is the small-vessel venulitis isolated to the skin. However, certain patients may develop a fulminant systemic necrotizing vasculitis.

Secondary vasculitis has also been observed in association with *ulcerative colitis*, *congenital deficiencies of various complement components*, *sarcoidosis*, *primary biliary cirrhosis*, *α_1-antitrypsin deficiency*, and *intestinal bypass surgery*.

FURTHER READING

Buttgereit F et al: Polymyalgia rheumatica and giant cell arteritis: A systematic review. JAMA 315:2442, 2016.

Fauci AS et al: Wegener's granulomatosis: Prospective clinical and therapeutic experience with 85 patients for 21 years. Ann Intern Med 98:76, 1983.

Finkielman JD et al: Antiproteinase 3 antineutrophil cytoplasmic antibodies and disease activity in Wegener granulomatosis. Ann Intern Med 147:611, 2007.

Guillevin L et al: Churg-Strauss syndrome. Clinical study and long-term follow-up of 96 patients. Medicine (Baltimore) 78:26, 1999.

Hoffman GS et al: Wegener granulomatosis: An analysis of 158 patients. Ann Intern Med 16:488, 1992.

Jayne D et al: A randomized trial of maintenance therapy for vasculitis associated with antineutrophil cytoplasmic autoantibodies. N Engl J Med 349:36, 2003.

Kerr GS et al: Takayasu arteritis. Ann Intern Med 120:919, 1994.

Langford CA et al: A randomized, double-blind trial of abatacept (CTLA-4Ig) for the treatment of giant cell arteritis. Arthritis Rheumatol 69:837, 2017.

Langford CA et al: A randomized, double-blind trial of abatacept (CTLA-4Ig) for the treatment of Takayasu arteritis. Arthritis Rheumatol 69:846, 2017.

Pagnoux C et al: Clinical features and outcomes in 348 patients with polyarteritis nodosa: A systematic retrospective study of patients diagnosed between 1963 and 2005 and entered into the French Vasculitis Study Group Database. Arthritis Rheum 62:616, 2010.

Stone JH et al: Rituximab versus cyclophosphamide for ANCA-associated vasculitis. N Engl J Med 363:221, 2010.

Stone JH et al: Trial of tocilizumab in giant-cell arteritis. N Engl J Med 377:317, 2017.

Wechsler ME et al: Mepolizumab or placebo for eosinophilic granulomatosis with polyangiitis. N Engl J Med 376:1921, 2017.

Weyand CM, Goronzy JJ: Clinical practice. Giant-cell arteritis and polymyalgia rheumatica. N Engl J Med 371:50, 2014.

357 Behçet's Syndrome

Haralampos M. Moutsopoulos

DEFINITION, INCIDENCE, AND PREVALENCE

Behçet's syndrome is a multisystem disorder presenting with recurrent oral and genital ulcerations as well as ocular involvement. The diagnosis is clinical and based on internationally agreed diagnostic criteria **(Table 357-1)**.

The syndrome affects young males and females from the Mediterranean region, the Middle East, and the Far East. The frequency of Behçet's syndrome increases from north to south Europe. Males and females are affected equally. Males often have more severe disease. The syndrome is rare in Sub-Saharan Africa.

PATHOGENESIS

The etiology and pathogenesis of this syndrome remain obscure. The disease appears to be in the crossroads of autoinflammatory and autoimmune disorders. The main pathologic lesion is systemic perivasculitis with early neutrophil infiltration and endothelial swelling. In some patients, diffuse inflammatory disease, involving all layers of large vessels and resulting to formation of pseudoaneurysms suggests vasculitis of vasa vasorum. Apart from activated neutrophils, increased numbers of infiltrating T_H1, T_H17, cytotoxic CD8+, and γδ T cells are observed, supporting a link between innate and adaptive autoreactive immune response. Circulating autoantibodies against α-enolase of endothelial cells, selenium binding protein, and anti-*Saccharomyces cerevisiae* antibodies have been observed, but their pathogenic role remains unclear. A recent genome-wide association study confirmed the known association of Behçet's syndrome with HLA-B*51 and identified a second, independent association within the major histocompatibility complex (MHC) class I region. In addition, an association with interleukin (IL) 10 and the IL-23R–IL-12RB2 locus was also observed. Interestingly, the disease-associated IL-10 variant was correlated with diminished mRNA expression and low protein production.

CLINICAL FEATURES

The recurrent oral aphthous ulcerations are a sine qua non for the diagnosis. The ulcers are usually painful, shallow or deep with a central yellowish necrotic base, appear singly or in crops, and can be located anywhere in the oral cavity. Small ulcers, <10 mm in diameter, are seen in 85% of patients, whereas large or herpetiform lesions are less frequent. The ulcers persist for 1–2 weeks and subside without leaving scars. The genital ulcers are less common but more specific, are painful, do not affect the glans penis or urethra, and produce scrotal and vulvar scars.

TABLE 357-1 Diagnostic Criteria of Behçet's Syndrome

Recurrent oral ulceration plus two of the following:
Recurrent genital ulceration
Ocular lesions
Skin lesions
Positive Pathergy test

Skin involvement is observed in 80% of patients and includes folliculitis, erythema nodosum, acne-like rashes, and, infrequently, vasculitis, Sweet syndrome, and pyoderma gangrenosum. Nonspecific skin inflammatory reactivity to scratch or intradermal saline injection (pathergy test) is a specific manifestation.

Eye involvement with scarring and bilateral panuveitis is the most dreaded complication, since it occasionally progresses rapidly to blindness. Uveitis occurs in 10–15% of patients, primarily in males. It is usually present at the onset but may also develop within the first few years. In addition to iritis, posterior uveitis, retinal vessel occlusions, and optic neuritis can be rarely seen in some patients.

Non-deforming arthritis or arthralgias are seen in 50% of patients and affect mostly the knees and ankles. Enthesopathy, avascular necrosis, myalgia, and myositis can be also seen.

Superficial or deep peripheral vein thrombosis is seen in 30% of patients. Pulmonary emboli are a rare complication. The superior vena cava is obstructed occasionally. Arterial involvement occurs in <1–5% of patients and presents with aortitis or peripheral arterial aneurysm and arterial thrombosis. Pulmonary artery vasculitis presenting with dyspnea, cough, chest pain, hemoptysis, and infiltrates on chest roentgenograms has been reported in <1% of patients and should be differentiated from thromboembolic disease since it warrants immunosuppressive and not thrombolytic therapy.

Neurologic involvement (5–10%) appears mainly in the parenchymal form (80%); it is associated with brainstem involvement and has a grave prognosis (*central nervous system [CNS]-Behçet's syndrome*). Cerebral brain thrombosis is more common in female patients. IL-6 is persistently raised in cerebrospinal fluid of these patients. Cerebral venous thrombosis is most frequently observed in the superior sagittal and transverse sinuses and is associated with headache and increased intracranial pressure. Magnetic resonance imaging (MRI) and/or proton magnetic resonance spectroscopy (MRS) are very sensitive and should be employed if CNS-Behçet's syndrome is suspected.

Gastrointestinal involvement is seen more frequently in patients from Japan and consists of mucosal ulcerations of the gut, resembling Crohn's disease.

Epididymitis is seen in 5% of patients, whereas AA amyloidosis and glomerulonephritis are uncommon.

Laboratory findings are mainly nonspecific indices of inflammation, such as leukocytosis and elevated erythrocyte sedimentation rate, as well as C-reactive protein levels.

TREATMENT

Behçet's Syndrome

The severity of the syndrome usually abates with time. Apart from the patients with CNS-Behçet's syndrome and major vessel disease, the life expectancy seems to be normal and the only serious complication is blindness.

Mucous membrane involvement may respond to topical glucocorticoids in the form of mouthwash or paste. In more serious cases, thalidomide (100 mg/d) is effective. Recently it was shown that Apremilast, an inhibitor of phosphodiesterase-4, was effective for oral ulcers; however, the efficacy did not persist after withdrawal. Colchicine can be beneficial for the mucocutaneous manifestations and arthritis. Uveitis and CNS-Behçet's syndrome require systemic glucocorticoid therapy (prednisone, 1 mg/kg per day), azathioprine (2–3 mg/kg per day) and cyclosporine (2–5 mg/kg). Anti–tumor necrosis factor therapy is recommended in panuveitis refractory to immunosuppressive agents. Thrombophlebitis is treated with immunosuppressives. Pulse doses of cyclophosphamide are useful early in the course of the disease for pulmonary or peripheral arterial aneurysms.

FURTHER READING

HATEMI G et al: Behçet's syndrome: A critical digest of the 2014–2015 literature. Clin Exp Rheumatol 33(6 Suppl 94):S3, 2015.

358 Inflammatory Myopathies

Steven A. Greenberg, Anthony A. Amato

This chapter focuses on the major types of inflammatory myopathies (IM), including dermatomyositis (DM), polymyositis (PM), immune-mediated necrotizing myopathy (IMNM), antisynthetase syndrome (ASS), and inclusion body myositis (IBM) **(Table 358-1)**. Other IM include those caused by infection, eosinophilic myositis and granulomatous myositis. Of note, inflammatory cell infiltrates can also be occasionally seen in muscle biopsies in hereditary (e.g., muscular dystrophies, metabolic myopathies) and toxic myopathies.

Epidemiological studies suggest that the incidence of IM grouped together is greater than 4 cases per 100,000 with prevalence in the range of 14–32 per 100,000. Defining the actual incidence and prevalence of the individual myositides is limited however by the different diagnostic criteria employed in various epidemiological studies, increasing recognition of ASS, and frequent misdiagnosis of IBM and IMNM. Idiopathic PM without signs of an overlap syndrome is quite rare, while DM, IBM, and IMNM occur in roughly similar frequencies. DM can occur in children (juvenile DM), while IBM always occurs in adults and is the most common cause of myopathy in those aged >50. DM, PM, and ASS are more common in women, while IBM is more common in men.

DIAGNOSTIC APPROACH AND DIFFERENTIAL DIAGNOSIS

The approach to patients with suspected myopathy is detailed in **Chap. 441**. In any patient presenting with weakness, the first step is to localize the site of the lesion by history and clinical findings **(Chap. 21)**. Weakness could be caused by a process in the cerebral hemispheres, spinal cord **(Chap. 434)**, anterior horn cell **(Chap. 429)**, peripheral nerve **(Chaps. 438–439)**, neuromuscular junction **(Chap. 440)**, or muscle **(Chap. 441)**. Past medical history, medication use, and family history, combined with a detailed clinical examination and an appreciation for the pattern of muscle involvement (e.g., what muscles are weak and atrophic or hypertrophic as well as the presence of scapular winging, early contractures, sensory abnormalities, fasciculations, or rash) help differentiate myopathies from other neuromuscular disorders and the different types of myopathies from each other **(see Chap. 441)**. For example, atrophy with fasciculations suggest a neurogenic process such as amyotrophic lateral sclerosis, fatigable weakness on examination points to a neuromuscular junction defect such as myasthenia gravis, and concomitant sensory symptoms suggest a central process such as a spinal cord disorder or a polyneuropathy. Scapular winging, calf hypertrophy or atrophy, and early contractures before significant weakness develops would strongly suggest a muscular dystrophy, particularly if there is a positive family history. A heliotrope rash combined with Gottron papules and dilated nailfold capillaries is diagnostic for DM. The presence of atrophy and weakness of the flexor forearm muscles and quadriceps in a person aged >50 years is most likely IBM.

When the site of the lesion cannot be localized based on history and clinical examination alone, laboratory testing is required. Serum creatine kinase (CK) is the most sensitive laboratory marker of muscle destruction. Not all myopathies are associated with elevated CKs, but a markedly elevated CK (e.g., >2000 U/L) is almost always due to a myopathy. A slightly elevated CK can also be seen in neurogenic disorders, however. Myositis-associated and myositis-specific antibodies (MSA) help to distinguish subtypes of IM, as discussed below. Electromyography and nerve conductions studies (EMG/NCS) are useful in localizing the site of the lesion, but are less specific in helping to determine the actual cause of a myopathy. EMG can be useful at times in guiding what muscle to biopsy, especially if muscles typically biopsied are normal on clinical examination. Imaging skeletal muscle can be helpful in

TABLE 358-1 Inflammatory Myopathies: Clinical and Laboratory Features

DISORDER	SEX	AGE OF ONSET	RASH	PATTERN OF WEAKNESS	LABORATORY FEATURES	MUSCLE BIOPSY	CELLULAR INFILTRATE	RESPONSE TO IS THERAPY	COMMON ASSOCIATED CONDITIONS
DM	F > M	Childhood and Adult	Yes	Proximal > distal	Normal or increased CK (up to 50× normal or higher); Various MSAs (anti-MDA5, anti-TIF1, anti-Mi-2, anti-NXP2)	Perimysial and Perivascular Inflammation; IFN-1 regulated proteins (MHC-1, MxA), MAC deposition on capillaries	CD4+ Dendritic cells; B cells; macrophages	Yes	Myocarditis, ILD, Malignancy, Vasculitis, Other CTDs
PM	F > M	Adult	No	Proximal > distal	Increased CK (up to 50× normal or higher)	Endomysial and perivascular inflammation; ubiquitous expression of MHC-1	CD8+ T-cells; Macrophages; plasma cells	Yes	Myocarditis, ILD, Other CTDs
NM	M = F	Children and adults	No	Proximal > distal	Elevated CK (> 10× normal or higher); anti-HMGCR or anti-SRP antibodies	Necrotic muscle fibers; minimal inflammatory infiltrate	Macrophages in necrotic fibers undergoing phagocytosis	Yes	Malignancy, CTD, HMGCR antibody cases can be triggered by statin use
ASS	F > M	Children and adults	Sometimes	Proximal > distal	Elevated CK (>10× normal or higher); antisynthetase antibodies	Perimysial and Perivascular Inflammation; perimysial fragmentation with alkaline phosphatase staining; perimysial muscle damage with necrosis	CD4+ Dendritic cells; B cells; macrophages	Yes	Non-erosive arthritis, ILD, Raynaud phenomenon, mechanic hands, and fever
IBM	M > F	Older adults (>50 yrs)	No	Proximal and distal; predilection for: finger/ wrist flexors, knee extensors	Normal or mildly increased CK (usually <10× normal); anti-cN-1A antibodies; large granular lymphocytes on flow cytometry and reduced CD4/CD8 ratio with increased CD8 count	Endomysial and perivascular inflammation; ubiquitous expression of MHC-1; Rimmed Vacuoles; p62, LC3, TDP-43 aggregates; EM: 15–18 nm tubulofilaments; ragged red and COX negative fibers	CD8+ T-cells; Macrophages; plasma cells; myeloid dendritic cells; large granular lymphocytes	None or Minimal	Granular lymphocytic leukemia/ lymphocytosis, sarcoidosis, SICCA or Sjogren syndrome

Abbreviations: CK, creatine kinase; cN-1A, cytosolic 5′-nucleotidase 1A; CTDs, connective tissue diseases; COX, cytochrome oxidase; DM, dermatomyositis; F, female; g, immunoglobulin; IBM, inclusion body myositis; IFN-1, type 1 interferon; ILD, interstitial lung disease; IS, immunosuppressive; M, male; MAC, membrane attack complex; MDA5, melanoma differentiation antigen; MHC-1, major histocompatibility antigen 1; NCP2, nuclear matrix protein 2 (NXP2); NM, necrotizing myopathy; PM, polymyositis; TIF1, transcriptional intermediary factor 1.

Source: Modified from AA Amato, JA Russell (eds): *Neuromuscular Disorders*, 2nd ed. New York, McGraw-Hill Education; 2016, Table 33-1, p. 824, with permission.

assessing muscle involvement, revealing fatty replacement, atrophy, or edema within muscle or surround fascia.

A muscle biopsy is usually required to definitively distinguish one myopathy from another. The different forms of IM can have distinctive histopathological abnormalities as discussed below. In a patient with a classic DM rash, a muscle or skin biopsy can be performed, but an argument can also be made that biopsy is unnecessary—particularly if the patient also has an MSA specific for DM. However, a muscle biopsy should be performed in every case of suspected PM to exclude IBM (if not clinically apparent) and other causes of myopathy. Diagnosis of IMNM is by definition based upon histological findings. It is important to biopsy a muscle that is clinically affected but not too weak (e.g., Medical Research Council grade 4 out of 5 in strength), otherwise one may just see end-stage muscle. A biopsy should always be coordinated with an experienced muscle histopathology laboratory.

Patients with severe muscle pain, subjective weakness, and fatigue with normal strength and function on examination are not likely to have an IM. Polymyalgia rheumatica should be considered in older individuals with an elevated ESR or CRP, but normal CK and EMG. Fibromyalgia is likely in patients with a normal laboratory workup. In general, a muscle biopsy is not indicated unless there is objective weakness, an abnormal EMG, or elevated CK.

SPECIFIC DISORDERS

DERMATOMYOSITIS

Clinical Features DM manifests with symmetric, proximal greater than distal weakness along with a characteristic rash that includes the heliotrope rash (erythematous discoloration of eyelids with periorbital edema), Gottron sign (erythematous rash over the extensor surfaces of joints such as the knuckles, elbows, knees, and ankles), Gottron papules (raised erythematous rash over knuckles) **(Fig. 358-1)**, V-sign (rash on the sun-exposed anterior neck and chest), shawl sign over the back of the neck and shoulders, nail bed telengiectasiae, and subcutaneous calcium deposits. The weakness and rash usually accompany one another but can be separated by several months. Furthermore, there is a spectrum of involvement such that some patients continue to manifest only with a rash (amyopathic DM), while others may present mainly with weakness and little or no visible skin changes. Patients can also complain of myalgias, arthralgias, dysphagia, and dysarthria. Cutaneous disease activity is highly relevant in DM; in comparison to other debilitating skin diseases including cutaneous lupus erythematosus, psoriasis, and atopic dermatitis, skin symptoms in DM patients are associated with an overall reduction in life quality. Pruritus can be especially debilitating. Dyspnea can occur from ventilatory muscle

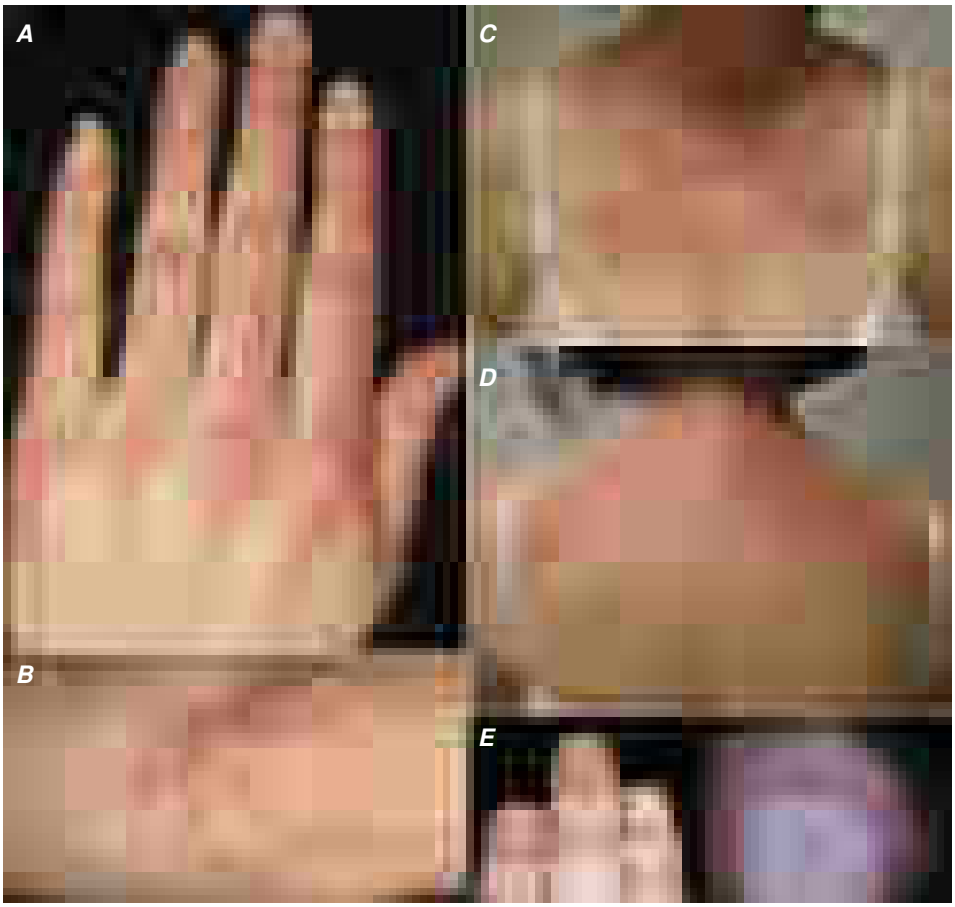


FIGURE 358-1 Cutaneous manifestations of dermatomyositis. ***A.*** Macular erythema plaques (Gottron sign) and erythematous papules (Gottron papules) on extensor surface of fingers and ***B.*** elbow. ***C.*** Macular erythema plaques over anterior neck and chest (V-sign) and ***D***. the posterior neck, shoulder and upper back (Shawl sign). ***E***. Nail bed changes with dilated capillaries.

weakness or intrinsic pulmonary problems including interstitial lung disease (ILD), bronchopneumonia and alveolitis. Pulmonary manifestations are often associated with antisynthetase antibodies; myositis associated with the ASS can be considered a distinct disorder (discussed below). DM can present in children (juvenile DM) or in adults. There is a higher risk for malignancy in adult onset cases, ~15% within the first 2–3 years.

Laboratory Features Serum CK levels are elevated in 70–80% of patients; in 10% of those with normal CK, serum aldolase may be increased. Antinuclear antibodies can be positive but are a nonspecific finding. DM is associated with several MSA targeting melanoma differentiation antigen 5 (MDA5), transcriptional intermediary factor 1 (TIF1), Mi-2 and nuclear matrix protein 2 (NXP2). These antibodies are usually associated with characteristic clinical features. For example, anti-MDA5 antibodies are associated with amyopathic DM with severe palmar rash, digital ulcers, and rapidly progressive ILD. Anti-TIF1 (or p155) antibodies and anti-NXP2 antibodies are associated with an increased risk of cancer, while anti-Mi-2 antibodies are often associated with more benign DM and a favorable response to treatment.

EMG of weak muscles shows increased insertional and spontaneous activity in the form of positive sharp waves and fibrillation potentials, or complex repetitive discharges along with early recruitment of small amplitude, short duration, polyphasic motor units. These findings are non-specific and can be seen in other myopathies. Skeletal muscle magnetic resonance imaging (MRI muscle) reveals edema in affected muscles, and sometimes more specific findings of abnormalities of fascia suggesting fasciitis.

Histopathology and Pathogenesis The characteristic histopathological abnormality on muscle biopsy is perifasicular atrophy **(Fig. 358-2*A*)**; however, this is present in perhaps only 50% of patients. Immunohistochemical staining for myxovirus resistance protein A (MxA) is diagnostically more sensitive and highly specific **(Fig. 358-2*B*)**. The inflammatory cell infiltrate is predominantly perivascular and in the perimysium and is composed primarily of macrophages, B cells, and plasmacytoid dendritic cells (PDCs). Skin biopsies reveal cell-poor interface dermatitis, which is analogous to the perifasicular atrophy in that the basal layer of kerotinocytes are most damaged; the inflammatory infiltrate is typically absent or minimal, and when present is located mainly at the border zone of the dermis and epidermis.

The pathogenesis of DM was traditionally attributed to an antibody-mediated attack on endothelial cells, followed by complement-mediated destruction of capillaries and watershed ischemia of muscle fibers. However, recent studies suggest that this is not likely the case. Immunoglobulin deposition is largely absent on endothelial cells, and complement deposition may be a secondary phenomenon. There is increasing evidence that the microvasculopathy, skin, and muscle damage associated with DM is primarily due to toxicity from type I interferon-mediated pathways, most likely IFN-β.

Prognosis In the absence of malignancy, prognosis is generally favorable in patients with DM, with 5-year survival rates ranging from 70 to 93%. Poor prognostic features are increased age, associated ILD, cardiac disease, and late or previous inadequate treatment.

POLYMYOSITIS

Clinical Features PM is a heterogenous group of disorders that usually presents with symmetric and proximal weakness that worsens

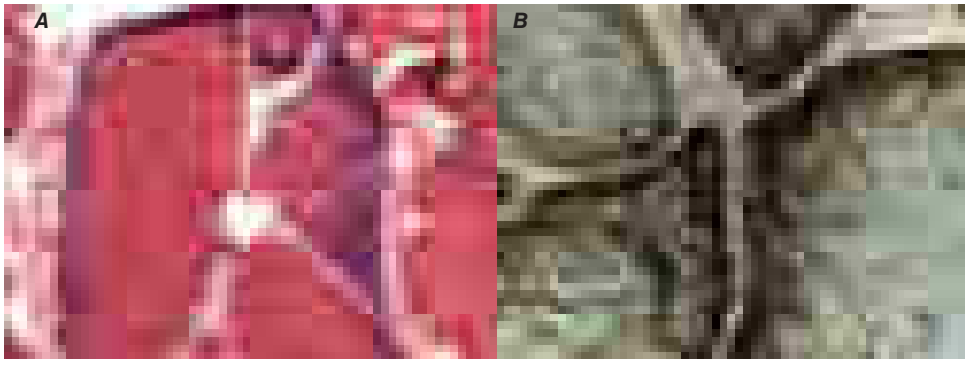


FIGURE 358-2 Perifascicular atrophy and myxovirus resistance protein A (MxA) expression in dermatomyositis. ***A.*** Perifascicular myofibers (*black arrows*) bordering on disrupted perimysial connective tissue are atrophic and basophilic on hematoxylin and eosin (H&E) stains. ***B***. Perifascicular myofibers (*white arrows*) show intense staining for MxA protein along a gradient from superficial to deep; all capillaries show intense MxA expression (*white arrowheads*).

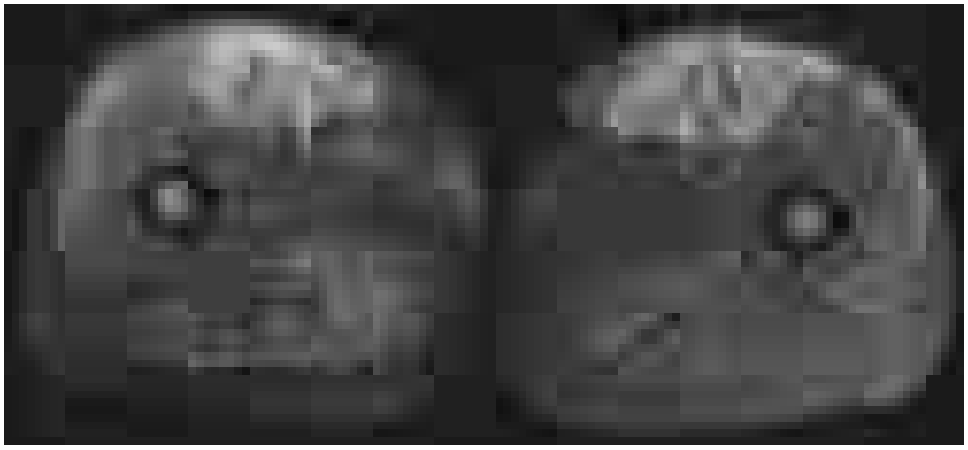


FIGURE 358-3 Skeletal muscle MRI with short T1 inversion recovery (STIR) imaging in polymyositis. MRI of the thigh demonstrates bright signal indicative of edema/inflammation, particularly in the rectus femoris muscle. This contrasts with MRI in IBM in which there is more selective involvement of the vastus lateralis and medialis with relative sparing of the rectus femoris (see Figure 358-7 ***F*** and ***G***).

over several weeks to months. As with DM, there can be associated heart, lung, and joint involvement as well as an increased risk of cancer. Some epidemiological studies suggest that the risk of cancer in PM is less than that in DM, but these older series likely included patients with IBM and dystrophies with inflammation who were misdiagnosed as having PM.

Laboratory Features CK levels are always elevated in uncontrolled PM. A normal CK should alert clinicians to the possibility of IBM. As in DM, EMG and skeletal muscle imaging can be abnormal, but the findings are not specific **(Fig. 358-3)**.

Histopathology and Pathogenesis Because PM is a heterogeneous category, muscle pathology varies substantially. Most often, patients with non-specific inflammatory cells present in perimysial more often than endomysial locations have been categorized as PM. A small minority of patients have mononuclear inflammatory infiltrate that surround fibers with sarcolemmal major histocompatibility (MHC-I) expression **(Fig. 358-4)**. There is debate as to whether true invasion of myofibers occurs in PM, or rather always indicates IBM. The inflammatory infiltrate predominantly consists of CD8+ T cells and macrophages located in the endomysial, perimysial and perivascular regions. As PM is heterogeneous, its varied forms of pathogenesis are poorly understood.

Prognosis Most patients with PM improve with immunotherapies, but usually require life-long treatment. Some retrospective studies suggest that PM does not respond as well as DM to these therapies.

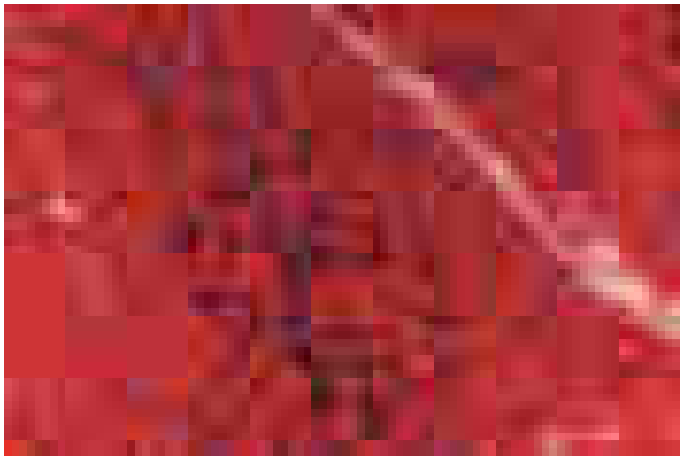


FIGURE 358-4 Pathology of polymyositis. Muscle biopsy demonstrates endomysial infiltrates surrounding non-necrotic muscle fibers.

However, many of these older series of "PM" likely included patients who actually had IMNM, IBM, or other myopathies (including muscular dystrophies) that do not respond to immunotherapies. As in DM, poor prognostic features are cancer, increased age, lung or cardiac involvement, and late or previously inadequate treatment.

OVERLAP SYNDROMES

The term "overlap syndrome" is applied when DM or PM is associated with other well-defined connective tissue diseases (CTDs) such as scleroderma, mixed connective tissue disease (MCTD), Sjögren syndrome, systemic lupus erythematosus (SLE), or rheumatoid arthritis. As in DM and PM, the myositis associated with these overlap syndromes is usually responsive to immunotherapies.

IMMUNE-MEDIATED NECROTIZING MYOPATHY

Clinical Features IMNM is characterized by the acute or insidious onset of symmetric, proximal more than distal weakness. Dysphagia, dysarthria, or myalgia may occur. Patients may have an underlying CTD (usually scleroderma or MCTD), cancer (paraneoplastic necrotizing myopathy), or it may be idiopathic. There are at least two distinct forms of IMNM associated with specific autoantibodies (anti-HMGCR and anti-signal recognition particle [SRP]). HMGCR myopathy can be seen in patients receiving statins, inhibitors of 3-hydroxy-3-methylglutaryl-coenzyme reductase (HMGCR), particularly in patients aged >50 years. However, HMGCR myopathy can develop in children and young adults without a history of statin use and can mimic a limb girdle muscular dystrophy. Unlike the more common "toxic" myopathy associated with statin use, HMGCR myopathy does not improve when statins are discontinued. SRP myopathies are notable for the presence of anti-SRP antibodies and a typically subacute, aggressive, and relatively refractory course.

Laboratory Features CK levels are markedly elevated (usually >10 × normal) in IMNM. As mentioned, IMNM can be associated with anti-HMGCR or anti-SRP antibodies. EMG and skeletal muscle imaging findings are non-specifically abnormal.

Histopathology and Pathogenesis Muscle biopsies reveal multifocal necrotic and regenerating muscle fibers with a paucity of inflammatory cells **(Fig. 358-5)**. However, some patients with HMGCR myopathy have endomysial, macrophage-predominant infiltrates similar to what is seen in PM. Overexpression of MHC-I and membrane attack complex (MAC) may be evident on sarcolemma of non-necrotic fibers and MAC deposition on capillaries. The pathogenesis of IMNM is poorly understood.

Prognosis IMNM is generally much more difficult to treat than either DM or PM and aggressive immunotherapy is usually required. The progressive course despite immunotherapy and marked weakness with atrophy can lead to a misdiagnosis of a limb-girdle muscular dystrophy. There may be an increased incidence of cancer in patients with HMGCR myopathy, thus patients should undergo a malignancy workup.

ANTISYNTHETASE SYNDROME

Clinical Features The presence of myositis, non-erosive arthritis, ILD, Raynaud phenomenon, mechanic hands, and fever associated with antibodies against aminoacyl-tRNA synthetase constitute the ASS. Some patients have an erythematous rash and muscle biopsies share histopathological features of DM, which likely accounts for many of these patients being classified as having DM.

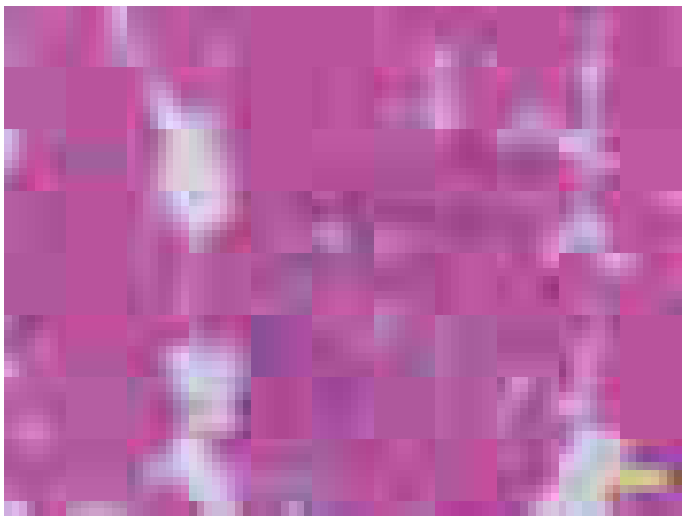


FIGURE 358-5 Pathology of immune-mediated necrotizing myopathy. Muscle biopsy demonstrates scattered necrotic fibers with inflammatory infiltrate confined to those fibers undergoing myophagocytosis along with a few regenerating fibers.

Laboratory Features Antibodies against aminoacyl-tRNA synthetases are the most common MSA, present in 25–35% of patients with myositis. The most common aminoacyl-tRNA synthetase antibody is anti-Jo-1. CK is usually elevated in patients with ASS and myositis. Those with ILD demonstrate reduced forced vital capacity and diffusion capacity on pulmonary function tests. Spiral chest CT scans are best at demonstrating the honeycomb pattern of ILD. Skeletal muscle MRI and EMG show abnormalities similar to DM, PM, and IMNM.

Histopathology and Pathogenesis Muscle biopsies demonstrate a predilection for perimysial damage including perimysial fragmentation and staining with alkaline phosphatase (Fig. 358-6), PDCs and macrophages in the perimysium and around blood vessels, and MAC deposition on capillaries. Also similar to DM there is perifascicular muscle fiber damage, but with AAS there is more perifascicular muscle fiber necrosis compared to DM in which perifascicular atrophy is more prominent. MHC-1 and MAC deposits on muscle fibers may be seen on sarcolemma of perifascicular muscle fibers.

Prognosis Most patients respond to treatment although responses are less complete than for DM and PM; ILD can be particularly refractory to treatment. Unlike DM, PM, and IMNM, there does not appear to be an increased risk of malignancy.

INCLUSION BODY MYOSITIS

Clinical Features IBM usually manifests in patients over the age of 50 years and is slightly more common in men than women. It is associated with slowly progressive weakness and muscle atrophy that has a predilection for early involvement of the wrist and finger flexors in the arms and quadriceps in the legs (Fig. 358-7). Weakness is often asymmetric. Dysphagia is common and rarely can be the presenting feature. These clinical features can help distinguish IBM from PM and other forms of myopathy. The mean duration from onset of symptoms to use of wheelchair or scooter is ~15 years. There is no known increased risk of malignancy.

Laboratory Features CK levels can be normal or only slightly elevated (usually <10 times normal). Antibodies targeting cytosolic 5′-nucleotidase 1A (cN-1A) are detected in the blood in a third to more than two-thirds of IBM patients, and is a highly specific diagnostic biomarker for IBM among patients with myopathy. Other blood biomarkers for IBM include the presence of an abnormal population of large granular lymphocytes on flow cytometry and a reduced CD4/CD8 ratio with an increased CD8 count. Needle EMG may demonstrate large amplitude, long duration motor unit potentials that can be misinterpreted as neurogenic but reflect the chronicity of the myopathy. Muscle MRI may show a predilection for involvement of the flexor digitorum profundus in the arms and the vastus medialis and lateralis muscles with sparing of the rectus femoris muscle.

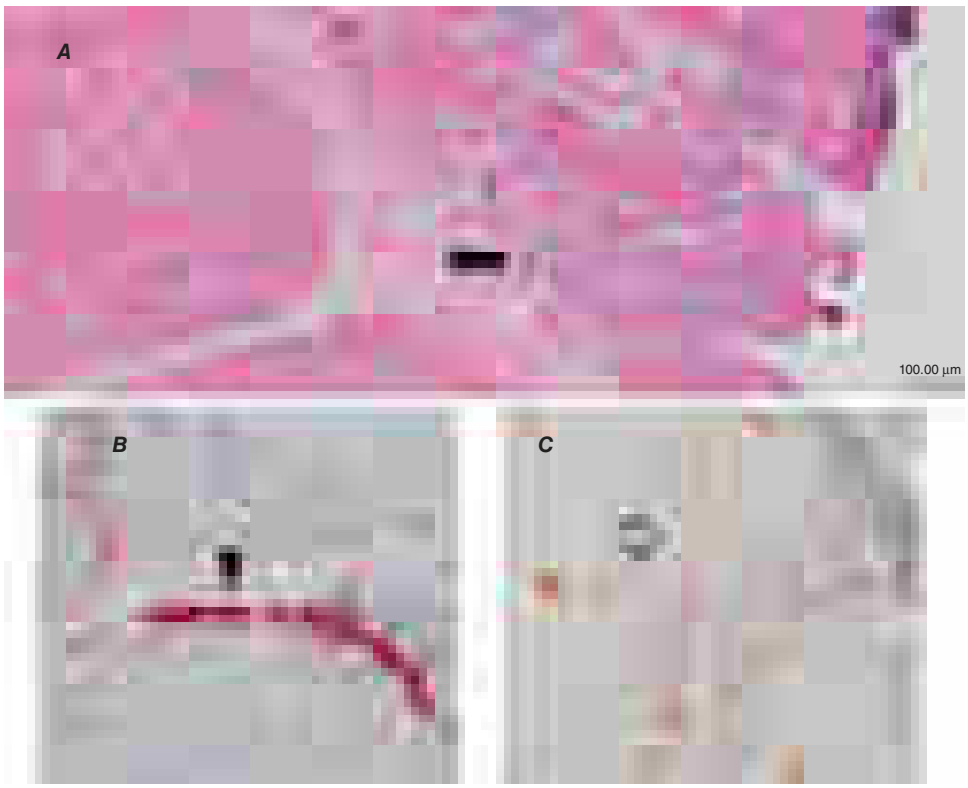


FIGURE 358-6 Pathology of myositis with anti-Jo-1 antibodies (antisynthetase syndrome). *A.* Perifascicular/perimysial muscle fiber atrophy and necrosis (*thin arrow*) associated with perimysial connective tissue is edematous and fragmented in appearance (*thick arrow*), H&E stain. ***B.*** The perimysial connective tissue intensely stains red with alkaline phosphatase stain (*arrowhead*). ***C.*** Immunostaining demonstrates deposition of membrane attack complex (MAC) deposits on the sarcolemma of non-necrotic perifascicular muscle fibers (*open arrow*).

Histopathology and Pathogenesis Muscle biopsies demonstrate endomysial inflammatory infiltrates predominantly composed of CD8+ T cells and macrophages surrounding and invading non-necrotic muscle fibers, MHC-1 expression on the sarcolemma, fibers with rimmed vacuoles, cytochrome oxidase (COX) negative fibers, and inclusions on light or electron microscopy (Fig. 358-8). The inclusions contain beta-sheet misfolded proteins (amyloid), but are difficult to appreciate with routine Congo red stain (they are seen on frozen but not paraffin sections). Immunostaining for p62 appears to be the most sensitive stain for detection of these inclusions. Importantly, rimmed vacuoles may not be seen in as many as 20–30% of muscle biopsies. In such cases, the presence of mitochondrial abnormalities (ragged red and COX negative fibers) and immunostaining demonstrating p62 inclusions are helpful in distinguishing IBM from PM (aside from the clinical pattern of muscle weakness).

The pathogenesis of IBM is poorly understood. The marked adaptive immune system abnormalities related to T cell inflammation and the presence of a relatively specific autoantibody against a muscle protein indicate an autoimmune attack on muscle. The

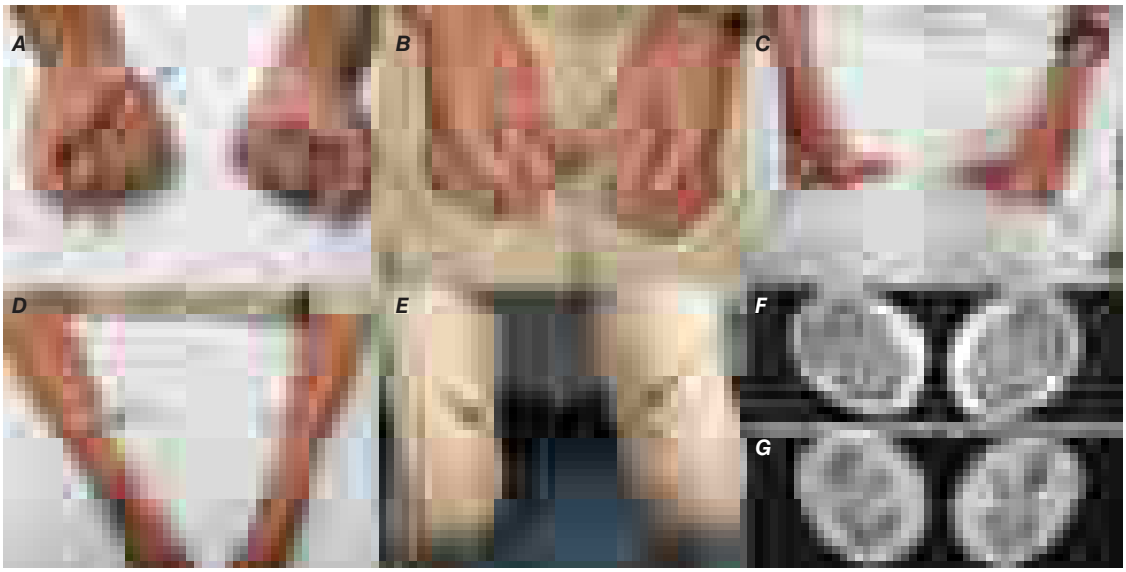


FIGURE 358-7 Muscle manifestations of inclusion body myositis (*A*-*C*). Finger flexor weakness can be (***A***) subtle and multifocal (*black arrows*), (***B***) moderate, or (***C***) severe. Note even with complete paralysis of deep and superficial finger flexors, metacarpophalangeal joint flexion (*arrows*) is often maintained due to preservation of lumbricals. ***D***. Ventral forearm atrophy (*arrows*). ***E.*** Atrophy of medial thighs due to loss of vastus medialis (*arrows*). ***F.*** Early IBM, with relatively preserved vastus medialis (*arrows*), in contrast to (***G***) advanced IBM with marked fibrous replacement of vastus medialis (*arrows*).

chronic and highly inflammatory environment within muscles in IBM may alter protein synthesis and degradation pathways in part via aberrant immunoproteasome expression. Additional histologic features, typically referred to as "degenerative," include aggregation of various proteins including markers of endoplasmic reticulum (ER) stress and autophagy (e.g., p62 and LC3). Involvement of ER stress and autophagy have also been observed in other autoimmune diseases, such as primary biliary cholangitis (PBC), inflammatory bowel disease, and ankylosing spondylitis, some of which can be highly refractory to immunotherapy.

Prognosis The myopathy is slowly progressive, and is not typically responsive to immunotherapies. Most patients require a scooter or wheelchair within 10–15 years of onset of symptoms. Life expectancy is not significantly altered in IBM.

TREATMENT OF THE IM (TABLE 358-2)

DM, PM, ASS, and IMNM are typically responsive to immunotherapy. High dose glucocorticoids (i.e., starting dose of prednisone 0.75–1.0 mg/kg per day) is considered the first-line treatment. There is uncertainty regarding when to start second-line agents (e.g., methotrexate, azathioprine, mycophenolate, immunoglobulin, or rituximab). The clinician must weigh with the patient the increased risks of immunosuppression versus possible benefits (e.g., faster improvement, steroid-sparing effect and/or avoidance of the morbidities associated with long-term glucocorticoid use). We usually start a second line agent (usually methotrexate) with glucocorticoids in patients with severe weakness or other organ system involvement (e.g., myocarditis, ILD), those with increased risk of steroid complications (e.g., diabetics, osteoporosis, or postmenopausal women), and patients with IMNM who are known to have difficult to treat myositis. In those in whom we initiate treatment with prednisone alone, a second-line agent is added in patients who fail to significantly improve after 2–4 months of treatment or in those who cannot be tapered to a low dose of prednisone.

Most patients with IMNM do not respond to prednisone alone or even prednisone plus a second-line agent in combination. Many require triple therapy with prednisone, methotrexate, and intravenous immunoglobulin (IVIG), and if this fails, rituximab. Recent reports suggest that anti-HMGCR myopathy may respond to monotherapy with IVIG, and a large multicenter clinical trial to test this approach is currently being organized.

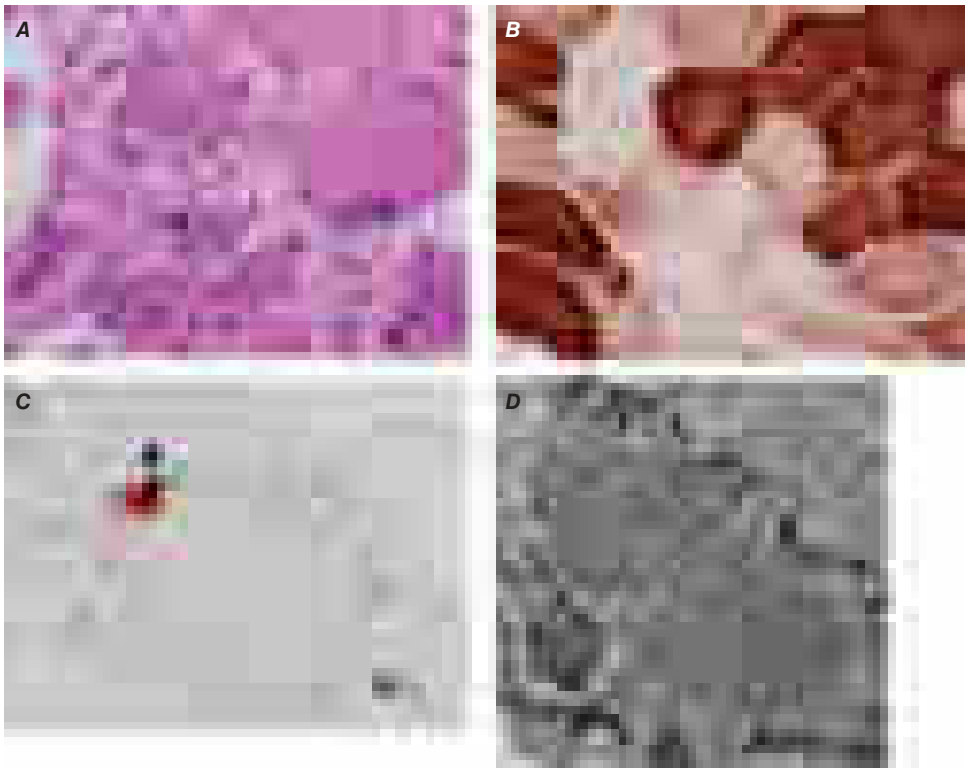


FIGURE 358-8 Pathology of inclusion body myositis. *A.* Scattered muscle fibers with rimmed vacuoles and rare fibers with eosinophilic inclusions (*arrow*), H&E stain. ***B.*** Cytochrome oxidase stain demonstrates an increased number of pale staining or COX negative muscle fibers. ***C.*** Cytoplasmic inclusions stain positive with p62 within a muscle fiber (*thick arrow*). ***D.*** Electromicroscopy reveals 15–21 nm tubulofilamentous inclusions within a myonucleus.

TABLE 358-2 Immunotherapies for Inflammatory Myopathies

THERAPY	ROUTE	DOSE	SIDE EFFECTS	MONITOR
Prednisone	Oral	0.75–1.5 mg kg/d to start	Hypertension, fluid and weight gain, hyperglycemia, hypokalemia, cataracts, gastric irritation, osteoporosis, infection, aseptic femoral necrosis	Weight, blood pressure, serum glucose/potassium, cataract formation
Methylprednisone	Intravenous	1 g in 100 mL/normal saline over 1–2 h, daily or every other day for 3–6 doses	Arrhythmia, flushing, dysgeusia, anxiety, insomnia, fluid and weight gain, hyperglycemia, hypokalemia, infection	Heart rate, blood pressure, serum glucose/potassium
Azathioprine	Oral	2–3 mg/kg per day; single a.m. dose	Flu-like illness, hepatotoxicity, pancreatitis, leukopenia, macrocytosis, neoplasia, infection, teratogenicity	Blood count, liver enzymes
Methotrexate	Oral	7.5–20 mg weekly, single or divided doses; one day a week dosing	Hepatotoxicity, pulmonary fibrosis, infection, neoplasia, infertility, leukopenia, alopecia, gastric irritation, stomatitis, teratogenicity	Liver enzymes, blood count
	Subcutaneously	20–50 mg weekly; one day a week dosing	Same as oral	Same as p.o.
Cyclophosphamide	Oral Intravenous	1.5–2 mg/kg per day; single a.m. dose 0.5 to 1.0 g/m^2 per month × 6–12 months	Bone marrow suppression, infertility, hemorrhagic cystitis, alopecia, infections, neoplasia, teratogenicity	Blood count, urinalysis
Cyclosporine	Oral	4–6 mg/kg per day, split into two daily doses	Nephrotoxicity, hypertension, infection, hepatotoxicity, hirsutism, tremor, gum hyperplasia, teratogenicity	Blood pressure, creatinine/BUN, liver enzymes, cyclosporine levels
Tacrolimus	Oral	0.1–0.2 mg/kg per d in two divided doses	Nephrotoxicity, hypertension, infection, hepatotoxicity, hirsutism, tremor, gum hyperplasia, teratogenicity	Blood pressure, creatinine/BUN, liver enzymes, tacrolimus levels
Mycophenolate mofetil	Oral	Adults (1 g BID to 1.5 g BID) Children (600 mg/m^2 per dose BID (no >1 g/d in patients with renal failure)	Bone marrow suppression, hypertension, tremor, diarrhea, nausea, vomiting, headache, sinusitis, confusion, amblyopia, cough, teratogenicity, infection, neoplasia	Blood count
Intravenous Immunoglobulin	Intravenous	2 g/kg over 2–5 days; then 1-g/kg every 4–8 weeks as needed	Hypotension, arrhythmia, diaphoresis, flushing, nephrotoxicity, headache, aseptic meningitis, anaphylaxis, stroke	Heart rate, blood pressure, creatinine/BUN
Rituximab	Intravenous	A course is typically 750 mg-meter squared (up to 1 g) and repeated in 2 weeks Courses are then repeated usually every 6–18 months	Infusion reactions (as per IVIG), infection, progressive multifocal leukoencephalopathy	Some check B-cell count prior to subsequent courses (but this may not be warranted)

Abbreviations: BUN, blood urea nitrogen; CBC, complete blood count; Cr, creatinine; ECG, electrocardiogram; GI, gastrointestinal; IgA, immunoglobulin A; IVIG, intravenous immunoglobulin; LFT, liver function test; m^2, body surface area; mo: month(s); PFT, pulmonary function test; PRES, posterior reversible encephalopathy syndrome.

Source: Modified from AA Amato, JA Russell (eds): *Neuromuscular Disorders*, 2nd ed. New York, McGraw-Hill Education; 2016, Table 33-8, p. 859, with permission.

Unfortunately, IBM does not typically respond to any known immunotherapies. The mainstay of treatment is physical and occupational therapy to improve function and swallowing therapy (and sometimes esophageal dilation or cricopharyngeal myotomy) in those with dysphagia.

GENERAL GUIDELINES FOR USE OF SPECIFIC IMMUMOTHERAPIES

Glucocorticoids Treatment is initiated with prednisone (0.75–1.5 mg/kg up to 100 mg) administered as a daily morning single-dose (the most common dose used in adults is 60 mg daily). In patients with severe weakness or comorbidities (e.g., ILD, myocarditis), treatment with a short course of intravenous methylprednisolone (1 g daily for 3 days) is recommended prior to starting oral glucocorticoids. Patients are generally maintained on high-dose prednisone until strength normalizes or until improvement in strength has reached a plateau (usually 3–6 months). Subsequently, prednisone can be tapered by 5 mg every 2–4 weeks. Once the dose is reduced to 20 mg every day or every other day, the taper is slowed to 2.5 mg every 2–4 weeks. The goal is to taper prednisone to ≤10 mg daily. Although most patients improve, the response may not be complete and many will require at least a small dose of prednisone or a second-line agent to have a sustained remission. Serum CK levels are monitored; however, dose adjustments of prednisone and other immunotherapies are primarily based on the objective clinical examination and not the CK levels or the patients' subjective response. When no response is noted after an adequate trial of high-dose prednisone, alternative diagnoses (e.g., IBM or an inflammatory muscular dystrophy) and a repeat muscle biopsy should be considered.

Relapse of the myositis needs to be distinguished from steroid myopathy. Features suggesting a steroid myopathy include weakness developing while on high dosage, a normal serum CK, clinical features of steroid excess such as ecchymoses and "moon facies," and absence of muscle membrane irritability on EMG. By contrast, patients experiencing relapse of myositis may become weaker during the prednisone taper, have increasing serum CK levels, and display abnormal spontaneous activity on EMG.

SECOND-LINE THERAPIES

Methotrexate Methotrexate is usually the second-line treatment of choice because most authorities believe it works faster than other agents. An oral dose of 5 or 7.5 mg/week is initiated, and then gradually increased as needed up to 25 mg/week. If there is no improvement after 1 month of 25 mg/week of oral methotrexate, a switch to weekly parenteral (usually subcutaneous) methotrexate is the next step, with dose escalation by 5 mg weekly; only rarely is a dose >35 mg/week used. The major side effects of methotrexate are alopecia, stomatitis, ILD, teratogenicity, oncogenicity, risk of infection, and pulmonary fibrosis, along with bone marrow, renal, and liver toxicity. Patients are concomitantly treated with folate or folinic acid.

Azathioprine A recommended initial dose is 50 mg/day in adults, which can be increased by 50 mg every 2 weeks up to 2–3 mg/kg per d.

Approximately 12% of patients develop a systemic reaction characterized by fever, abdominal pain, nausea, vomiting, and anorexia that requires discontinuation of the drug. The major practical limitation of azathioprine is that 6–18 months of treatment is usually required before benefit can be seen. Patients can be prescreened for thiopurine methyltransferase (TPMT) deficiency that is associated with severe bone marrow toxicity from this drug.

Mycophenolate Mofetil This drug inhibits the proliferation of T and B lymphocytes by blocking purine synthesis. It appears to be effective in different forms of myositis, and is the second-line treatment of choice for myositis patients with ILD. The starting dose is 1.0 g twice daily and can be increased to 3 g daily in divided doses, if necessary. Mycophenylate is excreted through the kidneys; therefore, the dose should be decreased (no >1 g/d total dose) in patients with renal insufficiency. An advantage of mycophenylate compared to other immunosuppressive agents is the lack of renal or hepatic toxicity.

Intravenous Immunoglobulin IVIG is used in patients refractory to prednisone and at least one second-line immunosuppressive agent, although recent reports suggest that it may be the treatment of choice and effective as a monotherapy in anti-HMCCR myopathy. A dose of 2 g/kg is divided over 2–5 days, and repeat infusions given at monthly intervals for at least 3 months. Subsequently, intervals can be lengthened or dosage decreased: 2 g/kg every 2 months or 1 g/kg per month.

Rituximab Rituximab is a monoclonal antibody directed against CD20+ B-cells. A large randomized controlled trial found no benefit, but there were flaws in the study design. Most authorities feel that rituximab can be beneficial in some patients who are refractory to prednisone and at least one of the other second-line agents. The typical dosage is 750 mg/m^2 (up to 1 g) IV with a second infusion 2 weeks later, and repeat courses (375 mg/m^2 as a single infusion or with a second infusion 2 weeks apart) every 6–18 months as needed.

GLOBAL ISSUES

There is a lack of epidemiological data regard to the incidence and prevalence of various subtypes of IM throughout the world. Complicating the issue is disease awareness and the inability to obtain and process muscle biopsies and MSAs, particularly in less developed countries. Nevertheless each of these disorders occurs throughout the world. The specific environmental triggers and genetic risk factors are likely variable. Interestingly, a report out of Japan found that 28% of IBM patients had evidence of exposure to hepatitis C, which was much higher than seen in the western hemisphere and also more common than seen in PM and healthy population controls in Japan. HIV-associated PM and IBM are more commonly encountered in areas endemic for HIV and recent studies suggest most of these "PM" cases turn out to have IBM and can develop symptoms at an earlier age (e.g., in the 30s). Pyomyositis and parasitic myositis are clearly more common in the tropics. The prevalence of different types of cancers vary in different parts of the world, an important consideration with respect to paraneoplastic myositis seen in DM, PM, and IMNM. For example, nasopharyngeal cancer is particularly common in Asia, thus assessment for this type of cancer should be considered in the workup of patients from high-risk regions.

FURTHER READING

Amato AA, Russell JA (eds): *Neuromuscular Disorders*, 2nd ed. New York, McGraw-Hill Education, 2016, pp. 827–871.

Greenberg SA et al: Association of inclusion body myositis with T cell large granular lymphocytic leukaemia. Brain 139:1348, 2016.

Huard C et al: Correlation of cutaneous disease activity with type 1 interferon gene signature and interferon beta in dermatomyositis. Br J Dermatol 176:1224, 2017.

Larman HB et al: Cytosolic 5′-nucleotidase 1A autoimmunity in sporadic inclusion body myositis. Ann Neurol 73:408, 2013.

Lundberg IE et al: Diagnosis and classification of idiopathic inflammatory myopathies. J Intern Med 280:39, 2016.

Mammen AL et al: Autoantibodies against 3-hydroxy-3-methylglutaryl-coenzyme A reductase in patients with statin-associated autoimmune myopathy. Arthritis Rheum 63:713, 2011.

Pluk H et al: Autoantibodies to cytosolic 5′-nucleotidase 1A in inclusion body myositis. Ann Neurol 73:397, 2013.

Rose MR, ENMC IBM Working Group: 188th ENMC International Workshop: Inclusion Body Myositis, 2–4 December 2011, Naarden, The Netherlands. Neuromuscul Disord 23:1044, 2013.

Watanabe Y et al: Clinical features and prognosis in anti-SRP and anti-HMGCR necrotising myopathy. J Neurol Neurosurg Psychiatry 87:1038, 2016.

359 Relapsing Polychondritis

Carol A. Langford

Relapsing polychondritis is an uncommon disorder of unknown cause characterized by inflammation of cartilage predominantly affecting the ears, nose, and laryngotracheobronchial tree. Other manifestations include scleritis, neurosensory hearing loss, polyarthritis, cardiac abnormalities, skin lesions, and glomerulonephritis. Relapsing polychondritis has been estimated to have an incidence of 3.5 per million population per year. The peak age of onset is between the ages of 40 and 50 years, but relapsing polychondritis may affect children and the elderly. It is found in all races, and both sexes are equally affected. No familial tendency is apparent. A significantly higher frequency of HLA-DR4 has been found in patients with relapsing polychondritis than in healthy individuals. A predominant subtype allele(s) of HLA-DR4 was not found. Approximately 30% of patients with relapsing polychondritis will have another rheumatologic disorder, the most frequent being systemic vasculitis, followed by rheumatoid arthritis, and systemic lupus erythematosus (SLE). Nonrheumatic disorders have also been associated with relapsing polychondritis (Table 359-1). In most cases, these disorders antedate the appearance of relapsing polychondritis, usually by months or years; however, in other instances, the onset of relapsing polychondritis can accompany disease presentation.

PATHOLOGY AND PATHOPHYSIOLOGY

The earliest abnormality of hyaline and elastic cartilage noted histologically is a focal or diffuse loss of basophilic staining indicating depletion of proteoglycan from the cartilage matrix. Inflammatory infiltrates are found adjacent to involved cartilage and consist predominantly

TABLE 359-1 Disorders Associated with Relapsing Polychondritis[a]

Systemic vasculitis
Rheumatoid arthritis
Systemic lupus erythematosus
Overlapping connective tissue disease
Spondyloarthritides
Behçet's disease
Polymyalgia rheumatica
Primary biliary cirrhosis
Pulmonary fibrosis
Hashimoto's thyroiditis
Grave's disease
Crohn's disease
Ulcerative colitis
Myelodysplastic syndrome

[a]Systemic vasculitis is the most common association, followed by rheumatoid arthritis and systemic lupus erythematosus.

Source: Modified from CJ Michet et al: Ann Intern Med 104:74, 1986.

of mononuclear cells and occasional plasma cells. In acute disease, polymorphonuclear white cells may also be present. Destruction of cartilage begins at the outer edges and advances centrally. There is lacunar breakdown and loss of chondrocytes. Degenerating cartilage is replaced by granulation tissue and later by fibrosis and focal areas of calcification. Small loci of cartilage regeneration may be present. Immunofluorescence studies have shown immunoglobulins and complement at sites of involvement. Extracellular granular material observed in the degenerating cartilage matrix by electron microscopy has been interpreted to be enzymes, immunoglobulins, or proteoglycans.

The accumulating data strongly suggest that both humoral and cell-mediated immunity play an important role in the pathogenesis of relapsing polychondritis. Immunoglobulin and complement deposits are found at sites of inflammation. In addition, antibodies to type II collagen and to matrilin-1 and immune complexes are detected in the sera of some patients. The possibility that an immune response to type II collagen may be important in the pathogenesis is supported experimentally by the occurrence of auricular chondritis in rats immunized with type II collagen. Antibodies to type II collagen are found in the sera of these animals, and immune deposits are detected at sites of ear inflammation. Humoral immune responses to type IX and type XI collagen, matrilin-1, and cartilage oligomeric matrix protein have been demonstrated in some patients. In a study, rats immunized with matrilin-1 were found to develop severe inspiratory stridor and swelling of the nasal septum. The rats had severe inflammation with erosions of the involved cartilage, which was characterized by increased numbers of CD4+ and CD8+ T cells in the lesions. The cartilage of the joints and ear pinna was not involved. All had IgG antibodies to matrilin-1. Matrilin-1 is a noncollagenous protein present in the extracellular matrix in cartilage. It is present in high concentrations in the trachea and is also present in the nasal septum but not in articular cartilage. A subsequent study demonstrated serum anti-matrilin-1 antibodies in ~13% of patients with relapsing polychondritis; ~70% of these patients had respiratory symptoms. Cell-mediated immunity may also be operative in causing tissue injury, since lymphocyte transformation can be demonstrated when lymphocytes of patients are exposed to cartilage extracts. T cells specific for type II collagen have been found in some patients, and CD4+ T cells have been observed at sites of cartilage inflammation.

CLINICAL MANIFESTATIONS

The onset of relapsing polychondritis is frequently abrupt, with the appearance of one or two sites of cartilaginous inflammation. The pattern of cartilaginous involvement and the frequency of episodes vary widely among patients. Noncartilaginous presentations may also occur. Systemic inflammatory features such as fever, fatigue, and weight loss occur and may precede the clinical signs of relapsing polychondritis by several weeks. Relapsing polychondritis may go unrecognized for several months or even years in patients who only initially manifest intermittent joint pain and/or swelling, or who have unexplained eye inflammation, hearing loss, valvular heart disease, or pulmonary symptoms.

Auricular chondritis is the most frequent presenting manifestation of relapsing polychondritis, occurring in 40% of patients and eventually affecting about 85% of patients **(Table 359-2)**. One or both ears are involved, either sequentially or simultaneously. Patients experience the sudden onset of pain, tenderness, and swelling of the cartilaginous portion of the ear **(Fig. 359-1)**. This typically involves the pinna of the ears, sparing the earlobes because they do not contain cartilage. The overlying skin has a beefy red or violaceous color. Prolonged or recurrent episodes lead to cartilage destruction and result in a flabby or droopy ear. Swelling may close off the eustachian tube or the external auditory meatus, either of which can impair hearing. Inflammation of the internal auditory artery or its cochlear branch produces hearing loss, vertigo, ataxia, nausea, and vomiting. Vertigo is almost always accompanied by hearing loss.

Approximately 61% of patients will develop nasal involvement, with 21% having this at the time of presentation. Patients may experience nasal stuffiness, rhinorrhea, and epistaxis. The bridge of the nose and surrounding tissue become red, swollen, and tender and may collapse, producing a saddle nose deformity **(Fig. 359-2)**. In some patients, nasal deformity develops insidiously without overt inflammation. Saddle nose is observed more frequently in younger patients, especially in women.

TABLE 359-2 Clinical Manifestations of Relapsing Polychondritis

CLINICAL FEATURE	PRESENTING	CUMULATIVE
	Frequency, %	
Auricular chondritis	43	89
Arthritis	32	72
Nasal chondritis	21	61
Ocular inflammation	18	59
Laryngotracheal symptoms	23	55
Reduced hearing	7	40
Saddle nose deformity	11	25
Cutaneous	4	25
Laryngotracheal stricture	15	23
Vasculitis	2	14
Elevated creatinine	7	13
Aortic or mitral regurgitation	0	12

Source: Modified from PD Kent et al: Curr Opin Rheumatol 16:56, 2004.

Joint involvement is the presenting manifestation in relapsing polychondritis in approximately one-third of patients and may be present for several months before other features appear. Eventually, more than one-half of the patients will have arthralgias or arthritis. The arthritis is usually asymmetric and oligo- or polyarticular, and it involves both large and small peripheral joints. An episode of arthritis lasts from a few days to several weeks and resolves spontaneously without joint erosion or deformity. Attacks of arthritis may not be temporally related to other manifestations of relapsing polychondritis. Joint fluid has been reported to be noninflammatory. In addition to peripheral joints, inflammation may involve the costochondral, sternomanubrial, and sternoclavicular cartilages. Destruction of these cartilages may result in a pectus excavatum deformity or even a flail anterior chest wall.

Eye manifestations occur in more than one-half of patients and include conjunctivitis, episcleritis, scleritis, iritis, uveitis, and keratitis. Ocular inflammation can be severe and visually threatening. Other manifestations include eyelid and periorbital edema, proptosis, optic neuritis, extraocular muscle palsies, retinal vasculitis, and renal vein occlusion.

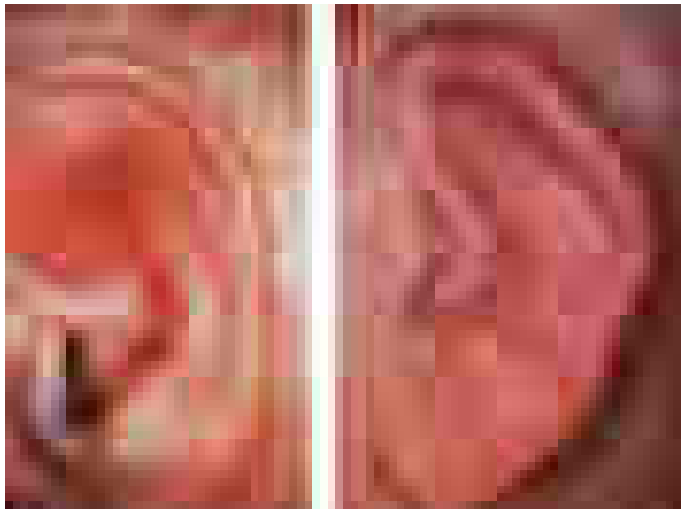


FIGURE 359-1 ***Left.*** The pinna is erythematous, swollen, and tender. Not shown is the ear lobule that is spared as there is no underlying cartilage. ***Right.*** The pinna is thickened and deformed. The destruction of the underlying cartilage results in a floppy ear. *(Reprinted from the Clinical Slide Collection on the Rheumatic Diseases, © 2018 American College of Rheumatology. Used by permission of the American College of Rheumatology.)*

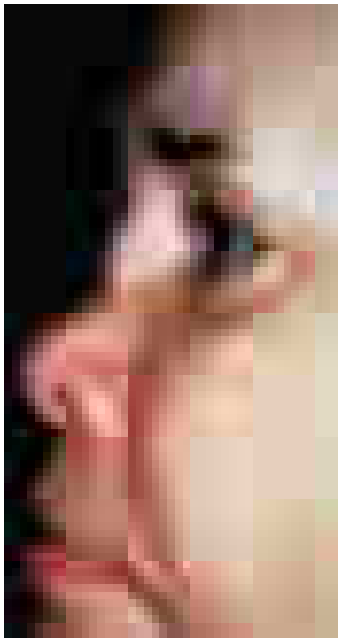


FIGURE 359-2 Saddle nose results from destruction and collapse of the nasal cartilage. *(Reprinted from the Clinical Slide Collection on the Rheumatic Diseases, © 2018 American College of Rheumatology. Used by permission of the American College of Rheumatology.)*

Laryngotracheobronchial involvement occurs in ~50% of patients and is among the most serious manifestations of relapsing polychondritis. Symptoms include hoarseness, a nonproductive cough, and tenderness over the larynx and proximal trachea. Mucosal edema, strictures, and/or collapse of laryngeal or tracheal cartilage may cause stridor and life-threatening airway obstruction necessitating tracheostomy. Involvement can extend into the lower airways resulting in tracheobronchomalacia. Collapse of cartilage in bronchi leads to pneumonia and, when extensive, to respiratory insufficiency.

Cardiac valvular regurgitation occurs in about 5–10% of patients and is due to progressive dilation of the valvular ring or to destruction of the valve cusps. Aortic regurgitation occurs in about 7% of patients, with the mitral and other heart valves being affected less often. Other cardiac manifestations include pericarditis, myocarditis, coronary vasculitis, and conduction abnormalities. Aneurysms of the proximal, thoracic, or abdominal aorta may occur even in the absence of active chondritis and occasionally rupture.

Renal disease occurs in about 10% of patients. The most common renal lesions include mesangial expansion or segmental necrotizing glomerulonephritis, which have been reported to have small amounts of electron-dense deposits in the mesangium where there is also faint deposition of C3 and/or IgG or IgM. Tubulointerstitial disease and IgA nephropathy have also been reported.

Approximately 25% of patients have skin lesions, which can include purpura, erythema nodosum, erythema multiforme, angioedema/urticaria, livedo reticularis, and panniculitis.

Features of vasculitis are seen in up to 25% of patients and can affect any size vessel. Large vessel vasculitis may present with aortic aneurysms, and medium vessel disease may affect the coronary, hepatic, mesenteric, or renal arteries or vessel supplying nerves. Skin vessel disease and involvement of the postcapillary venules can also occur. A variety of primary vasculitides have also been reported to occur in association with relapsing polychondritis **(Chap. 356)**. One specific overlap is the "MAGIC" syndrome (*m*outh *a*nd *g*enital ulcers with *i*nflamed *c*artilage) in which patients present with features of both relapsing polychondritis and Behçet's disease **(Chap. 357)**.

LABORATORY FINDINGS AND DIAGNOSTIC IMAGING

There are no laboratory features that are diagnostic for relapsing polychondritis. Mild leukocytosis and normocytic, normochromic anemia are often present. Eosinophilia is observed in 10% of patients. The erythrocyte sedimentation rate and C-reactive protein are usually elevated. Rheumatoid factor and antinuclear antibody tests are occasionally positive in low titers, and complement levels are normal. Antibodies to type II collagen are present in fewer than one-half of the patients and are not specific. Circulating immune complexes may be detected, especially in patients with early active disease. Elevated levels of γ globulin may be present. Antineutrophil cytoplasmic antibodies (ANCA), either cytoplasmic (cANCA) or perinuclear (pANCA), are found in some patients with active disease. However, on target antigen–specific testing, there are only occasional reports of positive myeloperoxidase-ANCA, and proteinase 3-ANCA are very rarely found in relapsing polychondritis.

The upper and lower airways can be evaluated by imaging techniques such as computed tomography and magnetic resonance imaging (MRI). Bronchoscopy provides direct visualization of the airways but can be a high-risk procedure in patients with airway compromise. Pulmonary function testing with flow-volume loops can show inspiratory and/or expiratory obstruction. Imaging can also be useful to detect extracartilaginous disease. The chest film may show widening of the ascending or descending aorta due to an aneurysm, and cardiomegaly when aortic insufficiency is present. MRI can assess aortic aneurysmal dilatation. Electrocardiography and echocardiography can be useful in further evaluating for cardiac features of disease.

DIAGNOSIS

Diagnosis is based on recognition of the typical clinical features. Biopsies of the involved cartilage from the ear, nose, or respiratory tract will confirm the diagnosis but are only necessary when clinical features are not typical. Diagnostic criteria were suggested in 1976 by McAdam et al and modified by Damiani and Levine in 1979. These criteria continue to be generally used in clinical practice. McAdam et al proposed the following: (1) recurrent chondritis of both auricles; (2) nonerosive inflammatory arthritis; (3) chondritis of nasal cartilage; (4) inflammation of ocular structures, including conjunctivitis, keratitis, scleritis/episcleritis, and/or uveitis; (5) chondritis of the laryngeal and/or tracheal cartilages; and (6) cochlear and/or vestibular damage manifested by neurosensory hearing loss, tinnitus, and/or vertigo. The diagnosis is certain when three or more of these features are present along with a positive biopsy from the ear, nasal, or respiratory cartilage. Damiani and Levine later suggested that the diagnosis could be made when one or more of the above features and a positive biopsy were present, when two or more separate sites of cartilage inflammation were present that responded to glucocorticoids or dapsone, or when three or more of the above features were present.

The differential diagnosis of relapsing polychondritis is centered around its sites of clinical involvement. Patients with granulomatosis with polyangiitis (Wegener's) may have a saddle nose and tracheal involvement but can be distinguished by the primary inflammation occurring in the mucosa at these sites, the absence of auricular involvement, and the presence of pulmonary parenchymal disease. Patients with Cogan's syndrome have interstitial keratitis and vestibular and auditory abnormalities, but this syndrome does not involve the respiratory tract or ears. Reactive arthritis may initially resemble relapsing polychondritis because of oligoarticular arthritis and eye involvement, but it is distinguished in time by the appearance of urethritis and typical mucocutaneous lesions and the absence of nose or ear cartilage involvement. Rheumatoid arthritis may initially suggest relapsing polychondritis because of arthritis and eye inflammation. The arthritis in rheumatoid arthritis, however, is erosive and symmetric. In addition, rheumatoid factor titers are usually high compared with those in relapsing polychondritis, and anti-cyclic citrullinated peptide is usually not seen. Bacterial infection of the pinna may be mistaken for relapsing polychondritis but differs by usually involving only one ear, including the earlobe. Auricular cartilage may also be damaged by trauma or frostbite. Nasal destructive disease and auricular abnormalities can also be seen in patients using cocaine adulterated with levamisole. Ear involvement in this setting differs from relapsing polychondritis by typically manifesting as purpuric plaques with necrosis extending to the pinna, which does not contain cartilage.

TREATMENT

Relapsing Polychondritis

In patients with active chondritis, prednisone, 40–60 mg/d, is often effective in suppressing disease activity; it is tapered gradually once disease is controlled. In some patients, prednisone can be stopped, whereas in others, low doses in the range of 5–10 mg/d are required for continued suppression of disease. Dapsone 50–100 mg/d has been effective for cartilage inflammation and joint features in some patients; however, its use can be limited by the complication of hemolytic anemia and other side effects as well as contraindications in glucose-6-phosphate dehydrogenase (G6PD) deficiency and pregnancy. Other immunosuppressive drugs such as cyclophosphamide, methotrexate, azathioprine, or cyclosporine should be reserved for patients who have severe organ-threatening disease, fail to respond to prednisone, or require high doses to control disease activity. Patients with significant ocular inflammation often require intraocular glucocorticoids as well as high doses of prednisone. There are a small number of reports on the use of tumor necrosis factor antagonists, rituximab (anti-CD20), and tocilizumab (anti-interleukin 6 receptor), which are too few in number to assess efficacy. Heart valve replacement or repair of an aortic aneurysm may be necessary. When airway obstruction is severe, tracheostomy is required. Stents may be necessary in patients with tracheobronchial collapse.

PATIENT OUTCOME, PROGNOSIS, AND SURVIVAL

The course of relapsing polychondritis is highly variable. Some patients experience inflammatory episodes lasting from a few days to several weeks that then subside spontaneously or with treatment. Attacks may recur at intervals varying from weeks to months. In other patients, the disease has a chronic, smoldering course that may be severe. In one study, the 5-year estimated survival rate was 74% and the 10-year survival rate was 55%. About one-half of the deaths could be attributed to relapsing polychondritis or complications of treatment. Airway complications accounted for 10% of all fatalities although higher rates have been reported in other series. In general, patients with more widespread disease have a worse prognosis.

FURTHER READING

Chopra R et al: Relapsing polychondritis. Rheum Dis Clin North Am 39:263, 2013.

Ernst A et al: Relapsing polychondritis and airway involvement. Chest 135:1024, 2009.

Kent PD et al: Relapsing polychondritis. Curr Opin Rheumatol 16:56, 2004.

Longo L et al: Relapsing polychondritis: A clinical update. Autoimmun Rev 15:539, 2016.

Trentham DE, Le CH: Relapsing polychondritis. Ann Intern Med 129:114, 1998.

360 Sarcoidosis

Robert P. Baughman, Elyse E. Lower

DEFINITION

Sarcoidosis is an inflammatory disease characterized by the presence of noncaseating granulomas. The disease is often multisystem and requires the presence of involvement in two or more organs for a specific diagnosis. The finding of granulomas is not specific for sarcoidosis, and other conditions known to cause granulomas must be ruled out. These conditions include mycobacterial and fungal infections, malignancy, and environmental agents such as beryllium. Although sarcoidosis can affect virtually every organ of the body, the lung is most commonly affected. Other organs commonly affected are the liver, skin, and eye. The clinical outcome of sarcoidosis varies, with remission occurring in over one-half of patients within a few years of diagnosis; however, the remaining patients may develop a chronic disease that lasts for decades.

ETIOLOGY

Despite multiple investigations, the cause of sarcoidosis remains unknown. Currently, the most likely etiology is an infectious or noninfectious environmental agent that triggers an inflammatory response in a genetically susceptible host. Among the possible infectious agents, careful studies have shown a much higher incidence of *Propionibacter acnes* in the lymph nodes of sarcoidosis patients compared to controls. An animal model has shown that *P. acnes* can induce a granulomatous response in mice similar to sarcoidosis. Others have demonstrated the presence of a mycobacterial protein (*Mycobacterium tuberculosis* catalase-peroxidase [mKatG]) in the granulomas of some sarcoidosis patients. This protein is very resistant to degradation and may represent the persistent antigen in sarcoidosis. Immune response to this and other mycobacterial proteins has been documented by another laboratory. These studies suggest that a mycobacterium similar to *M. tuberculosis* could be responsible for sarcoidosis. The mechanism exposure/infection with such agents has been the focus of other studies. Environmental exposures to insecticides and mold have been associated with an increased risk for disease. In addition, health care workers appear to have an increased risk. Also, sarcoidosis in a donor organ has occurred after transplantation into a sarcoidosis patient. Some authors have suggested that sarcoidosis is not due to a single agent but represents a particular host response to multiple agents. Some studies have been able to correlate the environmental exposures to genetic markers. These studies have supported the hypothesis that a genetically susceptible host is a key factor in the disease.

INCIDENCE, PREVALENCE, AND GLOBAL IMPACT

Sarcoidosis is seen worldwide, with the highest prevalence reported in the Nordic population. In the United States, the disease has been reported more commonly in African Americans than whites, with the ratio of African Americans to whites ranging from 3:1 to 17:0. In the United Sates, women are more susceptible than men. The higher incidence in African Americans may have been influenced by the fact that African Americans seem to develop more extensive and chronic pulmonary disease. Because most sarcoidosis clinics are run by pulmonologists, a selection bias may have occurred. Worldwide, the prevalence of the disease varies from 20–60 per 100,000 for many groups such as Japanese, Italians, and American whites. Higher rates occur in Ireland and Nordic countries. In one closely observed community in Sweden, the lifetime risk for developing sarcoidosis was 3%.

Sarcoidosis often occurs in young, otherwise healthy adults. It is uncommon to diagnose the disease in someone aged <18 years. However, it has become clear that a second peak in incidence develops around age 60. In a study of nearly 30,000 sarcoidosis patients in the United States, the median age at diagnosis was 55.

Although most cases of sarcoidosis are sporadic, a familial form of the disease exists. At least 5% of patients with sarcoidosis will have a family member with sarcoidosis. Sarcoidosis patients who are Irish or African American seem to have a two to three times higher rate of familial disease.

PATHOPHYSIOLOGY AND IMMUNOPATHOGENESIS

The granuloma is the pathologic hallmark of sarcoidosis. A distinct feature of sarcoidosis is the local accumulation of inflammatory cells. Extensive studies in the lung using bronchoalveolar lavage (BAL) have demonstrated that the initial inflammatory response is an influx of T helper cells. In addition, there is an accumulation of activated monocytes. **Figure 360-1** is a proposed model for sarcoidosis. Using the HLA-CD4 complex, antigen-presenting cells present an unknown antigen to the helper T cell. Studies have clarified that specific HLA

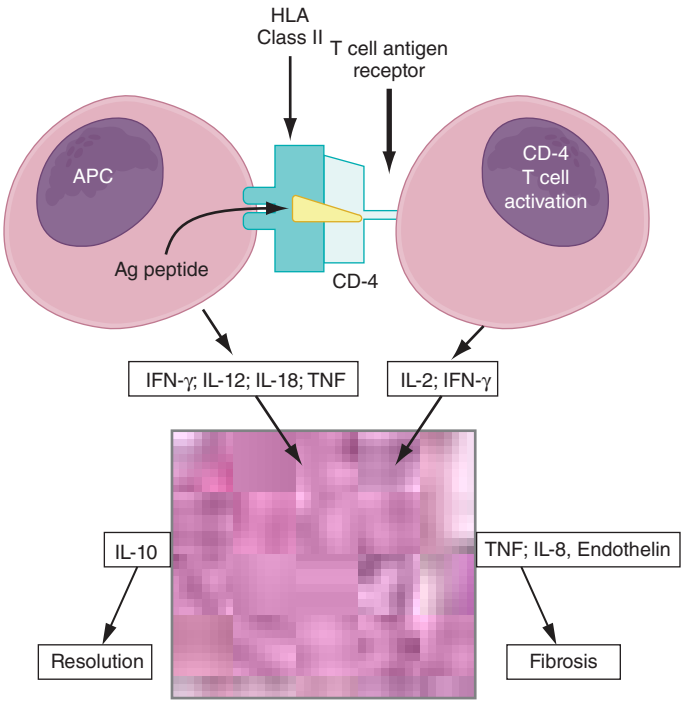


FIGURE 360-1 Schematic representation of initial events of sarcoidosis. The antigen-presenting cell and helper T cell complex leads to the release of multiple cytokines. This forms a granuloma. Over time, the granuloma may resolve or lead to chronic disease, including fibrosis. APC, antigen-presenting cell; HLA, human leukocyte antigen; IFN, interferon; IL, interleukin; TNF, tumor necrosis factor.

haplotypes such as HLA-DRB1*1101 are associated with an increased risk for developing sarcoidosis. In addition, different HLA haplotypes are associated with different clinical outcomes.

The macrophage/helper T cell cluster leads to activation with the increased release of several cytokines. These include interleukin (IL)-2 released from the T cell and interferon γ and tumor necrosis factor (TNF) released by the macrophage. The T cell is a necessary part of the initial inflammatory response. In advanced, untreated HIV infection, patients who lack helper T cells rarely develop sarcoidosis. In contrast, several reports confirm that sarcoidosis becomes unmasked as HIV-infected individuals receive antiretroviral therapy, with subsequent restoration of their immune system. In contrast, treatment of established pulmonary sarcoidosis with cyclosporine, a drug that downregulates helper T cell responses, seems to have little impact on sarcoidosis.

The granulomatous response of sarcoidosis can resolve with or without therapy. However, in at least 20% of patients with sarcoidosis, a chronic form of the disease develops. This persistent form of the disease is associated with increased levels in blood and/or BAL of IL-8, IL-17, and CXCL9. Also, studies have reported that patients with this chronic form of disease release excessive amounts of TNF in areas of inflammation. Specific gene signatures have been associated with more severe disease, such as cardiac, neurologic, and fibrotic pulmonary disease.

At diagnosis the natural history of the disease may be difficult to predict. One form of the disease, *Löfgren's syndrome*, consists of erythema nodosum and hilar adenopathy on chest roentgenogram. In some cases, periarticular arthritis may be identified without erythema nodosum. Löfgren's syndrome is associated with a good prognosis, with >90% of patients experiencing disease resolution within 2 years. Recent studies have demonstrated that the HLA-DRB1*03 was found in two-thirds of Scandinavian patients with Löfgren's syndrome. More than 95% of those patients who were HLA-DRB1*03 positive had resolution of their disease within 2 years, whereas nearly one-half of the remaining patients had disease for >2 years. It remains to be determined whether these observations can be applied to a non-Scandinavian population.

CLINICAL MANIFESTATIONS

The presentation of sarcoidosis ranges from patients who are asymptomatic to those with organ failure. It is unclear how often sarcoidosis is asymptomatic. In countries where routine chest roentgenogram screening is performed, 20–30% of pulmonary cases are detected in asymptomatic individuals. The inability to screen for other asymptomatic forms of the disease would suggest that as many as one-third of sarcoidosis patients are asymptomatic.

Respiratory complaints including cough and dyspnea are the most common presenting symptoms. In many cases, the patient presents with a 2- to 4-week history of these symptoms. Unfortunately, due to the nonspecific nature of pulmonary symptoms, the patient may see physicians for up to a year before a diagnosis is confirmed. For these patients, the diagnosis of sarcoidosis is usually only suggested when a chest roentgenogram is performed.

Symptoms related to cutaneous and ocular disease are the next two most common complaints. Skin lesions are often nonspecific. However, because these lesions are readily observed, the patient and treating physician are often led to a diagnosis. In contrast to patients with pulmonary disease, patients with cutaneous lesions are more likely to be diagnosed within 6 months of symptoms.

Nonspecific constitutional symptoms include fatigue, fever, night sweats, and weight loss. Fatigue is perhaps the most common constitutional symptom that affects these patients. Given its insidious nature, patients are usually not aware of the association with their sarcoidosis until their disease resolves.

The overall incidence of sarcoidosis at the time of diagnosis and eventual common organ involvement are summarized in Table 360-1. Over time, skin, eye, and neurologic involvement seem more apparent. In the United States, the frequency of specific organ involvement appears to be affected by age, race, and gender. For example, eye disease is more common among African Americans. Under the age of 40, it occurs more frequently in women. However, in those diagnosed over the age of 40, eye disease is more common in men.

LUNG

Lung involvement occurs in >90% of sarcoidosis patients. The most commonly used method for detecting lung disease is still the chest roentgenogram. Figure 360-2 illustrates the chest roentgenogram from a sarcoidosis patient with bilateral hilar adenopathy. Although the computed tomography (CT) scan has changed the diagnostic approach to interstitial lung disease, the CT scan is not usually considered a monitoring tool for patients with sarcoidosis. Figure 360-3 demonstrates some of the characteristic CT features, including peribronchial thickening and reticular nodular changes, which are predominantly subpleural. The peribronchial thickening seen on CT scan seems to explain the high yield of granulomas from bronchial biopsies performed for diagnosis.

Although the CT scan is more sensitive, the standard scoring system described by Scadding in 1961 for chest roentgenograms remains the preferred method of characterizing the chest involvement. Stage 1 is hilar adenopathy alone (Fig. 360-2), often with right paratracheal

TABLE 360-1 Frequency of Common Organ Involvement and Lifetime Risk[a]

	PRESENTATION, %[b]	FOLLOW-UP, %[c]
Lung	95	94
Skin	24	43
Eye	12	29
Extrathoracic lymph node	15	16
Liver	12	14
Spleen	7	8
Neurologic	5	16
Cardiac	2	3

[a]Patients could have more than one organ involved. [b]From ACCESS study of 736 patients evaluated within 6 months of diagnosis. [c]From follow-up of 1024 sarcoidosis patients seen at the University of Cincinnati Interstitial Lung Disease and Sarcoidosis Clinic from 2002 to 2006.

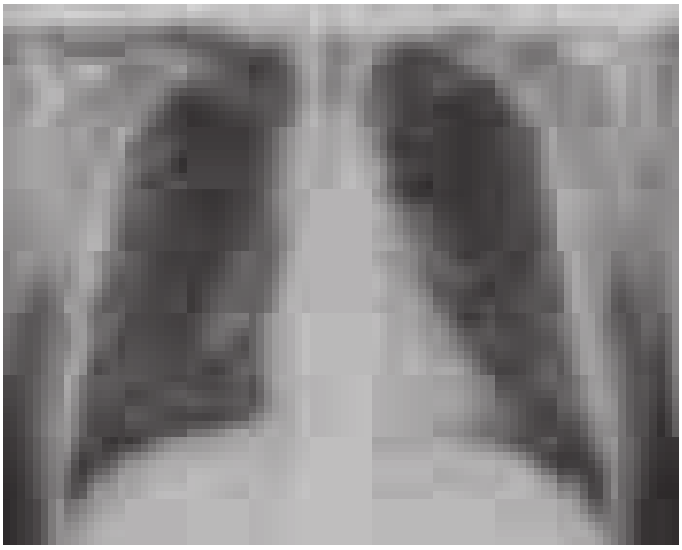


FIGURE 360-2 Posterior-anterior chest roentgenogram demonstrating bilateral hilar adenopathy, stage 1 disease.

involvement. Stage 2 is a combination of adenopathy plus infiltrates, whereas stage 3 reveals infiltrates alone. Stage 4 consists of fibrosis. Usually the infiltrates in sarcoidosis are predominantly an upper lobe process. Only in a few noninfectious diseases is an upper lobe predominance noted. In addition to sarcoidosis, the differential diagnosis of upper lobe disease includes hypersensitivity pneumonitis, silicosis, and Langerhans cell histiocytosis. For infectious diseases, tuberculosis and *Pneumocystis* pneumonia can often present as upper lobe diseases.

Lung volumes, mechanics, and diffusion are all useful in evaluating interstitial lung diseases such as sarcoidosis. The diffusion of carbon monoxide (DL_{CO}) is the most sensitive test for an interstitial lung disease. Reduced lung volumes are a reflection of the restrictive lung disease seen in sarcoidosis. However, a third of the patients presenting with sarcoidosis still have lung volumes within the normal range, despite abnormal chest roentgenograms and dyspnea.

Approximately one-half of sarcoidosis patients present with obstructive disease, reflected by a reduced ratio of forced vital capacity expired in 1 second (FEV_1/FVC). Cough is a very common symptom. Airway involvement causing varying degrees of obstruction underlies the cough in most sarcoidosis patients. Airway hyperreactivity, as determined by methacholine challenge, will be positive in some of these patients. A few patients with cough will respond to traditional bronchodilators as the only form of treatment. In some cases, high-dose inhaled glucocorticoids alone are useful. Airway obstruction can be due to large airway stenosis, which can become fibrotic and unresponsive to anti-inflammatory therapy.

Pulmonary arterial hypertension is reported in at least 5% of sarcoidosis patients. Either direct vascular involvement or the consequence of fibrotic changes in the lung can lead to pulmonary arterial hypertension. In sarcoidosis patients with end-stage fibrosis awaiting lung transplant, 70% will have pulmonary arterial hypertension. This is a much higher incidence than that reported for other fibrotic lung diseases. In less advanced, but still symptomatic, patients, pulmonary arterial hypertension has been noted in up to 50% of the cases. Because sarcoidosis-associated pulmonary arterial hypertension may respond to therapy, evaluation for this should be considered in persistently dyspneic patients.

FIGURE 360-3 High-resolution computed tomography scan of chest demonstrating patchy reticular nodularity, including areas of confluence.

SKIN

Skin involvement is eventually identified in over a third of patients with sarcoidosis. The classic cutaneous lesions include erythema nodosum, maculopapular lesions, hyper- and hypopigmentation, keloid formation, and subcutaneous nodules. A specific complex of involvement of the bridge of the nose, the area beneath the eyes, and the cheeks is referred to as *lupus pernio* **(Fig. 360-4)** and is diagnostic for a chronic form of sarcoidosis.

In contrast, erythema nodosum is a transient rash that can be seen in association with hilar adenopathy and uveitis (Löfgren's syndrome). Erythema nodosum is more common in women and in certain self-described demographic groups including whites and Puerto Ricans. In the United States, the other manifestations of skin sarcoidosis, especially lupus pernio, are more common in African Americans than whites.

The maculopapular lesions from sarcoidosis are the most common chronic form of the disease **(Fig. 360-5)**. These are often overlooked by the patient and physician, because they are chronic and not painful. Initially, these lesions are usually purplish papules and are often indurated. They can become confluent and infiltrate large areas of the skin. With treatment, the color and induration may fade. Because these lesions are caused by noncaseating granulomas, the diagnosis of sarcoidosis can be readily made by a skin biopsy.

EYE

The frequency of ocular manifestations for sarcoidosis varies depending on race. In Japan, >70% of sarcoidosis patients develop ocular

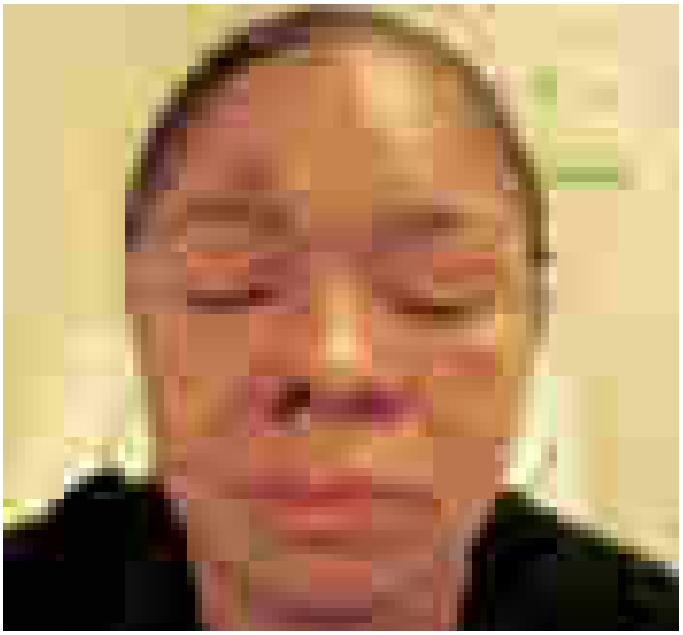


FIGURE 360-4 Chronic inflammatory lesions around nose, eyes, and cheeks, referred to as lupus pernio.

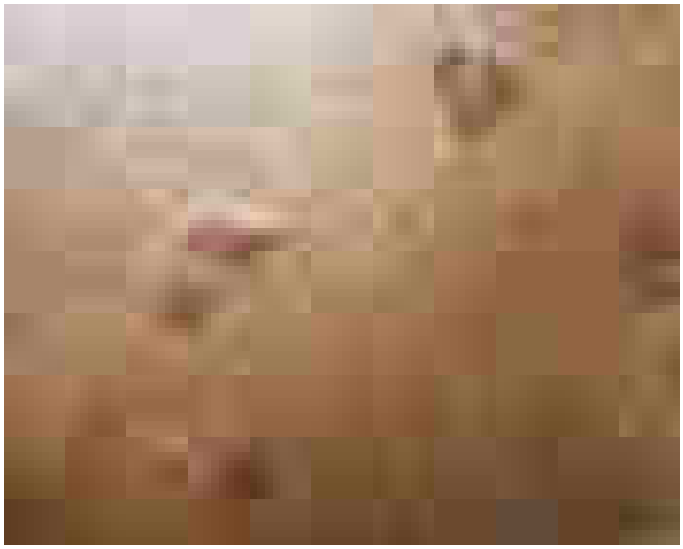


FIGURE 360-5 Maculopapular lesions on the trunk of a sarcoidosis patient.

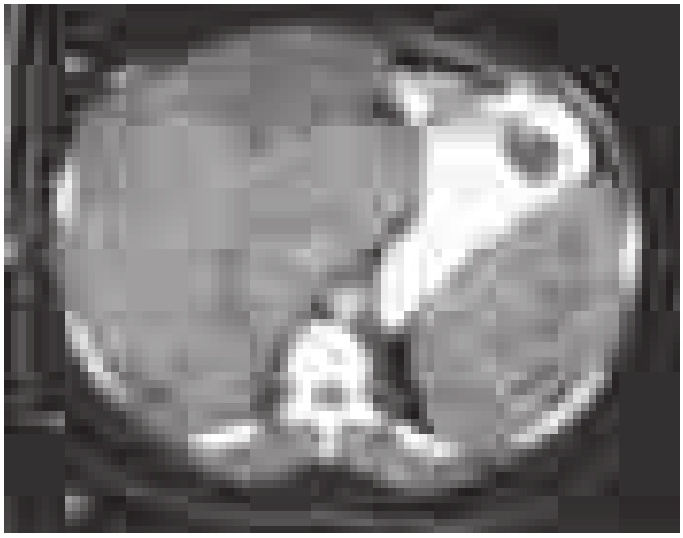


FIGURE 360-6 Computed tomography scan of the abdomen after oral and intravenous contrast. The stomach is compressed by the enlarged spleen. Within the spleen, areas of hypo- and hyperdensity are identified.

disease, whereas in the United States only 30% have eye disease, with problems more common in African Americans than whites. Although the most common manifestation is an anterior uveitis, over a quarter of patients will have inflammation at the posterior of the eye, including retinitis and pars planitis. Although symptoms such as photophobia, blurred vision, and increased tearing can occur, some asymptomatic patients still have active inflammation. Initially asymptomatic patients with ocular sarcoidosis can eventually develop blindness. Therefore, it is recommended that all patients with sarcoidosis receive a dedicated ophthalmologic examination. Sicca is seen in over one-half of the chronic sarcoidosis patients. Dry eyes appear to be a reflection of prior lacrimal gland disease. Although the patient may no longer have active inflammation, the dry eyes may require natural tears or other lubricants.

LIVER

Using biopsies to detect granulomatous disease, liver involvement can be identified in over one-half of sarcoidosis patients. However, using liver function studies, only 20–30% of patients will have evidence of liver involvement. The most common abnormality of liver function is an elevation of the alkaline phosphatase level, consistent with an obstructive pattern. In addition, elevated transaminase levels can occur. An elevated bilirubin level is a marker for more advanced liver disease. Overall, only 5% of sarcoidosis patients have sufficient symptoms from their liver disease to require specific therapy. Although symptoms can be due to hepatomegaly, more frequently symptoms result from extensive intrahepatic cholestasis leading to portal hypertension. In this case, ascites and esophageal varices can occur. It is rare that a sarcoidosis patient will require a liver transplant, because even the patient with cirrhosis due to sarcoidosis can respond to systemic therapy.

BONE MARROW AND SPLEEN

One or more bone marrow manifestations can be identified in many sarcoidosis patients. The most common hematologic problem is lymphopenia, which is a reflection of sequestration of the lymphocytes into the areas of inflammation. Anemia occurs in 20% of patients, and leukopenia is less common. Bone marrow examination will reveal granulomas in about a third of patients. Although splenomegaly can be detected in 5–10% of patients, splenic biopsy reveals granulomas in 60% of patients. The CT scan can be relatively specific for sarcoidosis involvement of the spleen **(Fig. 360-6)**. Both bone marrow and spleen involvement are more common in African Americans than whites. Although these manifestations alone are rarely an indication for therapy, on rare occasion, splenectomy may be indicated for massive symptomatic splenomegaly or profound pancytopenia. Nonthoracic lymphadenopathy can occur in up to 20% of patients.

CALCIUM METABOLISM

Hypercalcemia and/or hypercalciuria occur in about 10% of sarcoidosis patients. It is more common in whites than African Americans and in men. The mechanism of abnormal calcium metabolism is increased production of 1,25-dihydroxyvitamin D by the granuloma itself. The 1,25-dihydroxyvitamin D causes increased intestinal absorption of calcium, leading to hypercalcemia with a suppressed parathyroid hormone (PTH) level **(Chap. 403)**. Increased exogenous vitamin D from diet or sunlight exposure may exacerbate this problem. Serum calcium should be determined as part of the initial evaluation of all sarcoidosis patients, and a repeat determination may be useful during the summer months with increased sun exposure. In patients with a history of renal calculi, a 24-h urine calcium measurement should be obtained. If a sarcoidosis patient with a history of renal calculi is to be placed on calcium supplements, a follow-up 24-h urine calcium level should be measured.

RENAL DISEASE

Direct kidney involvement occurs in <5% of sarcoidosis patients. It is associated with granulomas in the kidney itself and can lead to nephritis. However, hypercalcemia is the most likely cause of sarcoidosis-associated renal disease. In 1–2% of sarcoidosis patients, acute renal failure may develop as a result of hypercalcemia. Successful treatment of hypercalcemia with glucocorticoids and other therapies often improves but usually does not totally resolve the renal dysfunction.

NERVOUS SYSTEM

Neurologic disease is reported in 5–10% of sarcoidosis patients and appears to be of equal frequency across all ethnic groups. Any part of the central or peripheral nervous system can be affected. The presence of granulomatous inflammation is often visible on magnetic resonance imaging (MRI) studies. The MRI with gadolinium enhancement may demonstrate space-occupying lesions, but the MRI can be negative due to small lesions or the effect of systemic therapy in reducing the inflammation. The cerebral spinal fluid (CSF) findings include lymphocytic meningitis with a mild increase in protein. The CSF glucose is usually normal but can be low. Certain areas of the nervous system are more commonly affected in neurosarcoidosis. These include cranial nerve involvement, basilar meningitis, myelopathy, and anterior hypothalamic disease with associated diabetes insipidus **(Chap. 374)**. Seizures and cognitive changes also occur. Of the cranial nerves, seventh nerve paralysis can be transient and mistaken for Bell's palsy (idiopathic seventh nerve paralysis). Because this form of neurosarcoidosis often resolves within weeks and may not recur, it may have occurred prior to a definitive diagnosis of sarcoidosis. Optic neuritis is another cranial nerve manifestation of sarcoidosis. This manifestation is more chronic

and usually requires long-term systemic therapy. It can be associated with both anterior and posterior uveitis. Differentiating between neurosarcoidosis and multiple sclerosis can be difficult at times. Optic neuritis can occur in both diseases. In some patients with sarcoidosis, multiple enhancing white matter abnormalities may be detected by MRI, suggesting multiple sclerosis. In such cases, the presence of meningeal enhancement or hypothalamic involvement suggests neurosarcoidosis, as does evidence of extraneurologic disease such as pulmonary or skin involvement, which also suggests sarcoidosis. Because the response of neurosarcoidosis to glucocorticoids and cytotoxic therapy is different from that of multiple sclerosis, differentiating between these disease entities is important.

CARDIAC

The presence of cardiac involvement is influenced by race. Although over a quarter of Japanese sarcoidosis patients develop cardiac disease, only 5% of sarcoidosis patients in the United States and Europe develop symptomatic cardiac disease. However, there is no apparent racial predilection between whites and African Americans. Cardiac disease, which usually presents as either congestive heart failure or cardiac arrhythmias, results from infiltration of the heart muscle by granulomas. Diffuse granulomatous involvement of the heart muscle can lead to profound dysfunction with left ventricular ejection fractions <10%. Even in this situation, improvement in the ejection fraction can occur with systemic therapy. Arrhythmias can also occur with diffuse infiltration or with more patchy cardiac involvement. If the atrioventricular (AV) node is infiltrated, heart block can occur, which can be detected by routine electrocardiography. Ventricular arrhythmias and sudden death due to ventricular tachycardia are common causes of death. Arrhythmias are best detected using 24-h ambulatory monitoring, and electrophysiology studies may be negative. Other screening tests for cardiac disease include routine electrocardiography and echocardiography. The confirmation of cardiac sarcoidosis is usually performed with either MRI or positron emission tomography (PET) scanning. Because ventricular arrhythmias are usually multifocal due to patchy multiple granulomas in the heart, ablation therapy is not useful. Patients with significant ventricular arrhythmias should be considered for an implanted defibrillator, which appears to have reduced the rate of death in cardiac sarcoidosis. Although systemic therapy can be useful in treating the arrhythmias, patients may still have malignant arrhythmias up to 6 months after starting successful treatment, and the risk for recurrent arrhythmias occurs whenever medications are tapered.

MUSCULOSKELETAL SYSTEM

Direct granulomatous involvement of bone and muscle can be documented by radiography (x-ray, MRI, PET scan [Fig. 360-7], or gallium scan) or confirmed by biopsy in about 10% of sarcoidosis patients. However, a larger percentage of sarcoidosis patients complain of myalgias and arthralgias. These complaints are similar to those reported by patients with other inflammatory diseases, including chronic infections such as mononucleosis. Fatigue associated with sarcoidosis may be overwhelming for many patients. Recent studies have demonstrated a link between fatigue and small peripheral nerve fiber disease in sarcoidosis.

OTHER ORGAN INVOLVEMENT

Although sarcoidosis can affect any organ of the body, rarely does it involve the breast, testes, ovary, or stomach. Because of the rarity of involvement, a mass in one of these areas requires a biopsy to rule out other diseases including cancer. For example, in a study of breast problems in female sarcoidosis patients, a breast lesion was more likely to be a granuloma from sarcoidosis than from breast cancer. However, findings on the physical examination or mammogram cannot reliably differentiate between these lesions. More importantly, as women with sarcoidosis age, breast cancer becomes more common. Therefore, it is recommended that routine screening including mammography be performed along with other imaging studies (ultrasound, MRI) or biopsy as clinically indicated.

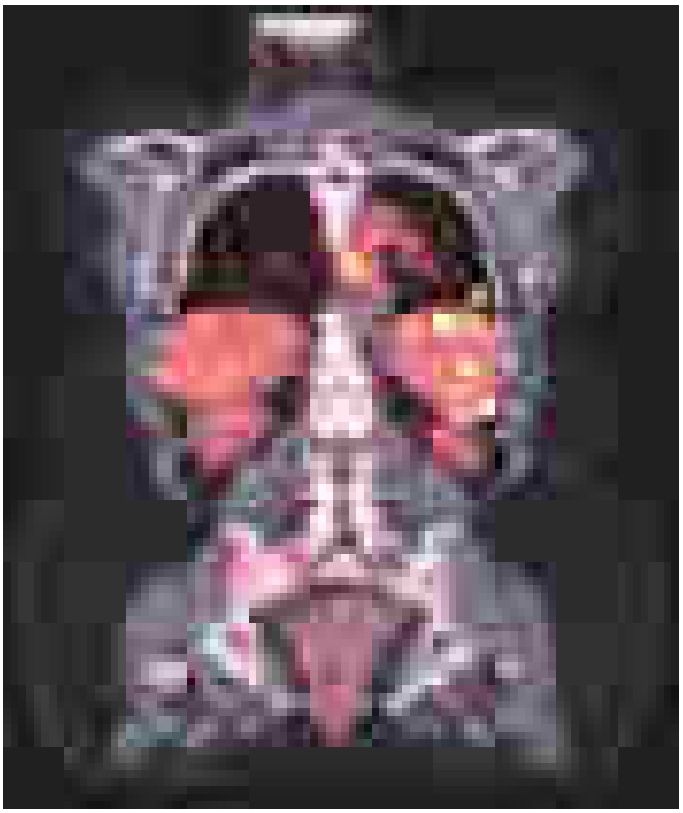


FIGURE 360-7 Positron emission tomography and computed tomography scan merged demonstrating increased activity in spleen, ribs, and spine of patient with sarcoidosis.

COMPLICATIONS

Sarcoidosis is usually a self-limited, non-life-threatening disease. However, organ-threatening disease can occur. These complications can include blindness, paraplegia, or renal failure. Death from sarcoidosis occurs in about 5% of patients seen in sarcoidosis referral clinics. The usual causes of death related to sarcoidosis are from lung, cardiac, neurologic, or liver involvement. In respiratory failure, an elevation of the right atrial pressure is a poor prognostic finding. Lung complications can also include infections such as mycetoma, which can subsequently lead to massive bleeding. In addition, the use of immunosuppressive agents can increase the incidence of serious infections.

LABORATORY FINDINGS

The chest roentgenogram remains the most commonly used tool to assess lung involvement in sarcoidosis. As noted above, the chest roentgenogram classifies involvement into four stages, with stages 1 and 2 having hilar and paratracheal adenopathy. The CT scan has been used increasingly in evaluating interstitial lung disease. In sarcoidosis, the presence of adenopathy and a nodular infiltrate is not specific for sarcoidosis. Adenopathy up to 2 cm can be seen in other inflammatory lung diseases such as idiopathic pulmonary fibrosis. However, adenopathy >2 cm in the short axis supports the diagnosis of sarcoidosis over other interstitial lung diseases.

The PET scan has increasingly replaced gallium-67 scanning to identify areas of granulomatous disease in the chest and other parts of the body (Fig. 360-7). Both tests can be used to identify potential areas for biopsy. Cardiac PET scanning has also proved useful in assessing cardiac sarcoidosis. The identification of hypermetabolic activity may be due to the granulomas from sarcoidosis and not to disseminated malignancy.

MRI has also proved useful in the assessment of extrapulmonary sarcoidosis. Gadolinium enhancement has been demonstrated in areas of inflammation in the brain, heart, and bone. MRI scans may detect asymptomatic lesions. Like PET scan, MRI changes appear similar to those seen with malignancy and infection. In some cases, biopsy may be necessary to determine the cause of the radiologic abnormality.

Serum levels of angiotensin-converting enzyme (ACE) can be helpful in the diagnosis of sarcoidosis. However, the test has somewhat low sensitivity and specificity. Elevated levels of ACE are reported in 60% of patients with acute disease and only 20% of patients with chronic disease. Although there are several causes for mild elevation of ACE, including diabetes, elevations of >50% of the upper limit of normal are seen in only a few conditions including sarcoidosis, leprosy, Gaucher's disease, hyperthyroidism, and disseminated granulomatous infections such as miliary tuberculosis. Because the ACE level is determined by a biologic assay, the concurrent use of an ACE inhibitor such as lisinopril will lead to a very low ACE level.

DIAGNOSIS

The diagnosis of sarcoidosis requires both compatible clinical features and pathologic findings. Because the cause of sarcoidosis remains elusive, the diagnosis cannot be made with 100% certainty. Nevertheless, the diagnosis can be made with reasonable certainty based on history and physical features along with laboratory and pathologic findings.

Patients are usually evaluated for possible sarcoidosis based on two scenarios **(Fig. 360-8)**. In the first scenario, a patient may undergo a biopsy revealing a noncaseating granuloma in either a pulmonary or an extrapulmonary organ. If the clinical presentation is consistent with sarcoidosis and there is no alternative cause for the granulomas identified, then the patient is felt to have sarcoidosis.

In the second scenario, signs or symptoms suggesting sarcoidosis such as the presence of bilateral adenopathy may be present in an otherwise asymptomatic patient or a patient with uveitis or a rash consistent with sarcoidosis. At this point, a diagnostic procedure should be performed. For the patient with a compatible skin lesion, a skin biopsy should be considered. Other biopsies to consider could include liver, extrathoracic lymph node, or muscle. In some cases, a biopsy of the affected organ may not be easy to perform (such as a brain or spinal cord lesion). In other cases, such as an endomyocardial biopsy, the likelihood of a positive biopsy is low. Because of the high rate of pulmonary involvement in these cases, the lung may be easier to approach by bronchoscopy. During the bronchoscopy, a transbronchial biopsy, bronchial biopsy, or transbronchial needle aspirate can be performed. The endobronchial ultrasonography-guided (EBUS) transbronchial needle aspirate can assist in diagnosing sarcoidosis in patients with mediastinal adenopathy (stage 1 or 2 radiographic pulmonary disease), whereas transbronchial biopsy has a higher diagnostic yield for those with only parenchymal lung disease (stage 3). These tests are complementary and may be performed together.

If the biopsy reveals granulomas, an alternative diagnosis such as infection or malignancy must be excluded. Bronchoscopic washings can be sent for cultures for fungi and tuberculosis. For the pathologist, the more tissue that is provided, the more comfortable is the diagnosis of sarcoidosis. A needle aspirate may be adequate in an otherwise classic case of sarcoidosis, but may be insufficient in a patient in whom lymphoma or fungal infection is a likely alternative diagnosis. Because granulomas can be seen on the edge of a lymphoma, the presence of a few granulomas from a needle aspirate may not be sufficient to clarify the diagnosis. Mediastinoscopy provides a larger sample to confirm the presence or absence of lymphoma in the mediastinum. Alternatively, for most patients, evidence of extrathoracic disease (e.g., eye involvement) may further support the diagnosis of sarcoidosis.

For patients with negative pathology, positive supportive tests may increase the likelihood of the diagnosis of sarcoidosis. These tests include an elevated ACE level, which can also be elevated in other granulomatous diseases but not in malignancy. A positive PET scan can support the diagnosis if multiple organs are affected. A BAL is often performed during the bronchoscopy. An increase in the percentage of lymphocytes supports the diagnosis of sarcoidosis. The use of the lymphocyte markers CD4 and CD8 can be used to determine the CD4/CD8 ratio of these increased lymphocytes in the BAL fluid. A ratio of >3.5 is strongly supportive of sarcoidosis but is less sensitive than an increase in lymphocytes alone. Although in general, an increase in BAL lymphocytes is supportive of the diagnosis, other conditions must be considered.

Supportive findings, when combined with commonly associated but nondiagnostic clinical features of the disease, improve the diagnostic probability of sarcoidosis. These clinical features include uveitis, renal stones, hypercalcemia, seventh cranial nerve paralysis, or erythema nodosum. The presence of one or more of these features in a patient suspected of having sarcoidosis increases the probability of sarcoidosis.

The *Kviem-Siltzbach procedure* is a specific diagnostic test for sarcoidosis. An intradermal injection of specially prepared tissue derived from the spleen of a known sarcoidosis patient is biopsied 4–6 weeks after injection. If noncaseating granulomas are seen, this is highly specific for the diagnosis of sarcoidosis. Unfortunately, there is no commercially available Kviem-Siltzbach reagent, and some locally prepared batches have lower specificity. Thus, this test is of historic interest and is rarely used in current clinical practice.

Because the diagnosis of sarcoidosis can never be certain, over time other features may arise that lead to an alternative diagnosis. Conversely, evidence for new organ involvement may eventually confirm the diagnosis of sarcoidosis.

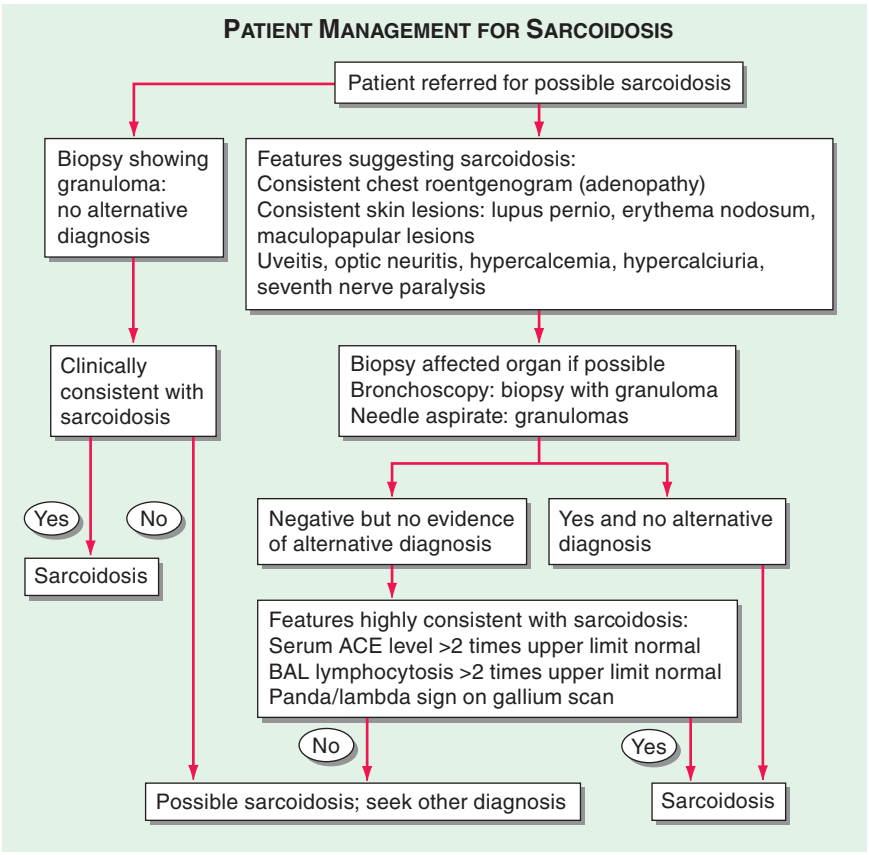


FIGURE 360-8 Proposed approach to management of patient with possible sarcoidosis. Presence of one or more of these features supports the diagnosis of sarcoidosis: uveitis, optic neuritis, hypercalcemia, hypercalciuria, seventh cranial nerve paralysis, diabetes insipidus. ACE, angiotensin-converting enzyme; BAL, bronchoalveolar lavage.

PROGNOSIS

The risk of death or loss of organ function remains low in sarcoidosis. Poor outcomes usually occur in patients who present with advanced disease in whom treatment seems to have little impact. In these cases, irreversible fibrotic changes have frequently occurred. Over the past 20 years, the

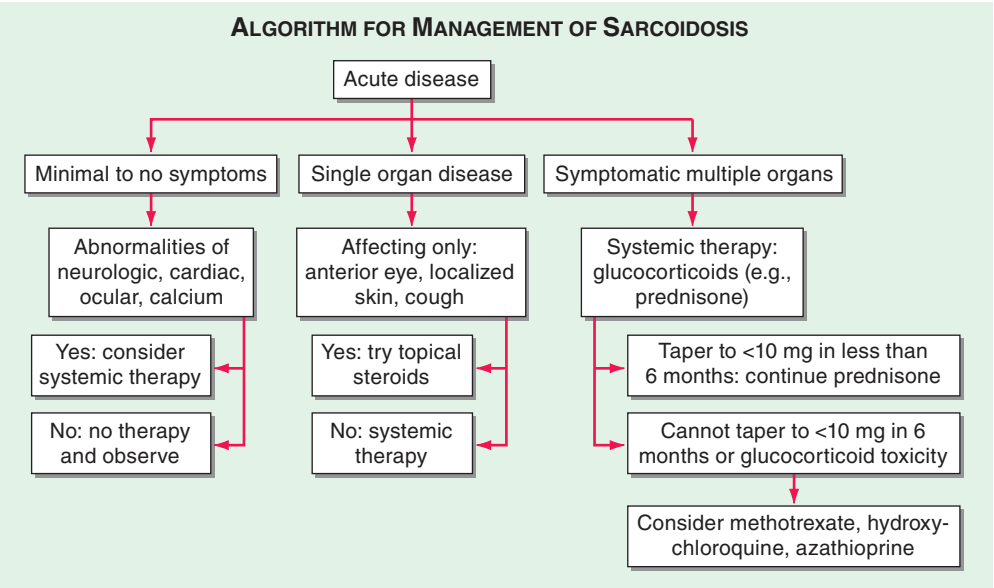


FIGURE 360-9 The management of acute sarcoidosis is based on level of symptoms and extent of organ involvement. In patients with mild symptoms, no therapy may be needed unless specified manifestations are noted.

reported mortality from sarcoidosis has increased in the United States and England. Whether this is due to heightened awareness of the chronic nature of this disease or to other factors such as more widespread immunosuppressive therapy usage remains unclear.

For the majority of patients, initial presentation occurs during the granulomatous phase of the disease as depicted in Fig. 360-1. It is clear that many patients resolve their disease within 2–5 years. These patients are felt to have acute, self-limiting sarcoidosis. However, there is a form of the disease that does not resolve within the first 2–5 years. These chronic patients can be identified at presentation by certain risk factors at presentation such as fibrosis on chest roentgenogram, presence of lupus pernio, bone cysts, cardiac or neurologic disease (except isolated seventh nerve paralysis), and presence of renal calculi due to hypercalciuria. In several studies, patients who required glucocorticoids for any manifestation of their disease in the first 6 months of presentation had a >50% chance of having chronic disease. In contrast, <10% of patients who require no systemic therapy in the first 6 months required chronic therapy.

TREATMENT

Sarcoidosis

Indications for therapy should be based on symptoms or presence of organ- or life-threatening disease, including disease involving the eye, heart, or nervous system. The patient with asymptomatic elevated liver function tests or an abnormal chest roentgenogram probably does not benefit from treatment. However, these patients should be monitored for evidence of progressive, symptomatic disease.

One approach to therapy is summarized in Figs. 360-9 and 360-10. We have divided the approach into treating acute versus chronic disease. For acute disease, no therapy remains a viable option for patients with no or mild symptoms. For symptoms confined to only one organ, topical therapy is preferable. For multi-organ disease or disease too extensive for topical therapy, an approach to systemic therapy is outlined. Glucocorticoids remain the drugs of choice for this disease. However, the decision to continue to treat with glucocorticoids or

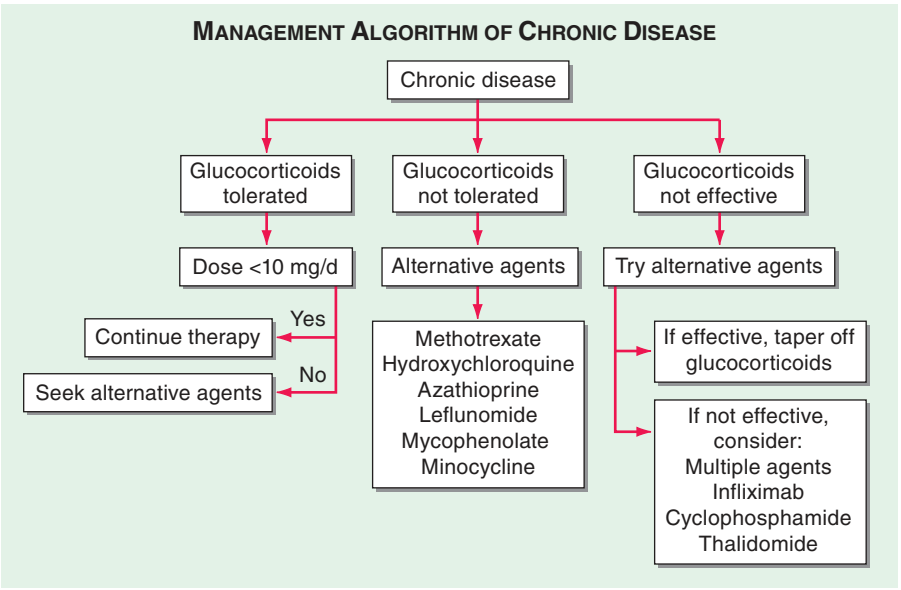


FIGURE 360-10 Approach to chronic disease is based on whether glucocorticoid therapy is tolerated or not.

TABLE 360-2 Commonly Used Drugs to Treat Sarcoidosis

DRUG	INITIAL DOSE	MAINTENANCE DOSE	MONITORING	TOXICITY	SUPPORT THERAPY[a]	SUPPORT MONITORING[a]
Prednisone	20–40 mg qd	Taper to 5–10 mg	Glucose, blood pressure, bone density	Diabetes, osteoporosis	A: Acute pulmonary D: Extrapulmonary	
Hydroxychloroquine	200–400 mg qd	400 mg qd	Eye examination q6–12 mo	Ocular	B: Some forms of disease	D: Routine eye examination
Methotrexate	10 mg qwk	2.5–15 mg qwk	CBC, renal, hepatic q2mo	Hematologic, nausea, hepatic, pulmonary	B: Steroid sparing C: Some forms chronic disease	D: Routine hematologic, renal, and hepatic monitoring
Azathioprine	50–150 mg qd	50–200 mg qd	CBC, renal q2mo	Hematologic, nausea	C: Some forms chronic disease	D: Routine hematologic monitoring
Infliximab	3–5 mg/kg q2wk for 2 doses	3–10 mg/kg q4–8 wk	Initial PPD	Infections, allergic reaction, carcinogen	A: Chronic pulmonary disease	B: Caution in patients with latent tuberculosis or advanced congestive heart failure

[a]Grade A: supported by at least two double-blind randomized control trials; grade B: supported by prospective cohort studies; grade C: supported primarily by two or more retrospective studies; grade D: only one retrospective study or based on experience in other diseases.

Abbreviations: CBC, complete blood count; PPD, purified protein derivative test for tuberculosis.

Source: Adapted from RP Baughman, O Selroos: Evidence-based approach to treatment of sarcoidosis, in PG Gibson et al (eds): *Evidence-Based Respiratory Medicine*. Oxford, BMJ Books Blackwell, 2005, pp 491–508.

to add steroid-sparing agents depends on the tolerability, duration, and dosage of glucocorticoids. Table 360-2 summarizes the dosage and monitoring of several commonly used drugs. According to the available trials, evidence-based recommendations are made. Most of these recommendations are for pulmonary disease because most of the trials were performed only in pulmonary disease. Treatment recommendations for extrapulmonary disease are usually similar with a few modifications. For example, the dosage of glucocorticoids is usually higher for neurosarcoidosis and lower for cutaneous disease. There was some suggestion that higher doses would be beneficial for cardiac sarcoidosis, but one study found that initial prednisone doses >40 mg/d were associated with a worse outcome because of toxicity.

Systemic therapies for sarcoidosis are usually immunosuppressive including glucocorticoids, cytotoxics, or biologics. Although most patients receive glucocorticoids as their initial systemic therapy, toxicity associated with prolonged therapy often leads to steroid-sparing alternatives. The antimalarial drugs such as hydroxychloroquine are more effective for skin than pulmonary disease. Minocycline may also be useful for cutaneous sarcoidosis. For pulmonary and other extrapulmonary disease, cytotoxic agents that include methotrexate, azathioprine, leflunomide, mycophenolate, and cyclophosphamide are often used. The most widely studied cytotoxic agent has been methotrexate. This agent works in approximately two-thirds of sarcoidosis patients, regardless of the disease manifestation. In one retrospective study comparing methotrexate to azathioprine, both drugs were equally effective. However, methotrexate was associated with significantly less toxicity. As noted in Table 360-2, specific guidelines for monitoring therapy have been recommended. Cytokine modulators such as thalidomide and pentoxifylline have also been used in a limited number of cases.

The biologic anti-TNF agents have recently been studied in sarcoidosis, with prospective randomized trials completed for etanercept, golibmumab, and infliximab. Etanercept has a limited role as a steroid-sparing agent. Golibmumab was not significantly different than placebo in treating chronic pulmonary disease. However, this may have been due to the relatively low dose of golibmumab studied. Infliximab significantly improved lung function when administered to glucocorticoid and cytotoxic pretreated patients with chronic disease. The difference in response between etanercept and infliximab is similar to that observed in Crohn's disease, where infliximab is effective and etanercept is not. However, there is a higher risk for reactivation of tuberculosis with infliximab compared to etanercept. The differential response rate could be explained by differences in mechanism of action because etanercept is a TNF receptor antagonist and infliximab is a monoclonal antibody against TNF. In contrast to etanercept, infliximab also binds to TNF on the surface of some cells that release TNF, which leads to cell lysis. This effect has been documented in Crohn's disease. Adalimumab is a humanized monoclonal anti-TNF antibody that also appears effective for sarcoidosis when dosed at higher strengths, as recommended for the treatment of Crohn's disease. The role of the newer therapeutic agents for sarcoidosis is still evolving. However, these targeted therapies confirm that TNF may be an important target, especially in the treatment of chronic disease. However, these agents are not a panacea, because sarcoidosis-like disease has occurred in patients treated with anti-TNF agents for nonsarcoidosis indications.

FURTHER READING

Baughman RP, Selroos O: Evidence-based approach to treatment of sarcoidosis, in *Evidence-Based Respiratory Medicine*, PG Gibson et al (eds). Oxford, BMJ Books Blackwell, 2005, pp 491–508.

Baughman RP et al: Established and experimental medical therapy of pulmonary sarcoidosis. Eur Respir J 41:1424, 2013.

Iannuzzi MC et al: Sarcoidosis. N Engl J Med 357:2153, 2007.

Judson MA et al: Safety and efficacy of ustekinumab or golimumab in patients with chronic sarcoidosis. Eur Respir J 44:1296, 2014.

Newman LS et al: A case control etiologic study of sarcoidosis: Environmental and occupational risk factors. Am J Respir Crit Care Med 170:1324, 2004.

Valeyre D et al: Sarcoidosis. Lancet 383:1155, 2013.

361 IgG4-Related Disease

John H. Stone

IgG4-related disease (IgG4-RD) is a fibroinflammatory condition characterized by a tendency to form tumefactive lesions. The clinical manifestations of this disease, however, are protean, as IgG4-RD can affect virtually any organ system. Commonly affected organs are the biliary tree, major salivary glands, periorbital tissues, kidneys, lungs, lymph nodes, and retroperitoneum. In addition, IgG4-RD involvement of the meninges, aorta, prostate, thyroid, pericardium, skin, and other organs is well described. The disease is believed to affect the brain parenchyma, the joints, the bone marrow, and the bowel mucosa only rarely.

The clinical features of IgG4-RD are numerous, but the pathologic findings are consistent across all affected organs. These findings include a lymphoplasmacytic infiltrate with a high percentage of

IgG4-positive plasma cells; a characteristic pattern of fibrosis termed "storiform" (from the Latin *storea*, for "woven mat"); a tendency to target blood vessels, particularly veins, through an obliterative process ("obliterative phlebitis"); and a mild to moderate tissue eosinophilia.

IgG4-RD encompasses a number of conditions previously regarded as separate, organ-specific entities. A condition once known as "lymphoplasmacytic sclerosing pancreatitis" became the paradigm of IgG4-RD in 2000, when Japanese investigators recognized that these patients had elevated serum concentrations of IgG4. This form of sclerosing pancreatitis is now termed type 1 (IgG4-related) autoimmune pancreatitis (AIP). By 2003, extrapancreatic disease manifestations had been identified in patients with type 1 AIP, and descriptions of IgG4-RD other organs followed. *Mikulicz's disease*, once considered to be a subset of Sjögren's syndrome that affected the lacrimal, parotid, and submandibular glands, is now known to be one of the most common presentations of IgG4-RD. Similarly, the steroid-responsive subset of primary sclerosing cholangitis is explained by the fact that such patients actually have a separate disease, that is, IgG4-related sclerosing cholangitis. In this manner, the understanding of IgG4-RD has extended to include nearly every specialty of medicine.

CLINICAL FEATURES

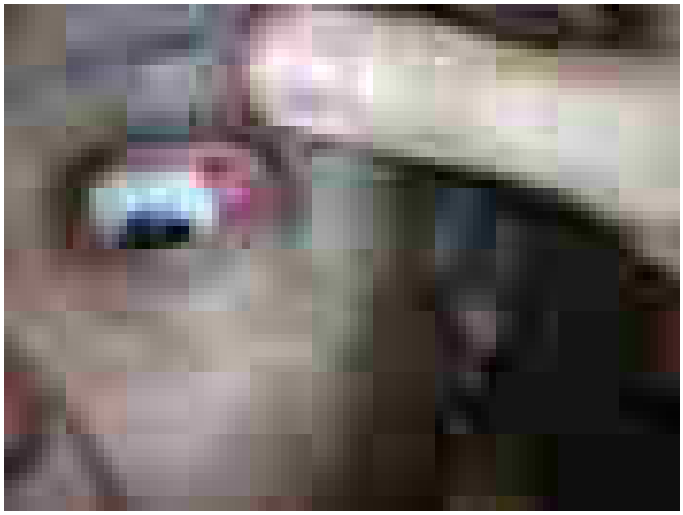
The major organ lesions are summarized in Table 361-1. IgG4-RD usually presents subacutely, and even in the setting of multi-organ disease most patients do not have fevers or dramatic elevations of C-reactive protein levels. Some patients, however, experience substantial weight loss over periods of months. Clinically apparent disease can evolve over months, years, or even decades before the manifestations within a given organ becomes sufficiently severe to bring the patient to medical attention. Some patients have disease that is marked by the appearance and then resolution or temporary improvement in symptoms within a particular organ. Other patients accumulate new organ involvement as their disease persists in previously affected organs. Many patients with IgG4-RD are misdiagnosed as having other conditions, particularly malignancies, or their findings are attributed initially to nonspecific inflammation. The disorder is often identified incidentally through radiologic findings or unexpectedly in pathology specimens.

Multiorgan disease may be evident at diagnosis but can also evolve over months to years. Some patients have disease confined to a single organ for many years. Others have either known or subclinical organ involvement at the same time as the major clinical feature. Patients with type 1 AIP may have their major disease focus in the pancreas; however, thorough evaluations by history, physical examination, blood tests, and cross-sectional imaging may demonstrate lacrimal gland enlargement, sialoadenitis, lymphadenopathy, a variety of pulmonary findings, tubulointerstitial nephritis, hepatobiliary disease, aortitis, retroperitoneal fibrosis, or other organ involvement. Spontaneous improvement, sometimes leading to clinical resolution of certain organ system manifestations, is reported in a small percentage of patients.

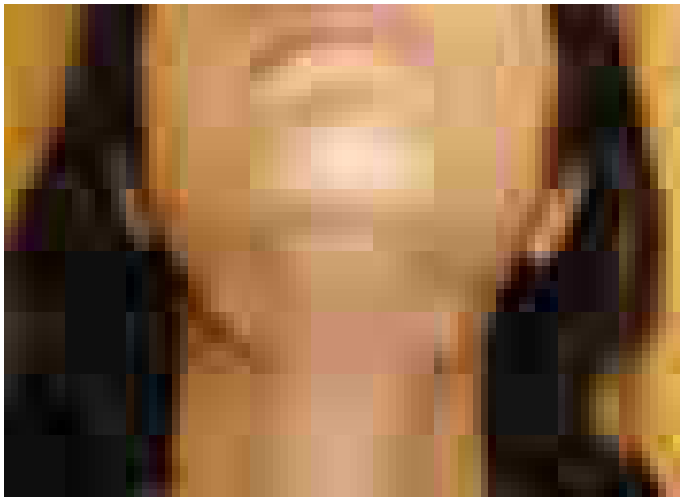
Two common characteristics of IgG4-RD are allergic disease and the tendency to form tumefactive lesions that mimic malignancies (Fig. 361-1). Many IgG4-RD patients have allergic features such as atopy, eczema, asthma, nasal polyps, sinusitis, and modest peripheral eosinophilia. IgG4-RD also appears to account for a significant proportion of tumorous swellings—pseudotumors—in many organ systems. Some patients undergo major surgeries (e.g., Whipple procedures or thyroidectomy) for the purpose of resecting malignancies before the correct diagnosis is identified. Frequent sites of pseudotumors are the major salivary glands, lacrimal glands, lungs, and kidneys; however, nearly all organs have been affected with this manifestation.

TABLE 361-1 Organ Manifestations of IgG4-Related Disease

ORGAN	MAJOR CLINICAL FEATURES
Orbits and periorbital tissues	Painless eyelid or periocular tissue swelling; orbital pseudotumor; dacryoadenitis; dacryocystitis; orbital myositis; and mass lesions extending into the pterygopalatine fossa and infiltrating along the trigeminal nerve
Ears, nose, and sinuses	Allergic phenomena (nasal polyps, asthma, allergic rhinitis, peripheral eosinophilia); nasal obstruction, rhinorrhea, anosmia, chronic sinusitis; occasional bone-destructive lesions
Salivary glands	Submandibular and/or parotid gland enlargement (isolated bilateral submandibular gland involvement more common); minor salivary glands sometimes involved
Meninges	Headache, radiculopathy, cranial nerve palsies, or other symptoms resulting from spinal cord compression; tendency to form mass lesions; magnetic resonance imaging shows marked thickening and enhancement of dura
Hypothalamus and pituitary	Clinical syndromes resulting from involvement of the hypothalamus and pituitary, e.g., anterior pituitary hormone deficiency, central diabetes insipidus, or both; imaging reveals thickened pituitary stalk or mass formation on the stalk, swelling of the pituitary gland, or mass formation within the pituitary
Lymph nodes	Generalized lymphadenopathy or localized disease adjacent to a specific affected organ; the lymph nodes involved are generally 1–2 cm in diameter and nontender
Thyroid gland	Riedel's thyroiditis; fibrosing variant of Hashimoto's thyroiditis
Lungs	Asymptomatic finding on lung imaging; cough, hemoptysis, dyspnea, pleural effusion, or chest discomfort; associated with parenchymal lung involvement, pleural disease, or both; four main clinical lung syndromes: inflammatory pseudotumor, paravertebral mass often extending over several vertebrae, central airway disease, localized or diffuse interstitial pneumonia; pleural lesions have severe, nodular thickening of the visceral or parietal pleura with diffuse sclerosing inflammation, sometimes associated with pleural effusion
Aorta	Asymptomatic finding on radiologic studies; surprise finding at elective aortic surgery; aortic dissection; clinicopathologic syndromes described include lymphoplasmacytic aortitis of thoracic or abdominal aorta, aortic dissection, periaortitis and periarteritis, and inflammatory abdominal aneurysm
Retroperitoneum	Backache, lower abdominal pain, lower extremity edema, hydronephrosis from ureteral involvement, asymptomatic finding on radiologic studies. Classic radiologic appearance is peri-aortic inflammation extending caudally to involve the iliac vessels.
Kidneys	Tubulointerstitial nephritis; membranous glomerulonephritis in a small minority; asymptomatic tumoral lesions, typically multiple and bilateral, are sometimes detected on radiologic studies; renal involvement strongly associated with hypocomplementemia
Pancreas	Type 1 autoimmune pancreatitis, presenting as mild abdominal pain; weight loss; acute, obstructive jaundice, mimicking adenocarcinoma of the pancreas (including a pancreatic mass); between 20 and 50% of patients present with acute glucose intolerance; imaging shows diffuse (termed "sausage-shaped pancreas") or segmental pancreatic enlargement, with loss of normal lobularity; a mass often raises the suspicion of malignancy
Biliary tree and liver	Obstructive jaundice associated with autoimmunity in most cases; weight loss; steatorrhea; abdominal pain; and new-onset diabetes mellitus; mimicker of primary sclerosing cholangitis and cholangiocarcinoma
Other organs involved	Gallbladder, liver (mass), breast (pseudotumor), prostate (prostatism), pericardium (constrictive pericarditis), mesentery (sclerosing mesenteritis), mediastinum (fibrosing mediastinitis), skin (erythematous or flesh-colored papules), peripheral nerve (perineural inflammation)



A



B

FIGURE 361-1 A major clinical feature of IgG4-related disease is its tendency to form tumefactive lesions. Shown here are mass lesions of the lacrimal glands (***A***) and the submandibular glands (***B***).

IgG4-RD often causes major morbidity and can lead to organ failure; however, its general pattern is to cause damage in a subacute manner. Destructive bone lesions in the sinuses, head, and middle ear spaces that mimic granulomatous polyangiitis (formerly Wegener's granulomatosis) occur rarely in IgG4-RD, but less aggressive lesions are the rule in most organs. In regions such as the retroperitoneum, substantial fibrosis often occurs before the diagnosis is established, leading to ureteral entrapment, hydronephrosis, postobstructive uropathy, and renal atrophy. The chronic pain often associated with IgG4-related retroperitoneal fibrosis probably, results from the encasement of peripheral nerves by the inflammatory process. Undiagnosed or undertreated IgG4-related cholangitis can lead to hepatic failure within months. Similarly, IgG4-related aortitis, believed to be the cause of a substantial minority of inflammatory aortitis cases, can cause aneurysms and dissections. Substantial renal dysfunction and even renal failure can ensue from IgG4-related tubulointerstitial nephritis, and renal atrophy is a frequent sequel to this disease complication. IgG4-related membranous glomerulonephropathy, a less common renal manifestation than tubulointerstitial nephritis, must be distinguished from idiopathic membranous glomerulonephropathy.

SEROLOGIC FINDINGS

The majority of patients with IgG4-RD have elevated serum IgG4 concentrations; however, the range of elevation varies widely. Serum concentrations of IgG4 as high as 30 or 40 times the upper limit of normal sometimes occur, usually in patients with disease that affects multiple organ systems simultaneously. Approximately 30% of patients have normal serum IgG4 concentrations despite classic histopathologic and immunohistochemical findings. Such patients tend to have disease that affects fewer organs. Patients with IgG4-related retroperitoneal fibrosis have a high likelihood of normal serum IgG4 concentrations, perhaps because the process has advanced to a fibrotic stage by the time the diagnosis is considered.

The correlation between serum IgG4 concentrations and disease activity and the need for treatment is imperfect. Serum IgG4 concentrations typically decline swiftly with the institution of therapy but often do not normalize completely. Patients can achieve clinical remissions yet have persistently elevated serum IgG4 concentrations. Rapidly rising serum IgG4 concentrations may identify patients at greatest risk for clinical flares. Monitoring of serial IgG4 concentrations identifies early relapse in some patients; however, the temporal relationship between modest IgG4 elevations and the need for clinical treatment is poor. Clinical relapses occur in some patients despite persistently normal IgG4 concentrations.

IgG4 concentrations in serum are usually measured by nephelometry assays. These assays can lead to reports of spuriously low IgG4 values because of the prozone effect. This effect can be corrected by dilution of the serum sample in the laboratory. The prozone effect should be considered when the results of serologic testing for IgG4 concentrations are at odds with clinical features that strongly suggest IgG4-RD.

EPIDEMIOLOGY

The typical patient with IgG4-RD is a middle-aged to elderly man. This epidemiology stands in stark contrast to that of many classic autoimmune conditions, which tend to affect young women. Studies of AIP patients in Japan indicate that the male-to-female ratio in that disease subset is on the order of 3:1. Even more striking, male predominance has been reported in IgG4-related tubulointerstitial nephritis and IgG4-related retroperitoneal fibrosis. Among IgG4-RD manifestations that involve organs of the head and neck, however, the sex ratio may be closer to 1:1.

PATHOLOGY

The key histopathology characteristics of IgG4-RD are a dense lymphoplasmacytic infiltrate (Fig. 361-2) that is organized in a storiform pattern, obliterative phlebitis, and a mild to moderate eosinophilic infiltrate.

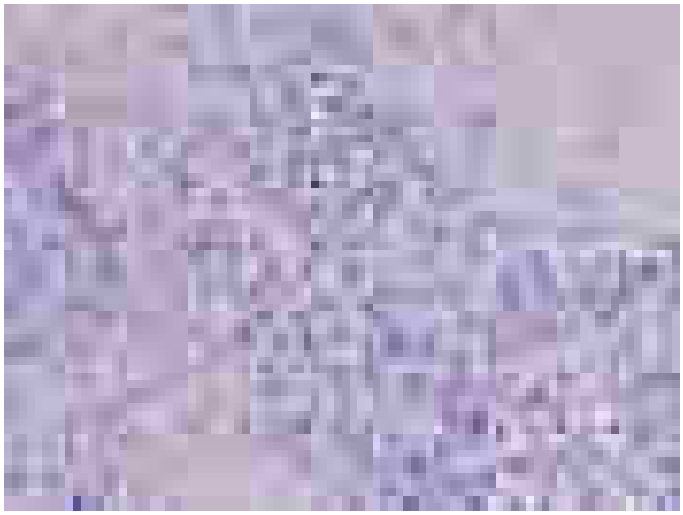


FIGURE 361-2 Hallmark histopathology characteristics of IgG4-related disease (IgG4-RD) are a dense lymphoplasmacytic infiltrate and a mild to moderate eosinophilic infiltrate. The cellular inflammation is often encased in a distinctive type of fibrosis termed "storiform," which often has a basket weave pattern. Abundant fibroblasts and strands of fibrosis accompany the lymphoplasmacytic infiltrate and eosinophils in this figure. This biopsy was taken from a nodular lesion on the cheek; however, the findings are identical to the pathology found in the pancreas, kidneys, lungs, salivary glands, and other organs affected by IgG4-RD.

Lymphoid follicles and germinal centers are frequently observed. The infiltrate tends to aggregate around ductal structures when it affects glands such as the lacrimal, submandibular, and parotid glands or the pancreas. The inflammatory lesion often aggregates into tumefactive masses that destroy the involved tissue.

Obliterative arteritis is observed in some organs, particularly the lung; however, venous involvement is more common (and is indeed a hallmark of IgG4-RD). Several histopathology features are uncommon in IgG4-RD and, when detected, mitigate against the diagnosis of IgG4-RD. These include intense neutrophilic infiltration, leukocytoclasis, granulomatous inflammation, multinucleated giant cells, and fibrinoid necrosis.

The inflammatory infiltrate is composed of an admixture of B and T lymphocytes. B cells are typically organized in germinal centers. Plasma cells staining for CD19, CD138, and IgG4 appear to radiate from the germinal centers. In contrast, the T cells, usually CD4+, are distributed more diffusely throughout the lesion and generally represent the most abundant cell type. Fibroblasts, histiocytes, and eosinophils can all be observed in moderate numbers. Some biopsy samples are particularly enriched with eosinophils. In other samples, particularly from long-standing cases, fibrosis predominates.

The histologic appearance of IgG4-RD, although highly characteristic, requires immunohistochemical confirmation of the diagnosis with IgG4 immunostaining. IgG4-positive plasma cells predominate within the lesion, but plasma cells containing immunoglobulins from each subclass can be found. The number of IgG4-positive plasma cells can be quantified by either counting the number of cells per high-power field (HPF) or by calculating the ratio of IgG4- to IgG-bearing plasma cells. Tissue fibrosis predominates in the latter phases of organ involvement, and in this relatively acellular phase of inflammation, both the IgG4:-total IgG ratio and the pattern of tissue fibrosis are more important than the number of IgG4-positive cells per HPF in establishing the diagnosis.

PATHOPHYSIOLOGY

Despite the emphasis of IgG4 in the name of this disease, the IgG4 molecule is not believed to play a direct role in the pathophysiology of disease within most organs. The IgG4 molecule can undergo Fab exchange, a phenomenon in which the two halves of the molecule dissociate from each other and reassociate with dissimilar hemi-molecules originating from other IgG4 molecules. Partly as a result of this Fab exchange, IgG4 antibodies do not bind antigen tightly. Moreover, the molecules have low affinities for Fc receptors and C1q and are regarded generally as noninflammatory immunoglobulins. The low affinities for Fc receptors and C1q impair the ability of IgG4 antibodies to induce phagocyte activation, antibody-dependent cellular cytotoxicity, and complement-mediated damage. It is possible that the role of IgG4 in this disease is actually as a counterregulatory mechanism rather than part of the primary inflammatory process.

Next-generation sequencing studies of CD4+ effector T cells have demonstrated a unique CD4+ cytotoxic T cell. This cell, also found in abundance at tissue sites of disease, makes interferon gamma, T-cell growth factor-beta, and interleukin-1, all of which may contribute to the storiform fibrosis found in this condition. The cells also elaborate perforin, granzyme A and B, and granulysin, products capable of inducing cytotoxicity. The pronounced oligoclonal expansion of this CD4+ cytotoxic T cell at tissue sites suggests that this cell is a major disease driver.

Oligoclonal expansions of plasmablasts are also present within the blood of patients with IgG4-RD. Continuous antigen presentation by B cells and plasmablasts may support this cell, which in turn produces profibrotic cytokines and other molecules, thereby directly mediating tissue injury.

TREATMENT

Not every disease manifestation of IgG4-RD requires immediate treatment because the disease takes an indolent form in many patients. IgG4-related lymphadenopathy, for example, can be asymptomatic for years, without evolution to other disease manifestations. Thus, watchful waiting is prudent in some cases. Vital organ involvement must be treated aggressively, however, because IgG4-RD can lead to serious organ dysfunction and failure. Aggressive disease can lead quickly to end-stage liver disease, permanent impairment of pancreatic function, renal atrophy, aortic dissection or aneurysms, and destructive lesions in the sinuses and nasopharynx.

Glucocorticoids are the first line of therapy. Treatment regimens, extrapolated from experience with the management of type 1 AIP, generally begin with 40 mg/d of prednisone, with tapering to discontinuation or maintenance doses of 5 mg/d within 2 or 3 months. Although the clinical response to glucocorticoids is usually swift and striking, prolonged steroid-free remissions are uncommon and the risk of steroid-induced morbidity in this middle-aged to elderly patient population is high, particularly those with baseline comorbidities and pancreatic involvement by IgG4-RD. Few data exist to support the utility of conventional steroid-sparing agents in this disease.

For patients with relapsing or glucocorticoid-resistant disease, B cell depletion with rituximab is an excellent second-line therapy. Rituximab treatment (two doses of 1 g IV, separated by approximately 15 days) leads to a swift decline in serum IgG4 concentrations, suggesting that rituximab achieves its effects in part by preventing the repletion of short-lived plasma cells that produce IgG4. More important than its effects on IgG4 concentrations, however, may be the effect of B cell depletion on T cell function. Specific effects of rituximab on the CD4+ cytotoxic T cell described above have been documented in IgG4-RD. Rituximab may be an appropriate first-line therapy for some patients, particularly those at high risk for glucocorticoid toxicity and patients with immediately organ-threatening disease. The rapidly evolving understanding of the pathophysiology of IgG4-RD suggests several novel targeted approaches to treating the disease, some of which are in clinical trials. These novel targets generally involve interference with either B or T cell directly for the purpose of depletion or functional inhibition.

FURTHER READING

Khosroshahi A et al: International Consensus Guidance Statement on the Management and Treatment of IgG4-Related Disease. Arthritis Rheumatol 67:1688, 2015.

Mahajan V et al: IgG4-related disease. Ann Rev Pathol 9:315, 2014.

Mattoo H et al: *De novo* oligoclonal expansions of circulating plasmablasts in active and relapsing IgG4-related disease. J Allergy Clin Immunol 134:679, 2014.

Mattoo H et al: Clonal expansion of CD4+ cytotoxic T lymphocytes in IgG4-related disease. J Allergy Clin Immunol 138:825, 2016.

Wallace ZS et al: IgG4-related disease: Baseline clinical and laboratory features in 125 patients with biopsy-proven disease. Arthritis Rheumatol 67:2466, 2015.

362 Familial Mediterranean Fever and Other Hereditary Autoinflammatory Diseases

Daniel L. Kastner

Familial Mediterranean fever (FMF) is the prototype of a group of inherited diseases **(Table 362-1)** that are characterized by recurrent episodes of fever with serosal, synovial, or cutaneous inflammation and, in some individuals, the eventual development of systemic AA amyloidosis **(Chap. 108)**. Because of the relative infrequency of high-titer autoantibodies or antigen-specific T cells, the term *autoinflammatory* has been proposed to describe these disorders, rather than autoimmune. The innate immune system, with its myeloid effector cells and germline receptors for pathogen-associated molecular patterns and endogenous danger signals, plays a predominant role in the pathogenesis of the autoinflammatory diseases. Although the hereditary recurrent fevers

TABLE 362-1 The Hereditary Recurrent Fever Syndromes

	FMF	TRAPS	HIDS	MWS	FCAS	NOMID
Ethnicity	Jewish, Arab, Turkish, Armenian, Italian	Any ethnic group	Predominantly Dutch, northern European	Any ethnic group	Any ethnic group	Any ethnic group
Inheritance	Recessive[a]	Dominant	Recessive	Dominant	Dominant	Most commonly de novo mutations; somatic mosaicism in a significant minority
Gene/chromosome	*MEFV*/16p13.3	*TNFRSF1A*/12p13	*MVK*/12q24	*NLRP3*/1q44	*NLRP3*/1q44	*NLRP3*/1q44
Protein	Pyrin	p55 TNF receptor	Mevalonate kinase	NLRP3 (cryopyrin)	NLRP3 (cryopyrin)	NLRP3 (cryopyrin)
Attack length	1–3 days	Often >7 days	3–7 days	1–2 days	Minutes–3 days	Continuous, with flares
Serosa	Pleurisy, peritonitis; asymptomatic pericardial effusions	Pleurisy, peritonitis, pericarditis	Abdominal pain, but seldom peritonitis; pleurisy, pericarditis uncommon	Abdominal pain; pleurisy, pericarditis rare	Rare	Rare
Skin	Erysipeloid erythema	Centrifugally migrating erythema	Diffuse maculopapular rash; oral ulcers	Diffuse urticaria-like rash	Cold-induced urticaria-like rash	Diffuse urticaria-like rash
Joints	Acute monoarthritis; chronic hip arthritis (rare)	Acute monoarthritis, arthralgia	Arthralgia, oligoarthritis	Arthralgia, large joint oligoarthritis	Polyarthralgia	Epiphyseal, patellar overgrowth, clubbing
Muscle	Exercise-induced myalgia common; protracted febrile myalgia rare	Migratory myalgia	Uncommon	Myalgia common	Sometimes myalgia	Sometimes myalgia
Eyes, ears	Uncommon	Periorbital edema, conjunctivitis, rarely uveitis	Uncommon	Conjunctivitis, episcleritis, optic disc edema; sensorineural hearing loss	Conjunctivitis	Conjunctivitis, uveitis, optic disc edema, blindness, sensorineural hearing loss
CNS	Aseptic meningitis rare	Headache	Headache	Headache	Headache	Aseptic meningitis, seizures
Amyloidosis	Most common in M694V homozygotes	~15% of cases, most often cysteine mutations, T50M	Uncommon	~25% of cases	Uncommon	Late complication
Treatment	Oral colchicine prophylaxis, IL-1 inhibitors for refractory cases	Glucocorticoids, etanercept, IL-1 inhibitors	NSAIDs for fever; IL-1 inhibitors	Anakinra, rilonacept, canakinumab	Anakinra, rilonacept, canakinumab	Anakinra

[a]A substantial percentage of patients with clinical FMF have only a single demonstrable *MEFV* mutation on DNA sequencing.

Abbreviations: CNS, central nervous system; FCAS, familial cold autoinflammatory syndrome; FMF, familial Mediterranean fever; HIDS, hyperimmunoglobulinemia D with periodic fever syndrome; IL, interleukin; MWS, Muckle-Wells syndrome; NOMID, neonatal-onset multisystem inflammatory disease; NSAIDs, nonsteroidal anti-inflammatory drugs; TNF, tumor necrosis factor; TRAPS, TNF receptor-associated periodic syndrome.

comprise a major category of the autoinflammatory diseases, other inherited disorders of inflammation in which recurrent fever plays a less prominent role are now also considered to be autoinflammatory.

BACKGROUND AND PATHOPHYSIOLOGY

FMF was first recognized among Armenians, Arabs, Turks, and non-Ashkenazi (primarily North African and Iraqi) Jews. With the advent of genetic testing, FMF has been documented with increasing frequency among Ashkenazi Jews, Italians, and other Mediterranean populations, and occasional cases have been confirmed even in the absence of known Mediterranean ancestry. FMF is generally regarded as recessively inherited, but there is an increasing awareness of clear-cut clinical cases with only a single demonstrable genetic mutation, and, for certain relatively rare FMF mutations, there is strong evidence for dominant inheritance. Particularly in countries where families are small, a positive family history can only be elicited in ~50% of cases. DNA testing demonstrates carrier frequencies as high as 1:3 among affected populations, suggesting a heterozygote advantage.

The FMF gene encodes a 781-amino acid, ~95 kDa protein denoted *pyrin* (or *marenostrin*) that is expressed in granulocytes, eosinophils, monocytes, dendritic cells, and synovial and peritoneal fibroblasts. The N-terminal 92 amino acids of pyrin define a motif, the PYRIN domain, that mediates homotypic protein-protein interactions and has been found in several other proteins, including cryopyrin (NLRP3), which is mutated in three other recurrent fever syndromes. Through the interaction of its PYRIN domain with an intermediary adaptor protein, pyrin nucleates the formation of a macromolecular *pyrin inflammasome* to activate caspase-1 (interleukin [IL] 1β-converting enzyme), and thereby IL-1β secretion. Certain bacterial toxins that block leukocyte cytoskeletal assembly by inactivating RhoA GTPase trigger pyrin inflammasome activation as a part of the normal host defense; in FMF patients the threshold for pyrin inflammasome activation is reduced.

ACUTE ATTACKS

Febrile episodes in FMF may begin even in early infancy; 90% of patients have had their first attack by age 20. Typical FMF episodes generally last 24–72 h, with arthritic attacks tending to last somewhat longer. In some patients, the episodes occur with great regularity, but more often, the frequency of attacks varies over time, ranging from as often as once every few days to remissions lasting several years. Attacks are often unpredictable, although some patients relate them to physical exertion, emotional stress, or menses; pregnancy may be associated with remission.

If measured, fever is nearly always present throughout FMF attacks. Severe hyperpyrexia and even febrile seizures may be seen in infants, and fever is sometimes the only manifestation of FMF in young children.

Over 90% of FMF patients experience abdominal attacks at some time. Episodes range in severity from dull, aching pain and distention

with mild tenderness on direct palpation to severe generalized pain with absent bowel sounds, rigidity, rebound tenderness, and air-fluid levels on upright radiographs. Computed tomography (CT) scanning may demonstrate a small amount of fluid in the abdominal cavity. If such patients undergo exploratory laparotomy, a sterile, neutrophil-rich peritoneal exudate is present, sometimes with adhesions from previous episodes. Ascites is rare.

Pleural attacks are usually manifested by unilateral, sharp, stabbing chest pain. Radiographs may show atelectasis and sometimes an effusion. If performed, thoracentesis demonstrates an exudative fluid rich in neutrophils. After repeated attacks, pleural thickening may develop.

FMF arthritis is most frequent among individuals homozygous for the M694V mutation, which is especially common in the non-Ashkenazi Jewish population. Acute arthritis in FMF is usually monoarticular, affecting the knee, ankle, or hip, although other patterns can be seen. Large sterile effusions rich in neutrophils are frequent, without commensurate erythema or warmth. Even after repeated arthritic attacks, radiographic changes are rare. Before the advent of colchicine prophylaxis, chronic arthritis of the knee or hip was seen in ~5% of FMF patients with arthritis. Chronic sacroiliitis can occur in FMF irrespective of the HLA-B27 antigen, even in the face of colchicine therapy. In the United States, FMF patients are much more likely to have arthralgia than arthritis.

The most characteristic cutaneous manifestation of FMF is erysipelas-like erythema, a raised erythematous rash that most commonly occurs on the dorsum of the foot, ankle, or lower leg alone or in combination with abdominal pain, pleurisy, or arthritis. Biopsy demonstrates perivascular infiltrates of granulocytes and monocytes. This rash is seen most often in M694V homozygotes and is relatively rare in the United States.

Exercise-induced (nonfebrile) myalgia is common in FMF, and a small percentage of patients develop a protracted febrile myalgia that can last several weeks. Symptomatic pericardial disease is rare, although small pericardial effusions may be noted on echocardiography. Unilateral acute scrotal inflammation may occur in prepubertal boys. Aseptic meningitis has been reported in FMF, but the causal connection is controversial. Vasculitis, including Henoch-Schönlein purpura and polyarteritis nodosa **(Chap. 356)**, may be seen at increased frequency in FMF. The M694V FMF mutation has recently been shown to be a risk factor for Behçet's disease.

Laboratory features of FMF attacks are consistent with acute inflammation and include an elevated erythrocyte sedimentation rate, leukocytosis, thrombocytosis (in children), and elevations in C-reactive protein, fibrinogen, haptoglobin, and serum immunoglobulins. Transient albuminuria and hematuria may also be seen.

AMYLOIDOSIS

Before the advent of colchicine prophylaxis, systemic amyloidosis was a common complication of FMF. It is caused by deposition of a fragment of serum amyloid A, an acute-phase reactant, in the kidneys, adrenals, intestine, spleen, lung, and testes **(Chap. 108)**. Amyloidosis should be suspected in patients who have proteinuria between attacks; renal or rectal biopsy is used most often to establish the diagnosis. Risk factors include the M694V homozygous genotype, positive family history (independent of FMF mutational status), the *SAA1* genotype, male gender, noncompliance with colchicine therapy, and having grown up in the Middle East.

DIAGNOSIS

For typical cases, physicians experienced with FMF can often make the diagnosis on clinical grounds alone. Clinical criteria sets for FMF have been shown to have high sensitivity and specificity in parts of the world where the pretest probability of FMF is high. Genetic testing can provide a useful adjunct in ambiguous cases or for physicians not experienced in FMF. Most of the more severe disease-associated FMF mutations are in exon 10 of the gene. An updated list of mutations for FMF and other hereditary recurrent fevers can be found online at *http://fmf.igh.cnrs.fr/infevers/*.

Genetic testing has permitted a broadening of the clinical spectrum and geographic distribution of FMF and may be of prognostic value. Most studies indicate that M694V homozygotes have an earlier age of onset and a higher frequency of arthritis, rash, and amyloidosis. In contrast, the E148Q variant in exon 2 is quite common in certain Asian populations and is more likely to affect overall levels of inflammation than to cause clinical FMF. E148Q is sometimes found in *cis* with exon 10 mutations, which may complicate the interpretation of genetic test results. Only ~70% of patients with clinically typical FMF have two identifiable mutations in *trans*, consistent with the concept that FMF mutations are gain-of-function with regard to inflammasome activation, with a gene dosage effect. In those cases in which only a single mutation is identified, clinical judgment is very important, and sometimes a therapeutic trial of colchicine may help to confirm the diagnosis.

If a patient is seen during his or her first attack, the differential diagnosis may be broad, although delimited by the specific organ involvement. After several attacks the differential diagnosis may include the other hereditary recurrent fever syndromes (Table 362-1); the syndrome of periodic fever with aphthous ulcers, pharyngitis, and cervical adenopathy (PFAPA); systemic-onset juvenile rheumatoid arthritis or adult Still's disease; porphyria; hereditary angioedema; inflammatory bowel disease; and, in women, gynecologic disorders.

TREATMENT

Familial Mediterranean Fever

The treatment of choice for FMF is daily oral colchicine, which decreases the frequency and intensity of attacks and prevents the development of amyloidosis in compliant patients. Intermittent dosing at the onset of attacks is not as effective as daily prophylaxis and is of unproven value in preventing amyloidosis. The usual adult dose of colchicine is 1.2–1.8 mg/d, which causes substantial reduction in symptoms in two-thirds of patients and some improvement in >90%. Children may require lower doses, although not proportionately to body weight.

Common side effects of colchicine include bloating, abdominal cramps, lactose intolerance, and diarrhea. They can be minimized by starting at a low dose and gradually advancing as tolerated, splitting the dose, use of simethicone for flatulence, and avoidance of dairy products. If taken by either parent at the time of conception, colchicine may cause a small increase in the risk of trisomy 21 (Down's syndrome). In elderly patients with renal insufficiency, colchicine can cause a myoneuropathy characterized by proximal muscle weakness and elevation of the creatine kinase. Cyclosporine inhibits hepatic excretion of colchicine by its effects on the multidrug resistance 1 (MDR1) transport system, sometimes leading to colchicine toxicity in patients who have undergone renal transplantation for amyloidosis. Intravenous colchicine should generally not be administered to patients already taking oral colchicine, because severe, sometimes fatal, toxicity can occur in this setting.

For FMF patients who do not respond to colchicine or cannot tolerate therapeutic doses, injectable IL-1 inhibitors may be used. Based on a randomized, placebo-controlled phase III trial, the monoclonal anti-IL-1β antibody canakinumab recently received the Food and Drug Administration (FDA) approval for this indication. In a small randomized placebo-controlled trial, weekly subcutaneous rilonacept, a recombinant interleukin 1 (IL-1) receptor fusion protein, significantly reduced the frequency of attacks. There is also substantial anecdotal experience with daily subcutaneous anakinra, a recombinant IL-1 receptor antagonist, in preventing the acute attacks of FMF and, in some cases, reducing established amyloid deposits. Bone marrow transplantation has been suggested for refractory FMF, but the risk-benefit ratio is currently regarded as unacceptable.

OTHER HEREDITARY RECURRENT FEVERS

TNF RECEPTOR-ASSOCIATED PERIODIC SYNDROME

TNF receptor–associated periodic syndrome (TRAPS) is caused by dominantly inherited mutations in the extracellular domains of the 55-kDa TNF receptor (TNFR1, p55). Although originally described

in a large Irish family (and hence the name *familial Hibernian fever*), TRAPS has a broad ethnic distribution. TRAPS episodes often begin in childhood. The duration of attacks ranges from 1 to 2 days to as long as several weeks, and in severe cases symptoms may be nearly continuous. In addition to peritoneal, pleural, and synovial attacks similar to FMF, TRAPS patients frequently have ocular inflammation (most often conjunctivitis and/or periorbital edema), and a distinctive migratory myalgia with overlying painful erythema may be present. TRAPS patients generally respond better to glucocorticoids than to prophylactic colchicine. Untreated, ~15% develop amyloidosis. The diagnosis of TRAPS is based on the demonstration of a *TNFRSF1A* mutation in the presence of characteristic symptoms. Two particular variants, R92Q and P46L, are common in certain populations and may act more as functional polymorphisms than as disease-causing mutations. In contrast, pathogenic *TNFRSF1A* mutations, including a number of substitutions at highly conserved cysteine residues, are associated with intracellular TNFR1 misfolding, aggregation, and retention, with consequent ligand-independent kinase activation, mitochondrial reactive oxygen species production, and proinflammatory cytokine release. Etanercept, a TNF inhibitor, ameliorates TRAPS attacks, but the long-term experience with this agent has been less favorable. IL-1 inhibition has been beneficial in a large percentage of the patients in whom it has been used, and canakinumab recently received FDA approval for the treatment of TRAPS. Monoclonal anti-TNF antibodies should be avoided, because they may exacerbate TRAPS attacks.

HYPERIMMUNOGLOBULINEMIA D WITH PERIODIC FEVER SYNDROME

Hyperimmunoglobulinemia D with periodic fever syndrome (HIDS) is a recessively inherited recurrent fever syndrome found primarily in individuals of northern European ancestry. It is caused by mutations in mevalonate kinase (*MVK*), encoding an enzyme involved in the synthesis of cholesterol and nonsterol isoprenoids, including geranylgeranyl pyrophosphate. The latter compound is essential for proper localization of RhoA GTPase to the cell membrane, and the mislocalization of RhoA leads to its inactivation and the consequent activation of the pyrin inflammasome. HIDS attacks usually begin in infancy and last 3–5 days. Clinically distinctive features include painful cervical adenopathy, a diffuse maculopapular rash sometimes affecting the palms and soles, and aphthous ulcers; pleurisy is rare, as is amyloidosis. Although originally defined by the persistent elevation of serum IgD, disease activity is not related to IgD levels, and some patients with FMF or TRAPS may have modestly increased serum IgD. Moreover, occasional patients with *MVK* mutations and recurrent fever have normal IgD levels, while all patients with mutations have markedly elevated urinary mevalonate levels during their attacks. For these reasons, some have proposed renaming this disorder *mevalonate kinase deficiency* (MKD). Canakinumab was recently FDA-approved for the treatment of HIDS/MKD.

THE CRYOPYRINOPATHIES OR CRYOPYRIN-ASSOCIATED PERIODIC SYNDROMES

Three hereditary febrile syndromes, familial cold autoinflammatory syndrome (FCAS), Muckle-Wells syndrome (MWS), and neonatal-onset multisystem inflammatory disease (NOMID), are all caused by mutations in *NLRP3* (formerly known as *CIAS1*), the gene encoding cryopyrin (or NLRP3), and represent a clinical spectrum of disease. FCAS patients develop chills, fever, headache, arthralgia, conjunctivitis, and an urticaria-like rash in response to generalized cold exposure. In MWS, an urticarial rash is noted, but it is not usually induced by cold; MWS patients also develop fevers, abdominal pain, limb pain, arthritis, conjunctivitis, and, over time, sensorineural hearing loss. NOMID is the most severe of the three disorders, with chronic aseptic meningitis, a characteristic arthropathy, and rash. Like the FMF protein, pyrin, cryopyrin has an N-terminal PYRIN domain, allowing the formation of an *NLRP3 inflammasome* that mediates caspase-1 activation and IL-1β release. Peripheral blood leukocytes from patients with FCAS, MWS, and NOMID release increased amounts of IL-1β upon in vitro stimulation, relative to healthy controls. Macrophages from cryopyrin-deficient mice exhibit decreased IL-1β production in response to certain gram-positive bacteria, bacterial RNA, and monosodium urate crystals. Patients with all three cryopyrinopathies or cryopyrin-associated periodic syndromes (CAPS) show a dramatic response to injections of IL-1 inhibitors. Canakinumab and rilonacept are FDA-approved for the treatment of FCAS and MWS, while anakinra is approved for the treatment of NOMID.

Approximately one-third of patients with clinical manifestations of NOMID do not have germline mutations in *NLRP3*, but they have been found to be mosaic for somatic *NLRP3* mutations. Such patients also respond dramatically to IL-1 inhibition. Similarly, somatic mosaicism in *NLRP3* has been found in Schnitzler's syndrome, which presents in middle age with recurrent fever, urticarial rash, elevated acute phase reactants, monoclonal IgM gammopathy, and abnormal bone remodeling. IL-1 inhibition is the treatment of choice for Schnitzler's syndrome.

OTHER INHERITED AUTOINFLAMMATORY DISEASES

There are a number of other Mendelian autoinflammatory diseases in which recurrent fevers are not a prominent clinical sign but that involve abnormalities of innate immunity. The syndrome of pyogenic arthritis with pyoderma gangrenosum and acne (PAPA) is a dominantly inherited disorder that presents with episodes of sterile pyogenic monoarthritis often induced by trauma, severe pyoderma gangrenosum, and severe cystic acne usually beginning in puberty. It is caused by mutations in *PSTPIP1*, which encodes a pyrin-binding protein, and the arthritic manifestations often respond to IL-1 inhibition. Patients with the recessively inherited deficiency of the IL-1 receptor antagonist (DIRA) present with a generalized pustular rash and multifocal sterile osteomyelitis and show dramatic clinical responses to anakinra, the recombinant form of the protein they lack. IL-36 is another member of the IL-1 family of cytokines that is regulated by an endogenous receptor antagonist. The recessively inherited deficiency of the IL-36 receptor antagonist (DITRA) presents with episodes of generalized pustular psoriasis and dramatic systemic inflammation. Dominantly inherited gain-of-function mutations in *NLRC4* lead to increased IL-1 and IL-18 production and potentially life-threatening recurrent macrophage activation syndrome.

Whereas the aforementioned disorders all involve mutations in IL-1-related molecules, other autoinflammatory diseases are caused by mutations in other components of innate immunity. *Blau's syndrome* is caused by mutations in *CARD15* (also known as *NOD2*), which regulates nuclear factor κB activation. Blau's syndrome is characterized by granulomatous dermatitis, uveitis, and arthritis; distinct *CARD15* variants predispose to Crohn's disease. Recessive mutations in one or more components of the proteasome lead to excessive interferon signaling and the syndrome of chronic atypical neutrophilic dermatosis with lipodystrophy and elevated temperature (CANDLE), a severe form of generalized panniculitis. De novo gain-of-function mutations in *TMEM173*, encoding the stimulator of interferon genes (STING), cause severe vasculopathy and pulmonary fibrosis. Recessive loss-of-function mutations in *CERCR1*, encoding adenosine deaminase 2 (ADA2), cause a vasculopathy that can manifest as livedoid rash, early-onset lacunar strokes, or polyarteritis nodosa. Mutations in the gene encoding the A20 ubiquitin-modifying enzyme cause a Behçet's-like monogenic illness ("HA20") while mutations in a different deubiquitinase (OTULIN) cause a form of panniculitis ("otulipenia").

Finally, it should be noted that a number of common, genetically complex disorders are now sometimes considered autoinflammatory, because of evidence that components of the innate immune system, such as the inflammasome, may play a role in the pathogenesis. Two prominent examples are gout and atherosclerosis.

GLOBAL CONSIDERATIONS

All the disorders discussed in this chapter have been observed in multiple populations. However, as noted herein, FMF is most frequently observed in Mediterranean and Middle-Eastern populations and HIDS in Northern European populations, particularly the Dutch. A recessive founder mutation in *CECR1* is

particularly common in the Georgian Jewish population and is associated with polyarteritis nodosa.

FURTHER READING

Canna SW et al: An activating *NLRC4* inflammasome mutation causes autoinflammation with recurrent macrophage activation syndrome. Nat Genet 46:1140, 2014.

de Jesus AA, Goldbach-Mansky R: Newly recognized Mendelian disorders with rheumatic manifestations. Curr Opin Rheumatol 27:511, 2015.

Park YH et al: Pyrin inflammasome activation and RhoA signaling in the autoinflammatory diseases FMF and HIDS. Nat Immunol 17:914, 2016.

Stoffels M, Kastner DL: Old dogs, new tricks: Monogenic autoinflammatory disease unleashed. Annu Rev Genom Hum Genet 17:245, 2016.

Zhou Q et al: Early-onset stroke and vasculopathy associated with mutations in ADA2. N Engl J Med 370:911, 2014.

Section 3 Disorders of the Joints and Adjacent Tissues

363 Approach to Articular and Musculoskeletal Disorders

John J. Cush

Musculoskeletal complaints account for >315 million outpatient visits per year and >20% of all outpatient visits in the United States. The Centers for Disease Control and Prevention estimate that 54.4 million, or 1 in 5 adults) of the U.S. population has physician-diagnosed arthritis. While many patients will have self-limited conditions requiring minimal evaluation, reassurance, and symptomatic therapy, specific musculoskeletal presentations or their persistence may herald a more serious condition that requires further evaluation or laboratory testing to establish a diagnosis. The goal of the musculoskeletal evaluation is to formulate a differential diagnosis that leads to an accurate diagnosis and timely therapy, while avoiding excessive diagnostic testing and unnecessary treatment **(Table 363-1)**. There are several urgent conditions that must be diagnosed promptly to avoid significant morbid or mortal sequelae. These "red flag" diagnoses include septic arthritis, acute crystal-induced arthritis (e.g., gout), and fracture. Each may be suspected by its acute onset and monarticular or focal musculoskeletal pain.

The majority of individuals with musculoskeletal complaints can be diagnosed with a thorough history and a comprehensive physical and musculoskeletal examination. The initial encounter should determine whether the musculoskeletal complaint signals a red flag condition (septic arthritis, gout, or fracture) or not. The evaluation should ascertain if the complaint is (1) *articular* or *nonarticular* in origin, (2) *inflammatory* or *noninflammatory* in nature, (3) *acute* or *chronic* in duration, and (4) *localized (monarticular)* or *widespread* (*polyarticular*) in distribution.

With this approach, the musculoskeletal presentation can be characterized (e.g., acute inflammatory monarthritis or a chronic noninflammatory, nonarticular widespread pain) to narrow the diagnostic possibilities. However, some patients will not fit immediately into an established diagnostic category. Many musculoskeletal disorders resemble each other at the outset, and some may take weeks or months (but not years) to evolve into a recognizable diagnostic entity. This consideration should temper the desire to establish a definitive diagnosis at the first encounter.

TABLE 363-1 Evaluation of Patients with Musculoskeletal Complaints

Goals
Accurate diagnosis
Timely provision of therapy
Avoidance of unnecessary diagnostic testing
Identification of acute, focal/monarticular "red flag" conditions
Approach
Determine the chronology (acute vs chronic)
Determine the nature of the pathologic process (inflammatory vs noninflammatory)
Determine the extent of involvement (monarticular, polyarticular, focal, widespread)
Anatomic localization of complaint (articular vs nonarticular)
Consider the most common disorders first
Consider the need for diagnostic testing
Formulate a differential diagnosis

ARTICULAR VERSUS NONARTICULAR

The musculoskeletal evaluation must discriminate the anatomic origin(s) of the patient's complaint. For example, ankle pain can result from a variety of pathologic conditions involving disparate anatomic structures, including gouty arthritis, calcaneal fracture, Achilles tendinitis, plantar fasciitis, cellulitis, and peripheral or entrapment neuropathy. Distinguishing between articular and nonarticular conditions requires a careful and detailed examination. Articular structures include the synovium, synovial fluid, articular cartilage, intraarticular ligaments, joint capsule, and juxtaarticular bone. Nonarticular (or periarticular) structures, such as supportive extraarticular ligaments, tendons, bursae, muscle, fascia, bone, nerve, and overlying skin, may be involved in the pathologic process. Although musculoskeletal complaints are often ascribed to the joints, nonarticular disorders more frequently underlie such complaints. Distinguishing between these potential sources of pain may be challenging to the unskilled examiner. Articular disorders may be characterized by deep or diffuse pain, pain or limited range of motion on active and passive movement, and swelling (caused by synovial proliferation, effusion, or bony enlargement), crepitation, instability, "locking," or deformity. By contrast, nonarticular disorders tend to be painful on active, but not passive (or assisted), range of motion. Periarticular conditions often demonstrate point or focal tenderness in regions adjacent to articular structures, may radiate or be elicited with a specific movement or position, and have physical findings remote from the joint capsule. Moreover, nonarticular disorders seldom demonstrate swelling, crepitus, instability, or deformity of the joint itself.

INFLAMMATORY VERSUS NONINFLAMMATORY DISORDERS

In the course of a musculoskeletal evaluation, the examiner should determine the nature of the underlying pathologic process and whether inflammatory or noninflammatory findings exist. Inflammatory disorders may be infectious (*Neisseria gonorrhoeae* or *Mycobacterium tuberculosis*), crystal-induced (gout, pseudogout), immune-related (rheumatoid arthritis [RA], systemic lupus erythematosus [SLE]), reactive (rheumatic fever, reactive arthritis), or idiopathic. Inflammatory disorders may be identified by any of the four cardinal signs of inflammation (erythema, warmth, pain, or swelling), systemic symptoms (fatigue, fever, rash, weight loss), or laboratory evidence of inflammation (elevated erythrocyte sedimentation rate [ESR] or C-reactive protein [CRP], thrombocytosis, anemia of chronic disease, or hypoalbuminemia). Articular stiffness commonly accompanies chronic musculoskeletal disorders. The duration of stiffness may be prolonged (hours) with inflammatory disorders (such as RA or polymyalgia rheumatica) and improves with activity. By contrast, intermittent stiffness (also known as gel phenomenon) is typical of noninflammatory conditions (such as osteoarthritis [OA]), shorter in duration (<60 min), and is exacerbated by activity. Fatigue may be profound with inflammation (as seen in RA and polymyalgia rheumatica) but may also be a consequence of fibromyalgia (a noninflammatory disorder), chronic pain, poor sleep, depression, anemia, cardiac failure, endocrinopathy, or malnutrition.

Noninflammatory disorders may be related to trauma (rotator cuff tear), repetitive use (bursitis, tendinitis), degeneration or ineffective

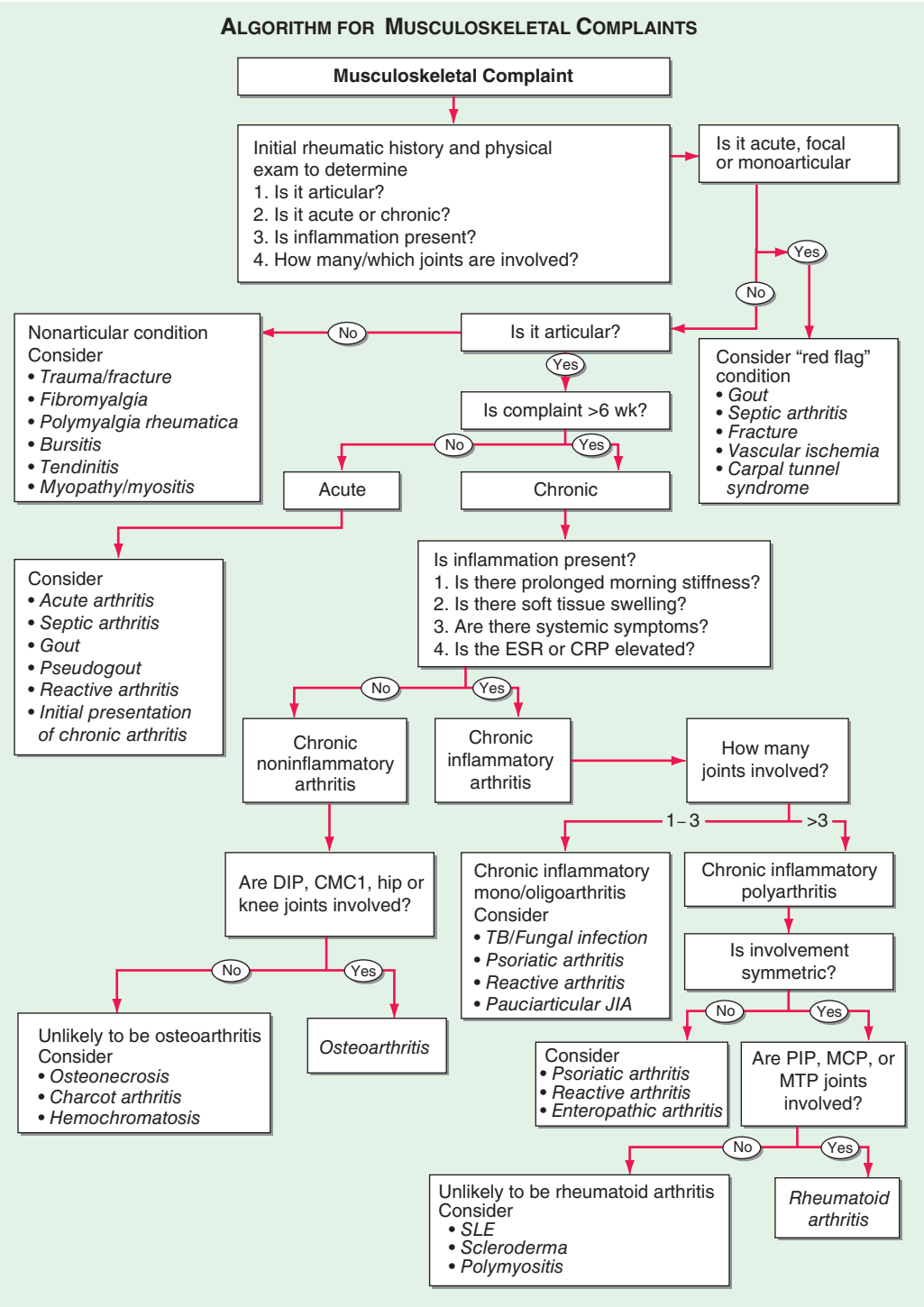


FIGURE 363-1 Algorithm for the diagnosis of musculoskeletal complaints. An approach to formulating a differential diagnosis (shown in italics). CMC, carpometacarpal; CRP, C-reactive protein; DIP, distal interphalangeal; ESR, erythrocyte sedimentation rate; JIA, juvenile idiopathic arthritis; MCP, metacarpophalangeal; MTP, metatarsophalangeal; PIP, proximal interphalangeal; PMR, polymyalgia rheumatica; SLE, systemic lupus erythematosus.

repair (OA), neoplasm (pigmented villonodular synovitis), or pain amplification (fibromyalgia). Noninflammatory disorders are often characterized by pain without synovial swelling or warmth, absence of inflammatory or systemic features, daytime, intermittent gel phenomena rather than prolonged morning stiffness, and normal (for age) or negative laboratory investigations.

Identification of the nature of the underlying process and the site of the complaint will enable the examiner to characterize the musculoskeletal presentation (e.g., acute inflammatory monarthritis, chronic noninflammatory, nonarticular widespread pain). By narrowing the diagnostic considerations, the examiner can assess the need for immediate diagnostic or therapeutic intervention or for continued observation. Figure 363-1 presents an algorithmic approach to the evaluation of patients with musculoskeletal complaints. This approach relies on clinical and historic features, rather than laboratory testing, to diagnose many common rheumatic disorders.

A simpler, alternative approach would consider the most commonly encountered complaints first, based on frequency in younger versus older populations. The most prevalent causes of musculoskeletal complaints are shown in Fig. 363-2. Because trauma, fracture, overuse syndromes, and fibromyalgia are among the most common causes of musculoskeletal pain, these should be considered during the initial encounter. If excluded, other frequently occurring disorders should be considered according to the patient's age. Hence, those aged <60 years

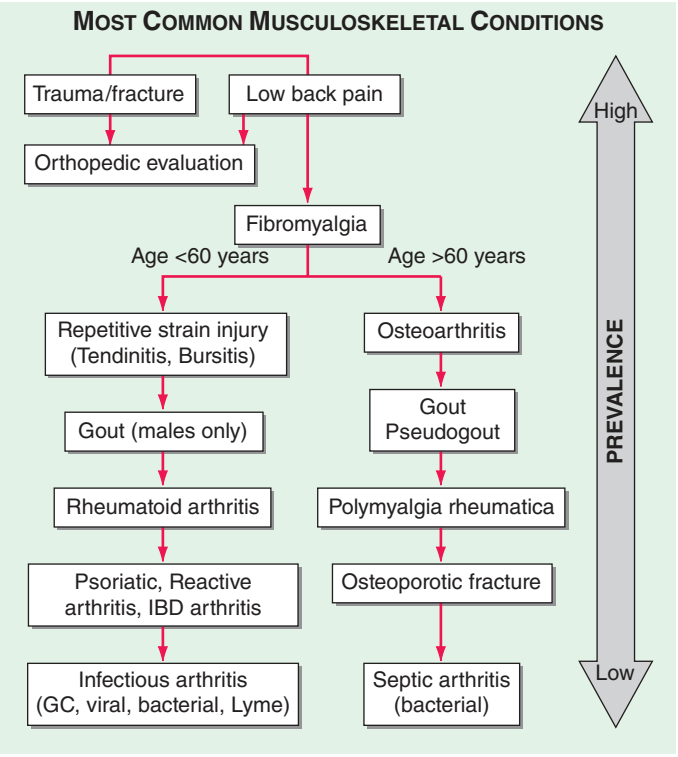


FIGURE 363-2 Algorithm for consideration of the most common musculoskeletal conditions. GC, gonococcal; IBD, inflammatory bowel disease.

are commonly affected by repetitive use/strain disorders, gout (men only), RA, spondyloarthritis, and uncommonly, infectious arthritis. Patients aged >60 years are frequently affected by OA, crystal (gout and pseudogout) arthritis, polymyalgia rheumatica, osteoporotic fracture, and uncommonly, septic arthritis. These conditions are between 10 and 100 times more prevalent than other serious autoimmune conditions, such as SLE, scleroderma, polymyositis, and vasculitis.

CLINICAL HISTORY

Historic features may reveal important clues to the diagnosis. Aspects of the patient profile, complaint chronology, extent of joint involvement, and precipitating factors can provide important information. Certain diagnoses are more frequent in different *age* groups. SLE and reactive arthritis occur more frequently in the young, whereas fibromyalgia and RA are frequent in middle age, and OA and polymyalgia rheumatica are more prevalent among the elderly. Diagnostic clustering is also evident when *sex* and *race* are considered. Gout, spondyloarthritis, and ankylosing spondylitis are more common in men, whereas RA, fibromyalgia, osteoporosis and lupus are more frequent in women. *Racial predilections* may be evident. Thus, polymyalgia rheumatica, giant cell arteritis, and granulomatosis with polyangiitis (GPA; formerly called Wegener's granulomatosis) commonly affect whites, whereas sarcoidosis and SLE more commonly affect African Americans. *Familial aggregation* is most common with ankylosing spondylitis, gout, and Heberden's nodes of OA.

The chronology of the complaint is an important diagnostic feature and can be divided into the *onset, evolution*, and *duration*. The onset of disorders such as septic arthritis or gout tends to be abrupt, whereas OA, RA, and fibromyalgia may have more indolent presentations. The patients' complaints may evolve differently and be classified as chronic (OA), intermittent (crystal or Lyme arthritis), migratory (rheumatic fever, gonococcal or viral arthritis), or additive (RA, psoriatic arthritis). Musculoskeletal disorders are typically classified as acute or chronic based on a symptom duration that is either <6 weeks or >6 weeks, respectively. Acute arthropathies tend to be infectious, crystal-induced, or reactive. Chronic conditions include noninflammatory or immunologic arthritides (e.g., OA, RA) and nonarticular disorders (e.g., fibromyalgia).

The *extent* or *distribution* of articular involvement is often informative. Articular disorders are classified based on the number of joints involved, as either *monarticular* (one joint), *oligoarticular* or *pauciarticular* (two or three joints), or *polyarticular* (four or more joints). Although crystal and infectious arthritis are often mono- or oligoarticular, OA and RA are polyarticular disorders. Nonarticular disorders may be classified as either focal or widespread. Complaints secondary to tendinitis or carpal tunnel syndrome are typically focal, whereas weakness and myalgia, caused by polymyositis or fibromyalgia, are more widespread in their presentation. Joint involvement in RA tends to be symmetric and polyarticular. By contrast, spondyloarthritis, reactive arthritis, gout, and sarcoid are often asymmetric and oligoarticular. OA and psoriatic arthritis may be either symmetric or asymmetric and oligo- or polyarticular. The upper extremities are frequently involved in RA and OA, whereas lower extremity arthritis is characteristic of reactive arthritis and gout at their onset. Involvement of the axial skeleton is common in OA and ankylosing spondylitis but is infrequent in RA, with the notable exception of the cervical spine.

The clinical history should also identify *precipitating events*, such as trauma (osteonecrosis, meniscal tear), drug administration **(Table 363-2)**, antecedent or intercurrent infection (rheumatic fever, reactive arthritis, hepatitis), or illnesses that may have contributed to the patient's complaint. Certain comorbidities may have musculoskeletal consequences. This is especially so for diabetes mellitus (carpal tunnel syndrome), renal insufficiency (gout), depression or insomnia (fibromyalgia), myeloma (low back pain), cancer (myositis), and osteoporosis (fracture) or when using certain drugs such as glucocorticoids (osteonecrosis, septic arthritis), diuretics or chemotherapy (gout) (Table 363-2).

Lastly, a thorough *rheumatic review of systems* may disclose useful diagnostic information. A variety of musculoskeletal disorders may be associated with systemic features such as fever (SLE, infection), rash (SLE, psoriatic arthritis), nail abnormalities (psoriatic or reactive arthritis), myalgias (fibromyalgia, statin- or drug-induced myopathy), or weakness (polymyositis, neuropathy). In addition, some conditions are associated with involvement of other organ systems including the eyes (Behçet's disease, sarcoidosis, spondyloarthritis), gastrointestinal tract (scleroderma, inflammatory bowel disease), genitourinary tract (reactive arthritis, gonococcemia), or nervous system (Lyme disease, vasculitis).

FIBROMYALGIA

Historically, syphilis and tuberculosis were labeled as the "great masqueraders" as their protean symptoms and potential for multi-organ involvement may result in delays in diagnosis and treatment. In the modern era, other serious diagnoses (including lupus, sarcoidosis, vasculitis and lymphoma) have also been labeled as great masqueraders. All of these are either uncommon or rare, and are overshadowed by the most common masquerader with musculoskeletal complaints—fibromyalgia. Fibromyalgia **(see Chap. 366)** is a pain amplification disorder unified by sleep disturbance, exaggerated pain and sensitivity (owing to lowered pain thresholds), and a multiplicity of symptoms with a paucity of abnormalities on clinical examination or laboratory testing. Tender "trigger points" are often found and include tenderness over the epicondyles, trochanteric bursae, anserine bursae, and muscles (gluteal, trapezius, supraspinatus) that often are misdiagnosed as other nonarticular conditions. Although fibromyalgia classically manifests as widespread aches and pains, presenting symptoms tend to be less specific, and only on further evaluation will the widespread noninflammatory features be disclosed. Fibromyalgia has numerous comorbidities including irritable bowel syndrome, dysmenorrhea, migraine, depression, anxiety, memory loss, non-anatomic paresthesia or dysesthesia, fatigue, myalgias, temporomandibular joint pain, and multiple chemical sensitivities. Fibromyalgia patients often see multiple specialists, are twice as likely to be hospitalized, and are plagued by polypharmacy. Fibromyalgia affects nearly 5 million Americans. Yet, fibromyalgia is frequently underrecognized or misdiagnosed as arthritis, lupus, multiple sclerosis, autoimmune disease, etc. Hence, patients are often referred to multiple consultants and are subjected to multiple investigations and even surgical interventions. Early consideration of this very common disorder can avert needless investigation, therapy, and concern for those afflicted (Fig. 363-2).

TABLE 363-2 Drug-Induced Musculoskeletal Conditions

Arthralgias	Quinidine, cimetidine, beta blockers, quinolones, chronic acyclovir, interferons, IL-2, nicardipine, vaccines, rifabutin, aromatase inhibitors, HIV protease inhibitors, DPP-4 inhibitors (sitagliptin, linagliptin, alogliptin), checkpoint inhibitors (ipilimumab, pembrolizumab, nivolumab)
Myalgias/myopathy	Glucocorticoids, penicillamine, hydroxychloroquine, AZT, lovastatin, simvastatin, atorvastatin, pravastatin, clofibrate, amiodarone, interferon, IL-2, alcohol, cocaine, paclitaxel, docetaxel, imatinib mesylate, colchicine, quinolones, cyclosporine, tacrolimus, protease inhibitors, checkpoint inhibitors
Tendon rupture/tendinitis	Quinolones, glucocorticoids, isotretinoin, statins, aromatase inhibitors, collagenase injections
Gout	Diuretics, aspirin, cytotoxics, cyclosporine, alcohol, moonshine, ethambutol, fructose-containing soft drinks
Drug-induced lupus	Hydralazine, procainamide, quinidine, phenytoin, carbamazepine, methyldopa, isoniazid, chlorpromazine, lithium, penicillamine, tetracyclines, TNF inhibitors, ACE inhibitors, ticlopidine, aromatase inhibitors
Drug-induced subacute lupus	Proton pump inhibitors, calcium channel blockers (diltiazem), ACE inhibitors, TNF inhibitors, terbinafine, interferons (α and β-1a), paclitaxel, docetaxel, gemcitabine, capecitabine, aromatase inhibitors, HCTZ
Osteonecrosis/Atypical fractures	Glucocorticoids, alcohol, radiation, bisphosphonates
Osteopenia	Glucocorticoids, chronic heparin, phenytoin, aromatase inhibitors, anti-androgen therapy, thiazolinediones
Psoriasis	TNF inhibitors, beta blockers, lithium, hydroxychloroquine, chloroquine, minocycline, ACE inhibitors, terbinafine
Scleroderma	Vinyl chloride, bleomycin, baricitinib, pentazocine, organic solvents, carbidopa, tryptophan, rapeseed oil
Raynaud's phenomenon	Cisplatin, bleomycin, beta blockers, clonidine, bromocriptine, ergot alkaloids, cocaine, methylphenidate, dextroamphetamine, phenteramine, interferon therapy
Vasculitis	Allopurinol, amphetamines, cocaine (often levamisole adulterated), cannabis, thiazides, penicillamine, propylthiouracil, montelukast, TNF inhibitors, hepatitis B vaccine, trimethoprim/sulfamethoxazole, minocycline, hydralazine

Abbreviations: ACE, angiotensin-converting enzyme; AZT, zidovudine; HCTZ, hydrochlorothiazide; IL-2, interleukin 2; TNF, tumor necrosis factor.

RHEUMATOLOGIC EVALUATION OF THE ELDERLY

The incidence of rheumatic diseases rises with age, such that 58% of those >65 years will have joint complaints. Musculoskeletal disorders in elderly patients are often not diagnosed because the signs and symptoms may be insidious, overlooked, or overshadowed by comorbidities. These difficulties are compounded by the diminished reliability of laboratory testing in the elderly, who often manifest nonpathologic abnormal results. For example, the ESR may be misleadingly elevated, and low-titer positive tests for rheumatoid factor (RF) and antinuclear antibodies (ANAs) may be seen in up to 15% of elderly patients. Although nearly all rheumatic disorders afflict the elderly, geriatric patients are particularly prone to OA, osteoporosis, osteoporotic fractures, gout, pseudogout, polymyalgia rheumatica, vasculitis, and drug-induced disorders (Table 363-2). The elderly should be approached in the same manner as other patients with musculoskeletal complaints, but with an emphasis on identifying the potential rheumatic consequences of medical comorbidities and therapies. The physical examination should identify the nature of the musculoskeletal complaint as well as coexisting diseases that may influence diagnosis and choice of treatment.

RHEUMATOLOGIC EVALUATION OF THE HOSPITALIZED PATIENT

Evaluation of a hospitalized patient with rheumatic complaints is often more complex owing to greater symptom severity, more acute presentations, and greater interplay of comorbidities. In patients with rheumatic disorders tend to be admitted for one of several reasons: (1) acute onset of inflammatory arthritis (possibly gout or septic arthritis); (2) undiagnosed systemic or febrile illness; (3) musculoskeletal trauma; (4) exacerbation or deterioration of an existing autoimmune disorder (e.g., SLE); or (5) new medical comorbidities (e.g., thrombotic event, lymphoma, infection) arising in patients with an established rheumatic disorder. Notably, rheumatic patients are seldom if ever admitted because of widespread pain or serologic abnormalities or for the initiation of new therapies.

Acute monarticular inflammatory arthritis may be a "red flag" condition (e.g., septic arthritis, gout, pseudogout) that will require arthrocentesis and, on occasion, hospitalization if infection is suspected. However, new-onset inflammatory polyarthritis will have a wider differential diagnosis (e.g., RA, hepatitis-related arthritis, chikungunya arthritis, serum sickness, drug-induced lupus, polyarticular septic arthritis) and may require targeted laboratory investigations rather than synovial fluid analyses. Patients with febrile, multisystem disorders will require exclusion of crystal, infectious, or neoplastic etiologies and an evaluation driven by the dominant symptom/finding with the greatest specificity. Conditions worthy of consideration may include gout or pseudogout, vasculitis (giant cell arteritis in the elderly or polyarteritis nodosa in younger patients), adult-onset Still's disease, SLE, antiphospholipid antibody syndrome, IgG4-related disease, and sarcoidosis. A preexisting rheumatic diagnosis (e.g., SLE, RA, ankylosing spondylitis) should be confirmed by careful history, examination and review of medical records, as this will influence the ensuing in-patient evaluation. It is important to note that when established rheumatic disease patients are admitted to the hospital, it is usually not for a medical problem related to their autoimmune disease, but rather because of either a comorbid condition or complication of drug therapy. Patients with chronic inflammatory disorders (e.g., RA, SLE, psoriasis) have an augmented risk of infection, cardiovascular events, and neoplasia.

Certain conditions, such as acute gout, can be precipitated in hospitalized patients by surgery, dehydration, or medications and should be considered when hospitalized patients are evaluated for the acute onset of a musculoskeletal condition. Lastly, overly aggressive and unfocused laboratory testing will often yield abnormal findings that are better explained by the patient's preexisting condition (chronic lung, renal, or liver disease) rather than a new inflammatory or autoimmune disorder (lupus, vasculitis).

PHYSICAL EXAMINATION

The goal of the physical examination is to ascertain the structures involved, the nature of the underlying pathology, the functional consequences of the process, and the presence of systemic or extraarticular manifestations. A knowledge of topographic anatomy is necessary to identify the primary site(s) of involvement and differentiate articular from nonarticular disorders. The musculoskeletal examination depends largely on careful inspection, palpation, and a variety of specific physical maneuvers to elicit diagnostic signs **(Table 363-3)**. Although most articulations of the appendicular skeleton can be examined in this manner, adequate inspection and palpation are not possible for many axial (e.g., zygapophyseal) and inaccessible (e.g., sacroiliac or hip) joints. For such joints, there is a greater reliance on specific maneuvers and imaging for assessment.

Examination of involved and uninvolved joints will determine whether *pain, warmth, erythema*, or *swelling* is present. The locale and level of pain elicited by palpation or movement should be quantified. One standard would be to count the number of tender joints on palpation of 28 easily examined joints (proximal interphalangeals [PIPs], metacarpophalangeals [MCPs], wrists, elbows, shoulders, and knees). Similarly, the number of swollen joints (0–28) can be counted and recorded. Careful examination should distinguish between true articular swelling (caused by bony hypertrophy, synovial effusion or proliferation), and nonarticular (or periarticular) involvement,

TABLE 363-3 Glossary of Musculoskeletal Terms

Crepitus	A palpable (less commonly audible) vibratory or crackling sensation elicited with joint motion; fine joint crepitus is common and often insignificant in large joints; coarse joint crepitus indicates advanced cartilaginous and degenerative changes (as in osteoarthritis)
Subluxation	Alteration of joint alignment such that articulating surfaces incompletely approximate each other
Dislocation	Abnormal displacement of articulating surfaces such that the surfaces are not in contact
Range of motion	For diarthrodial joints, the arc of measurable movement through which the joint moves in a single plane
Contracture	Loss of full movement resulting from a fixed resistance caused either by tonic spasm of muscle (reversible) or by fibrosis of periarticular structures (permanent)
Deformity	Abnormal shape or size resulting from bony hypertrophy, malalignment of articulating structures, or damage to periarticular supportive structures
Enthesitis	Inflammation of the entheses (tendinous or ligamentous insertions on bone)
Epicondylitis	Infection or inflammation involving an epicondyle

which usually extends beyond the normal joint margins. Synovial effusion can be distinguished from synovial hypertrophy or bony hypertrophy by palpation or specific maneuvers. For example, small to moderate knee effusions may be identified by the "bulge sign" or "ballottement of the patellae." Bursal effusions (e.g., effusions of the olecranon or prepatellar bursa) are often focal, periarticular, overlie bony prominences, and are fluctuant with defined borders. Joint *stability* can be assessed by stabilizing the proximal joint, by palpation, and by the application of manual stress to the distal appendage. *Subluxation* or *dislocation*, which may be secondary to traumatic, mechanical, or inflammatory causes, can be assessed by inspection and palpation. Joint *swelling* or *volume* can be assessed by palpation. Distention of the articular capsule usually causes pain and evident enlargement or fluctuance. The patient will attempt to minimize the pain by maintaining the joint in the position of least intraarticular pressure and greatest volume, usually partial flexion. For this reason, inflammatory effusions may give rise to flexion contractures. Clinically, this may be detected as fluctuant or "squishy" swelling in larger joints and grapelike compressibility in smaller joints. Inflammation may result in fixed flexion deformities or diminished range of motion—especially on extension, when intraarticular pressure is increased. Active and passive *range of motion* should be assessed in all planes, with contralateral comparison. A goniometer may be used to quantify the arc of movement. Each joint should be passively manipulated through its full range of motion (including, as appropriate, flexion, extension, rotation, abduction, adduction, lateral bending, inversion, eversion, supination, pronation, medial/lateral deviation, and plantar- or dorsiflexion). Extreme range of motion may be seen with hypermobility syndrome, with joint pain and connective tissue laxity, often associated with Ehlers-Danlos or Marfan's syndrome. Limitation of motion is frequently caused by inflammation, effusion, pain, deformity, contracture, or restriction from neuromyopathic causes. If passive motion exceeds active motion, a periarticular process (e.g., tendinitis, tendon rupture, or myopathy) should be considered. *Contractures* may reflect antecedent synovial inflammation or trauma. Minor joint *crepitus* is common during joint palpation and maneuvers but may indicate significant cartilage degeneration as it becomes coarser (e.g., OA). Joint *deformity* usually indicates a long-standing or aggressive pathologic process. Deformities may result from ligamentous destruction, soft tissue contracture, bony enlargement, ankylosis, erosive disease, subluxation, trauma, or loss of proprioception. Examination of the musculature will document strength, atrophy, pain, or spasm. Appendicular muscle weakness should be characterized as proximal or distal. Muscle strength should be assessed by observing the patient's performance (e.g., walking, rising from a chair, grasping, writing). Strength may also be graded on a 5-point scale: 0 for no movement; 1 for trace movement or twitch; 2 for movement with gravity eliminated; 3 for movement against gravity only; 4 for movement against gravity and resistance; and 5 for normal strength. The examiner should assess for often-overlooked nonarticular or periarticular involvement, especially when articular complaints are not supported by objective findings referable to the joint capsule. The identification of soft tissue/nonarticular pain will prevent unwarranted and often expensive additional evaluations. Specific maneuvers may reveal common nonarticular abnormalities, such as a carpal tunnel syndrome (which can be identified by Tinel's or Phalen's sign). Other examples of soft tissue abnormalities include olecranon bursitis, epicondylitis (e.g., tennis elbow), enthesitis (e.g., Achilles tendinitis), and tender trigger points associated with fibromyalgia.

APPROACH TO REGIONAL RHEUMATIC COMPLAINTS

Although all patients should be evaluated in a logical and thorough manner, many cases with focal musculoskeletal complaints are caused by commonly encountered disorders that exhibit a predictable pattern of onset, evolution, and localization; they can often be diagnosed immediately on the basis of limited historic information and selected maneuvers or tests. Although nearly every musculoskeletal complaint could be approached in this manner, the evaluation of four common involved anatomic regions—the hand, shoulder, hip, and knee—are reviewed here.

HAND PAIN

Focal or unilateral hand pain may result from trauma, overuse, infection, or a reactive or crystal-induced arthritis. By contrast, bilateral hand complaints commonly suggest a degenerative (e.g., OA), systemic, or inflammatory/immune (e.g., RA) etiology. The distribution or pattern of joint involvement is highly suggestive of certain disorders **(Fig. 363-3)**. Thus, OA (or degenerative arthritis) may manifest as distal interphalangeal (DIP) and PIP joint pain with bony hypertrophy sufficient to produce Heberden's and Bouchard's nodes, respectively. Pain, with or without bony swelling, involving the base of the thumb (first carpometacarpal joint) is also highly suggestive of OA. By contrast, RA tends to cause symmetric, polyarticular involvement of the PIP, MCP, intercarpal, and carpometacarpal joints (wrist) with pain and palpable synovial tissue hypertrophy. Psoriatic arthritis may mimic the pattern of joint involvement seen in OA (DIP and PIP joints), but can be distinguished by the presence of inflammatory signs (erythema, warmth, synovial swelling), with or without carpal involvement, nail pitting, or onycholysis. Whereas lateral or medial subluxations at the PIP or DIP joints are most likely due to inflammatory OA or psoriatic arthritis, dorsal or ventral deformities (swan neck or boutonnière deformities) are typical of RA. Hemochromatosis should be considered when degenerative changes (bony hypertrophy) are seen at the second and third MCP joints with associated radiographic chondrocalcinosis or episodic, inflammatory wrist arthritis.

Dactylitis manifests as soft tissue swelling of the whole digit and may have a sausage-like appearance. Common causes of dactylitis include psoriatic arthritis, spondyloarthritis, juvenile spondylitis, mixed connective tissue disease, scleroderma, sarcoidosis, and sickle cell disease. Soft tissue swelling over the dorsum of the hand and wrist may suggest an inflammatory extensor tendon tenosynovitis possibly caused by gonococcal infection, gout, or inflammatory arthritis (e.g., RA). Tenosynovitis is suggested by localized warmth, swelling, or pitting edema and may be confirmed when the soft tissue swelling tracks with tendon movement during flexion and extension of fingers, or when pain is induced while stretching the extensor tendon sheaths (flexing the digits distal to the MCP joints and maintaining the wrist in a fixed, neutral position).

Focal wrist pain localized to the radial aspect may be caused by de Quervain's tenosynovitis resulting from inflammation of the tendon sheath(s) involving the abductor pollicis longus or extensor pollicis

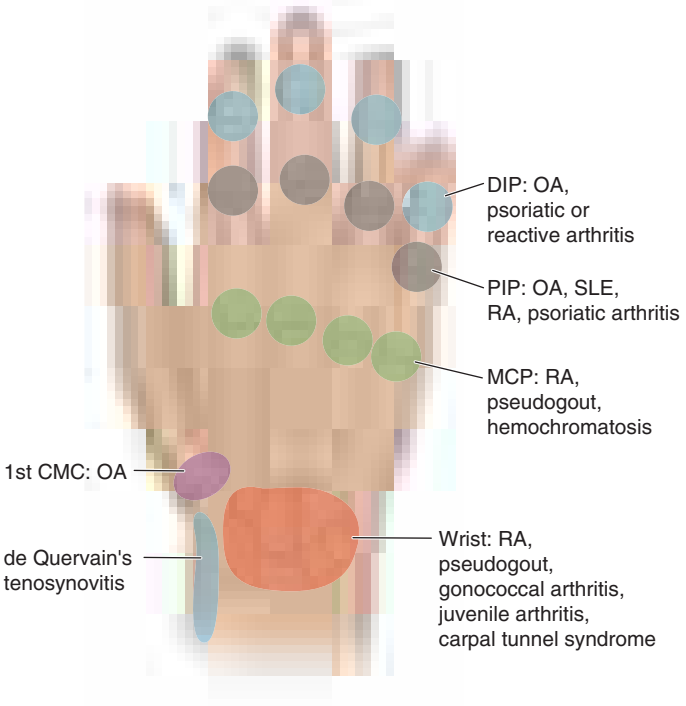


FIGURE 363-3 Sites of hand or wrist involvement and their potential disease associations. CMC, carpometacarpal; DIP, distal interphalangeal; MCP, metacarpophalangeal; OA, osteoarthritis; PIP, proximal interphalangeal; RA, rheumatoid arthritis; SLE, systemic lupus erythematosus. *(From JJ Cush et al: Evaluation of musculoskeletal complaints, in Rheumatology: Diagnosis and Therapeutics, 2nd ed, JJ Cush et al [eds]. Philadelphia, Lippincott Williams & Wilkins, 2005, pp 3–20. Used with permission from Dr. John J. Cush.)*

brevis (Fig. 363-3). This commonly results from overuse or follows pregnancy and may be diagnosed with Finkelstein's test. A positive result is present when radial wrist pain is induced after the thumb is flexed and placed inside a clenched fist and the patient actively deviates the hand downward with ulnar deviation at the wrist. Carpal tunnel syndrome is another common disorder of the upper extremity and results from compression of the median nerve within the carpal tunnel. Manifestations include pain in the wrist that may radiate with paresthesia to the thumb, second and third fingers, and radial half of the fourth finger and, at times, atrophy of thenar musculature. Carpal tunnel syndrome is commonly associated with pregnancy, edema, trauma, OA, inflammatory arthritis, and infiltrative disorders (e.g., amyloidosis). The diagnosis may be suggested by a positive Tinel's or Phalen's sign. With each test, paresthesia in a median nerve distribution is induced or increased by either "thumping" the volar aspect of the wrist (Tinel's sign) or pressing the extensor surfaces of both flexed wrists against each other (Phalen's sign). The low sensitivity and moderate specificity of these tests may require nerve conduction velocity testing to confirm a suspected diagnosis.

SHOULDER PAIN

During the evaluation of shoulder disorders, the examiner should carefully note any history of trauma, fibromyalgia, infection, inflammatory disease, occupational hazards, or previous cervical disease. In addition, the patient should be questioned as to the activities or movement(s) that elicit shoulder pain. While arthritis is suggested by pain on movement in all planes, pain with specific active motion suggests a periarticular (nonarticular) process. Shoulder pain may originate in the glenohumeral or acromioclavicular joints, subacromial (subdeltoid) bursa, periarticular soft tissues (e.g., fibromyalgia, rotator cuff tear/tendinitis), or cervical spine **(Fig. 363-4)**. Shoulder pain is referred frequently from the cervical spine but may also be referred from intrathoracic lesions (e.g., a Pancoast tumor) or from gallbladder, hepatic, or diaphragmatic disease. These same visceral causes may also manifest as focal scapular pain. Fibromyalgia should be suspected when glenohumeral pain is accompanied by diffuse periarticular (i.e., subacromial, bicipital) pain, tender points (i.e., trapezius or supraspinatus), and a sleep disturbance. The shoulder should be put through its full range of motion both actively and passively (with examiner assistance): forward flexion, extension, abduction, adduction, and internal and external rotation. Manual inspection of the periarticular structures will often provide important diagnostic information. Glenohumeral involvement is best detected by placing the thumb over the glenohumeral joint just medial and inferior to the coracoid process and applying pressure anteriorly while internally and externally rotating the humeral head. Pain localized to this region is indicative of glenohumeral pathology. Synovial effusion or tissue is seldom palpable but, if present, may suggest infection, RA, amyloidosis, or an acute tear of the rotator cuff. The examiner should apply direct manual pressure over the subacromial bursa that lies lateral to and immediately beneath the acromion (Fig. 363-4). Subacromial bursitis is a frequent cause of shoulder pain. Anterior to the subacromial bursa, the bicipital tendon traverses the bicipital groove. This tendon is best identified by palpating it in its groove as the patient rotates the humerus internally and externally. Direct pressure over the tendon may reveal pain indicative of bicipital tendinitis. Palpation of the acromioclavicular joint may disclose local pain, bony hypertrophy, or, uncommonly, synovial swelling. Whereas OA and RA commonly affect the acromioclavicular joint, OA seldom involves the glenohumeral joint, unless there is a traumatic or occupational cause.

Rotator cuff tendinitis or tear is a very common cause of shoulder pain. Nearly 30% of the elderly will have shoulder pain, with rotator cuff tendinitis or tear as a primary cause. The rotator cuff is formed by four tendons that attach the scapula to the proximal humerus (supraspinatus, infraspinatus, teres minor, and subscapularis tendons). Of these, the supraspinatus muscle is the most commonly damaged. Rotator cuff tendinitis is suggested by pain on active abduction (but not passive abduction), pain over the lateral deltoid muscle, night pain, and evidence of the impingement signs (pain with overhead arm activities). The Neer test for impingement is performed by the

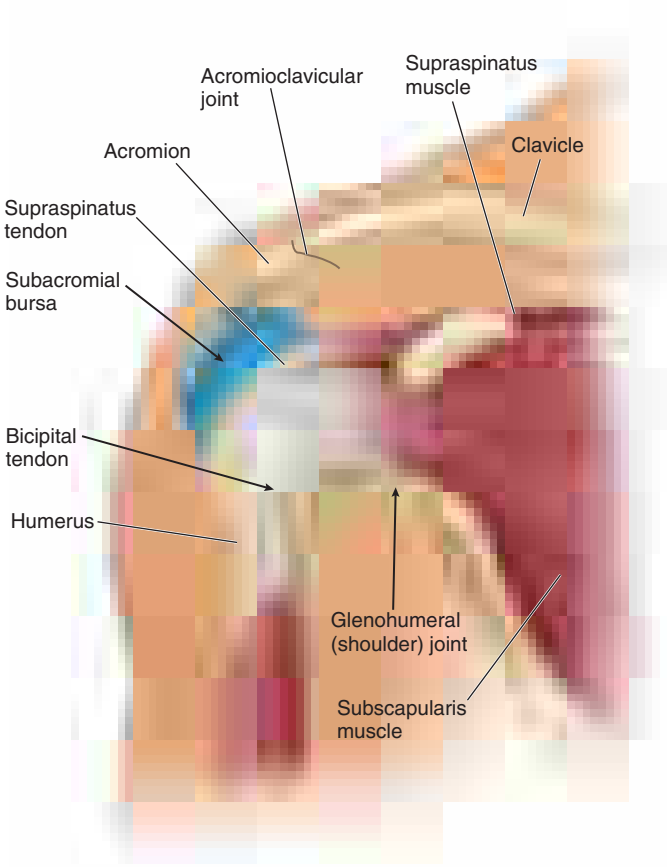


FIGURE 363-4 Origins of shoulder pain. The schematic diagram of the shoulder indicates, with arrows, the anatomic origins of shoulder pain.

examiner raising the patient's arm into forced flexion while stabilizing and preventing rotation of the scapula. A positive sign is present if pain develops before 180° of forward flexion. Tear of the rotator cuff is common in the elderly and often results from trauma; it may manifest in the same manner as tendinitis. The drop arm test is abnormal with supraspinatus pathology and is demonstrated by passive abduction of the arm to 90° by the examiner. If the patient is unable to hold the arm up actively or unable to lower the arm slowly without dropping, the test is positive. Tendinitis or tear of the rotator cuff is best confirmed by magnetic resonance imaging (MRI) or ultrasound.

KNEE PAIN

Knee pain may result from intraarticular (OA, RA) or periarticular (anserine bursitis, collateral ligament strain) processes or be referred from hip pathology. A careful history should delineate the chronology of the knee complaint and whether there are predisposing conditions, trauma, or medications that might underlie the complaint. For example, patellofemoral disease (e.g., OA) may cause anterior knee pain that worsens with climbing stairs. Observation of the patient's gait is also important. The knee should be carefully inspected in the upright (weight-bearing) and supine positions for swelling, erythema, malalignment, visible trauma, muscle wasting, and leg length discrepancy. The most common malalignment in the knee is *genu varum* (bowlegs) or *genu valgum* (knock-knees) resulting from asymmetric cartilage loss medially or laterally. Bony swelling of the knee joint commonly results from hypertrophic osseous changes seen with disorders such as OA and neuropathic arthropathy. Swelling caused by hypertrophy of the synovium or synovial effusion may manifest as a fluctuant, ballotable, or soft tissue enlargement in the suprapatellar pouch (suprapatellar reflection of the synovial cavity) or regions lateral and medial to the patella. Synovial effusions may also be detected by balloting the patella downward toward the femoral groove or by eliciting a "bulge sign." With the knee extended, the examiner should manually compress, or "milk," synovial fluid down from the suprapatellar pouch and lateral to the patellae. The application of manual pressure lateral to the patella may cause an observable shift in synovial fluid (bulge) to the medial aspect. The examiner should note that this maneuver is only effective in detecting small to moderate effusions (<100 mL). Inflammatory disorders such as RA, gout, pseudogout, and psoriatic arthritis may involve the knee joint and produce significant pain, stiffness, swelling, or warmth. A popliteal or *Baker's cyst* may be palpated with the knee partially flexed and is best viewed posteriorly with the patient standing and knees fully extended to visualize isolated or unilateral popliteal swelling or fullness.

Anserine bursitis is an often missed periarticular cause of knee pain in adults. The pes anserine bursa underlies the insertion of the conjoined tendons (sartorius, gracilis, semitendinosus) on the anteromedial proximal tibia and may be painful following trauma, overuse, or inflammation. It is often tender in patients with fibromyalgia, obesity, and knee OA. Other forms of bursitis may also present as knee pain. The prepatellar bursa is superficial and is located over the inferior portion of the patella. The infrapatellar bursa is deeper and lies beneath the patellar ligament before its insertion on the tibial tubercle.

Internal derangement of the knee may result from trauma or degenerative processes. Damage to the meniscal cartilage (medial or lateral) frequently presents as chronic or intermittent knee pain. Such an injury should be suspected when there is a history of trauma, athletic activity, or chronic knee arthritis, and when the patient relates symptoms of "locking" or "giving way" of the knee. With the knee flexed 90° and the patient's foot on the table, pain elicited during palpation over the joint line or when the knee is stressed laterally or medially may suggest a meniscal tear. A positive McMurray test may also indicate a meniscal tear. To perform this test, the knee is first flexed at 90°, and the leg is then extended while the lower extremity is simultaneously torqued medially or laterally. A painful click during inward rotation may indicate a lateral meniscus tear, and pain during outward rotation may indicate a tear in the medial meniscus. Lastly, damage to the cruciate ligaments should be suspected with acute onset of pain, possibly with swelling, a history of trauma, or a synovial fluid aspirate that is grossly bloody. Examination of the cruciate ligaments is best accomplished by eliciting a drawer sign. With the patient recumbent, the knee should be partially flexed and the foot stabilized on the examining surface. The examiner should manually attempt to displace the tibia anteriorly or posteriorly with respect to the femur. If anterior movement is detected, then anterior cruciate ligament damage is likely. Conversely, significant posterior movement may indicate posterior cruciate damage. Contralateral comparison will assist the examiner in detecting significant anterior or posterior movement.

HIP PAIN

The hip is best evaluated by observing the patient's gait and assessing range of motion. The vast majority of patients reporting "hip pain" localize their pain unilaterally to the posterior gluteal musculature **(Fig. 363-5)**. Such pain tends to radiate down the posterolateral aspect of the thigh and may or may not be associated with complaints of low back pain. This presentation frequently results from degenerative arthritis of the lumbosacral spine or disks and commonly follows a dermatomal distribution with involvement of nerve roots between L4 and S1. Sciatica is caused by impingement of the L4, L5, or S1 nerve (i.e., from a herniated disk) and manifests as unilateral neuropathic pain extending from the gluteal region down the posterolateral leg to the foot. Some individuals instead localize their "hip pain" laterally to the area overlying the trochanteric bursa. Because of the depth of this bursa, swelling and warmth are usually absent. Diagnosis of trochanteric bursitis or enthesitis can be confirmed by inducing point tenderness over the trochanteric bursa. Gluteal and trochanteric pain are common findings in fibromyalgia. Range of movement may be limited by pain. Pain in the hip joint is less common and tends to be located anteriorly, over the inguinal ligament; it may radiate medially to the groin. Uncommonly, iliopsoas bursitis may mimic true hip joint pain. Diagnosis of iliopsoas bursitis may be suggested by a history of trauma or inflammatory arthritis. Pain associated with iliopsoas bursitis is localized to the groin or anterior thigh and tends to worsen with hyperextension of the hip; many patients prefer to flex and externally rotate the hip to reduce the pain from a distended bursa.

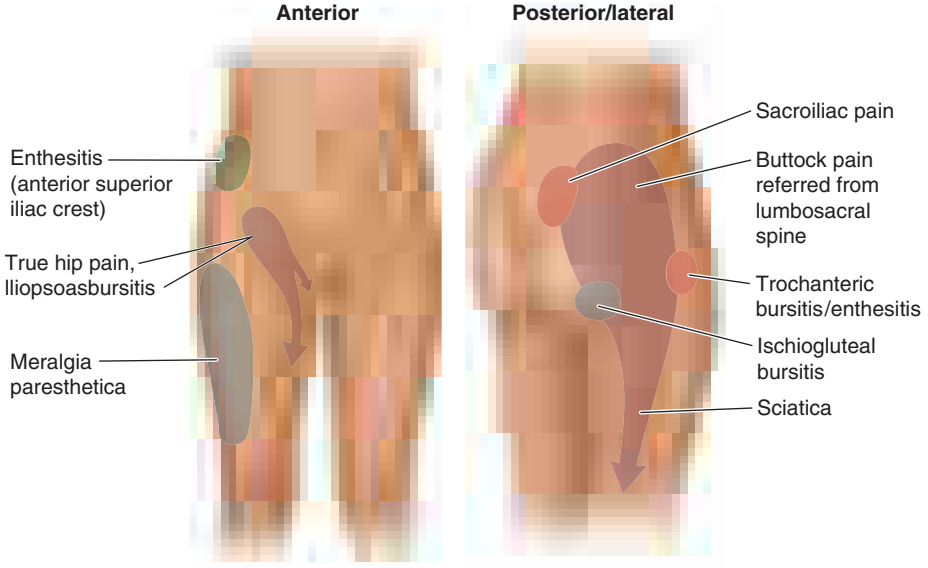


FIGURE 363-5 Origins of hip pain and dysesthesias. *(From JJ Cush et al: Evaluation of musculoskeletal complaints, in Rheumatology: Diagnosis and Therapeutics, 2nd ed, JJ Cush et al [eds]. Philadelphia, Lippincott Williams & Wilkins, 2005, pp 3–20. Used with permission from Dr. John J. Cush.)*

LABORATORY INVESTIGATIONS

The vast majority of musculoskeletal disorders can be logically diagnosed by a complete history and physical examination. An additional objective of the initial encounter is to determine whether additional investigations or immediate therapy is required. Additional evaluation is indicated with: (1) monarticular conditions; (2) traumatic or inflammatory conditions; (3) the presence of neurologic findings; (4) systemic manifestations; or (5) chronic symptoms (>6 weeks) and a lack of response to symptomatic measures. The extent and nature of the additional investigation should be dictated by the clinical features and suspected pathologic process. Laboratory tests should be used to confirm a specific clinical diagnosis and not be used to screen or evaluate patients with vague rheumatic complaints. Indiscriminate use of broad batteries of diagnostic tests and radiographic procedures is rarely a useful or cost-effective means to establish a diagnosis.

Besides a complete blood count, including a white blood cell (WBC) and differential count, the routine evaluation should include a determination of an acute-phase reactant such as the ESR or CRP, which can be useful in discriminating inflammatory from noninflammatory disorders. Both are inexpensive, easily obtained, and may be elevated with infection, inflammation, autoimmune disorders, neoplasia, pregnancy, renal insufficiency, advanced age, or hyperlipidemia. Extreme elevation of the acute-phase reactants (CRP, ESR) is seldom seen without evidence of serious illness (e.g., sepsis, pleuropericarditis, polymyalgia rheumatica, giant cell arteritis, adult Still's disease).

Serum uric acid determinations are useful in the diagnosis of gout and in monitoring the response to urate-lowering therapy. Uric acid, the end product of purine metabolism, is primarily excreted in the urine. Serum values range from 238 to 516 μmol/L (4.0–8.6 mg/dL) in men; the lower values (178–351 μmol/L [3.0–5.9 mg/dL]) seen in women are caused by the uricosuric effects of estrogen. Urinary uric acid levels are normally <750 mg per 24 h. Although hyperuricemia (especially levels >535 μmol/L [>9 mg/dL]) is associated with an increased incidence of gout and nephrolithiasis, levels may not correlate with the severity of articular disease. Uric acid levels (and the risk of gout) may be increased by inborn errors of metabolism (Lesch-Nyhan syndrome), disease states (renal insufficiency, myeloproliferative disease, psoriasis), or drugs (alcohol, cytotoxic therapy, thiazides). Although nearly all patients with gout will demonstrate hyperuricemia at some time during their illness, up to 50% of patients with an acute gouty attack will have normal serum uric acid levels. Monitoring serum uric acid is useful in assessing the response to urate-lowering therapy or chemotherapy, with the target goal being a serum urate <6 mg/dL.

Serologic tests for RF, cyclic anticitrullinated peptide (CCP or ACPA) antibodies, ANAs, complement levels, Lyme and antineutrophil cytoplasmic antibodies (ANCA), or antistreptolysin O (ASO) titer should be carried out only when there is clinical evidence to specifically suggest an associated diagnosis because these have poor predictive value when used for screening, especially when the pretest probability is low. For most of these, there is no value to repeated or serial serologic testing. Although 4–5% of a healthy population will have positive tests for RF and ANAs, only 1% and <0.4% of the population will have RA or SLE, respectively. IgM RF (autoantibodies against the Fc portion of IgG) is found in 80% of patients with RA and may also be seen in low titers in patients with chronic infections (tuberculosis, leprosy, hepatitis); other autoimmune diseases (SLE, Sjögren's syndrome); and chronic pulmonary, hepatic, or renal diseases. When considering RA, both serum RF and anti-CCP antibodies should be obtained as these are complementary. Both are comparably sensitive, but CCP antibodies are more specific than RF. In RA, the presence of anti-CCP and RF antibodies may indicate a greater risk for more severe, erosive polyarthritis. ANAs are found in nearly all patients with SLE and may also be seen in patients with other autoimmune diseases (polymyositis, scleroderma, antiphospholipid syndrome, Sjögren's syndrome), drug-induced lupus (Table 363-2), chronic liver or renal disorders, and advanced age. Positive ANAs are found in 5% of adults and in up to 14% of elderly or chronically ill individuals. The ANA test is very sensitive but poorly specific for lupus, as only 1–2% of all positive results will be caused by lupus alone. The interpretation of a positive ANA test may depend on the magnitude of the titer and the pattern observed by immunofluorescence microscopy **(Table 363-4)**. Diffuse and speckled patterns are least specific, whereas a peripheral, or rim, pattern (related to autoantibodies against double-strand [native] DNA) is highly specific and suggestive of lupus. Centromeric patterns are seen in patients with limited scleroderma (*c*alcinosis, *R*aynaud's phenomenon, *e*sophageal involvement, *s*clerodactyly, *t*elangiectasia [CREST] syndrome), primary biliary sclerosis, Sjögren's syndrome or thyroiditis and nucleolar patterns may be seen in patients with diffuse systemic sclerosis or inflammatory myositis.

TABLE 363-4 Antinuclear Antibody (ANA) Patterns and Clinical Associations

ANA PATTERN	ANTIGEN IDENTIFIED	CLINICAL CORRELATE
Diffuse	Deoxyribonucleoprotein	Nonspecific
	Histones	Drug-induced lupus, lupus
Peripheral (rim)	ds-DNA	50% of SLE (specific)
Speckled	U1-RNP	>90% of MCTD
	Sm	30% of SLE (specific)
	Ro (SS-A)	Sjögren's 60%, SCLE, neonatal lupus, ANA(–) lupus
	La (SS-B)	50% of Sjögren's, 15% lupus
	Scl-70	40% of diffuse scleroderma
	PM-1	Polymyositis (PM), dermatomyositis
	Jo-1	PM w/pneumonitis + arthritis
Nucleolar	RNA polymerase I, others	40% of PSS
Centromere	Kinetochore	75% CREST (limited scleroderma), PBC, Sjogren's, thyroiditis

Abbreviations: ANA, antinuclear antibody; CREST, *c*alcinosis, *R*aynaud phenomenon, *e*sophageal involvement, *s*clerodactyly, and *t*elangiectasia; MCTD, mixed connective tissue disease; PBC, primary biliary cirrhosis; PSS, progressive systemic sclerosis; SCLE, subacute cutaneous lupus erythematosus; SLE, systemic lupus erythematosus.

Aspiration and analysis of synovial fluid are always indicated in acute monarthritis or when an infectious or crystal-induced arthropathy is suspected. Synovial fluid may distinguish between noninflammatory and inflammatory processes by analysis of the appearance, viscosity, and cell count. Tests for synovial fluid glucose, protein, lactate dehydrogenase, lactic acid, or autoantibodies are not recommended because they have no diagnostic value. Normal synovial fluid is clear or a pale straw color and is viscous, primarily because of the high levels of hyaluronate. Noninflammatory synovial fluid is clear, viscous, and amber-colored, with a WBC count of <2000/μL and a predominance of mononuclear cells. The viscosity of synovial fluid is assessed by expressing fluid from the syringe one drop at a time. Normally, there is a stringing effect, with a long tail behind each synovial drop. Effusions caused by OA or trauma will have normal viscosity. Inflammatory fluid is turbid and yellow, with an increased WBC count (2000–50,000/μL) and a polymorphonuclear leukocyte predominance. Inflammatory fluid has reduced viscosity (no stringing), diminished hyaluronate, and little or no tail following each drop of synovial fluid. Such effusions are found in RA, gout, and other inflammatory arthritides. Septic fluid is opaque and purulent, with a WBC count usually >50,000/μL, a predominance of polymorphonuclear leukocytes (>75%), and low viscosity. Such effusions are typical of septic arthritis but may also occur with RA or gout. In addition, hemorrhagic synovial fluid may be seen with trauma, hemarthrosis, or neuropathic arthritis. An algorithm for synovial fluid aspiration and analysis is shown in **Fig. 363-6**. Synovial fluid should be analyzed immediately for appearance, viscosity, and cell count. Monosodium urate crystals (observed in gout) are seen by polarized microscopy and are long, needle-shaped, negatively birefringent, and usually intracellular. In chondrocalcinosis

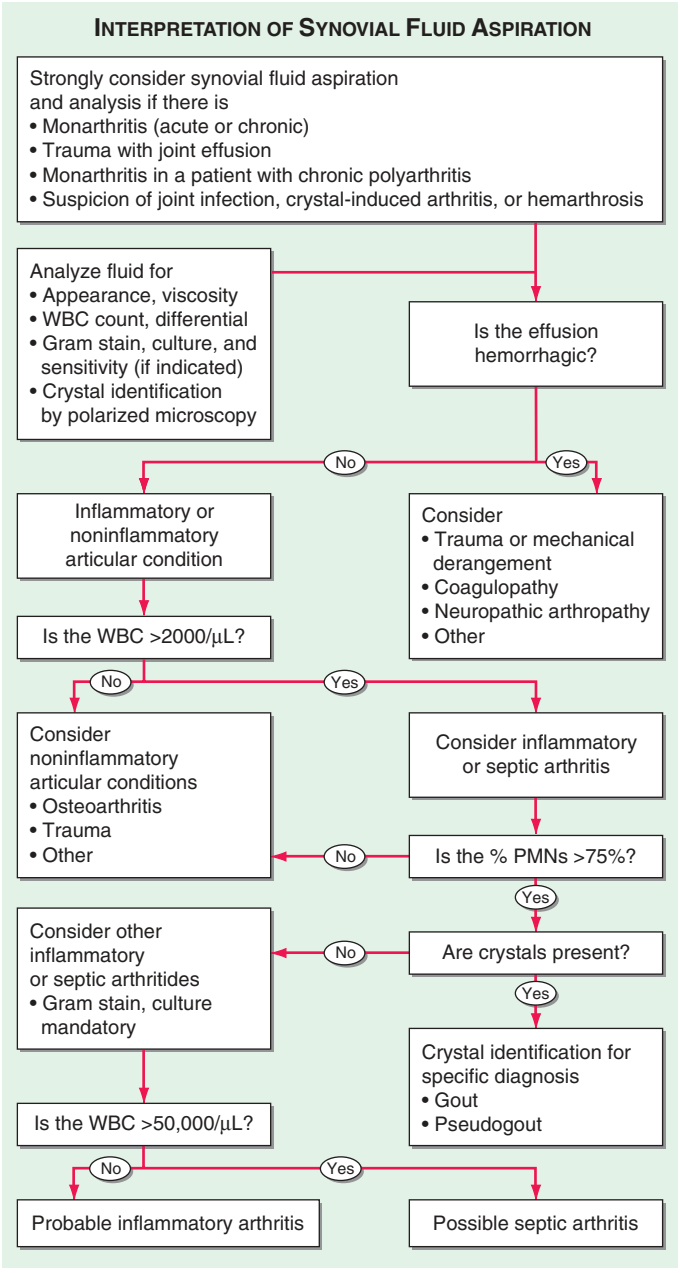


FIGURE 363-6 Algorithmic approach to the use and interpretation of synovial fluid aspiration and analysis. PMNs, polymorphonuclear (leukocytes); WBC, white blood cell count.

and pseudogout, calcium pyrophosphate dihydrate crystals are usually short, rhomboid-shaped, and positively birefringent. Whenever infection is suspected, synovial fluid should be Gram stained and cultured appropriately. If gonococcal arthritis is suspected, nucleic acid amplification tests should be used to detect either *Chlamydia trachomatis* or *N. gonorrhoeae* infection. Synovial fluid from patients with chronic monarthritis should also be cultured for *M. tuberculosis* and fungi. Last, it should be noted that crystal-induced arthritis and septic arthritis occasionally occur together in the same joint.

DIAGNOSTIC IMAGING IN JOINT DISEASES

Conventional radiography has been a valuable tool in the diagnosis and staging of articular disorders. Plain x-rays are most appropriate and cost effective when there is a history of trauma, suspected chronic infection, progressive disability, or monarticular involvement; when therapeutic alterations are considered; or when a baseline assessment is desired for what appears to be a chronic process. However, in acute inflammatory arthritis, early radiography is rarely helpful in establishing a diagnosis and may only reveal soft tissue swelling or juxtaarticular demineralization. As the disease progresses, calcification (of soft tissues, cartilage, or bone), joint space narrowing, erosions, bony ankylosis, new bone formation (sclerosis, osteophytes, or periostitis), or subchondral cysts may develop and suggest specific clinical entities. Consultation with a radiologist will help define the optimal imaging modality, technique, or positioning and prevent the need for further studies.

Additional imaging techniques may possess greater diagnostic sensitivity and facilitate early diagnosis in a limited number of articular disorders and in selected circumstances and are indicated when conventional radiography is inadequate or nondiagnostic **(Table 363-5)**. *Ultrasonography* is useful in the detection of soft tissue abnormalities,

TABLE 363-5 Diagnostic Imaging Techniques for Musculoskeletal Disorders

METHOD	IMAGING TIME, H	COST[a]	CURRENT INDICATIONS
Ultrasound	<1	++	Synovial (Baker's) cysts Rotator cuff tears Bursitis, tendinitis, tendon injury Enthesitis Carpal tunnel syndrome Urate or calcium pyrophosphate deposition on cartilage Early detection of synovial inflammation or erosions Ultrasound-guided injection/ arthrocentesis
Radionuclide scintigraphy			Metastatic bone survey Evaluation of Paget's disease
^{99m}Tc	1–4	++	Identifying occult arthritis in patients with undiagnosed polyarthralgia
^{111}In-WBC	24	+++	Acute infection Prosthetic infection Acute osteomyelitis
^{67}Ga	24–48	++++	Acute and chronic infection Acute osteomyelitis
Computed tomography (CT)	<1	+++	Herniated intervertebral disk Sacroiliitis Spinal stenosis Spinal trauma Osteoid osteoma Stress fracture
Dual-energy CT	<1	NA	Uric acid deposition Tophus localization
Magnetic resonance imaging	1/2–2	++++	Avascular necrosis Osteomyelitis Septic arthritis, infected prosthetic joints Early sacroiliitis Intraarticular derangement and soft tissue injury Derangements of axial skeleton and spinal cord Herniated intervertebral disk Pigmented villonodular synovitis Inflammatory and metabolic muscle pathology

[a]Relative cost for imaging study.

Abbreviations: NA, not commercially available; WBC, white blood cell.

such as tendinitis, tenosynovitis, enthesitis, bursitis, and entrapment neuropathies. Wider use, lower cost, better technology, and enhanced site-specific transducers now allow for wider use in outpatient care, especially for the evaluation of synovial (Baker's) cysts, rotator cuff tears, tendinitis and tendon injury, and crystal deposition on cartilage. Use of power Doppler allows for early detection of synovitis and bony erosions. *Radionuclide scintigraphy* is a very sensitive, but poorly specific, means of detecting inflammatory or metabolic alterations in bone or periarticular soft tissue structures (Table 363-5). Scintigraphy is best suited for total-body assessment (extent and distribution) of skeletal involvement (neoplasia, Paget's disease) and the assessment of patients with undiagnosed polyarthralgias, looking for occult arthritis. The use of scintigraphy has declined with greater use and declining cost of ultrasound and MRI. MRI has largely replaced scintigraphy in diagnosing osseous infection, neoplasia, inflammation, increased blood flow, bone remodeling, heterotopic bone formation, or avascular necrosis. Gallium scanning uses ^{67}Ga, which binds serum and cellular transferrin and lactoferrin and is preferentially taken up by neutrophils, macrophages, bacteria, and tumor tissue (e.g., lymphoma). As such, it is primarily used in the identification of occult infection or malignancy. Scanning with ^{111}In-labeled WBCs has been used to detect osteomyelitis and infectious or inflammatory arthritis. Despite their utility, ^{111}In-labeled WBC or ^{67}Ga scanning has largely been replaced by MRI, except when there is a suspicion of septic joint or prosthetic joint infections.

Computed tomography (*CT*) provides detailed visualization of the axial skeleton. Articulations previously considered difficult to visualize by radiography (e.g., zygapophyseal, sacroiliac, sternoclavicular, hip joints) can be effectively evaluated using CT. CT has been demonstrated to be useful in the diagnosis of low back pain syndromes (e.g., spinal stenosis vs herniated disk), sacroiliitis, osteoid osteoma, and stress fractures. Helical or spiral CT (with or without contrast angiography) is a novel technique that is rapid, cost effective, and sensitive in diagnosing pulmonary embolism or obscure fractures, often in the setting of initially equivocal findings. High-resolution CT can be advocated in the evaluation of suspected or established infiltrative lung disease (e.g., scleroderma or rheumatoid lung). The recent use of hybrid (positron emission tomography [PET] or single-photon emission CT [SPECT]) CT scans in metastatic evaluations has incorporated CT to provide better anatomic localization of scintigraphic abnormalities.

^{18}F-Fluorodeoxyglucose (FDG) is the most commonly used radiopharmaceutical in PET scanning. FDG-PET/CT scans have been seldom used in the evaluation of septic or inflammatory arthritis, but have also been useful in the evaluation of patients with fever of unknown origin or suspected large vessel vasculitis. Dual-energy CT (DECT) scanning, developed in urology to identify urinary calculi, has been a highly sensitive and specific method used to identify and quantify uric acid deposition in tissues **(Fig. 363-7)**.

MRI has significantly advanced the ability to image musculoskeletal structures. MRI has the advantages of providing multiplanar images with fine anatomic detail and contrast resolution **(Fig. 363-8)** that allows for the superior ability to visualize bone marrow and soft tissue periarticular structures. Although more costly with a longer procedural time than CT, the MRI has become the preferred technique when evaluating complex musculoskeletal disorders.

MRI can image fascia, vessels, nerve, muscle, cartilage, ligaments, tendons, pannus, synovial effusions, and bone marrow. Visualization of particular structures can be enhanced by altering the pulse sequence to produce either T1- or T2-weighted spin echo, gradient echo, or inversion recovery (including short tau inversion recovery [STIR]) images. Because of its sensitivity to changes in marrow fat, MRI is a sensitive but nonspecific means of detecting osteonecrosis, osteomyelitis, and marrow inflammation indicating overlying synovitis or osteitis (Fig. 363-8). Because of its enhanced soft tissue resolution, MRI is more sensitive than arthrography or CT in the diagnosis of soft tissue injuries (e.g., meniscal and rotator cuff tears); intraarticular derangements; marrow abnormalities (osteonecrosis, myeloma); and spinal cord or nerve root damage, synovitis, or cartilage damage or loss.

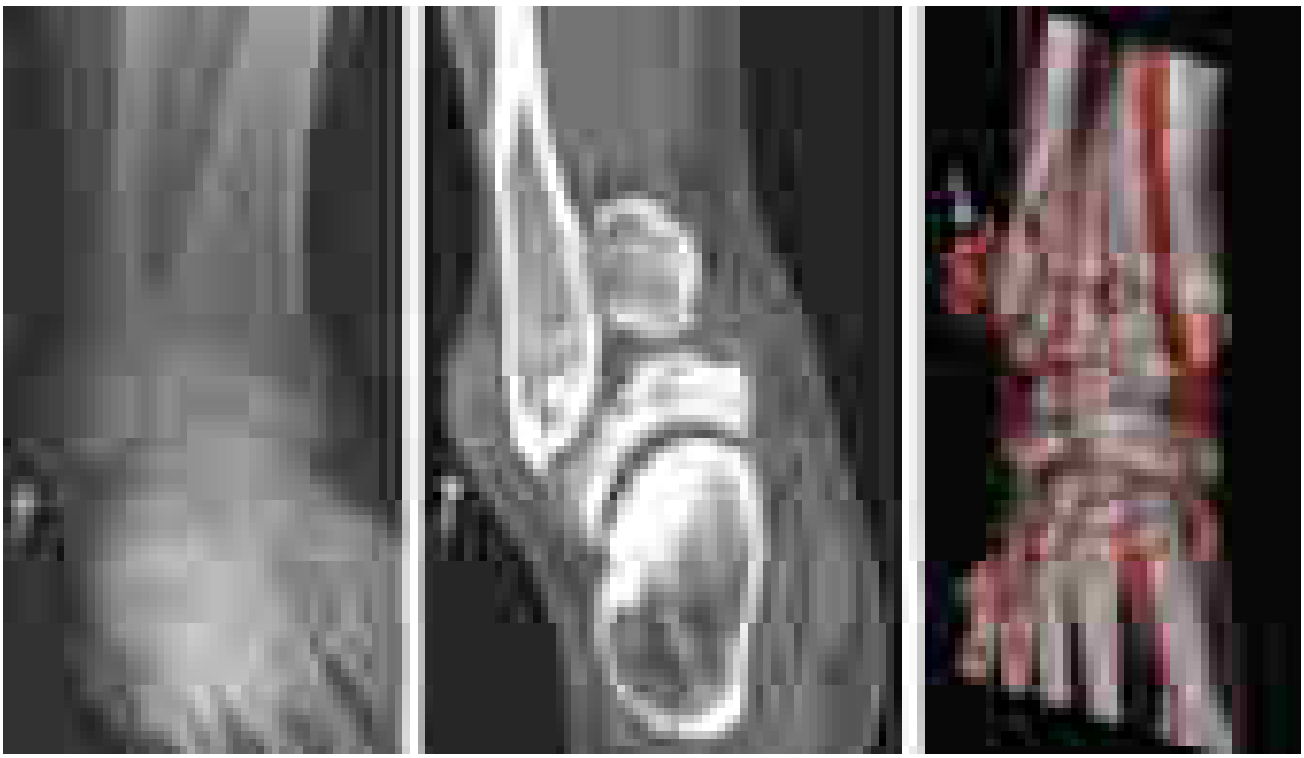


FIGURE 363-7 Dual-energy computed tomography (DECT) scan from a 45-year-old woman with right ankle swelling around the lateral malleolus. Three-dimensional volume-rendered coronal reformatted DECT image shows that the mass is composed of monosodium urate (*red*) in keeping with tophus (*arrow*). *(Used with permission from S Nicolaou et al: AJR 194:1072, 2010.)*

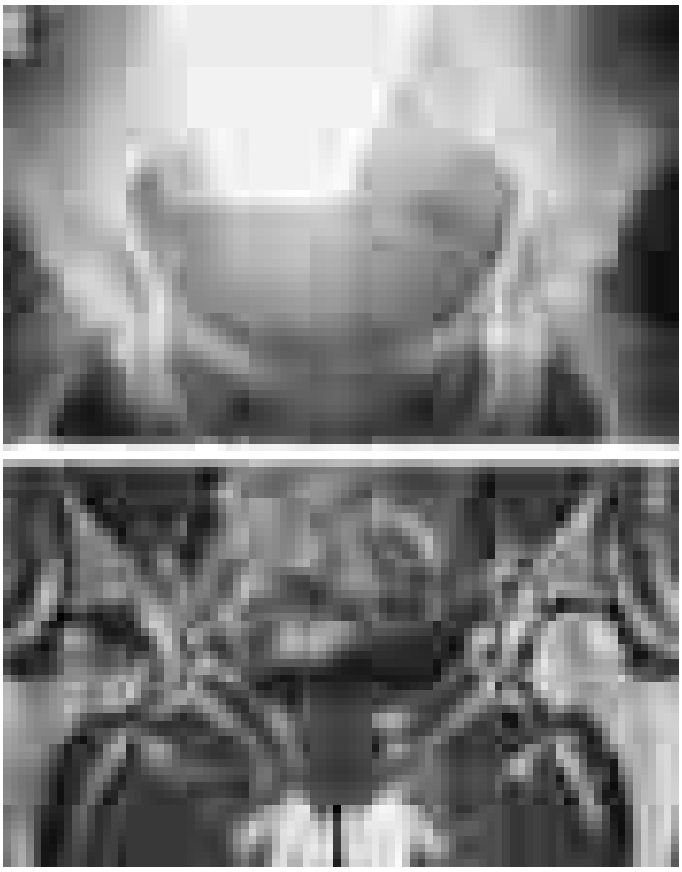


FIGURE 363-8 Superior sensitivity of magnetic resonance imaging (MRI) in the diagnosis of osteonecrosis of the femoral head. A 45-year-old woman receiving high-dose glucocorticoids developed right hip pain. Conventional x-rays (*top*) demonstrated only mild sclerosis of the right femoral head. T1-weighted MRI (*bottom*) demonstrated low-density signal in the right femoral head, diagnostic of osteonecrosis.

ACKNOWLEDGMENT

The author acknowledges the insightful contributions of Dr. Peter E. Lipsky to this chapter in previous editions.

FURTHER READING

BHAGAT M et al: Utility of immunologic testing in suspected rheumatologic disease. Curr Allergy Asthma Rep 14:405, 2014.

COZZANI E, DROSERA M: Serology of lupus erythematosus: Correlation between immunopathological features and clinical aspects. Autoimmune Dis 2014:321359, 2014.

CUSH JJ et al: Evaluation of musculoskeletal complaints. Available from *http://www.rheumaknowledgy.com/evaluation-of-musculoskeletal-complaints*. Accessed April 6, 2017.

DEFRODA SF et al: Differentiating radiculopathy from lower extremity arthropathy. Am J Med 129:1124. e1, 2016.

RUDWALEIT M et al: How to diagnose axial spondyloarthritis early. Ann Rheum Dis 63:535, 2004.

SIMPFENDORFER CS: Radiologic approach to musculoskeletal infections. Infect Dis Clin North Am 31:299, 2017.

364 Osteoarthritis

David T. Felson, Tuhina Neogi

Osteoarthritis (OA) is the most common type of arthritis. Its high prevalence, especially in the elderly, and its negative impact on physical function make it a leading cause of disability in the elderly. Because of the aging of Western populations and because obesity, a major risk factor, is increasing in prevalence, the occurrence of OA is on the rise.

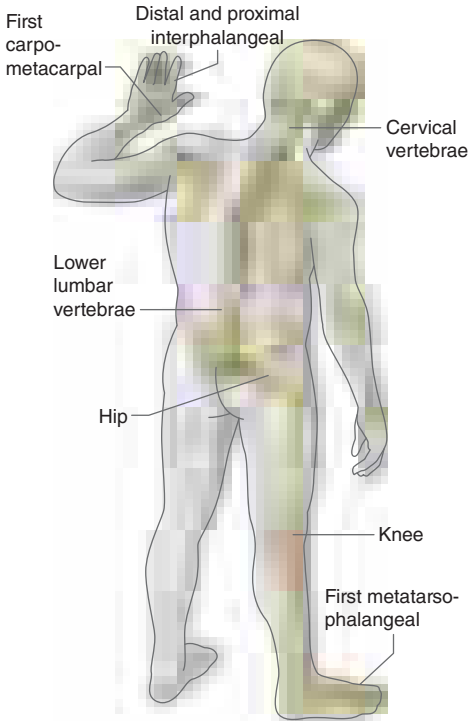


FIGURE 364-1 Joints commonly affected by osteoarthritis.

OA affects certain joints, yet spares others (Fig. 364-1). Commonly affected joints include the hip, knee, and first metatarsal phalangeal joint (MTP) and cervical and lumbosacral spine. In the hands, the distal and proximal interphalangeal joints and the base of the thumb are often affected. Usually spared are the wrist, elbow, and ankle. Our joints were designed, in an evolutionary sense, for brachiating apes, animals that still walked on four limbs. We thus develop OA in joints that were ill designed for human tasks such as pincer grip (OA in the thumb base) and walking upright (OA in knees and hips). Some joints, like the ankles, may be spared because their articular cartilage may be uniquely resistant to loading stresses.

OA can be diagnosed based on structural abnormalities or on the symptoms these abnormalities evoke. According to cadaveric studies, by elderly years, structural changes of OA are nearly universal. These include cartilage loss (seen as joint space loss on x-rays) and osteophytes. Many persons with x-ray evidence of OA have no joint symptoms, and although the prevalence of structural abnormalities is of interest in understanding disease pathogenesis, what matters more from a clinical perspective is the prevalence of symptomatic OA. Symptoms, usually joint pain, determine disability, visits to clinicians, and disease costs.

Symptomatic OA of the knee (pain on most days of a recent month plus x-ray evidence of OA in that knee) occurs in ~12% of persons age ≥60 in the United States and 6% of all adults age ≥30. Symptomatic hip OA is roughly one-third as common as disease in the knee. Although radiographic hand OA and the appearance of bony enlargement in affected hand joints (Fig. 364-2) are extremely common in older persons, most cases are often not symptomatic. Even so, symptomatic hand OA occurs in ~10% of elderly individuals and often produces measurable limitation in function.

The prevalence of OA rises strikingly with age, being uncommon in adults aged <40 and highly prevalent in those aged >60. It is also a disease that, at least in middle-aged and elderly persons, is much more common in women than in men.

X-ray evidence of OA is common in the lower back and neck, but back pain and neck pain have not been tied to findings of OA on x-ray. Thus, back pain and neck pain are treated separately (Chap. 14).

DEFINITION

OA is joint failure, a disease in which all structures of the joint have undergone pathologic change, often in concert. The pathologic sine qua non of disease is hyaline articular cartilage loss, present in a focal

FIGURE 364-2 Severe osteoarthritis of the hands affecting the distal interphalangeal joints (Heberden's nodes) and the proximal interphalangeal joints (Bouchard's nodes). There is no clear bony enlargement of the other common site in the hands, the thumb base.

and, initially, nonuniform manner. This is accompanied by increasing thickness and sclerosis of the subchondral bony plate, by outgrowth of osteophytes at the joint margin, by stretching of the articular capsule, by variable degrees of synovitis, and by weakness of muscles bridging the joint. In knees, meniscal degeneration is part of the disease. There are numerous pathways that lead to joint failure, but the initial step is often joint injury in the setting of a failure of protective mechanisms.

JOINT PROTECTIVE MECHANISMS AND THEIR FAILURE

Joint protectors include joint capsule and ligaments, muscle, sensory afferents, and underlying bone. Joint capsule and ligaments serve as joint protectors by providing a limit to excursion, thereby fixing the range of joint motion.

Synovial fluid reduces friction between articulating cartilage surfaces, thereby serving as a protector against friction-induced cartilage wear. This lubrication function depends on *hyaluronic acid* and on *lubricin*, a mucinous glycoprotein secreted by synovial fibroblasts whose concentration diminishes after joint injury and in the face of synovial inflammation.

The ligaments, along with overlying skin and tendons, contain mechanoreceptor sensory nerves. These mechanoreceptors fire at different frequencies throughout a joint's range of motion, providing feedback by way of the spinal cord to muscles and tendons. As a consequence, these muscles and tendons can assume the right tension at appropriate points in joint excursion to act as optimal joint protectors, anticipating joint loading.

Muscles and tendons that bridge the joint are key joint protectors. Focal stress across the joint is minimized by muscle contraction that decelerates the joint before impact and assures that when joint impact arrives, it is distributed broadly across the joint surface.

Failure of these joint protectors increases the risk of joint injury and OA. For example, in animals, OA develops rapidly when a sensory nerve to the joint is sectioned and joint injury induced. Similarly, in humans, Charcot's arthropathy, a severe and rapidly progressive OA, develops when minor joint injury occurs in the presence of posterior column peripheral neuropathy. Another example of joint protector failure is rupture of ligaments, a well-known cause of the early development of OA.

CARTILAGE AND ITS ROLE IN JOINT FAILURE

In addition to being a primary target tissue for disease, cartilage also functions as a joint protector. A thin rim of tissue at the ends of two opposing bones, cartilage is lubricated by synovial fluid to provide an almost frictionless surface across which these two bones move. The compressible stiffness of cartilage compared to bone provides the joint with impact-absorbing capacity.

The earliest changes of OA may occur in cartilage, and abnormalities there can accelerate disease development. The two major macromolecules in cartilage are type 2 collagen, which provides cartilage its tensile strength, and aggrecan, a proteoglycan macromolecule linked with hyaluronic acid, which consists of highly negatively charged glycosaminoglycans. In normal cartilage, type 2 collagen is woven tightly, constraining the aggrecan molecules in the interstices between collagen strands, forcing these highly negatively charged molecules into close proximity with one another. The aggrecan molecule, through electrostatic repulsion of its negative charges, gives cartilage its compressive stiffness. Chondrocytes, the cells within this avascular tissue, synthesize all elements of the matrix and produce enzymes that break down the matrix. Synovium and chondrocytes synthesize and release cytokines and growth factors, which provide feedback that modulates synthesis of matrix molecules **(Fig. 364-3)**. Cartilage matrix synthesis and catabolism are in a dynamic equilibrium influenced by the cytokine and growth factor environment. Mechanical and osmotic stress on chondrocytes induces these cells to alter gene expression and increase production of inflammatory cytokines and matrix-degrading enzymes.

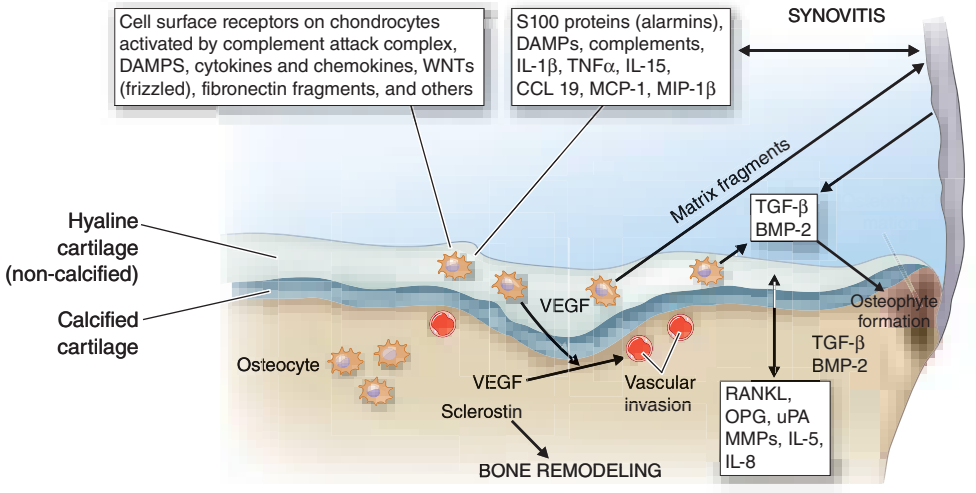


FIGURE 364-3 Selected factors involved in the osteoarthritic process including chondrocytes, bone, and synovium. Synovitis causes release of cytokines, alarmins, damage-associated molecular pattern (DAMP) molecules, and complement, which activate chondrocytes through cell surface receptors. Chondrocytes produce matrix molecules (collagen type 2, aggrecan) and the enzymes responsible for the degradation of the matrix (e.g., ADAMTS-5 and matrix metalloproteinases [MMPs]). Bone invasion occurs through the calcified cartilage, triggered by vascular endothelial growth factor (VEGF) and other molecules. IL, interleukin; TGF, transforming growth factor; TNF, tumor necrosis factor. *(From RF Loeser et al: Arthritis Rheum 64:1697, 2012.)*

While chondrocytes synthesize numerous enzymes, matrix metalloproteinases (MMP) (especially collagenases and ADAMTS-5) are critical enzymes in the breakdown of cartilage matrix. Both collagenase and aggrecanases act primarily in the territorial matrix surrounding chondrocytes; however, as the osteoarthritic process develops, their activities and effects spread throughout the matrix, especially in the superficial layers of cartilage.

The synovium, cartilage, and bone all influence disease development through cytokines, chemokines, and even complement activation (Fig. 364-3). These act on chondrocyte cell surface receptors and ultimately have transcriptional effects. Matrix fragments released from cartilage stimulate synovitis. Inflammatory cytokines such as interleukin 1β (IL-1β) and tumor necrosis factor α (TNF-α) induce chondrocytes to synthesize prostaglandin E_2 and nitric oxide, which have complex effects on matrix synthesis and degradation. At early stages in the matrix response to injury the net effect of cytokine stimulation may be matrix synthesis, but ultimately, the combination of effects on chondrocytes triggers matrix degradation. Enzymes in the matrix are held in check by activation inhibitors, including tissue inhibitor of metalloproteinase (TIMP). Growth factors are also part of this complex network, with Bone Morphogenetic Protein 2 (BMP-2) and transforming growth factor β (TGF-β) playing prominent roles in stimulating the development of osteophytes. Whereas healthy articular cartilage is avascular in part due to angiogenesis inhibitors present in cartilage, disease is characterized by the invasion of blood vessels into cartilage from underlying bone. This is influenced by vascular endothelial growth factor (VEGF) synthesis in the cartilage and bone. With these blood vessels come nerves that may bring nociceptive innervation.

Probably as a result of chronic oxidative damage, articular chondrocytes exhibit an age-related decline in synthetic capacity while maintaining the ability to produce proinflammatory mediators and matrix-degrading enzymes, findings characteristic of a senescent secretory phenotype. These chondrocytes are unable to maintain tissue homeostasis (such as after insults of a mechanical or inflammatory nature). Thus, with age, cartilage is easily damaged by minor sometimes unnoticed injuries, including those that are part of daily activities.

OA cartilage is characterized by gradual depletion of aggrecan, an unfurling of the tightly woven collagen matrix, and loss of type 2 collagen. With these changes comes increasing vulnerability of cartilage, which loses its compressive stiffness.

RISK FACTORS

Joint vulnerability and joint loading are the two major factors contributing to the development of OA. On the one hand, a vulnerable joint whose protectors are dysfunctional can develop OA with minimal levels of loading, perhaps even levels encountered during everyday activities. On the other hand, in a young joint with competent protectors, a major acute injury or long-term overloading is necessary to precipitate disease. Risk factors for OA can be understood in terms of their effect either on joint vulnerability or on loading **(Fig. 364-4)**.

SYSTEMIC RISK FACTORS

Age is the most potent risk factor for OA. Radiographic evidence of OA is rare in individuals aged <40; however, in some joints, such as the hands, OA occurs in >50% of persons aged >70. Aging increases joint vulnerability through several mechanisms. Whereas dynamic loading of joints stimulates cartilage matrix synthesis by chondrocytes in young cartilage, aged cartilage is less responsive to these stimuli. Partly because of this failure to synthesize matrix with loading, cartilage thins with age, and thinner cartilage experiences higher shear stress and is at greater risk of cartilage damage. Also, joint protectors fail more often with age. Muscles that bridge the joint become weaker with age and also respond less quickly to oncoming impulses. Sensory nerve input slows with age, retarding the feedback loop of mechanoreceptors to muscles and tendons related to their tension and position. Ligaments stretch with age, making them less able to absorb impulses. These factors work in concert to increase the vulnerability of older joints to OA.

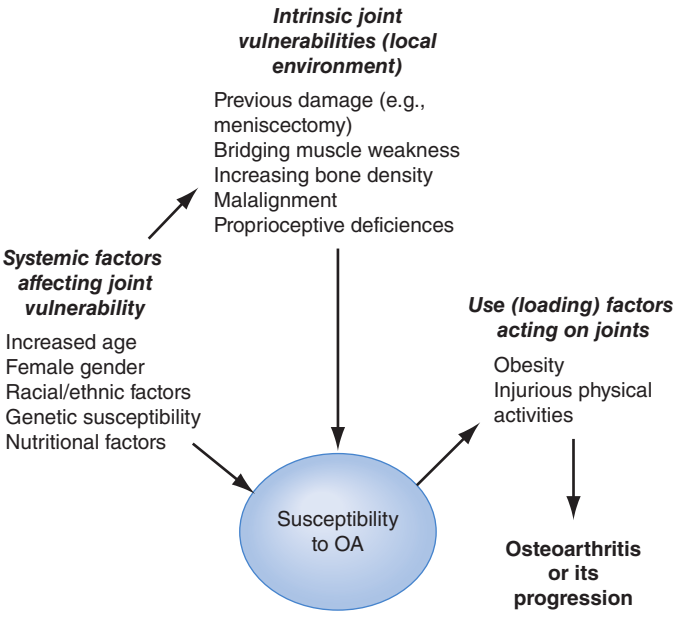


FIGURE 364-4 Risk factors for osteoarthritis (OA) either contribute to the susceptibility of the joint (systemic factors or factors in the local joint environment) or increase risk by the load they put on the joint. Usually a combination of loading and susceptibility factors is required to cause disease or its progression.

Older women are at high risk of OA in all joints, a risk that emerges as women reach their sixth decade. Although hormone loss with menopause may contribute to this risk, there is little understanding of the unique vulnerability of older women versus men to OA.

HERITABILITY AND GENETICS

OA is a highly heritable disease, but its heritability is joint specific. Fifty percent of the hand and hip OA in the community is attributable to inheritance, that is, to disease present in other members of the family. However, the heritable proportion of knee OA is at most 30%, with some studies suggesting no heritability at all. Whereas many people with OA have disease in multiple joints, this "generalized OA" phenotype is rarely inherited and is more often a consequence of aging.

Emerging evidence has identified genetic mutations that confer a high risk of OA, the best replicated is a polymorphism within the growth differentiation factor 5 (GDF5) gene. This polymorphism diminishes the quantity of GDF5; GDF5 has its main influence on joint shape which is likely to be the mechanism by which genes predisposing to OA increase risk of disease.

GLOBAL CONSIDERATIONS

With the aging of the populations, both the prevalence of OA and the amount of disability worldwide related to OA have been increasing especially in developed countries where many are living into old age. Hip OA is rare in China and in immigrants from China to the United States. However, OA in the knees is at least as common, if not more so, in Chinese than in whites from the United States, and knee OA represents a major cause of disability in China, especially in rural areas. Anatomic differences between Chinese and white hips may account for much of the difference in hip OA prevalence, with white hips having a higher prevalence of anatomic predispositions to the development of OA.

RISK FACTORS IN THE JOINT ENVIRONMENT

Some risk factors increase vulnerability of the joint through local effects on the joint environment. With changes in joint anatomy, for example, load across the joint is no longer distributed evenly across the joint surface, but rather shows an increase in focal stress. In the hip, three uncommon developmental abnormalities occurring in utero or in childhood, congenital dysplasia, Legg-Perthes disease, and slipped capital femoral epiphysis, leave a child with distortions of hip joint anatomy that often lead to OA later in life. Girls are predominantly affected by

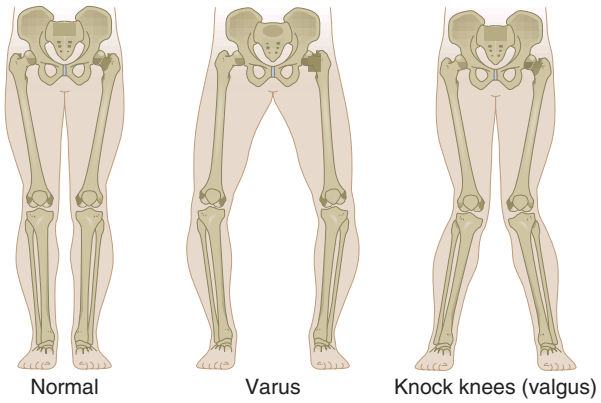


FIGURE 364-5 The two types of limb malalignment in the frontal plane: varus, in which the stress is placed across the medial compartment of the knee joint, and valgus, which places excess stress across the lateral compartment of the knee.

acetabular dysplasia, a mild form of congenital dislocation, whereas the other abnormalities more often affect boys. Depending on the severity of the anatomic abnormalities, hip OA occurs either in young adulthood (severe abnormalities) or middle age (mild abnormalities). Femoroacetabular impingement can develop during adolescence. It is a clinical syndrome in which anatomic abnormalities of the femoral head and/or the acetabulum result in abnormal contact between the two bones especially during hip flexion and rotation, leading to cartilage and labral damage and hip pain and ultimately in later life to possible hip OA.

Major injuries to a joint also can produce anatomic abnormalities that leave the joint susceptible to OA. For example, a fracture through the joint surface often causes OA in joints in which the disease is otherwise rare such as the ankle and the wrist. Avascular necrosis can lead to collapse of dead bone at the articular surface, producing anatomic irregularities and subsequent OA.

Tears of ligamentous and fibrocartilaginous structures that protect the joints, such as the meniscus in the knee and the labrum in the hip, can lead to premature OA. Meniscal tears increase with age and when chronic are often asymptomatic but lead to adjacent cartilage damage and accelerated OA. Even injuries in which the affected person never received a diagnosis may increase risk of OA. For example, in the Framingham Study subjects, men with a history of major knee injury, but no surgery, had a 3.5-fold increased risk for subsequent knee OA.

Another source of anatomic abnormality is malalignment across the joint **(Fig. 364-5)**. This factor has been best studied in the knee, which is the fulcrum of the longest lever arm in the body. Varus (bowlegged) knees with OA are at exceedingly high risk of cartilage loss in the medial or inner compartment of the knee, whereas valgus (knock-kneed) malalignment predisposes to rapid cartilage loss in the lateral compartment. Malalignment causes this effect by increasing stress on a focal area of cartilage, which then breaks down. There is evidence that malalignment in the knee not only causes cartilage loss but leads to underlying bone damage, producing bone marrow lesions seen on magnetic resonance imaging (MRI). Malalignment in the knee often produces such a substantial increase in focal stress within the knee (as evidenced by its destructive effects on subchondral bone) that severely malaligned knees may be destined to progress regardless of the status of other risk factors.

Weakness in the quadriceps muscles bridging the knee increases the risk of the development of painful OA in the knee.

The role of bone in serving as a shock absorber for impact load is not well understood, but persons with increased bone density are at high risk of OA, suggesting that the resistance of bone to impact during joint use may play a role in disease development.

LOADING FACTORS

Obesity Three to six times body weight is transmitted across the knee during single-leg stance. Any increase in weight may be multiplied by this factor to reveal the excess force across the knee in overweight persons during walking. Obesity is a well-recognized and potent risk factor for the development of knee OA and, less so, for hip OA. Obesity precedes the development of disease and is not just a consequence of the inactivity present in those with disease. It is a stronger risk factor for disease in women than in men, and in women, the relationship of weight to the risk of disease is linear, so that with each pound increase in weight, there is a commensurate increase in risk. Weight loss in women lowers the risk of developing symptomatic disease. Not only is obesity a risk factor for OA in weight-bearing joints, but obese persons have more severe symptoms from the disease.

Obesity's effect on the development and progression of disease is mediated mostly through the increased loading in weight-bearing joints that occurs in overweight persons. However, a modest association of obesity with an increased risk of hand OA suggests that systemic products of adipose tissue such as adipokines may affect disease risk also.

Repeated Use of Joint and Exercise There are two categories of repetitive joint use, occupational use and leisure time physical activities. Workers performing repetitive tasks as part of their occupations for many years are at high risk of developing OA in the joints they use repeatedly. For example, farmers are at high risk for hip OA, and miners have high rates of OA in knees and spine. Workers whose jobs require regular knee bending or lifting or carrying heavy loads have a high rate of knee OA. One reason why workers may get disease is that during long days at work, their muscles may gradually become exhausted, no longer serving as effective joint protectors.

It is widely recommended for people to adopt an exercise-filled lifestyle, and long-term studies of exercise suggest no consistent association of exercise with OA risk in the majority of persons. However, persons who already have injured joints may put themselves at greater risk by engaging in certain types of exercise. For example, persons who have already sustained major knee injuries are at increased risk of progressive knee OA as a consequence of running. In addition, compared to nonrunners, elite runners (professional runners and those on Olympic teams) have high risks of both knee and hip OA. Lastly, although recreational runners are not at increased risk of knee OA, studies suggest that they have a modest increased risk of disease in the hip.

PATHOLOGY

The pathology of OA provides evidence of the involvement of many joint structures in disease. Cartilage initially shows surface fibrillation and irregularity. As disease progresses, focal erosions develop there, and these eventually extend down to the subjacent bone. With further progression, cartilage erosion down to bone expands to involve a larger proportion of the joint surface, even though OA remains a focal disease with nonuniform loss of cartilage **(Fig. 364-6)**.

FIGURE 364-6 Pathologic changes of osteoarthritis in a toe joint. Note the nonuniform loss of cartilage (*arrowhead* vs *solid arrow*), the increased thickness of the subchondral bone envelope (*solid arrow*), and the osteophyte (*open arrow*). *(© 2018 American College of Rheumatology. Used with permission.)*

After an injury to cartilage, chondrocytes undergo mitosis and clustering. Although the metabolic activity of these chondrocyte clusters is high, the net effect of this activity is to promote proteoglycan depletion in the matrix surrounding the chondrocytes. This is because the catabolic activity is greater than the synthetic activity. As disease develops, collagen matrix becomes damaged, the negative charges of proteoglycans get exposed, and cartilage swells from ionic attraction to water molecules. Because in damaged cartilage proteoglycans are no longer forced into close proximity, cartilage does not bounce back after loading as it did when healthy, and cartilage becomes vulnerable to further injury. Chondrocytes at the basal level of cartilage undergo apoptosis.

With loss of cartilage comes alteration in subchondral bone. Stimulated by growth factors and cytokines, osteoclasts and osteoblasts in the subchondral bony plate, just underneath cartilage, become activated. Bone formation produces a thickening and stiffness of the subchondral plate that occurs even before cartilage ulcerates. Trauma to bone during joint loading may be the primary factor driving this bone response, with healing from injury (including microcracks) producing stiffness. Small areas of osteonecrosis usually exist in joints with advanced disease. Bone death may also be caused by bone trauma with shearing of microvasculature, leading to a cutoff of vascular supply to some bone areas.

At the margin of the joint, near areas of cartilage loss, osteophytes form. These start as outgrowths of new cartilage, and with neurovascular invasion from the bone, this cartilage ossifies. Osteophytes are an important radiographic hallmark of OA.

The synovium produces lubricating fluids that minimize shear stress during motion. In healthy joints, the synovium consists of a single discontinuous layer filled with fat and containing two types of cells, macrophages and fibroblasts, but in OA, it can sometimes become edematous and inflamed. There is a migration of macrophages from the periphery into the tissue, and cells lining the synovium proliferate. Inflammatory cytokines and alarmins secreted by the synovium activate chondrocytes to produce enzymes which accelerate destruction of matrix.

Additional pathologic changes occur in the capsule, which stretches, becomes edematous, and can become fibrotic.

The pathology of OA is not identical across joints. In hand joints with severe OA, for example, there are often cartilage erosions in the center of the joint probably produced by bony pressure from the opposite side of the joint.

Basic calcium phosphate and calcium pyrophosphate dihydrate crystals are present microscopically in most joints with end-stage OA. Their role in osteoarthritic cartilage is unclear, but their release from cartilage into the joint space and joint fluid likely triggers synovial inflammation, which can, in turn, produce release of cytokines and trigger nociceptive stimulation.

SOURCES OF PAIN

Because cartilage is aneural, cartilage loss in a joint is not accompanied by pain. Thus, pain in OA likely arises from structures outside the cartilage. Innervated structures in the joint include the synovium, ligaments, joint capsule, muscles, and subchondral bone. Most of these are not visualized by the x-ray, and the severity of x-ray changes in OA correlates poorly with pain severity. However, in later stages of OA, loss of cartilage integrity that is accompanied by neurovascular invasion may contribute to pain.

Based on MRI studies in osteoarthritic knees comparing those with and without pain and on studies mapping tenderness in unanesthetized joints, likely sources of pain include synovial inflammation, joint effusions, and bone marrow edema. Modest synovitis develops in many but not all osteoarthritic joints. The presence of synovitis on MRI is correlated with the presence and severity of knee pain. Capsular stretching from fluid in the joint stimulates nociceptive fibers there, inducing pain. Increased focal loading as part of the disease not only damages cartilage but probably also injures the underlying bone. As a consequence, bone marrow edema appears on the MRI; histologically, this edema signals the presence of microcracks and scar, which are the consequences of trauma. These lesions may stimulate bone nociceptive fibers.

Pain may arise from outside the joint also, including bursae near the joints. Common sources of pain near the knee are anserine bursitis and iliotibial band syndrome.

The pathologic changes of OA may eventually lead to alterations in nervous system signaling. Specifically, peripheral nociceptors can become more responsive to sensory input, known as peripheral sensitization, and there can also be an increase in central ascending nociceptive pathway activity, known as central sensitization. Individuals with OA may also have insufficient descending inhibitory modulation. Some individuals may be genetically predisposed to becoming sensitized; however, regardless of the etiology, these nervous system alterations are associated with more severe pain severity, and may contribute to the presence of allodynia and hyperalgesia in patients with OA.

CLINICAL FEATURES

Joint pain from OA is primarily activity-related in the early stages of the disease. Pain comes on either during or just after joint use and then gradually resolves. Examples include knee or hip pain with going up or down stairs, pain in weight-bearing joints when walking, and, for hand OA, pain when cooking. Early in disease, pain is episodic, triggered often by overactive use of a diseased joint, such as a person with knee OA taking a long run and noticing a few days of pain thereafter. As disease progresses, the pain becomes continuous and even begins to be bothersome at night. Stiffness of the affected joint may be prominent, but morning stiffness is usually brief (<30 min).

In knees, buckling may occur, in part, from weakness of muscles crossing the joint. Mechanical symptoms, such as buckling, catching, or locking, could also signify internal derangement, like an anterior cruciate ligament or meniscal tear; however, these symptoms, which are common in persons with knee OA need to be further evaluated only if they develop after an acute knee injury. In the knee, pain with activities requiring knee flexion, such as stair climbing and arising from a chair, often emanates from the patellofemoral compartment of the knee, which does not actively articulate until the knee is bent ~35°.

OA is the most common cause of chronic knee pain in persons aged >45, but the differential diagnosis is long. Inflammatory arthritis is likely if there is prolonged morning stiffness and many other joints are affected. Bursitis occurs commonly around knees and hips. A physical examination should focus on whether tenderness is over the joint line (at the junction of the two bones around which the joint is articulating) or outside of it. Anserine bursitis, medial and distal to the knee, is an extremely common cause of chronic knee pain that may respond to a glucocorticoid injection. Prominent nocturnal pain in the absence of end-stage OA merits a distinct workup. For hip pain, OA can be detected by loss of internal rotation on passive movement, and pain isolated to an area lateral to the hip joint usually reflects the presence of trochanteric bursitis.

No blood tests are routinely indicated for workup of patients with OA unless symptoms and signs suggest inflammatory arthritis. Examination of the synovial fluid is often more helpful diagnostically than an x-ray. If the synovial fluid white count is >1000/μL, inflammatory arthritis or gout or pseudogout is likely, the latter two being also identified by the presence of crystals.

X-rays are indicated to evaluate the possibility of OA only when joint pain and physical findings are not typical of OA or if pain persists after inauguration of treatment effective for OA. In OA, radiographic findings **(Fig. 364-7)** correlate poorly with the presence and severity of pain. Further, in both knees and hips, radiographs may be normal in early disease as they are insensitive to cartilage loss and other early findings.

Although MRI may reveal the extent of pathology in an osteoarthritic joint, it is not indicated as part of the diagnostic workup. Findings such as meniscal tears and cartilage and bone lesions occur not only in most patients with OA in the knee, but also in most older persons without joint pain. MRI findings almost never warrant a change in therapy.

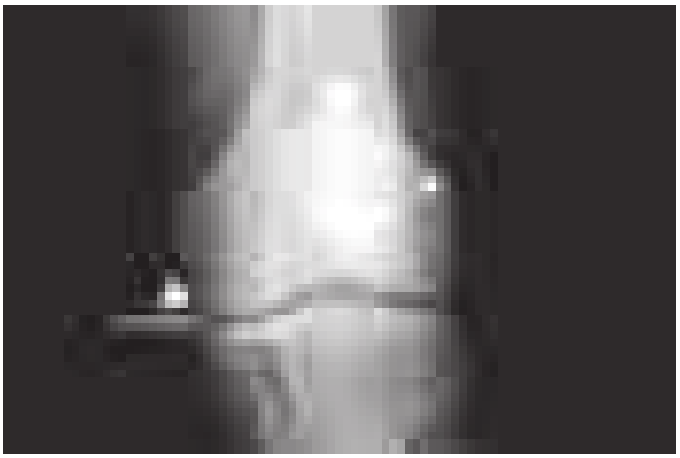


FIGURE 364-7 X-ray of knee with medial osteoarthritis. Note the narrowed joint space on medial side of the joint only (*white arrow*), the sclerosis of the bone in the medial compartment providing evidence of cortical thickening (*black arrow*), and the osteophytes in the medial femur (*white wedge*).

TREATMENT

Osteoarthritis

The goals of the treatment of OA are to alleviate pain and minimize loss of physical function. To the extent that pain and loss of function are consequences of inflammation, of weakness across the joint, and of laxity and instability, the treatment of OA involves addressing each of these impairments. Comprehensive therapy consists of a multimodality approach including nonpharmacologic and pharmacologic elements.

Patients with mild and intermittent symptoms may need only reassurance or nonpharmacologic treatments. Patients with ongoing, disabling pain are likely to need both nonpharmacotherapy and pharmacotherapy.

Treatments for knee OA have been more completely evaluated than those for hip and hand OA or for disease in other joints. Thus, although the principles of treatment are identical for OA in all joints, we shall focus below on the treatment of knee OA, noting specific recommendations for disease in other joints, especially when they differ from those for the knee.

NONPHARMACOTHERAPY

Because OA is a mechanically driven disease, the mainstay of treatment involves altering loading across the painful joint and improving the function of joint protectors, so they can better distribute load across the joint. Ways of lessening focal load across the joint include:

1. avoiding painful activities as these are usually activities that overload the joint;
2. improving the strength and conditioning of muscles that bridge the joint, so as to optimize their function; and
3. unloading the joint, either by redistributing load within the joint with a brace or a splint or by unloading the joint during weight bearing with a cane or a crutch.

The simplest treatment for many patients is to avoid activities that precipitate pain. For example, for the middle-aged patient whose long-distance running brings on symptoms of knee OA, a less demanding form of weight-bearing activity may alleviate all symptoms. For an older person whose daily walks up and down hills bring on knee pain, routing these away from hills might eliminate symptoms.

Since the loading effect of each pound of weight is multiplied across the knee three- to sixfold, each pound of weight loss may have a commensurate multiplier effect, unloading both knees and hips and probably relieving pain in those joints.

In hand joints affected by OA, splinting, by limiting motion, often minimizes pain for patients with involvement especially in the base of the thumb. Weight-bearing joints such as knees and hips can be unloaded by using a cane in the hand opposite the affected joint for partial weight bearing. A physical therapist can help teach the patient how to use the cane optimally, including ensuring that its height is optimal for unloading. Crutches or walkers can serve a similar beneficial function.

Exercise Osteoarthritic pain in knees or hips during weight bearing results in lack of activity and poor mobility, and because OA is so common, the inactivity that results increases the risk of cardiovascular disease and obesity. Aerobic capacity is poor in most elders with symptomatic knee OA, worse than others of the same age.

Weakness in muscles that bridge osteoarthritic joints is multifactorial in etiology. First, there is a decline in strength with age. Second, with limited mobility comes disuse muscle atrophy. Third, patients with painful knee or hip OA alter their gait so as to lessen loading across the affected joint, and this further diminishes muscle use. Fourth, "arthrogenous inhibition" may occur, whereby contraction of muscles bridging the joint is inhibited by a nerve afferent feedback loop emanating in a swollen and stretched joint capsule; this prevents maximal attainment of voluntary maximal strength. Because adequate muscle strength and conditioning are critical to joint protection, weakness in a muscle that bridges a diseased joint makes the joint more susceptible to further damage and pain. The degree of weakness correlates strongly with the severity of joint pain and the degree of physical limitation. One of the cardinal elements of the treatment of OA is to improve the functioning of muscles surrounding the joint.

Trials in knee and hip OA have shown that exercise lessens pain and improves physical function. Most effective exercise regimens consist of aerobic and/or resistance training, the latter of which focuses on strengthening muscles across the joint. Exercises are likely to be effective especially if they train muscles for the activities a person performs daily. Activities that increase pain in the joint should be avoided, and the exercise regimen needs to be individualized to optimize effectiveness. Range-of-motion exercises, which do not strengthen muscles, and isometric exercises that strengthen muscles, but not through range of motion, are unlikely to be effective by themselves. Low-impact exercises, including water aerobics and water resistance training, are often better tolerated by patients than exercises involving impact loading, such as running or treadmill exercises. A patient should be referred to an exercise class or to a therapist who can create an individualized regimen. In addition to conventional exercise regimens, tai chi may be effective for knee OA. However, there is no strong evidence that patients with hand OA benefit from therapeutic exercise.

Adherence over the long term is the major challenge to an exercise prescription. In trials involving patients with knee OA, who are engaged in exercise treatment, from a third to over half of patients stopped exercising by 6 months. Less than 50% continued regular exercise at 1 year. The strongest predictor of a patient's continued exercise is a previous personal history of successful exercise. Physicians should reinforce the exercise prescription at each clinic visit, help the patient recognize barriers to ongoing exercise, and identify convenient times for exercise to be done routinely. The combination of exercise with calorie restriction and weight loss is especially effective in lessening pain.

Correction of Malalignment Malalignment in the frontal plane (varus-valgus) markedly increases the stress across the joint, which can lead to progression of disease and to pain and disability (Fig. 364-5). Correcting varus-valgus malalignment, either surgically or with bracing, may relieve pain in persons whose knees are malaligned. However, correcting malalignment is often very challenging. Fitted braces that straighten varus knees by putting valgus stress across the knee can be effective. Unfortunately, many patients are unwilling to wear a realigning knee brace; in addition, in patients with obese legs, braces may slip with usage and lose their realigning effect. Braces are indicated for willing patients who can learn to put them on correctly and on whom they do not slip.

Pain from the patellofemoral compartment of the knee can be caused by tilting of the patella or patellar malalignment with the patella riding laterally in the femoral trochlear groove. Using a patellar brace to realign the patella, or tape to pull the patella back into the trochlear sulcus or reduce its tilt, has been shown, when compared to control in clinical trials, to lessen patellofemoral pain. However, patients may find it difficult to apply tape, and skin irritation from the tape is common and like realigning braces, patellar braces may slip.

Although their effect on malalignment is questionable, neoprene sleeves pulled up to cover the knee lessen pain and are easy to use and popular among patients. The explanation for their therapeutic effect on pain is unclear.

In patients with knee OA, acupuncture produces modest pain relief compared to placebo needles and may be an adjunctive treatment.

PHARMACOTHERAPY

Although nonpharmacologic approaches to therapy constitute its mainstay, pharmacotherapy serves an important adjunctive role in OA treatment for symptom management. Available drugs are administered using oral, topical, and intraarticular routes. To date, there are no available drugs that alter the disease process itself.

Acetaminophen, Nonsteroidal Anti-Inflammatory Drugs (NSAIDs), and Cyclooxygenase-2 (COX-2) Inhibitors Acetaminophen (paracetamol) is the initial analgesic of choice for patients with OA in knees, hips, or hands, even though its treatment effect in OA is small **(Table 364-1)**. For a minority of patients, it is adequate to control symptoms, in which case more toxic drugs such as NSAIDs can be avoided.

NSAIDs are the most popular drugs to treat osteoarthritic pain. They can be administered either topically or orally. In clinical trials, oral NSAIDs produce ~30% greater improvement in pain than high-dose acetaminophen. Occasional patients treated with NSAIDs experience dramatic pain relief, whereas others experience little improvement. Initially, NSAIDs should be administered topically or taken orally on an "as needed" basis because side effects are less frequent with low intermittent doses. If occasional medication use is insufficiently effective, then daily treatment may be indicated, with an anti-inflammatory dose selected (Table 364-1). Patients should be reminded to take low-dose aspirin and ibuprofen or naproxen at different times to eliminate a drug interaction.

NSAIDs taken orally have substantial and frequent side effects, the most common of which is upper gastrointestinal (GI) toxicity, including dyspepsia, nausea, bloating, GI bleeding, and ulcer disease. Thirty to forty percent of patients experience upper GI side effects so severe as to require discontinuation of medication. To minimize the risk of nonsteroidal-related GI side effects, patients should take NSAIDs after food; if risk is high, patients should take a gastroprotective agent, such as a proton pump inhibitor. Certain oral agents are safer to the stomach than others, including nonacetylated salicylates and nabumetone. Major NSAID-related GI side effects can occur in patients who do not complain of upper GI symptoms. In one study of patients hospitalized for GI bleeding, 81% had no premonitory symptoms.

Because of the increased rates of cardiovascular events associated with conventional NSAIDs such as diclofenac, many of these drugs are not appropriate long-term treatment choices for older persons with OA, especially those at high risk of heart disease or stroke. The American Heart Association has identified rofecoxib and all other COX-2 inhibitors as putting patients at high risk, although low doses of celecoxib (≤200 mg/d of celecoxib) are not associated with an elevation of risk. The only conventional NSAID that appears safe from a cardiovascular perspective is naproxen, but it does have GI toxicity.

There are other common side effects of NSAIDs, including the tendency to develop edema because of prostaglandin inhibition of afferent blood supply to glomeruli in the kidneys and, for similar reasons, a predilection toward reversible renal insufficiency. Blood pressure may increase modestly in some NSAID-treated patients. Oral NSAIDs should not be used in patients with stage IV or V renal disease and should be used with caution in those with stage III disease.

NSAIDs can be placed into a gel or topical solution with another chemical modality that enhances penetration of the skin barrier creating a topical NSAID. When absorbed through the skin, plasma concentrations are an order of magnitude lower than with the same amount of drug administered orally or parenterally. However, when these drugs are administered topically in proximity to a superficial joint (knees, hands, but not hips), the drug can be found in joint tissues such as the synovium and cartilage. Trial results have varied but generally have found that topical NSAIDs are slightly less efficacious than oral agents, but have far fewer GI and systemic side effects. Unfortunately, topical NSAIDs often cause local skin

TABLE 364-1 Pharmacologic Treatment for Osteoarthritis

TREATMENT	DOSAGE	COMMENTS
Acetaminophen	Up to 1 g tid	Prolongs half-life of warfarin. Make sure patient is not taking other treatments containing acetaminophen to avoid hepatic toxicity.
Oral NSAIDs and COX-2 inhibitors		Take with food. Increased risk of myocardial infarction and stroke for some NSAIDs and especially COX-2 inhibitors. High rates of gastrointestinal side effects, including ulcers and bleeding, occur. Patients at high risk for gastrointestinal side effects should also take either a proton pump inhibitor or misoprostol.[a] There is an increase in gastrointestinal side effects or bleeding when taken with acetylsalicylic acid. Can also cause edema and renal insufficiency.
Naproxen	375–500 mg bid	
Salsalate	1500 mg bid	
Ibuprofen	600–800 mg 3–4 times a day	
Celecoxib	100–200 mg qd	
Topical NSAIDs		Rub onto joint. Few systemic side effects. Skin irritation common.
Diclofenac Na 1% gel	4 g qid (for knees, hands)	
Opiates	Various	Common side effects include dizziness, sedation, nausea or vomiting, dry mouth, constipation, urinary retention, and pruritus. Respiratory and central nervous system depression can occur.
Capsaicin	0.025–0.075% cream 3–4 times a day	Can irritate mucous membranes.
Intraarticular injections		
Steroids		
Hyaluronans	Varies from 3 to 5 weekly injections depending on preparation	Mild to moderate pain at injection site. Controversy exists regarding efficacy.

[a]Patients at high risk include those with previous gastrointestinal events, persons ≥60 years, and persons taking glucocorticoids. Trials have shown the efficacy of proton pump inhibitors and misoprostol in the prevention of ulcers and bleeding. Misoprostol is associated with a high rate of diarrhea and cramping; therefore, proton pump inhibitors are more widely used to reduce NSAID-related gastrointestinal symptoms.

Abbreviations: COX-2, cyclooxygenase-2; NSAIDs, nonsteroidal anti-inflammatory drugs.

Source: Adapted from DT Felson: N Engl J Med 354:841, 2006.

irritation where the medication is applied, inducing redness, burning, or itching (see Table 364-1).

Intraarticular Injections: Glucocorticoids and Hyaluronic Acid Because synovial inflammation is likely to be a major cause of pain in patients with OA, local anti-inflammatory treatments administered intraarticularly may be effective in ameliorating pain, at least temporarily. Glucocorticoid injections provide such efficacy, but response is variable, with some patients having little relief of pain, whereas others experience pain relief lasting several months. Synovitis, a major cause of joint pain in OA, may abate after an injection, and this correlates with the reduction in knee pain severity. Glucocorticoid injections are useful to get patients over acute flares of pain, but their effects usually last less than 3 months. Repeated injections may cause minor amounts of cartilage loss with probably unimportant clinical consequences.

Hyaluronic acid injections can be given for treatment of symptoms in knee and hip OA, but there is controversy as to whether they have efficacy versus placebo (Table 364-1).

Other Classes of Drugs and Nutraceuticals For patients with symptomatic knee or hip OA who have not had an adequate response to the treatments above and are either unwilling to undergo or are not candidates for total joint arthroplasty, opioid analgesics have shown modest efficacy and can be tried. Opioid management plans and patient selection are critical. Another option is the use of duloxetine, which has demonstrated modest efficacy in OA.

Recent guidelines recommend against the use of glucosamine or chondroitin for OA. Large publicly supported trials have failed to show that, compared with placebo, these compounds relieve pain in persons with disease.

Optimal nonsurgical therapy for OA is often achieved by trial and error, with each patient having idiosyncratic responses to specific treatments. Placebo (or contextual) effects may account for 50% of more of treatment effects in OA and certain modes of treatment delivery including intraarticular injections have greater contextual effects than others such as pills. When medical therapies have failed and the patient has an unacceptable reduction in their quality of life and ongoing pain and disability, then at least for knee and hip OA, total joint arthroplasty is indicated.

SURGERY

For knee OA, several operations are available. Arthroscopic debridement and lavage have diminished in popularity after randomized trials evaluating this operation have showed that its efficacy is no greater than that of sham surgery for relief of pain or disability. Although arthroscopic meniscectomy is indicated for acute meniscal tears in which symptoms such as locking and acute pain are clearly related temporally to a knee injury that produced the tear, recent trials show that doing a partial meniscectomy in persons with OA and a symptomatic meniscal tear does not relieve knee pain or improve function or even lead to resolution of catching or locking of the knee.

For patients with knee OA isolated to the medial compartment, operations to realign the knee to lessen medial loading can relieve pain. These include a high tibial osteotomy, in which the tibia is broken just below the tibial plateau and realigned so as to load the lateral, nondiseased compartment, or a unicompartmental replacement with realignment. Each surgery may provide the patient with years of pain relief before a total knee replacement is required.

Ultimately, when the patient with knee or hip OA has failed nonsurgical treatment modalities with limitations of pain or function that compromise the quality of life, the patients with reasonable expectations and readiness for surgery should be referred for total knee or hip arthroplasty. These are highly efficacious operations that relieve pain and improve function in the vast majority of patients, although rates of success are higher for hip than knee replacement. Currently, failure rates for both are ~1% per year, although these rates are higher in obese patients. The chance of surgical success is greater in centers where at least 25 such operations are performed yearly or with surgeons who perform multiple operations annually. The timing of knee or hip replacement is critical. If the patient suffers for many years until their functional status has declined substantially, with considerable muscle weakness, postoperative functional status may not improve to a level achieved by others who underwent operation earlier in their disease course.

Cartilage Regeneration Chondrocyte transplantation has not been found to be efficacious in OA, perhaps because OA includes pathology of joint mechanics, which is not corrected by chondrocyte transplants. Similarly, abrasion arthroplasty (chondroplasty) has not been well studied for efficacy in OA, but it produces fibrocartilage in place of damaged hyaline cartilage. Both of these surgical attempts to regenerate and reconstitute articular cartilage may be more likely to be efficacious early in disease when joint malalignment and many of the other noncartilage abnormalities that characterize OA have not yet developed.

FURTHER READING

Felson D: Safety of nonsteroidal antiinflammatory drugs. N Engl J Med 375:2595, 2016.

Glyn-Jones S et al: Osteoarthritis. Lancet 386:376, 2015.

McAlindon TE et al: Effect of intra-articular triamcinolone vs saline on knee cartilage volume and pain in patients with knee osteoarthritis: A randomized clinical trial. JAMA 317:1967, 2017.

McAlindon TE et al: OARSI guidelines for the non-surgical management of knee osteoarthritis. Osteoarthritis Cartilage 22:363, 2014.

Neogi T et al: Sensitivity and sensitisation in relation to pain severity in knee osteoarthritis: Trait or state? Ann Rheum Dis 74:682, 2015.

O'Neill TW et al: Synovial tissue volume: A treatment target in knee osteoarthritis (OA). Ann Rheum Dis 75:84, 2016.

365 Gout and Other Crystal-Associated Arthropathies

H. Ralph Schumacher[1], Lan X. Chen

The use of polarizing light microscopy during synovial fluid analysis in 1961 by McCarty and Hollander and the subsequent application of other crystallographic techniques, such as electron microscopy, energy-dispersive elemental analysis, and x-ray diffraction, have allowed investigators to identify the roles of different microcrystals, including monosodium urate (MSU), calcium pyrophosphate (CPP), calcium apatite (apatite), and calcium oxalate (CaOx), in inducing acute or chronic arthritis or periarthritis. The clinical events that result from deposition of MSU, CPP, apatite, and CaOx have many similarities but also have important differences. Because of often similar clinical presentations, the need to perform synovial fluid analysis to distinguish the type of crystal involved must be emphasized. Polarized light microscopy alone can identify most typical crystals; apatite, however, is an exception. Aspiration and analysis of effusions are also important to assess the possibility of infection. Apart from the identification of specific microcrystalline materials or organisms, synovial fluid characteristics in crystal-associated diseases are nonspecific, and synovial fluid can be inflammatory or noninflammatory. Without crystal identification, these diseases can be confused with rheumatoid or other types of arthritis. A list of possible musculoskeletal manifestations of crystal-associated arthritis is shown in **Table 365-1**.

GOUT

Gout is a metabolic disease that most often affects middle-aged to elderly men and postmenopausal women. It results from an increased body pool of urate with hyperuricemia. It typically is characterized

[1]Deceased

TABLE 365-1 Musculoskeletal Manifestations of Crystal-Induced Arthritis

Acute mono- or polyarthritis	Destructive arthropathies
Bursitis	Chronic inflammatory arthritis
Tendinitis	Spinal arthritis
Enthesitis	Peculiar type of osteoarthritis
Tophaceous deposits	Carpal tunnel syndrome

by episodic acute arthritis or chronic arthritis caused by deposition of MSU crystals in joints and connective tissue tophi and the risk for deposition in kidney interstitium or uric acid nephrolithiasis **(Chap. 410)**.

ACUTE AND CHRONIC ARTHRITIS

Acute arthritis is the most common early clinical manifestation of gout. Usually, only one joint is affected initially, but polyarticular acute gout can occur in subsequent episodes. The metatarsophalangeal joint of the first toe often is involved, but tarsal joints, ankles, and knees also are affected commonly. Especially in elderly patients or in advanced disease, finger joints may be involved. Inflamed Heberden's or Bouchard's nodes may be a first manifestation of gouty arthritis. The first episode of acute gouty arthritis frequently begins at night with dramatic joint pain and swelling. Joints rapidly become warm, red, and tender, with a clinical appearance that often mimics that of cellulitis. Early attacks tend to subside spontaneously within 3–10 days, and most patients have intervals of varying length with no residual symptoms until the next episode. Several events may precipitate acute gouty arthritis: dietary excess, trauma, surgery, excessive ethanol ingestion, hypouricemic therapy, and serious medical illnesses such as myocardial infarction and stroke.

After many acute mono- or oligoarticular attacks, a proportion of gouty patients may present with a chronic nonsymmetric synovitis, causing potential confusion with rheumatoid arthritis **(Chap. 351)**. Less commonly, chronic gouty arthritis will be the only manifestation, and, more rarely, the disease will manifest only as periarticular tophaceous deposits in the absence of synovitis. Women represent only 5–20% of all patients with gout. Most women with gouty arthritis are postmenopausal and elderly, have osteoarthritis and arterial hypertension that causes mild renal insufficiency, and usually are receiving diuretics. Premenopausal gout is rare. Kindreds of precocious gout in young women caused by decreased renal urate clearance and renal insufficiency have been described.

Laboratory Diagnosis Even if the clinical appearance strongly suggests gout, the presumptive diagnosis ideally should be confirmed by needle aspiration of acutely or chronically involved joints or tophaceous deposits. Acute septic arthritis, several of the other crystalline- associated arthropathies, palindromic rheumatism, and psoriatic arthritis may present with similar clinical features. During acute gouty attacks, needle-shaped MSU crystals typically are seen both intracellularly and extracellularly **(Fig. 365-1)**. With compensated polarized light, these crystals are brightly birefringent with negative elongation. Synovial fluid leukocyte counts are elevated from 2000 to 60,000/μL. Effusions appear cloudy due to the increased numbers of leukocytes. Large amounts of crystals occasionally produce a thick pasty or chalky joint fluid. Bacterial infection can coexist with urate crystals in synovial fluid; if there is any suspicion of septic arthritis, joint fluid must be cultured.

MSU crystals also can often be demonstrated in the first metatarsophalangeal joint and in knees not acutely involved with gout. Arthrocentesis of these joints is a useful technique to establish the diagnosis of gout between attacks.

Serum uric acid levels can be normal or low at the time of an acute attack, as inflammatory cytokines can be uricosuric and effective initiation of hypouricemic therapy can precipitate attacks. This limits the value of serum uric acid determinations for the diagnosis of gout. Nevertheless, serum urate levels are almost always elevated at some time and are important to use to follow the course of hypouricemic therapy.

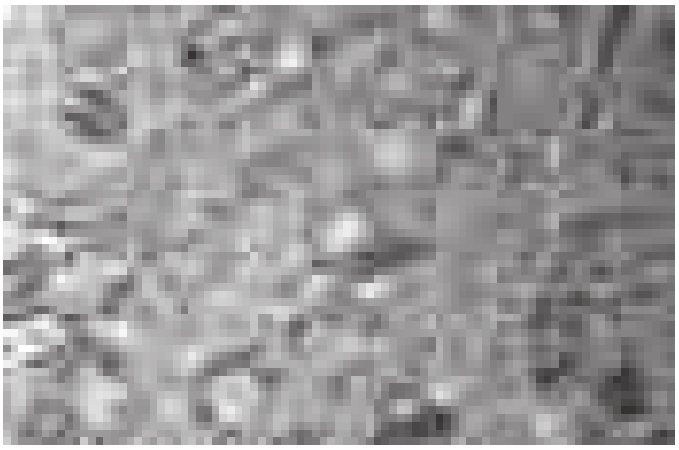


FIGURE 365-1 Extracellular and intracellular monosodium urate crystals, as seen in a fresh preparation of synovial fluid, illustrate needle- and rod-shaped crystals. These crystals are strongly negative birefringent crystals under compensated polarized light microscopy; 400×.

A 24-h urine collection for uric acid can, in some cases, be useful in assessing the risk of stones, elucidating overproduction or underexcretion of uric acid, and deciding whether it may be appropriate to use a uricosuric therapy **(Chap. 410)**. Excretion of >800 mg of uric acid per 24 h on a regular diet suggests that causes of overproduction of purine should be considered. Urinalysis, serum creatinine, hemoglobin, white blood cell (WBC) count, liver function tests, and serum lipids should be obtained because of possible pathologic sequelae of gout and other associated diseases requiring treatment and as baselines because of possible adverse effects of gout treatment.

Radiographic Features Cystic changes, well-defined erosions with sclerotic margins (often with overhanging bony edges), and soft tissue masses are characteristic radiographic features of advanced chronic tophaceous gout. Ultrasound may aid earlier diagnosis by showing a double contour sign overlying the articular cartilage. Dual-energy computed tomography (CT) can show specific features establishing the presence of urate crystals.

TREATMENT

Gout

ACUTE GOUTY ARTHRITIS

The mainstay of treatment during an acute attack is the administration of anti-inflammatory drugs such as nonsteroidal anti-inflammatory drugs (NSAIDs), colchicine, or glucocorticoids. NSAIDs are used most often in individuals without complicating comorbid conditions. Both colchicine and NSAIDs may be poorly tolerated and dangerous in the elderly and in the presence of renal insufficiency and gastrointestinal disorders. Ice pack applications and rest of the involved joints can be helpful. Colchicine given orally is a traditional and effective treatment if used early in an attack. Useful regimens are one 0.6-mg tablet given every 8 h with subsequent tapering or 1.2 mg followed by 0.6 mg in 1 h with subsequent day dosing depending on response. This is generally better tolerated than the formerly advised higher dose regimens. The drug must be at least temporarily discontinued promptly at the first sign of loose stools, and symptomatic treatment must be given for the diarrhea. Intravenous colchicine has been taken off the market. NSAIDs given in full anti-inflammatory doses are effective in ~90% of patients, and the resolution of signs and symptoms usually occurs in 5–8 days. The most effective drugs are any of those with a short half-life and include indomethacin, 25–50 mg tid; naproxen, 500 mg bid; ibuprofen, 800 mg tid; diclofenac, 50 mg tid; and celecoxib 800 mg followed by 400 mg 12 h later, then 400 mg bid.

Glucocorticoids given as an intramuscular injection or orally, for example, prednisone, 30–50 mg/d as the initial dose and gradually

tapered with the resolution of the attack, can be effective in polyarticular gout. For a single joint or a few involved joints, intraarticular triamcinolone acetonide, 20–40 mg, or methylprednisolone, 25–50 mg, have been effective and well tolerated. Based on recent evidence on the essential role of the inflammasome and interleukin 1β (IL-1β) in acute gout, daily anakinra has been used when other treatments have failed or were contraindicated.

HYPOURICEMIC THERAPY

Ultimate control of gout requires correction of the basic underlying defect: the hyperuricemia. Attempts to normalize serum uric acid to <300–360 μmol/L (5.0–6.0 mg/dL) to prevent recurrent gouty attacks and eliminate tophaceous deposits are critical and entail a commitment to hypouricemic regimens and medications that generally are required for life. Hypouricemic drug therapy should be considered when, as in most patients, the hyperuricemia cannot be corrected by simple means (control of body weight, low-purine diet, increase in liquid intake, limitation of ethanol use, decreased use of fructose-containing foods and beverages, and avoidance of diuretics). The decision to initiate hypouricemic therapy usually is made taking into consideration the number of acute attacks (urate lowering may be cost-effective after two attacks), serum uric acid levels (progression is more rapid in patients with serum uric acid >535 μmol/L [>9.0 mg/dL]), the patient's willingness to commit to lifelong therapy, or the presence of uric acid stones. Urate-lowering therapy should be initiated in any patient who already has tophi or chronic gouty arthritis. Uricosuric agents such as probenecid can be used in patients with good renal function who underexcrete uric acid, with <600 mg in a 24-h urine sample. Urine volume should be maintained by ingestion of 1500 mL of water every day. Probenecid can be started at a dose of 250 mg twice daily and increased gradually as needed up to 3 g per day to achieve and maintain a serum uric acid level of <6 mg/dL. Probenecid is generally not effective in patients with serum creatinine levels >177 μmol/L (2 mg/dL). These patients may require allopurinol or benzbromarone (not available in the United States). Benzbromarone is another uricosuric drug that is more effective in patients with chronic kidney disease. Lesinurad is a newer uricosuric; however, it is approved only in patients already on a xanthine oxidase inhibitor as an adjuvant at 200 mg per day. Some agents used to treat common comorbidities, including losartan, fenofibrate, and amlodipine, have some mild uricosuric effects.

The xanthine oxidase inhibitor allopurinol is by far the most commonly used hypouricemic agent and is the best drug to lower serum urate in overproducers, urate stone formers, and patients with renal disease. It can be given in a single morning dose, usually 100 mg initially and increasing up to 800 mg if needed. In patients with chronic renal disease, the initial allopurinol dose should be lower and adjusted depending on the serum creatinine concentration; for example, with a creatinine clearance of 10 mL/min, one generally would use 100 mg daily. Doses can be increased gradually to reach the target urate level of less than 6 mg/dL. Toxicity of allopurinol has been recognized increasingly in patients who use thiazide diuretics, in patients allergic to penicillin and ampicillin, and in Asians expressing *HLA-B*58:01*. The most serious side effects include life-threatening toxic epidermal necrolysis, systemic vasculitis, bone marrow suppression, granulomatous hepatitis, and renal failure. Patients with mild cutaneous reactions to allopurinol can reconsider the use of a uricosuric agent, undergo an attempt at desensitization to allopurinol, or take febuxostat, a new, chemically unrelated specific xanthine oxidase inhibitor. Febuxostat is approved in the United States at 40 or 80 mg once a day and does not require dose adjustment in mild to moderate renal disease. Pegloticase is a pegylated uricase, available for patients who do not tolerate or fail full doses of other treatments. It is given intravenously usually at 8 mg every 2 weeks and can dramatically lower serum uric acid in up to 50% of such patients.

Urate-lowering drugs are generally not initiated during acute attacks, but after the patient is stable and low-dose colchicine has been initiated to decrease the risk of the flares that often, without anti-inflammatory treatment, occur with urate lowering. Colchicine anti-inflammatory prophylaxis in doses of 0.6 mg one to two times daily should be given along with the hypouricemic therapy until the patient is normouricemic and without gouty attacks for 6 months or as long as tophi are present. Colchicine should not be used in dialysis patients and is given in lower doses to the patients with renal disease or with P glycoprotein or CYP3A4 inhibitors such as clarithromycin that can increase toxicity of colchicine.

CALCIUM PYROPHOSPHATE DEPOSITION (CPPD) DISEASE

PATHOGENESIS

The deposition of CPP crystals in articular tissues is most common in the elderly, occurring in 10–15% of persons age 65–75 years and 30–50% of those >85 years. In most cases, this process is asymptomatic, and the cause of CPPD is uncertain. Because >80% of patients are >60 years and 70% have preexisting joint damage from other conditions, it is likely that biochemical changes in aging or diseased cartilage favor crystal nucleation. In patients with CPPD arthritis, there is increased production of inorganic pyrophosphate and decreased levels of pyrophosphatases in cartilage extracts. Mutations in the *ANKH* gene, as described in both familial and sporadic cases, can increase elaboration and extracellular transport of pyrophosphate. The increase in pyrophosphate production appears to be related to enhanced activity of ATP pyrophosphohydrolase and 5′-nucleotidase, which catalyze the reaction of ATP to adenosine and pyrophosphate. This pyrophosphate could combine with calcium to form CPP crystals in matrix vesicles or on collagen fibers. There are decreased levels of cartilage glycosaminoglycans that normally inhibit and regulate crystal nucleation. High activities of transglutaminase enzymes also may contribute to the deposition of CPP crystals.

Release of CPP crystals into the joint space is followed by the phagocytosis of those crystals by monocyte-macrophages and neutrophils, which respond by releasing chemotactic and inflammatory substances and, as with MSU crystals, activating the inflammasome.

A minority of patients with CPPD arthropathy have metabolic abnormalities or hereditary CPP disease **(Table 365-2)**. These associations suggest that a variety of different metabolic products may enhance CPP crystal deposition either by directly altering cartilage or by inhibiting inorganic pyrophosphatases. Included among these conditions are hyperparathyroidism, hemochromatosis, hypophosphatasia, and hypomagnesemia. The presence of CPPD arthritis in individuals aged <50 years should lead to consideration of these metabolic disorders (Table 365-2) and inherited forms of disease, including those identified in a variety of ethnic groups. Genomic DNA studies performed on different kindreds have shown a possible location of genetic defects on chromosome 8q or on chromosome 5p in a region that expresses the gene of the membrane pyrophosphate channel (*ANKH* gene). Investigation of younger patients with CPPD should include inquiry for evidence of familial aggregation and evaluation of serum calcium, phosphorus, alkaline phosphatase, magnesium, iron, and transferrin.

TABLE 365-2 Conditions Associated with Calcium Pyrophosphate Crystal Deposition Disease

Aging
Disease-associated
Primary hyperparathyroidism
Hemochromatosis
Hypophosphatasia
Hypomagnesemia
Chronic gout
Postmeniscectomy
Gitelman's syndrome
Epiphyseal dysplasias

CLINICAL MANIFESTATIONS

CPPD arthropathy may be asymptomatic, acute, subacute, or chronic or may cause acute synovitis superimposed on chronically involved joints. Acute CPPD arthritis originally was termed *pseudogout* by McCarty and co-workers because of its striking similarity to gout. Other clinical manifestations of CPPD include (1) association with or enhancement of peculiar forms of osteoarthritis; (2) induction of severe destructive disease that may radiographically mimic neuropathic arthritis; (3) production of chronic symmetric synovitis that is clinically similar to rheumatoid arthritis; (4) intervertebral disk and ligament calcification with restriction of spine mobility, the crowned dens syndrome, or spinal stenosis (most commonly seen in the elderly); and (5) rarely periarticular tophus-like nodules.

The knee is the joint most frequently affected in CPPD arthropathy. Other sites include the wrist, shoulder, ankle, elbow, and hands. The temporomandibular joint may be involved. Clinical and radiographic evidence indicates that CPPD deposition is polyarticular in at least two-thirds of patients. When the clinical picture resembles that of slowly progressive osteoarthritis, diagnosis may be difficult. Joint distribution may provide important clues suggesting CPPD disease. For example, primary osteoarthritis less often involves metacarpophalangeal, wrist, elbow, shoulder, or ankle joints. If radiographs or ultrasound reveal punctate and/or linear radiodense deposits within fibrocartilaginous joint menisci or articular hyaline cartilage (*chondrocalcinosis*), the diagnostic likelihood of CPPD disease is further increased. *Definitive diagnosis* requires demonstration of typical rhomboid or rodlike crystals (generally weakly positively birefringent or nonbirefringent with polarized light) in synovial fluid or articular tissue **(Fig. 365-2)**. In the absence of joint effusion or indications to obtain a synovial biopsy, chondrocalcinosis is presumptive of CPPD. One exception is chondrocalcinosis due to CaOx in some patients with chronic renal failure.

Acute attacks of CPPD arthritis may be precipitated by trauma. Rapid diminution of serum calcium concentration, as may occur in severe medical illness or after surgery (especially parathyroidectomy), can also lead to attacks.

In as many as 50% of cases, episodes of CPPD-induced inflammation are associated with low-grade fever and, on occasion, temperatures as high as 40°C (104°F). In such cases, synovial fluid analysis with microbial cultures is essential to rule out the possibility of infection. In fact, infection in a joint with any microcrystalline deposition process can lead to crystal shedding and subsequent synovitis from both crystals and microorganisms. The leukocyte count in synovial fluid in acute CPPD can range from several thousand cells to 100,000 cells/μL, with the mean being about 24,000 cells/μL and the predominant cell being the neutrophil. CPP crystals may be seen inside tissue fragments and fibrin clots and in neutrophils (Fig. 365-2). CPP crystals may coexist with MSU and apatite in some cases.

FIGURE 365-2 Intracellular and extracellular calcium pyrophosphate (CPP) crystals, as seen in a fresh preparation of synovial fluid, illustrate rectangular, rod-shaped, and rhomboid crystals that are weakly positively or nonbirefringent crystals (compensated polarized light microscopy; 400×).

TREATMENT

CPPD Disease

Untreated acute attacks may last a few days to as long as a month. Treatment by rest, joint aspiration, and NSAIDs or by intraarticular glucocorticoid injection may result in more rapid return to prior status. For patients with frequent recurrent attacks, daily prophylactic treatment with low doses of colchicine may be helpful in decreasing the frequency of the attacks. Severe polyarticular attacks usually require short courses of glucocorticoids or an IL-1β antagonist, anakinra. Unfortunately, there is no effective way to remove CPP deposits from cartilage and synovium. Uncontrolled studies suggest that the administration of NSAIDS (with a gastric protective agent if required), hydroxychloroquine, or even methotrexate may be helpful in controlling persistent synovitis. Patients with progressive destructive large-joint arthropathy may require joint replacement.

CALCIUM APATITE DEPOSITION DISEASE

PATHOGENESIS

Apatite is the primary mineral of normal bone and teeth. Abnormal accumulation of basic calcium phosphates, largely carbonate substituted apatite, can occur in areas of tissue damage (dystrophic calcification), hypercalcemic or hyperparathyroid states (metastatic calcification), and certain conditions of unknown cause **(Table 365-3)**. In chronic renal failure, hyperphosphatemia can contribute to extensive apatite deposition both in and around joints. Familial aggregation is rarely seen; no association with *ANKH* mutations has been described thus far. Apatite crystals are deposited primarily on matrix vesicles. Incompletely understood alterations in matrix proteoglycans, phosphatases, hormones, and cytokines probably can influence crystal formation.

Apatite aggregates are commonly present in synovial fluid in an extremely destructive chronic arthropathy of the elderly that occurs most often in the shoulders (Milwaukee shoulder) and in a similar process in hips, knees, and erosive osteoarthritis of fingers. Joint destruction is associated with damage to cartilage and supporting structures, leading to instability and deformity. Progression tends to be indolent. Symptoms range from minimal to severe pain and disability that may lead to joint replacement surgery. Whether severely affected patients represent an extreme synovial tissue response to the apatite crystals that are so common in osteoarthritis is uncertain. Synovial lining cell

TABLE 365-3 Conditions Associated with Apatite Deposition Disease

Aging
Osteoarthritis
Hemorrhagic shoulder effusions in the elderly (Milwaukee shoulder)
Destructive arthropathy
Tendinitis, bursitis
Tumoral calcinosis (sporadic cases)
Disease-associated
Hyperparathyroidism
Milk-alkali syndrome
Renal failure/long-term dialysis
Connective tissue diseases (e.g., systemic sclerosis, dermatomyositis, SLE)
Heterotopic calcification after neurologic catastrophes (e.g., stroke, spinal cord injury)
Heredity
Bursitis, arthritis
Tumoral calcinosis
Fibrodysplasia ossificans progressiva

Abbreviation: SLE, systemic lupus erythematosus.

or fibroblast cultures exposed to apatite (or CPP) crystals can undergo mitosis and markedly increase the release of prostaglandin E_2, various cytokines, and also collagenases and neutral proteases, underscoring the destructive potential of abnormally stimulated synovial lining cells.

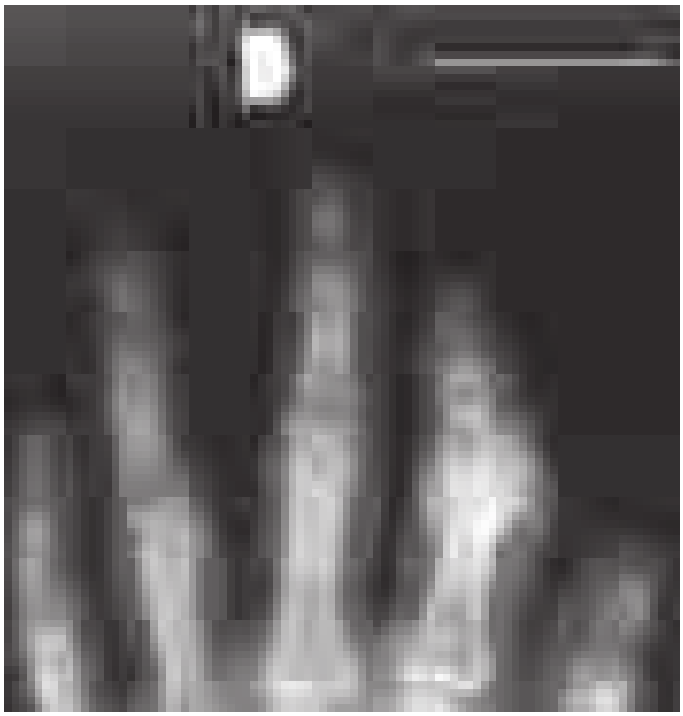
CLINICAL MANIFESTATIONS

Periarticular or articular deposits may occur and may be associated with acute reversible inflammation and/or chronic damage to the joint capsule, tendons, bursa, or articular surfaces. The most common sites of apatite deposition include bursae and tendons in and/or around the knees, shoulders, hips, and fingers. Clinical manifestations include asymptomatic radiographic abnormalities, acute synovitis, bursitis, tendinitis, and chronic destructive arthropathy. Although the true incidence of apatite arthritis is not known, 30–50% of patients with osteoarthritis have apatite microcrystals in their synovial fluid. Such crystals frequently can be identified in clinically stable osteoarthritic joints, but they are more likely to come to attention in persons experiencing acute or subacute worsening of joint pain and swelling. The synovial fluid leukocyte count in apatite arthritis is usually low (<2000/μL) despite dramatic symptoms, with predominance of mononuclear cells.

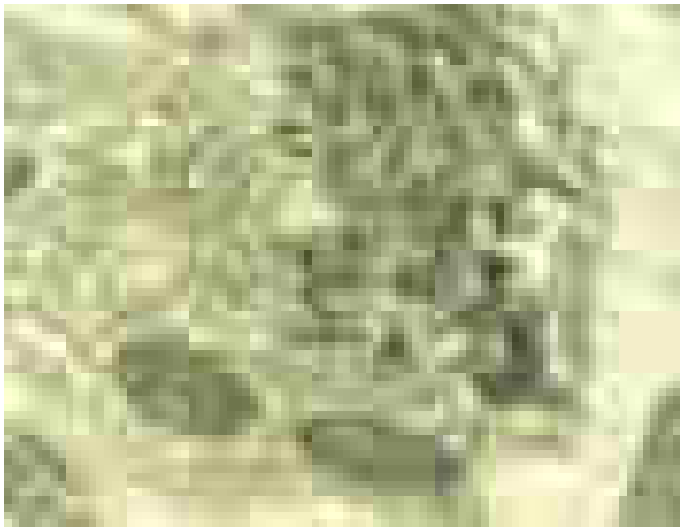
DIAGNOSIS

Intra- and/or periarticular calcifications with or without erosive, destructive, or hypertrophic changes may be seen on radiographs **(Fig. 365-3)**. They should be distinguished from the linear calcifications typical of CPPD.

Definitive diagnosis of apatite arthropathy, also called basic calcium phosphate disease, depends on identification of crystals from synovial fluid or tissue (Fig. 365-3). Individual crystals are very small and can be seen only by electron microscopy. Clumps of crystals may appear as 1- to 20-μm shiny intra- or extracellular nonbirefringent globules or aggregates that stain purplish with Wright's stain and bright red with alizarin red S. Tetracycline binding and other investigative techniques are under consideration as labeling alternatives. Absolute identification depends on electron microscopy with energy-dispersive elemental analysis, x-ray diffraction, infrared spectroscopy, or Raman microspectroscopy, but these techniques usually are not required in clinical diagnosis.



A



B

FIGURE 365-3 ***A.*** Radiograph showing calcification due to apatite crystals surrounding an eroded joint. ***B.*** An electron micrograph demonstrates dark needle-shaped apatite crystals within a vacuole of a synovial fluid mononuclear cell (30,000×).

TREATMENT

Calcium Apatite Deposition Disease

Treatment of apatite arthritis or periarthritis is nonspecific. Acute attacks of bursitis or synovitis may be self-limiting, resolving in days to several weeks. Aspiration of effusions and the use of either NSAIDs or oral colchicine for 2 weeks or intra- or periarticular injection of a depot glucocorticoid appear to shorten the duration and intensity of symptoms. Local injection of disodium ethylenediaminetetraacetic acid (EDTA) and SC anakinra have been suggested as effective in single studies of acute calcific tendinitis at the shoulder. Other reports have described that IV gamma globulin, rituximab, calcium channel blockers, or bisphosphonates may help diffuse calcinosis. Periarticular apatite deposits may be resorbed with resolution of attacks. Agents to lower serum phosphate levels may lead to resorption of deposits in renal failure patients receiving hemodialysis. In patients with underlying severe destructive articular changes, response to medical therapy is usually less rewarding.

CaOx DEPOSITION DISEASE

PATHOGENESIS

Primary oxalosis is a rare hereditary metabolic disorder **(Chap. 413)**. Enhanced production of oxalic acid may result from at least two different enzyme defects, leading to hyperoxalemia and deposition of CaOx crystals in tissues. Nephrocalcinosis and renal failure are typical results. Acute and/or chronic CaOx arthritis, periarthritis, and bone disease may complicate primary oxalosis during later years of illness.

Secondary oxalosis is more common than the primary disorder. In chronic renal disease, CaOx deposits have long been recognized in visceral organs, blood vessels, bones, and cartilage and are now known to be one of the causes of arthritis in chronic renal failure. Thus far, reported patients have been dependent on long-term hemodialysis or peritoneal dialysis **(Chap. 306)**, and many had received ascorbic acid supplements. Ascorbic acid is metabolized to oxalate, which is inadequately cleared in uremia and by dialysis. Such supplements and foods high in oxalate content usually are avoided in dialysis programs because of the risk of enhancing hyperoxalosis and its sequelae.

CLINICAL MANIFESTATIONS AND DIAGNOSIS

CaOx aggregates can be found in bone, articular cartilage, synovium, and periarticular tissues. From these sites, crystals may be shed, causing acute synovitis. Persistent aggregates of CaOx can, like apatite and CPP, stimulate synovial cell proliferation and enzyme release, resulting in progressive articular destruction. Deposits have been documented in fingers, wrists, elbows, knees, ankles, and feet.

FIGURE 365-4 Bipyramidal and small polymorphic calcium oxalate crystals from synovial fluid are a classic finding in calcium oxalate arthropathy (ordinary light microscopy; 400×).

Clinical features of acute CaOx arthritis may not be distinguishable from those due to urate, CPP, or apatite. Radiographs may reveal chondrocalcinosis or soft tissue calcifications. CaOx-induced synovial effusions are usually noninflammatory, with <2000 leukocytes/μL, or mildly inflammatory. Neutrophils or mononuclear cells can predominate. CaOx crystals have a variable shape and variable birefringence to polarized light. The most easily recognized forms are bipyramidal, have strong birefringence **(Fig. 365-4)**, and stain with alizarin red S.

TREATMENT

Calcium Oxalate Deposition Disease

Treatment of CaOx arthropathy with NSAIDs, colchicine, intra-articular glucocorticoids, and/or an increased frequency of dialysis has produced only slight improvement. In primary oxalosis, liver transplantation has induced a significant reduction in crystal deposits **(Chap. 413)**.

FURTHER READING

DALBETH N et al: Survey definitions of gout for epidemiological studies: Comparison with crystal identification as the gold standard. Arthritis Care Res (Hoboken) doi: 10.1002/acr.22896 2016 March 25 (Epub ahead of print).

OGDIE A et al: Imaging modalities for the classification of gout: Systematic literature review and meta-analysis. Ann Rheum Dis 74:1868, 2015.

RODDY E et al: Concordance of the management of chronic gout in a UK primary-care population with the EULAR gout recommendations. Ann Rheum Dis 66:1311, 2007.

SCHUMACHER HR JR, REGINATO AJ: *Atlas of Synovial Fluid Analysis and Crystal Identification*. Philadelphia: Lea & Febiger, 1991.

366 Fibromyalgia

Leslie J. Crofford

DEFINITION

Fibromyalgia (FM) is characterized by chronic widespread musculoskeletal pain and tenderness. Although FM is defined primarily as a pain syndrome, patients also commonly report associated neuropsychological symptoms of fatigue, unrefreshing sleep, cognitive dysfunction, anxiety, and depression. Patients with FM have an increased prevalence of other syndromes associated with pain and fatigue, including chronic fatigue syndrome **(Chap. 442)**, temporomandibular disorder, chronic headaches, irritable bowel syndrome, interstitial cystitis/painful bladder syndrome, and other pelvic pain syndromes. Available evidence implicates the central nervous system as key to maintaining pain and other core symptoms of FM and related conditions. The presence of FM is associated with substantial negative consequences for physical and social functioning.

EPIDEMIOLOGY

In clinical settings, a diagnosis of FM is made in ~2% of the population and is far more common in women than in men, with a ratio of ~9:1. However, in population-based survey studies worldwide, the prevalence rate is ~2–5%, with a female-to-male ratio of only 2–3:1 and with some variability depending on the method of ascertainment. The prevalence data are similar across socioeconomic classes. Cultural factors may play a role in determining whether patients with FM symptoms seek medical attention; however, even in cultures in which secondary gain is not expected to play a significant role, the prevalence of FM remains in this range.

CLINICAL MANIFESTATIONS

Pain and Tenderness At presentation, patients with FM most commonly report "pain all over." These patients have pain that is typically both above and below the waist on both sides of the body and involves the axial skeleton (neck, back, or chest). The pain attributable to FM is poorly localized, difficult to ignore, severe in its intensity, and associated with a reduced functional capacity. For a diagnosis of FM, pain should have been present most of the day on most days for at least 3 months.

The pain of FM is associated with tenderness and increased evoked pain sensitivity. In clinical practice, this elevated sensitivity may be identified by pain induced by the pressure of a blood pressure cuff or skin roll tenderness. More formally, an examiner may complete a tender-point examination in which the examiner uses the thumbnail to exert pressure of ~4 kg/m^2 (or the amount of pressure leading to blanching of the tip of the thumbnail) on well-defined musculotendinous sites **(Fig. 366-1)**. Previously, the classification criteria of the American College of Rheumatology required that 11 of 18 sites be perceived as painful for a diagnosis of FM. In practice, tenderness is a continuous variable, and strict application of a categorical threshold for diagnostic specifics is not necessary. Newer criteria eliminate the need for identification of tender points and focus instead on clinical symptoms of widespread or multi-site pain and neuropsychological symptoms. The newer criteria perform well in a clinical setting in comparison to the older, tender-point criteria. However, it appears that when the new criteria are applied to populations, the result is an increase in prevalence of FM and a change in the sex ratio (see "Epidemiology," earlier).

Patients with FM often have peripheral pain generators that are thought to serve as triggers for the more widespread pain attributed to central nervous system factors. Potential pain generators such as arthritis, bursitis, tendinitis, neuropathies, and other inflammatory or degenerative conditions should be identified by history and physical examination. More subtle pain generators may include joint hypermobility and scoliosis. In addition, patients may have chronic myalgias triggered by infectious, metabolic, or psychiatric conditions that can serve as triggers for the development of FM. These conditions are often identified in the differential diagnosis of patients with FM, and a major challenge is to distinguish the ongoing activity of a triggering condition from FM that is occurring as a consequence of a comorbid condition and that should itself be treated.

Neuropsychological Symptoms In addition to widespread pain, FM patients typically report fatigue, stiffness, sleep disturbance, cognitive dysfunction, anxiety, and depression. These symptoms are present to varying degrees in most FM patients but are not present in every patient or at all times in a given patient. Relative to pain, such symptoms may, however, have an equal or even greater impact on function and quality of life. Fatigue is highly prevalent in patients under primary care who ultimately are diagnosed with FM. Pain, stiffness,

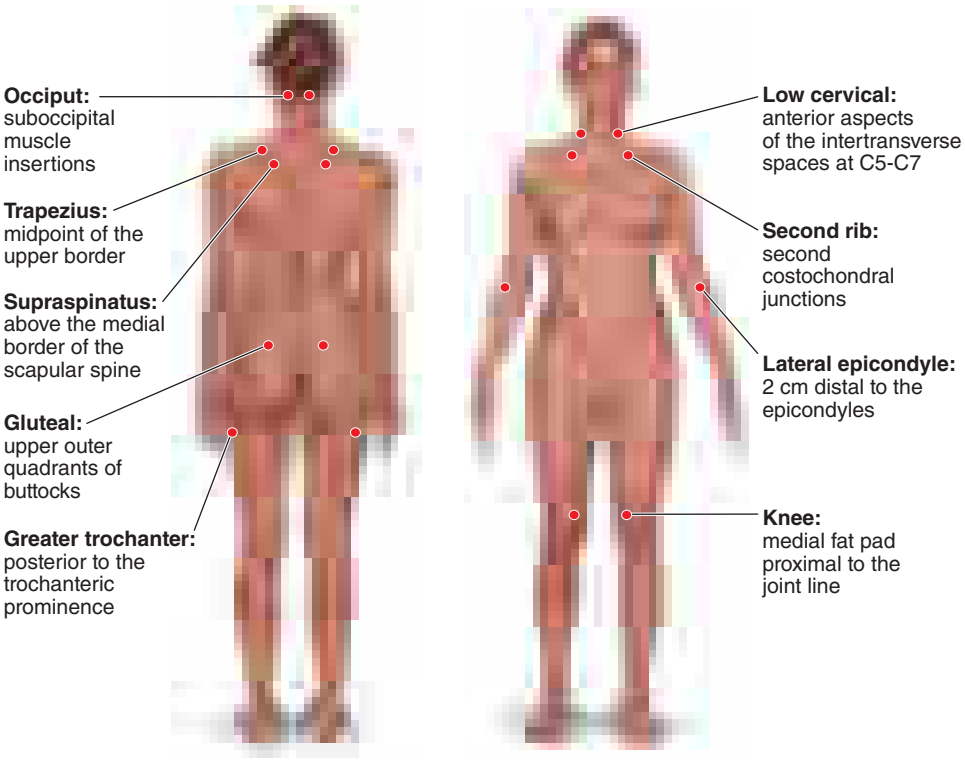


FIGURE 366-1 Tender-point assessment in patients with fibromyalgia. *(Figure created using data from F Wolfe et al: Arthritis Care Res 62:600, 2010.)*

and fatigue often are worsened by exercise or unaccustomed activity (postexertional malaise). The sleep complaints include difficulty falling asleep, difficulty staying asleep, and early-morning awakening. Regardless of the specific complaint, patients awake feeling unrefreshed. Patients with FM may meet criteria for restless legs syndrome and sleep-disordered breathing; frank sleep apnea can also be documented. Cognitive issues are characterized as difficulties with attention or concentration, problems with word retrieval, and short-term memory loss. Studies have demonstrated altered cognitive function in these domains in patients with FM, though speed of processing is age-appropriate. Symptoms of anxiety and depression are common, and the lifetime prevalence of mood disorders in patients with FM approaches 80%. Although depression is neither necessary nor sufficient for the diagnosis of FM, it is important to screen for major depressive disorders by querying for depressed mood and anhedonia. Analysis of genetic factors that are likely to predispose to FM reveals shared neurobiologic pathways with mood disorders, providing the basis for comorbidity (see later in this chapter).

Overlapping Syndromes Because FM can overlap in presentation with other chronic pain conditions, review of systems often reveals headaches, facial/jaw pain, regional myofascial pain particularly involving the neck or back, and arthritis. Visceral pain involving the gastrointestinal tract, bladder, and pelvic or perineal region is often present as well. Patients may or may not meet defined criteria for specific syndromes. It is important for patients to understand that shared pathways may mediate symptoms and treatment strategies effective for one condition may help with global symptom management.

Comorbid Conditions FM is often comorbid with chronic musculoskeletal, infectious, metabolic, or psychiatric conditions. Whereas FM affects only 2–5% of the general population, it occurs in ≥20% of patients with degenerative or inflammatory rheumatic disorders, likely because these conditions serve as peripheral pain generators to alter central pain-processing pathways. Similarly, chronic infectious, metabolic, or psychiatric diseases associated with musculoskeletal pain can mimic FM and/or serve as a trigger for the development of FM. It is particularly important for clinicians to be sensitive to pain management of these comorbid conditions so that when FM emerges—characterized by pain outside the boundaries of what could reasonably be explained by the triggering condition, development of neuropsychological symptoms, or tenderness on physical examination—treatment of central pain processes will be undertaken as opposed to a continued focus on treatment of peripheral or inflammatory causes of pain.

Psychosocial Considerations Symptoms of FM often have their onset and are exacerbated during periods of perceived stress. This pattern may reflect an interaction among central stress physiology, vigilance or anxiety, and central pain-processing pathways. An understanding of current psychosocial stressors will aid in patient management, as many factors that exacerbate symptoms cannot be addressed by pharmacologic approaches. Furthermore, there is a high prevalence of exposure to previous interpersonal and other forms of violence in patients with FM and related conditions. If posttraumatic stress disorder is an issue, the clinician should be aware of it and consider treatment options.

Functional Impairment It is crucial to evaluate the impact of FM symptoms on function and role fulfillment. In defining the success of a management strategy, improved function is a key measure. Functional assessment should include physical, mental, and social domains. Recognition of the ways in which role functioning falls short will be helpful in the establishing treatment goals.

DIFFERENTIAL DIAGNOSIS

Because musculoskeletal pain is such a common complaint, the differential diagnosis of FM is broad. Table 366-1 lists some of the more common conditions that should be considered. Patients with inflammatory causes for widespread pain should be identifiable on the basis of specific history, physical findings, and laboratory or radiographic tests.

LABORATORY OR RADIOGRAPHIC TESTING

Routine laboratory and radiographic tests yield normal results in FM. Thus diagnostic testing is focused on exclusion of other diagnoses and evaluation for pain generators or comorbid conditions (Table 366-2). Most patients with new chronic widespread pain should be assessed for the most common entities in the differential diagnosis. Radiographic testing should be used sparingly and only for diagnosis of inflammatory arthritis. After the patient has been evaluated thoroughly, repeat testing is discouraged unless the symptom complex changes. Particularly to be discouraged is magnetic resonance imaging (MRI) of the spine unless there are features suggesting inflammatory spine disease or neurologic symptoms.

GENETICS AND PHYSIOLOGY

As in most complex diseases, it is likely that a number of genes contribute to vulnerability to the development of FM. To date, these genes appear to be in pathways controlling pain and stress responses. Some of the genetic underpinnings of FM are shared across other chronic pain conditions. Genes associated with metabolism, transport, and receptors of serotonin and other monoamines have been implicated in FM and overlapping conditions. Genes associated with other pathways involved in pain transmission have also been described as vulnerability factors for FM. Taken together, the pathways in which polymorphisms have been identified in FM patients further implicate

TABLE 366-1 Common Conditions in the Differential Diagnosis of Fibromyalgia

Inflammatory
Polymyalgia rheumatica
Inflammatory arthritis: rheumatoid arthritis, spondyloarthritides
Connective tissue diseases: systemic lupus erythematosus, Sjögren's syndrome
Infectious
Hepatitis C
HIV infection
Lyme disease
Parvovirus B19 infection
Epstein-Barr virus infection
Noninflammatory
Degenerative joint/spine/disk disease
Myofascial pain syndromes
Bursitis, tendinitis, repetitive strain injuries
Endocrine
Hypo- or hyperthyroidism
Hyperparathyroidism
Neurologic Diseases
Multiple sclerosis
Neuropathic pain syndromes
Psychiatric Disease
Major depressive disorder
Drugs
Statins
Aromatase inhibitors

central factors in mediation of the physiology that leads to the clinical manifestations of FM.

Psychophysical testing of patients with FM has demonstrated altered sensory afferent pain processing and impaired descending noxious inhibitory control leading to hyperalgesia and allodynia. Functional MRI and other research imaging procedures clearly demonstrate activation of the brain regions involved in the experience of pain in response to stimuli that are innocuous in study participants without FM. Pain perception in FM patients is influenced by the emotional and cognitive dimensions, such as catastrophizing and perceptions of control, providing a solid basis for recommendations for cognitive and behavioral treatment strategies.

Studies have indicated that some patients meeting criteria for FM may have a small fiber neuropathy. Other studies have identified alterations in expressed gene signatures in peripheral blood. These early studies raise the possibility that confirmatory diagnostic testing could be developed to assist in the diagnosis of FM.

TABLE 366-2 Laboratory and Radiographic Testing in Patients with Fibromyalgia Symptoms

Routine
Erythrocyte sedimentation rate (ESR) or C-reactive protein (CRP)
Complete blood count (CBC)
Thyroid-stimulating hormone (TSH)
Guided by History and Physical Examination
Complete metabolic panel
Antinuclear antibody (ANA)
Anti-SSA (anti–Sjögren's syndrome A) and anti-SSB
Rheumatoid factor and anti–cyclic citrullinated peptide (anti-CCP)
Creatine phosphokinase (CPK)
Viral (e.g., Hepatitis C, HIV) and bacterial (e.g., Lyme) serologies
Spine and joint radiographs

Source: LM Arnold et al: J Women's Health 21:231, 2012; MA Fitzcharles et al: J Rheumatol 40:1388, 2013.

APPROACH TO THE PATIENT

Fibromyalgia

FM is common and has an extraordinary impact on the patient's function and health-related quality of life. Optimal management requires prompt diagnosis and assessment of pain, function, and psychosocial context. Physicians and other health professionals can be helpful in managing some of the symptoms and impact of FM. Developing a partnership with patients is essential for improving the outcome of FM, with a goal of understanding the factors involved, implementing a treatment strategy, and choosing appropriate nonpharmacologic and pharmacologic treatments.

TREATMENT

Fibromyalgia

NONPHARMACOLOGIC TREATMENT

Patients with chronic pain, fatigue, and other neuropsychological symptoms require a framework for understanding the symptoms that have such an important impact on their function and quality of life. Explaining the genetics, triggers, and physiology of FM can be an important adjunct in relieving associated anxiety and in reducing the overall cost of health care resources. In addition, patients must be educated regarding expectations for treatment. The physician should focus on improved function and quality of life rather than elimination of pain. Illness behaviors, such as frequent physician visits, should be discouraged and behaviors that focus on improved function strongly encouraged.

Treatment strategies should include physical conditioning, with encouragement to begin at low levels of aerobic exercise and to proceed with slow but consistent advancement. Patients who have been physically inactive or who report postexertional malaise may do best in supervised or water-based programs at the start. Strength training may be recommended after patients reach their aerobic goals. Meditative movement therapies, such as qigong, yoga, or Tai Chi, may also be helpful. Other defined physical therapies such as acupuncture or hydrotherapy may also be considered. Exercise programs are helpful in reducing tenderness and enhancing self-efficacy. Cognitive-behavioral strategies to improve sleep hygiene and reduce illness behaviors can also be helpful in management.

PHARMACOLOGIC APPROACHES

It is essential for the clinician to treat any comorbid triggering condition and to clearly delineate for the patient the treatment goals for each medication. For example, glucocorticoids or nonsteroidal anti-inflammatory drugs may be useful for management of inflammatory triggers but are not effective against FM-related symptoms. At present, the treatment approaches that have proved most successful in FM patients target afferent or descending pain pathways. Table 366-3 lists the drugs with demonstrated effectiveness. It should be emphasized that strong opioid analgesics are to be avoided in patients with FM. These agents have no demonstrated efficacy in FM and are associated with adverse effects that can worsen both symptoms and function. Tramadol, an opioid with mild serotonin-noradrenaline reuptake inhibitor activity has been studied in this population with indication of efficacy. Use of single agents to treat multiple symptom domains is strongly encouraged. For example, if a patient's symptom complex is dominated by pain and sleep disturbance, use of an agent that exerts both analgesic and sleep-promoting effects is desirable. These agents include cyclobenzaprine, sedating antidepressants such as amitriptyline, and alpha-2-delta ligands such as gabapentin and pregabalin. For patients whose pain is associated with fatigue, anxiety, or depression, drugs that have both analgesic and antidepressant/anxiolytic effects, such as duloxetine or milnacipran, may be the best first choice.

TABLE 366-3 Pharmacologic Agents Effective for Treatment of Fibromyalgia

Muscle relaxant
Cyclobenzaprine
Antidepressants: balanced serotonin–norepinephrine reuptake inhibitors
Amitryptiline[a]
Duloxetine[b,c]
Milnacipran[b,c]
Anticonvulsants: ligand of the alpha-2-delta subunit of voltage-gated calcium channels
Pregabalin[b]
Analgesic
Tramadol

[a]RA Moore et al: Cochrane Database Syst Rev 12:CD008242, 2012. [b]Approved by the U.S. Food and Drug Administration. [c]W Hauser et al: Cochrane Database Syst Rev 1: CD010292, 2013.

Source: GJ Macfarlane et al: EULAR revised recommendations for the management of fibromyalgia. Ann Rheum Dis 76:318, 2017.

FURTHER READING

Bennett RM et al: Criteria for the diagnosis of fibromyalgia: Validation of the modified 2010 preliminary American College of Rheumatology criteria and the development of alternative criteria. Arthritis Care Res 66:1364, 2014.

Clauw DJ: Fibromyalgia: A clinical review. JAMA 311:1547, 2014.

Macfarlane GJ et al: EULAR revised recommendations for the management of fibromyalgia. Ann Rheum Dis 76:318, 2017.

367 Arthritis Associated with Systemic Disease, and Other Arthritides

Carol A. Langford, Brian F. Mandell

ARTHRITIS ASSOCIATED WITH SYSTEMIC DISEASE

ARTHROPATHY OF ACROMEGALY

Acromegaly is the result of excessive production of growth hormone by an adenoma in the anterior pituitary gland **(Chap. 373)**. The excessive secretion of growth hormone along with insulin-like growth factor I stimulates proliferation of cartilage, periarticular connective tissue, and bone, resulting in several musculoskeletal problems, including osteoarthritis, back pain, muscle weakness, and carpal tunnel syndrome.

Osteoarthritis is a common feature, most often affecting the knees, shoulders, hips, and hands. Single or multiple joints may be affected. Hypertrophy of cartilage initially produces radiographic widening of the joint space. The newly synthesized cartilage is abnormally susceptible to fissuring, ulceration, and destruction. Ligamental laxity of joints further contributes to the development of osteoarthritis. Cartilage degrades, the joint space narrows, and subchondral sclerosis and osteophytes develop. Joint examination reveals crepitus and laxity. Joint fluid is noninflammatory. Calcium pyrophosphate dihydrate crystals are found in the cartilage in some cases of acromegaly arthropathy and, when shed into the joint, can elicit attacks of pseudogout. Chondrocalcinosis may be observed on radiographs. Back pain is extremely common, perhaps as a result of spine hypermobility. Spine radiographs show normal or widened intervertebral disk spaces, hypertrophic anterior osteophytes, and ligamental calcification. The latter changes are similar to those observed in patients with diffuse idiopathic skeletal hyperostosis. Dorsal kyphosis in conjunction with elongation of the ribs contributes to the development of the barrel chest seen in acromegalic patients. The hands and feet become enlarged as a result of soft tissue proliferation. The fingers are thickened and have spadelike distal tufts. One-third of patients have a thickened heel pad. Approximately 25% of patients exhibit Raynaud's phenomenon. Carpal tunnel syndrome occurs in about half of patients. The median nerve is compressed by excess connective tissue in the carpal tunnel. Patients with acromegaly may develop proximal muscle weakness, which is thought to be caused by the effect of growth hormone on muscle. Serum muscle enzyme levels and electromyographic findings are normal. Muscle biopsy specimens contain muscle fibers of varying size without inflammation.

ARTHROPATHY OF HEMOCHROMATOSIS

Hemochromatosis is a disorder of iron storage. Absorption of excessive amounts of iron from the intestine leads to iron deposition in parenchymal cells, which results in impairment of organ function **(Chap. 407)**. Symptoms of hemochromatosis usually begin between the ages of 40 and 60 but can appear earlier. Arthropathy, which occurs in 20–40% of patients, usually begins after the age of 50 and may be the first clinical feature of hemochromatosis. The arthropathy is an osteoarthritis-like disorder affecting the small joints of the hands and later the larger joints, such as knees, ankles, shoulders, and hips. The second and third metacarpophalangeal joints of both hands are often the first and most prominent joints affected; this clinical picture may provide an important clue to the possibility of hemochromatosis because these joints are not predominantly affected by "routine" osteoarthritis. Patients experience some morning stiffness and pain with use of involved joints. The affected joints are enlarged and mildly tender. Radiographs show narrowing of the joint space, subchondral sclerosis, subchondral cysts, and juxtaarticular proliferation of bone. Hooklike osteophytes are seen in up to 20% of patients; although they are regarded as a characteristic feature of hemochromatosis, they can also occur in osteoarthritis and are not disease specific. The synovial fluid is noninflammatory. The synovium shows mild to moderate proliferation of iron-containing lining cells, fibrosis, and some mononuclear cell infiltration. In approximately half of patients, there is evidence of calcium pyrophosphate deposition disease, and some patients late in the course of disease experience episodes of acute pseudogout **(Chap. 365)**. An early diagnosis is suggested by high serum transferrin saturation, which is more sensitive than ferritin elevation.

Iron may damage the articular cartilage in several ways. Iron catalyzes superoxide-dependent lipid peroxidation, which may play a role in joint damage. In animal models, ferric iron has been shown to interfere with collagen formation and increase the release of lysosomal enzymes from cells in the synovial membrane. Iron inhibits synovial tissue pyrophosphatase in vitro and therefore may inhibit pyrophosphatase in vivo, resulting in chondrocalcinosis.

TREATMENT

Arthropathy of Hemochromatosis

The treatment of hemochromatosis is repeated phlebotomy. Unfortunately, this treatment has little effect on established arthritis, which, along with chondrocalcinosis, may progress. Symptom-based treatment of the arthritis consists of administration of acetaminophen and nonsteroidal anti-inflammatory drugs (NSAIDs), as tolerated. Acute pseudogout attacks are treated with high doses of an NSAID or a short course of glucocorticoids. Hip or knee total joint replacement has been successful in advanced disease.

HEMOPHILIC ARTHROPATHY

Hemophilia is a sex-linked recessive genetic disorder characterized by the absence or deficiency of factor VIII (hemophilia A, or classic hemophilia) or factor IX (hemophilia B, or Christmas disease) **(Chap. 112)**. Hemophilia A constitutes 85% of cases. Spontaneous hemarthrosis is a common problem with both types of hemophilia and can lead to a deforming arthritis. The frequency and severity of hemarthrosis are

related to the degree of clotting factor deficiency. Hemarthrosis is not common in other disorders of coagulation such as von Willebrand disease, factor V deficiency, warfarin therapy, or thrombocytopenia.

Hemarthrosis occurs after 1 year of age, when a child begins to walk and run. In order of frequency, the joints most commonly affected are the knees, ankles, elbows, shoulders, and hips. Small joints of the hands and feet are occasionally involved.

In the initial stage of arthropathy, hemarthrosis produces a warm, tensely swollen, and painful joint. The patient holds the affected joint in flexion and guards against any movement. Blood in the joint remains liquid because of the absence of intrinsic clotting factors and the absence of tissue thromboplastin in the synovium. The synovial blood is resorbed over a period of ≥1 week, with the precise interval depending on the size of the hemarthrosis. Joint function usually returns to normal or baseline in ~2 weeks. Low-grade temperature elevation may accompany hemarthrosis, but a fever >38.3°C (101°F) warrants concern about infection.

Recurrent hemarthrosis may result in chronic arthritis. The involved joints remain swollen, and flexion deformities develop. Joint motion may be restricted and function severely limited. Restricted joint motion or laxity with subluxation is a feature of end-stage disease.

Bleeding into muscle and soft tissue also causes musculoskeletal dysfunction. When bleeding into the iliopsoas muscle occurs, the hip is held in flexion because of the pain, resulting in a hip flexion contracture. Rotation of the hip is preserved, which distinguishes this problem from hemarthrosis or other causes of hip synovitis. Expansion of the hematoma may place pressure on the femoral nerve, resulting in femoral neuropathy. Hemorrhage into a closed compartment space, such as the calf or the volar compartment in the forearm, can result in muscle necrosis, neuropathy, and flexion deformities of the ankles, wrists, and fingers. When bleeding involves periosteum or bone, a painful pseudotumor forms. These pseudotumors occur distal to the elbows or knees in children and improve with treatment of hemophilia. Surgical removal is indicated if the pseudotumor continues to enlarge. In adults, pseudotumors develop in the femur and pelvis and are usually refractory to treatment. When bleeding occurs in muscle, cysts may develop within the muscle. Needle aspiration of a cyst is contraindicated because this procedure can induce further bleeding; however, if the cyst becomes secondarily infected, drainage may be necessary (after factor repletion).

Septic arthritis is rare in hemophilia and is difficult to distinguish from acute hemarthrosis on physical examination. If there is serious suspicion of an infected joint, the joint should be aspirated immediately, the fluid cultured, and treatment with broad-spectrum antibiotics administered, with coverage for microorganisms including *Staphylococcus*, until culture results become available. Clotting-factor deficiency should be corrected before arthrocentesis to minimize the risk of traumatic bleeding.

Radiographs of joints reflect the stage of disease. In early stages, there is only capsule distention; later, juxtaarticular osteopenia, marginal erosions, and subchondral cysts develop. Late in the disease, the joint space is narrowed and there is bony overgrowth similar to that in osteoarthritis.

TREATMENT

Hemarthrosis

The treatment of musculoskeletal bleeding is initiated with the immediate infusion of factor VIII or IX at the first sign of joint or muscle hemorrhage. Patients who have developed factor inhibitors are at elevated risk for joint damage and may benefit from receiving recombinant activated factor VII or activated prothrombin complex concentrate. The joint should be rested in a position of forced extension, as tolerated, to avoid contracture. Analgesia should be provided; nonselective NSAIDs, which can diminish platelet function, should be avoided if possible. Selective cyclooxygenase-2 inhibitors do not interfere with platelet function, although cardiovascular and gastrointestinal risks must still be weighed. Synovectomy—open or arthroscopic—may be attempted in patients with chronic symptomatic synovial proliferation and recurrent hemarthrosis, although hypertrophied synovium is highly vascular and subject to bleeding. Both types of synovectomy reduce the number of hemarthroses. Open surgical synovectomy, however, is associated with some loss of range of motion. Both require aggressive prophylaxis against bleeding. Radiosynovectomy with either yttrium 90 silicate or phosphorus 31 colloid has been effective and may be attempted when surgical synovectomy is not practical. Total joint replacement is indicated for severe joint destruction and incapacitating pain.

ARTHROPATHIES ASSOCIATED WITH HEMOGLOBINOPATHIES

Sickle Cell Disease Sickle cell disease **(Chap. 94)** is associated with several musculoskeletal abnormalities **(Table 367-1)**. Children aged <5 years may develop diffuse swelling, tenderness, and warmth of the hands and feet lasting 1–3 weeks. This condition, referred to as *sickle cell dactylitis* or *hand-foot syndrome*, has also been observed in sickle cell thalassemia. Dactylitis is believed to result from infarction of the bone marrow and cortical bone leading to periostitis and soft tissue swelling. Radiographs show periosteal elevation, subperiosteal new-bone formation, and areas of radiolucency and increased density involving the metacarpals, metatarsals, and proximal phalanges. These bone changes disappear after several months. The syndrome leaves little or no residual damage. Because hematopoiesis ceases in the small bones of the hands and feet with age, the syndrome is rarely seen after age 5.

Sickle cell crisis is associated with periarticular pain and occasionally with joint effusions. The joint and periarticular area are warm and tender. Knees and elbows are most often affected, but other joints can be involved. Joint effusions are usually noninflammatory. Acute synovial infarction can cause a sterile effusion with high neutrophil counts in synovial fluid. Synovial biopsies have shown mild lining-cell proliferation and microvascular thrombosis with infarctions. Scintigraphic studies have shown decreased marrow uptake adjacent to the involved joint. The treatment for sickle cell crisis is detailed in **Chap. 94**.

Patients with sickle cell disease seem predisposed to osteomyelitis, which commonly involves the long tubular bones **(Chap. 126)**; *Salmonella* is a particularly common cause **(Chap. 160)**. Radiographs of the involved site initially show periosteal elevation, with subsequent disruption of the cortex. Treatment of the infection results in healing of the bone lesion. In addition, sickle cell disease is associated with bone infarction resulting from vasoocclusion secondary to the sickling of red cells. Bone infarction also occurs in hemoglobin sickle cell disease and sickle cell thalassemia **(Chap. 94)**. The bone pain in sickle cell crisis is due to infarction of bone and bone marrow. In children, infarction of the epiphyseal growth plate interferes with normal growth of the affected extremity. Radiographically, infarction of the bone cortex results in periosteal elevation and irregular thickening of the bone cortex. Infarction in the bone marrow leads to lysis, fibrosis, and new bone formation. Clinical distinction between osteomyelitis and bone infarctions can be difficult; imaging can be helpful.

Avascular necrosis of the head of the femur occurs in ~5% of patients. It also occurs in the humeral head and less commonly in the distal femur, tibial condyles, distal radius, vertebral bodies, and other juxtaarticular sites. Irregularity of the femoral head and other articular surfaces often results in degenerative joint disease. Radiography of the affected joint may show patchy radiolucency and density followed by flattening of the bone. MRI is a sensitive technique for detecting early avascular necrosis as well as bone infarction elsewhere. Total hip replacement and placement of prostheses in other joints may improve function and relieve joint pain in these patients.

TABLE 367-1 Musculoskeletal Abnormalities in Sickle Cell Disease

Sickle cell dactylitis	Avascular necrosis
Joint effusions in sickle cell crises	Bone changes secondary to marrow hyperplasia
Osteomyelitis	Septic arthritis
Infarction of bone	Gouty arthritis
Infarction of bone marrow	

Septic arthritis is occasionally encountered in sickle cell disease (Chap. 125). Multiple joints may be infected. Joint infection may result from bacteremia due to splenic dysfunction or from contiguous osteomyelitis. The more common microorganisms include *Staphylococcus aureus*, *Streptococcus*, and *Salmonella*. *Salmonella* does not cause septic arthritis as frequently as it causes osteomyelitis. Acute gouty arthritis is uncommon in sickle cell disease, even though 40% of patients are hyperuricemic. However, it may occur in patients generally not expected to get gout (young patients, female patients). Hyperuricemia is due to overproduction of uric acid secondary to increased red cell turnover as well as suboptimal renal excretion. Attacks may be polyarticular, and diagnostic arthrocentesis should be performed to distinguish infection from gout or synovial infarction.

The bone marrow hyperplasia in sickle cell disease results in widening of the medullary cavities, thinning of the cortices, and coarse trabeculations and central cupping of the vertebral bodies. These changes are also seen to a lesser degree in hemoglobin sickle cell disease and sickle cell thalassemia. In normal individuals red marrow is located mostly in the axial skeleton, but in sickle cell disease red marrow is found in the bones of the extremities and even in the tarsal and carpal bones. Vertebral compression may lead to dorsal kyphosis, and softening of the bone in the acetabulum may result in protrusio acetabuli.

Thalassemia A congenital disorder of hemoglobin synthesis, β thalassemia is characterized by impaired production of β chains (Chap. 94). Bone and joint abnormalities occur in β thalassemia, being most common in the major and intermedia groups. In one study, ~50% of patients with β thalassemia had evidence of symmetric ankle arthropathy characterized by a dull aching pain that was aggravated by weight bearing. The onset came most often in the second or third decade of life. The degree of ankle pain in these patients varied. Some patients experienced self-limited ankle pain that occurred only after strenuous physical activity and lasted several days or weeks. Other patients had chronic ankle pain that became worse with walking. Symptoms eventually abated in a few patients. Compression of the ankle, calcaneus, or forefoot was painful in some patients. Synovial fluid from two patients was noninflammatory. Radiographs of the ankle showed osteopenia, widened medullary spaces, thin cortices, and coarse trabeculations—findings that are largely the result of bone marrow expansion. The joint space was preserved. Specimens of bone from three patients revealed osteomalacia, osteopenia, and microfractures. Increased numbers of osteoblasts as well as increased foci of bone resorption were present on the bone surface. Iron staining was found in the bone trabeculae, in osteoid, and in the cement line. Synovium showed hyperplasia of lining cells, which contained deposits of hemosiderin. This arthropathy was considered to be related to the underlying bone pathology. The role of iron overload or abnormal bone metabolism in the pathogenesis of this arthropathy is not known. The arthropathy was treated with analgesics and splints. Patients also received transfusions to decrease hematopoiesis and bone marrow expansion.

In patients with β-thalassemia major and β-thalassemia intermedia, other joints are also involved, including the knees, hips, and shoulders. Acquired hemochromatosis with arthropathy has been described in a patient with thalassemia. Gouty arthritis and septic arthritis can occur. Avascular necrosis is not a feature of thalassemia because there is no sickling of red cells leading to thrombosis and infarction.

β-Thalassemia minor (also known as β-*thalassemia trait*) is likewise associated with joint manifestations. Chronic seronegative oligoarthritis affecting predominantly ankles, wrists, and elbows has been described; the affected patients had mild persistent synovitis without large effusions or joint erosions. Recurrent episodes of acute asymmetric arthritis have also been reported; episodes last <1 week and may affect the knees, ankles, shoulders, elbows, wrists, and metacarpal phalangeal joints. The mechanism underlying this arthropathy is unknown. Treatment with NSAIDs is not particularly effective.

MUSCULOSKELETAL DISORDERS ASSOCIATED WITH HYPERLIPIDEMIA

(See also Chap. 400) Musculoskeletal or cutaneous manifestations may be the first clinical indication of a specific hereditary disorder of lipoprotein metabolism. Patients with familial hypercholesterolemia (previously referred to as *type II hyperlipoproteinemia*) may have recurrent migratory polyarthritis involving the knees and other large peripheral joints and, to a lesser degree, peripheral small joints. Pain ranges from moderate to incapacitating. The involved joints can be warm, erythematous, swollen, and tender. Arthritis usually has a sudden onset, lasts from a few days to 2 weeks, and does not cause joint damage. Episodes may suggest acute gout attacks. Several attacks occur per year. Synovial fluid from involved joints is not inflammatory and contains few white cells and no crystals. Joint involvement may actually represent inflammatory periarthritis or peritendinitis and not true arthritis. The recurrent, transient nature of the arthritis may suggest rheumatic fever, especially because patients with hyperlipoproteinemia may have an elevated erythrocyte sedimentation rate and elevated antistreptolysin O titers (the latter being quite common). Attacks of tendinitis, including the large Achilles and patellar tendons, may come on gradually and last only a few days or may be acute as described above. Patients may be asymptomatic between attacks. Achilles tendinitis and other joint manifestations often precede the appearance of xanthomas and may be the first clinical indication of hyperlipoproteinemia. Attacks of tendinitis may follow treatment with a lipid-lowering drug. Over time, patients may develop tendinous xanthomas in the Achilles, patellar, and extensor tendons of the hands and feet. Xanthomas have also been reported in the peroneal tendon, the plantar aponeurosis, and the periosteum overlying the distal tibia. These xanthomas are located within tendon fibers. Tuberous xanthomas are soft subcutaneous masses located over the extensor surfaces of the elbows, knees, and hands as well as on the buttocks. They appear during childhood in homozygous patients and after the age of 30 in heterozygous patients. Patients with elevated plasma levels of very-low-density lipoprotein (VLDL) and triglycerides (previously referred to as *type IV hyperlipoproteinemia*) may also have a mild inflammatory arthritis affecting large and small peripheral joints, usually in an asymmetric pattern, with only a few joints involved at a time. The onset of arthritis usually comes in middle age. Arthritis may be persistent or recurrent, with episodes lasting a few days or weeks. Some patients may experience severe joint pain or morning stiffness. Joint tenderness and periarticular hyperesthesia may also be present, as may synovial thickening. Joint fluid is usually noninflammatory and without crystals but may have increased white blood cell counts with predominantly mononuclear cells. Radiographs may show juxtaarticular osteopenia and cystic lesions. Large bone cysts have been noted in a few patients. Xanthoma and bone cysts are also observed in other lipoprotein disorders. The pathogenesis of arthritis in patients with familial hypercholesterolemia or with elevated levels of VLDL and triglycerides is not well understood. NSAIDs or analgesics usually provide adequate relief of symptoms when used on an as-needed basis.

Patients may improve clinically as they are treated with lipid-lowering agents; however, patients treated with an HMG-CoA reductase inhibitor may experience myalgias, and a few patients develop myopathy, myositis, or even rhabdomyolysis. Patients who develop myositis during statin therapy may be susceptible to this adverse effect because of an underlying muscle disorder and should be reevaluated after discontinuation of the drug. Testing for anti-HMGCR autoantibodies in patients with elevated muscle enzymes on treatment may identify patients with statin-induced necrotizing autoimmune myopathy. Myositis has also been reported with the use of niacin (Chap. 358) but is less common than myalgias.

Musculoskeletal syndromes have not clearly been associated with the more common mixed hyperlipidemias seen in general practice.

OTHER ARTHRITIDES

NEUROPATHIC JOINT DISEASE

Neuropathic joint disease (Charcot joint) is a progressive destructive arthritis associated with loss of pain sensation, proprioception, or both. Normal muscular reflexes that modulate joint movement are impaired. Without these protective mechanisms, joints are subjected to repeated trauma, resulting in progressive cartilage and bone damage. Today,

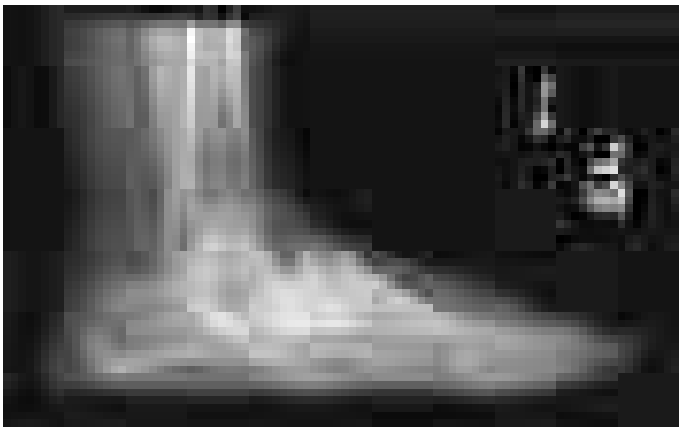


FIGURE 367-1 Charcot arthropathy associated with diabetes mellitus. Lateral foot radiograph demonstrating complete loss of the arch due to bony fragmentation and dislocation in the midfoot. *(Courtesy of Andrew Neckers, MD, and Jean Schils, MD; with permission.)*

diabetes mellitus is the most frequent cause of neuropathic joint disease **(Fig. 367-1)**. A variety of other disorders are associated with neuropathic arthritis, including tabes dorsalis, leprosy, yaws, syringomyelia, meningomyelocele, congenital indifference to pain, peroneal muscular atrophy (Charcot-Marie-Tooth disease), and amyloidosis. An arthritis resembling neuropathic joint disease has been reported in patients who have received intraarticular glucocorticoid injections, but this is a rare complication and was not observed in one series of patients with knee osteoarthritis who received intraarticular glucocorticoid injections every 3 months for 2 years. The distribution of joint involvement depends on the underlying neurologic disorder **(Table 367-2)**. In tabes dorsalis, the knees, hips, and ankles are most commonly affected; in syringomyelia, the glenohumeral joint, elbow, and wrist; and in diabetes mellitus, the tarsal and tarsometatarsal joints.

PATHOLOGY AND PATHOPHYSIOLOGY

The pathologic changes in the neuropathic joint are similar to those found in the severe osteoarthritic joint. There is fragmentation and eventual loss of articular cartilage with eburnation of the underlying bone. Osteophytes are found at the joint margins. With more advanced disease, erosions are present on the joint surface. Fractures, devitalized bone, intraarticular loose bodies, and microscopic fragments of cartilage and bone may be present.

At least two underlying mechanisms are believed to be involved in the pathogenesis of neuropathic arthritis. An abnormal autonomic nervous system is thought to be responsible for the dysregulated blood flow to the joint with subsequent resorption of bone. Loss of bone, particularly in the diabetic foot, may be the initial finding. With the loss of deep pain, proprioception, and protective neuromuscular reflexes, the joint is subjected to repeated microtrauma, resulting in ligamental tears and bone fractures. The injury that follows frequent intraarticular glucocorticoid injections is thought to be due to the analgesic effect of glucocorticoids, leading to overuse of an already damaged joint; the result is accelerated cartilage damage, although steroid-induced cartilage damage is more common in some other animal species than in humans. It is not understood why only a few patients with neuropathy develop clinically evident neuropathic arthritis.

CLINICAL MANIFESTATIONS

Neuropathic joint disease usually begins in a single joint and then becomes apparent in other joints, depending on the underlying neurologic disorder. The involved joint becomes progressively enlarged as a result of bony overgrowth and synovial effusion. Loose bodies may be palpated in the joint cavity. Joint instability, subluxation, and crepitus occur as the disease progresses. Neuropathic joints may develop rapidly, and a totally disorganized joint with multiple bony fragments may evolve within weeks or months. The amount of pain experienced by the patient is less than would be anticipated from the degree of joint damage. Patients may experience sudden joint pain from intraarticular fractures of osteophytes or condyles.

Neuropathic arthritis is encountered most often in patients with diabetes mellitus, with an incidence of ~0.5%. The onset of disease usually comes at an age of ≥50 years in a patient who has had diabetes for several years, but exceptions occur. The tarsal and tarsometatarsal joints are most often affected, with the metatarsophalangeal and talotibial joints next most commonly involved. The knees and spine are occasionally involved. Patients often attribute the onset of foot pain to antecedent trauma such as twisting of the foot. Neuropathic changes may develop rapidly after a foot fracture or dislocation. The foot and ankle are often swollen. Downward collapse of the tarsal bones leads to convexity of the sole, referred to as a "rocker foot." Large osteophytes may protrude from the top of the foot. Calluses frequently form over the metatarsal heads and may lead to infected ulcers and osteomyelitis. The value of protective inserts and orthotics, as well as regular foot examination, cannot be overstated. Radiographs may show resorption and tapering of the distal metatarsal bones. The term *Lisfranc fracture-dislocation* is sometimes used to describe the destructive changes at the tarsometatarsal joints.

TABLE 367-2 Disorders Associated with Neuropathic Joint Disease

Diabetes mellitus	Amyloidosis
Tabes dorsalis	Leprosy
Meningomyelocele	Congenital indifference to pain
Syringomyelia	Peroneal muscular atrophy

DIAGNOSIS

The diagnosis of neuropathic arthritis is based on the clinical features and characteristic radiographic findings in a patient with underlying sensory neuropathy. The differential diagnosis of neuropathic arthritis depends upon the severity of the process and includes osteomyelitis, avascular necrosis, advanced osteoarthritis, stress fractures, and calcium pyrophosphate deposition disease. Radiographs in neuropathic arthritis initially show changes of osteoarthritis with joint space narrowing, subchondral bone sclerosis, osteophytes, and joint effusions; marked destructive and hypertrophic changes follow later. The radiographic findings of neuropathic arthritis may be difficult to differentiate from those of osteomyelitis, especially in the diabetic foot. The joint margins in a neuropathic joint tend to be distinct, while in osteomyelitis they are blurred. Imaging studies may be helpful, but cultures of tissue from the joint are often required to exclude osteomyelitis. MRI and bone scans using indium 111–labeled white blood cells or indium 111–labeled immunoglobulin G, which will show increased uptake in osteomyelitis but not in a neuropathic joint, may be useful. A technetium bone scan will not distinguish osteomyelitis from neuropathic arthritis, as increased uptake is observed in both. The joint fluid in neuropathic arthritis is noninflammatory; it may be xanthochromic or even bloody and may contain fragments of synovium, cartilage, and bone. The finding of calcium pyrophosphate dihydrate crystals supports the diagnosis of crystal-associated arthropathy. In the absence of such crystals, an increased number of leukocytes may indicate osteomyelitis.

TREATMENT

Neuropathic Joint Disease

The primary focus of treatment is to stabilize the joint. Treatment of the underlying disorder, even if successful, does not usually affect established joint disease. Braces and splints are helpful. Their use requires close surveillance, because patients may be unable to appreciate pressure from a poorly adjusted brace. In the diabetic patient, early recognition of Charcot foot and its treatment—prohibition of weight bearing by the foot for at least 8 weeks—may possibly prevent severe disease from developing. Fusion of an unstable joint may improve function and reduce pain, but nonunion is frequent, especially when immobilization of the joint is inadequate.

HYPERTROPHIC OSTEOARTHROPATHY AND CLUBBING

Hypertrophic osteoarthropathy (HOA) is characterized by clubbing of digits and, in more advanced stages, by periosteal new-bone formation and synovial effusions. HOA may be primary or familial and may begin in childhood. Secondary HOA is associated with intrathoracic malignancies, suppurative and some hypoxemic lung diseases, congenital heart disease, and a variety of other disorders. Clubbing is almost always a feature of HOA but can occur as an isolated manifestation (Fig. 367-2). The presence of clubbing in isolation may be congenital or represent either an early stage or one element in the spectrum of HOA. Isolated acquired clubbing has the same clinical significance as clubbing associated with periostitis.

Pathology and Pathophysiology of Acquired HOA In HOA, bone changes in the distal extremities begin as periostitis followed by new bone formation. At this stage, a radiolucent area may be observed between the new periosteal bone and the subjacent cortex. As the process progresses, multiple layers of new bone are deposited and become contiguous with the cortex, with consequent cortical thickening. The outer portion of the bone is laminated in appearance, with an irregular surface. Initially, the process of periosteal new-bone formation involves the proximal and distal diaphyses of the tibia, fibula, radius, and ulna and, less frequently, the femur, humerus, metacarpals, metatarsals, and phalanges. Occasionally, scapulae, clavicles, ribs, and pelvic bones are also affected. The adjacent interosseous membranes may become ossified. The distribution of bone manifestations is usually bilateral and symmetric. The soft tissue overlying the distal third of the arms and legs may be thickened. Proliferation of connective tissue occurs in the nail bed and volar pad of digits, giving the distal phalanges a clubbed appearance. Small blood vessels in the clubbed digits are dilated and have thickened walls. In addition, the number of arteriovenous anastomoses is increased.

Several theories have been suggested for the pathogenesis of HOA, but many have been disproved or have not explained the condition's development in all clinical disorders with which it is associated. Previously proposed neurogenic and humoral theories are no longer considered likely explanations for HOA. Studies have suggested a role for platelets in the development of HOA. It has been observed that megakaryocytes and large platelet particles present in the venous circulation are fragmented in their passage through normal lung. In patients with cyanotic congenital heart disease and in other disorders associated with right-to-left shunts, these large platelet particles bypass the lung and reach the distal extremities, where they can interact with endothelial cells. Platelet–endothelial cell activation in the distal portion of the extremities may result in the release of platelet-derived growth factor (PDGF) and other factors leading to the proliferation of connective tissue and periosteum. Stimulation of fibroblasts by PDGF and transforming growth factor β results in cell growth and collagen synthesis. Elevated plasma levels of von Willebrand factor antigen have been found in patients with both primary and secondary forms of HOA, indicating endothelial activation or damage. Abnormalities of collagen synthesis have been demonstrated in the involved skin of patients with primary HOA. Other factors are undoubtedly involved in the pathogenesis of HOA, and further studies are needed to elucidate this disorder.

FIGURE 367-2 Clubbing of the fingers. *(Reprinted from the Clinical Slide Collection on the Rheumatic Diseases, © 2018 American College of Rheumatology. Used by permission of the American College of Rheumatology.)*

Clinical Manifestations Primary or familial HOA, also referred to as *pachydermoperiostitis* or *Touraine-Solente-Golé syndrome*, usually begins insidiously at puberty. In a smaller proportion of patients, the onset comes in the first year of life. The disorder is inherited as an autosomal dominant trait with variable expression and is nine times more common among boys than among girls. Approximately one-third of patients have a family history of primary HOA.

Primary HOA is characterized by clubbing, periostitis, and unusual skin features. A small number of patients with this syndrome do not express clubbing. The skin changes and periostitis are prominent features of this syndrome. The skin becomes thickened and coarse. Deep nasolabial folds develop, and the forehead may become furrowed. Patients may have heavy-appearing eyelids and ptosis. The skin is often greasy, and there may be excessive sweating of the hands and feet. Patients may also experience acne vulgaris, seborrhea, and folliculitis. In a few patients, the skin over the scalp becomes very thick and corrugated, a feature that has been descriptively termed *cutis verticis gyrata*. The distal extremities, particularly the legs, become thickened as a consequence of the proliferation of new bone and soft tissue; when the process is extensive, the distal lower extremities resemble those of an elephant. The periostitis usually is not painful, which it can be in secondary HOA. Clubbing of the fingers may be extensive, producing large, bulbous deformities, and clumsiness. Clubbing also affects the toes. Patients may experience articular and periarticular pain, especially in the ankles and knees, and joint motion may be mildly restricted by periarticular bone overgrowth. Noninflammatory effusions occur in the wrists, knees, and ankles. Synovial hypertrophy is not found. Associated abnormalities observed in patients with primary HOA include hypertrophic gastropathy, bone marrow failure, female escutcheon, gynecomastia, and cranial suture defects. In patients with primary HOA, the symptoms disappear when adulthood is reached.

HOA secondary to an underlying disease occurs more frequently than primary HOA. It accompanies a variety of disorders and may precede clinical features of the associated disorder by months. Clubbing is more frequent than the full syndrome of HOA in patients with associated illnesses. Because clubbing evolves over months and is usually asymptomatic, it is often recognized first by the physician and not the patient. Patients may experience a burning sensation in their fingertips. Clubbing is characterized by widening of the fingertips, enlargement of the distal volar pad, convexity of the nail contour, and the loss of the normal 15° angle between the proximal nail and cuticle. The thickness of the digit at the base of the nail is greater than the thickness at the distal interphalangeal joint. An objective measurement of finger clubbing can be made by determining the diameter at the base of the nail and at the distal interphalangeal joint of all 10 digits. Clubbing is present when the sum of the individual digit ratios is >10. At the bedside, clubbing can be appreciated by having the patient place the dorsal surface of the distal phalanges of the fourth fingers together with the nails opposing each other. Normally, an open area is visible between the bases of the opposing fingernails; when clubbing is present, this open space is no longer visible. The base of the nail feels spongy when compressed, and the nail can be easily rocked on its bed. When clubbing is advanced, the finger may have a drumstick appearance, and the distal interphalangeal joint can be hyperextended. Periosteal involvement in the distal extremities may produce a burning or deep-seated aching pain. The pain, which can be quite incapacitating, is aggravated by dependency and relieved by elevation of the affected limbs. Pressure applied over the distal forearms and legs or gentle percussion of distal long bones like the tibia may be quite painful.

Patients may experience joint pain, most often in the ankles, wrists, and knees. Joint effusions may be present; usually, they are small and noninflammatory. The small joints of the hands are rarely affected.

TABLE 367-3 Disorders Associated with Hypertrophic Osteoarthropathy

Pulmonary
Bronchogenic carcinoma and other neoplasms
Lung abscesses, empyema, bronchiectasis
Chronic interstitial pneumonitis
Cystic fibrosis
Sarcoidosis
Gastrointestinal
Inflammatory bowel disease
Sprue
Neoplasms: esophagus, liver, bowel
Cardiovascular
Cyanotic congenital heart disease
Subacute bacterial endocarditis
Infected arterial grafts[a]
Aortic aneurysm[b]
Aneurysm of major extremity artery[a]
Patent ductus arteriosus[b]
Arteriovenous fistula of major extremity vessel[a]
Thyroid (thyroid acropachy)
Hyperthyroidism (Graves' disease)

[a]Unilateral involvement. [b]Bilateral lower-extremity involvement.

Severe joint or long bone pain may be the presenting symptom of an underlying lung malignancy and may precede the appearance of clubbing. In addition, the progression of HOA tends to be more rapid when associated with malignancies, most notably bronchogenic carcinoma. Noninflammatory but variably painful knee effusions may occur prior to the appearance of clubbing and symptoms of distal periostitis. Unlike primary HOA, secondary HOA does not commonly include excessive sweating and oiliness of the skin or thickening of the facial skin.

HOA occurs in 5–10% of patients with intrathoracic malignancies, the most common being bronchogenic carcinoma and pleural tumors **(Table 367-3)**. Lung metastases infrequently cause HOA. HOA is also seen in patients with intrathoracic infections, including lung abscesses, empyema, and bronchiectasis, but is uncommon in pulmonary tuberculosis. HOA may accompany chronic interstitial pneumonitis, sarcoidosis, and cystic fibrosis. In cystic fibrosis, clubbing is more common than the full syndrome of HOA. Other causes of clubbing include congenital heart disease with right-to-left shunts, bacterial endocarditis, Crohn's disease, ulcerative colitis, sprue, and neoplasms of the esophagus, liver, and small and large bowel. In patients who have congenital heart disease with right-to-left shunts, clubbing alone occurs more often than the full syndrome of HOA.

Unilateral clubbing has been found in association with aneurysms of major extremity arteries, with infected arterial grafts, and with arteriovenous fistulas of brachial vessels. Clubbing of the toes but not the fingers has been associated with an infected abdominal aortic aneurysm and patent ductus arteriosus. Clubbing of a single digit may follow trauma and has been reported in tophaceous gout and sarcoidosis. While clubbing occurs more commonly than the full syndrome in most diseases, periostitis in the absence of clubbing has been observed in the affected limb of patients with infected arterial grafts.

Hyperthyroidism (Graves' disease), treated or untreated, is occasionally associated with clubbing and periostitis of the bones of the hands and feet. This condition is referred to as *thyroid acropachy*. Periostitis may be asymptomatic and occurs in the midshaft and diaphyseal portion of the metacarpal and phalangeal bones. Significant hand-joint pain may occur; this pain may respond to successful therapy for thyroid dysfunction. The long bones of the extremities are seldom affected. Elevated levels of long-acting thyroid stimulator are found in the sera of these patients.

Laboratory Findings The laboratory abnormalities reflect the underlying disorder. The synovial fluid of involved joints has <500 white cells/μL, and the cells are predominantly mononuclear. Radiographs show a faint radiolucent line beneath the new periosteal bone along the shaft of long bones at their distal end. These changes are observed most frequently at the ankles, wrists, and knees. The ends of the distal phalanges may show osseous resorption. Radionuclide studies show pericortical linear uptake along the cortical margins of long bones that may precede any radiographic changes.

TREATMENT

Hypertrophic Osteoarthropathy

The treatment of HOA aims to identify the associated disorder and treat it appropriately. The symptoms and signs of HOA may disappear completely with removal of or effective chemotherapy for a tumor or with antibiotic therapy for a chronic pulmonary infection and drainage of the infected site. Vagotomy or percutaneous block of the vagus nerve leads to symptomatic relief in some patients. NSAIDs or analgesics may help control symptoms of HOA.

REFLEX SYMPATHETIC DYSTROPHY SYNDROME

The reflex sympathetic dystrophy syndrome is now referred to as *complex regional pain syndrome, type 1*, according to the new classification system of the International Association for the Study of Pain. This syndrome is characterized by pain and swelling, usually of a distal extremity, accompanied by vasomotor instability, trophic skin changes, and the rapid development of bony demineralization. Reflex sympathetic dystrophy syndrome, including its treatment, is covered in greater detail in **Chap. 432**.

TIETZE SYNDROME AND COSTOCHONDRITIS

Tietze syndrome is manifested by painful swelling of one or more costochondral articulations. The age of onset is usually before 40, and both sexes are affected equally. In most patients, only one joint is involved, usually the second or third costochondral joint. The onset of anterior chest pain may be sudden or gradual. The pain may radiate to the arms or shoulders and is aggravated by sneezing, coughing, deep inspirations, or twisting motions of the chest. The term *costochondritis* is often used interchangeably with *Tietze syndrome*, but some workers restrict the former term to pain of the costochondral articulations without swelling. Costochondritis is observed in patients aged >40 years; it tends to affect the third, fourth, and fifth costochondral joints, and occurs more often in women. Both syndromes may mimic cardiac or upper abdominal causes of pain. Rheumatoid arthritis, ankylosing spondylitis, and reactive arthritis may involve costochondral joints but are distinguished easily by their other clinical features. Other skeletal causes of anterior chest wall pain are xiphoidalgia and the slipping rib syndrome, which usually involves the tenth rib. Malignancies such as breast cancer, prostate cancer, plasma cell cytoma, and sarcoma can invade the ribs, thoracic spine, or chest wall and produce symptoms suggesting Tietze's syndrome. Patients with osteomalacia may have significant rib pain, with or without documented microfractures. These conditions should be distinguishable by radiography, bone scanning, vitamin D measurement, or biopsy. Analgesics, anti-inflammatory drugs, and local glucocorticoid injections usually relieve symptoms of costochondritis/Tietze's syndrome. Care should be taken to avoid overdiagnosing these syndromes in patients with acute chest pain syndromes; many patients will be tender to overly vigorous palpation of the costochondral joints.

MYOFASCIAL PAIN SYNDROME

Myofascial pain syndrome is characterized by multiple areas of localized musculoskeletal pain and tenderness in association with tender points. The pain is deep and aching and may be accompanied by a burning sensation. Myofascial pain may be regional and follow trauma, overuse, or prolonged static contraction of a muscle or muscle group, which may occur when an individual is reading or writing at a desk or working at a computer. In addition, this syndrome may be associated with underlying osteoarthritis of the neck or low back. Pain may be

referred from tender points to defined areas distant from the area of original tenderness. Palpation of the tender point reproduces or accentuates the pain. The tender points are usually located in the center of a muscle belly, but they can occur at other sites such as costosternal junctions, the xiphoid process, ligamentous and tendinous insertions, fascia, and fatty areas. Tender point sites in muscle have been described as feeling indurated and taut, and palpation may cause the muscle to twitch. These findings, however, have been shown not to be unique to myofascial pain syndrome: in a controlled study, they were also present in some "normal" subjects. Myofascial pain most often involves the posterior neck, low back, shoulders, and chest. Chronic pain in the muscles of the posterior neck may involve referral of pain from a tender point in the erector neck muscle or upper trapezius to the head, leading to persistent headaches that may last for days. Tender points in the paraspinal muscles of the low back may refer pain to the buttock. Pain may be referred down the leg from a tender point in the gluteus medius and can mimic sciatica. A tender point in the infraspinatus muscle may produce local and referred pain over the lateral deltoid and down the outside of the arm into the hand. Injection of a local anesthetic such as 1% lidocaine into the tender point site often results in at least transient pain relief. Another useful technique is first to spray an agent such as ethyl chloride from the tender point toward the area of referred pain and then to stretch the muscle. This maneuver may need to be repeated several times. Massage and application of ultrasound to the affected area also may be beneficial. Patients should be instructed in methods to prevent muscle stresses related to work and recreation. Posture and resting positions are important in preventing muscle tension. The prognosis in most patients is good. In some patients, regionally localized myofascial pain syndrome may seem to evolve into more generalized fibromyalgia **(Chap. 366)**. Abnormal or nonrestorative sleep is a common accompaniment in these patients and may need to be specifically addressed.

NEOPLASIAS AND ARTHRITIS

Primary tumors and tumor-like disorders of synovium are uncommon but should be considered in the differential diagnosis of monarticular joint disease. In addition, metastases to bone and primary bone tumors adjacent to a joint may produce joint symptoms.

Pigmented villonodular synovitis (*PVNS*) is characterized by the slowly progressive, exuberant, benign proliferation of synovial tissue, usually involving a single joint. The most common age of onset is in the third decade, and women are affected slightly more often than men. The cause of this disorder is unknown.

The synovium has a brownish color and numerous large, finger-like villi that fuse to form pedunculated nodules. There is marked hyperplasia of synovial cells in the stroma of the villi. Hemosiderin granules and lipids are found in the cytoplasm of macrophages and in the interstitial tissue. Multinucleated giant cells may be present. The proliferative synovium grows into the subsynovial tissue and invades adjacent cartilage and bone.

The clinical picture of PVNS is characterized by the insidious onset of persistent swelling and pain in affected joints, most commonly the knee. Other joints affected include the hips, ankles, calcaneocuboid joints, elbows, and small joints of the fingers or toes. The disease may also involve the common flexor sheath of the hands or fingers. Less often, tendon sheaths in the wrist, ankle, or foot may be involved. Symptoms of pain, a catching sensation, or stiffness may initially be mild and intermittent and may be present for years before the patient seeks medical attention. Radiographs may show joint space narrowing, erosions, and subchondral cysts. The diagnosis of PVNS is strongly suggested by gradient echo MRI, which reveals a synovial mass lesion of low signal intensity typical of tissue containing hemosiderin **(Fig. 367-3)**. The joint fluid contains blood and is dark red or almost black in color. Lipid-containing macrophages may be present in the fluid. The joint fluid may be clear if hemorrhage has not occurred. Some patients have polyarticular involvement.

The treatment for PVNS is complete synovectomy. With incomplete synovectomy, the villonodular synovitis recurs, and the rate of tissue growth may be faster than it was originally. Irradiation of the involved joint has been successful in some patients.

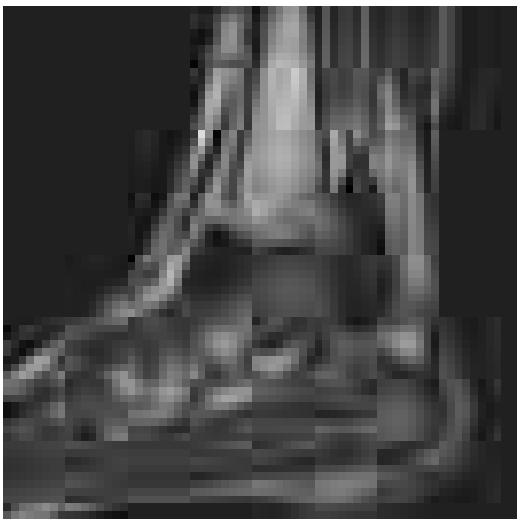


FIGURE 367-3 Pigmented villonodular synovitis. MRI gradient echo sagittal image showing a mass that abuts the neck of the talus with marked low signal typical of tissue containing hemosiderin. *(Courtesy of Donald Flemming, MD; with permission.)*

Synovial chondromatosis is a disorder characterized by multiple focal metaplastic growths of normal-appearing cartilage in the synovium or tendon sheath. Segments of cartilage break loose and continue to grow as loose bodies. When calcification and ossification of loose bodies occur, the disorder is referred to as *synovial osteochondromatosis*. The disorder is usually monarticular and affects young to middle-aged individuals. The knee is most often involved, followed by hip, elbow, and shoulder. Symptoms are pain, swelling, and decreased motion of the joint. Radiographs may show several rounded calcifications within the joint cavity. Treatment is synovectomy; however, as in PVNS, the tumor may recur.

Synovial sarcoma is a malignant neoplasm often found near a large joint of both upper and lower extremities, being more common in the lower extremity. It seldom arises within the joint itself. Synovial sarcomas constitute 10% of soft tissue sarcomas. The tumor is believed to arise from primitive mesenchymal tissue that differentiates into epithelial cells and/or spindle cells. Small foci of calcification may be present in the tumor mass. Synovial sarcoma occurs most often in young adults and is more common in men. The tumor presents as a slowly growing deep-seated mass near a joint, without much pain. The area of the knee is the most common site, followed by the foot, ankle, elbow, and shoulder. Other primary sites include the buttocks, abdominal wall, retroperitoneum, and mediastinum. The tumor spreads along tissue planes. The most common site of visceral metastasis is the lung. The diagnosis is made by biopsy. Treatment consists of wide resection of the tumor, including adjacent muscle and regional lymph nodes, followed by chemotherapy and radiation therapy. Amputation of the involved distal extremity may be required. Chemotherapy may be beneficial in some patients with metastatic disease. Isolated sites of pulmonary metastasis can be surgically removed. The 5-year survival rate with treatment is variable and depends on the staging of the tumor, ranging from ~25% to ≥60%. Synovial sarcomas tend to recur locally and metastasize to regional lymph nodes, lungs, and skeleton.

In addition to the rare direct metastases of solid cell tumors to the highly vascular synovium, neoplasia arising from nonarticular organ sites can affect joints in other ways. Acute leukemias in children can mimic juvenile inflammatory arthritis with severe joint pain and fever. In adults, chronic and acute myeloid leukemia can infiltrate the synovium in rare instances. The rarely occurring hairy cell leukemia has a peculiar tendency to cause episodic inflammatory oligoarthritis and tenosynovitis; these episodes are dramatic and mimic acute gout attacks. They respond to potent anti-inflammatory therapy with glucocorticoids; with remission of the leukemia, they may abate. Carcinomas can be associated with several paraneoplastic articular syndromes, including HOA (discussed above). Acute palmar fasciitis

with polyarthritis is a well-described but rare condition associated with certain cancers, mainly adenocarcinomas. Clinically, this syndrome is fairly abrupt in onset, with pain in the metacarpophalangeal and proximal interphalangeal joints of the hands and rapidly evolving contractures of the fingers due to thickening of the palmar (flexor) tendons. A similar syndrome can be seen in diabetics. Paraneoplastic arthritis has been described and may occur in several patterns: asymmetric disease predominantly affecting the lower extremity joints and symmetric polyarthritis with hand joint involvement. Tumors are often found after the onset of the arthritis, and many patients have a preceding period of malaise or weight loss. The onset is often acute, and patients tend to be older men. These features should raise the specter of an underlying malignancy (or a viral infection such as hepatitis C) as the cause of the arthritis. In one series, the symptoms resolved with successful therapy for the malignancy and did not recur with relapse of the malignancy. Dermatomyositis has a well-described association with neoplasms and may include joint pain and arthritis. Malignancy-associated arthritis may be responsive to NSAIDs and to treatment of the primary neoplasm.

Acknowledgment

This chapter represents a revised version of the chapter authored by Dr. Bruce C. Gilliland that appeared in previous editions of Harrison's. Dr. Gilliland passed away on February 17, 2007. He had been a contributor to Harrison's Principles of Internal Medicine since the 11th edition.

FURTHER READING

Aguilar C et al: Bone and joint disease in sickle cell disease. Hematol Oncol Clin North Am 19:929, 2005.

Botek G et al: Charcot neuroarthropathy: An often overlooked complication of diabetes. Cleve Clin J Med 77:593, 2010.

Guggenbuhl P et al: Miscellaneous non-inflammatory musculoskeletal conditions. Haemochromatosis: The bone and the joint. Best Pract Res Clin Rheumatol 25:649, 2011.

Kedar E, Gardner GC: Lipid-associated rheumatologic syndromes. Rheum Dis Clin North Am 39:481, 2013.

Killinger Z et al: Arthropathy in acromegaly. Rheum Dis Clin North Am 36:713, 2010.

Ottaviani S et al: Pigmented villonodular synovitis: A retrospective single-center study of 122 cases and review of the literature. Semin Arthritis Rheum 40:539, 2011.

Pineda C, Martínez-Lavín M: Hypertrophic osteoarthropathy: What a rheumatologist should know about this uncommon condition. Rheum Dis Clin North Am 39:383, 2013.

Vanderhave KL et al: Musculoskeletal care of the hemophiliac patient. J Am Acad Orthop Surg 20:553, 2012.

368 Periarticular Disorders of the Extremities

Carol A. Langford

Periarticular disorders are common musculoskeletal abnormalities that can affect people throughout a wide range of ages. This chapter discusses some of the more common periarticular disorders.

BURSITIS

Bursitis is inflammation of a bursa, which is a thin-walled sac lined with synovial tissue. The function of the bursa is to facilitate movement of tendons and muscles over bony prominences. Excessive frictional forces from overuse, trauma, systemic disease (e.g., rheumatoid arthritis, gout), or infection may cause bursitis. *Subacromial bursitis* (subdeltoid bursitis) is the most common form of bursitis. The subacromial bursa, which is contiguous with the subdeltoid bursa, is located between the undersurface of the acromion and the humeral head and is covered by the deltoid muscle. Bursitis often accompanies rotator cuff tendinitis. Another frequently encountered form is *trochanteric bursitis*, which involves the bursa around the insertion of the gluteus medius onto the greater trochanter of the femur. Patients experience pain over the lateral aspect of the hip and upper thigh and have tenderness over the posterior aspect of the greater trochanter. External rotation and resisted abduction of the hip elicit pain as will direct pressure applied to the bursa. *Olecranon bursitis* occurs over the posterior elbow, and when the area is acutely inflamed, infection or gout should be excluded by aspirating the bursa and performing a Gram stain and culture on the fluid as well as examining the fluid for urate crystals. *Achilles bursitis* involves the bursa located above the insertion of the tendon to the calcaneus and results from overuse and wearing tight shoes. *Retrocalcaneal bursitis* involves the bursa that is located between the calcaneus and posterior surface of the Achilles tendon. The pain is experienced at the back of the heel, and swelling appears on the medial and/or lateral side of the tendon. It occurs in association with spondyloarthritides, rheumatoid arthritis, gout, or trauma. *Ischial bursitis* affects the bursa separating the gluteus medius from the ischial tuberosity and develops from prolonged sitting and pivoting on hard surfaces. *Iliopsoas bursitis* affects the bursa that lies between the iliopsoas muscle and hip joint and is lateral to the femoral vessels. Pain is experienced over this area and is made worse by hip extension and flexion. *Anserine bursitis* is an inflammation of the sartorius bursa located over the medial side of the tibia just below the knee and under the conjoint tendon and is manifested by pain on climbing stairs. Tenderness is present over the insertion of the conjoint tendon of the sartorius, gracilis, and semitendinosus. *Prepatellar bursitis* occurs in the bursa situated between the patella and overlying skin and is caused by kneeling on hard surfaces. Gout or infection may also occur at this site. Bursitis is typically diagnosed by history and physical examination, but visualization by ultrasound may play a useful role in selected instances for diagnosis and directed guidance of glucocorticoid injection. Treatment of bursitis consists of prevention of any aggravating situation, rest of the involved part, administration of a nonsteroidal anti-inflammatory drug (NSAID) where appropriate for an individual patient, or local glucocorticoid injection.

ROTATOR CUFF TENDINITIS AND IMPINGEMENT SYNDROME

Tendinitis of the rotator cuff is the major cause of a painful shoulder and is currently thought to be caused by inflammation of the tendon(s). The rotator cuff consists of the tendons of the supraspinatus, infraspinatus, subscapularis, and teres minor muscles, and inserts on the humeral tuberosities. Of the tendons forming the rotator cuff, the supraspinatus tendon is the most often affected, probably because of its repeated impingement (*impingement syndrome*) between the humeral head and the undersurface of the anterior third of the acromion and coracoacromial ligament above as well as the reduction in its blood supply that occurs with abduction of the arm **(Fig. 368-1)**. The tendon of the infraspinatus and that of the long head of the biceps are less commonly involved. Subacromial bursitis also accompanies this syndrome. Symptoms can appear without a triggering cause or after injury or overuse, especially with activities involving elevation of the arm with some degree of forward flexion. Impingement syndrome occurs in persons participating in baseball, tennis, swimming, or occupations that require repeated elevation of the arm. Those aged >40 years are particularly susceptible. Patients complain of a dull aching in the shoulder, which may interfere with sleep. Severe pain is experienced when the arm is actively abducted into an overhead position. The arc between 60° and 120° is especially painful. Tenderness is present over the lateral aspect of the humeral head just below the acromion. NSAIDs, local glucocorticoid injection, and physical therapy may relieve symptoms. Surgical decompression of the subacromial space may be necessary in patients refractory to conservative treatment.

Patients may tear the supraspinatus tendon acutely by falling on an outstretched arm or lifting a heavy object. Symptoms are pain along with weakness of abduction and external rotation of the shoulder. Atrophy of the supraspinatus muscles develops. The diagnosis

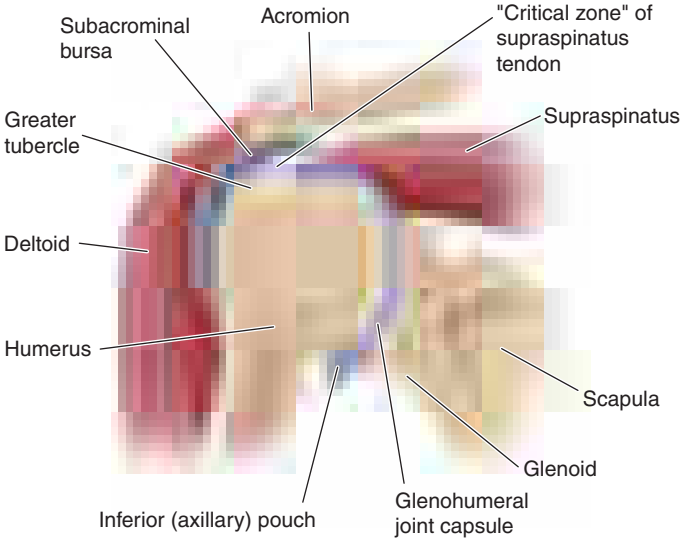


FIGURE 368-1 Coronal section of the shoulder illustrating the relationships of the glenohumeral joint, the joint capsule, the subacromial bursa, and the rotator cuff (supraspinatus tendon). *(From F Kozin, in Arthritis and Allied Conditions, 13th ed, WJ Koopman [ed]. Baltimore, Williams & Wilkins, 1997, with permission.)*

is established by ultrasound, magnetic resonance imaging (MRI), or arthrogram. Surgical repair may be necessary in patients who fail to respond to conservative measures. In patients with moderate-to-severe tears and functional loss, surgery is indicated.

CALCIFIC TENDINITIS

This condition is characterized by deposition of calcium salts, primarily hydroxyapatite, within a tendon. The exact mechanism of calcification is not known but may be initiated by ischemia or degeneration of the tendon. The supraspinatus tendon is most often affected because it is frequently impinged on and has a reduced blood supply when the arm is abducted. The condition usually develops after age 40. Calcification within the tendon may evoke acute inflammation, producing sudden and severe pain in the shoulder. However, it may be asymptomatic or not related to the patient's symptoms. Diagnosis of calcific tendonitis can be made by ultrasound or radiograph. Most cases are self-limited and respond to conservative therapy with physical therapy and/or NSAIDs. A subset of patients is refractory and requires ultrasound-guided percutaneous needle aspiration and lavage or surgery.

BICIPITAL TENDINITIS AND RUPTURE

Bicipital tendinitis, or tenosynovitis, is produced by friction on the tendon of the long head of the biceps as it passes through the bicipital groove. When the inflammation is acute, patients experience anterior shoulder pain that radiates down the biceps into the forearm. Abduction and external rotation of the arm are painful and limited. The bicipital groove is very tender to palpation. Pain may be elicited along the course of the tendon by resisting supination of the forearm with the elbow at 90° (Yergason's supination sign). Acute rupture of the tendon may occur with vigorous exercise of the arm and is often painful. In a young patient, it should be repaired surgically. Rupture of the tendon in an older person may be associated with little or no pain and is recognized by the presence of persistent swelling of the biceps produced by the retraction of the long head of the biceps. Surgery is usually not necessary in this setting.

DE QUERVAIN'S TENOSYNOVITIS

In this condition, inflammation involves the abductor pollicis longus and the extensor pollicis brevis as these tendons pass through a fibrous sheath at the radial styloid process. The usual cause is repetitive twisting of the wrist. It may occur in pregnancy, and it also occurs in mothers who hold their babies with the thumb outstretched. Patients experience pain on grasping with their thumb, such as with pinching. Swelling and tenderness are often present over the radial styloid process. The Finkelstein sign is positive, which is elicited by having the patient place the thumb in the palm and close the fingers over it. The wrist is then ulnarly deviated, resulting in pain over the involved tendon sheath in the area of the radial styloid. Treatment consists initially of splinting the wrist and an NSAID. When severe or refractory to conservative treatment, glucocorticoid injections can be very effective.

PATELLAR TENDINITIS

Tendinitis involves the patellar tendon at its attachment to the lower pole of the patella. Patients may experience pain when jumping during sports, going up stairs, or doing deep knee squats. Tenderness is noted on examination over the lower pole of the patella. Treatment consists of rest, icing, and NSAIDs, followed by strengthening and increasing flexibility.

DRUG-INDUCED TENDINOPATHIES

With the broadening range of available pharmacologic agents, the potential for drug-induced tendinopathies has become increasingly recognized. The drug classes most associated with tendinopathies include quinolones, glucocorticoids, aromatase inhibitors, and statins. Although any tendon can be affected, the tendons of the lower extremities are most often impacted, particularly the Achilles tendon. The pathophysiological mechanisms responsible for drug-induced tendinopathies remain unknown. Presenting features include pain and potentially swelling over the tendon, although some patients may first present with tendon rupture. Ultrasound and MRI can provide information on tendon structure and integrity in support of the diagnosis. When suspected, the potential agent should be withdrawn and not reintroduced where possible in the overall medical management of the patient. Tendon ruptures may require surgery.

ILIOTIBIAL BAND SYNDROME

The iliotibial band is a thick connective tissue that runs from the ilium to the fibula. Patients with iliotibial band syndrome most commonly present with aching or burning pain at the site where the band courses over the lateral femoral condyle of the knee; pain may also radiate up the thigh, toward the hip. Predisposing factors for iliotibial band syndrome include a varus alignment of the knee, excessive running distance, poorly fitted shoes, or continuous running on uneven terrain. Treatment consists of rest, NSAIDs, physical therapy, and addressing risk factors such as shoes and running surface. Glucocorticoid injection into the area of tenderness can provide relief, but running must be avoided for at least 2 weeks after the injection. Surgical release of the iliotibial band has been helpful in rare patients for whom conservative treatment has failed.

ADHESIVE CAPSULITIS

Often referred to as "frozen shoulder," adhesive capsulitis is characterized by pain and restricted movement of the shoulder, usually in the absence of intrinsic shoulder disease. Adhesive capsulitis may follow bursitis or tendinitis of the shoulder or be associated with systemic disorders such as chronic pulmonary disease, myocardial infarction, and diabetes mellitus. Prolonged immobility of the arm contributes to the development of adhesive capsulitis. Pathologically, the capsule of the shoulder is thickened, and a mild chronic inflammatory infiltrate and fibrosis may be present.

Adhesive capsulitis occurs more commonly in women aged >50 years. Pain and stiffness usually develop gradually but progress rapidly in some patients. Night pain is often present in the affected shoulder, and pain may interfere with sleep. The shoulder is tender to palpation, and both active and passive movements are restricted. Radiographs of the shoulder show osteopenia. The diagnosis is typically made by physical examination but can be confirmed if necessary by arthrography, in that only a limited amount of contrast material, usually <15 mL, can be injected under pressure into the shoulder joint.

In most patients, the condition improves spontaneously 1–3 years after onset. While pain usually improves, many patients are left with some limitation of shoulder motion. Early mobilization of the arm following an injury to the shoulder may prevent the development of

this disease. Physical therapy provides the foundation of treatment for adhesive capsulitis. Local injections of glucocorticoids and NSAIDs may also provide relief of symptoms. Slow but forceful injection of contrast material into the joint may lyse adhesions and stretch the capsule, resulting in improvement of shoulder motion. Manipulation under anesthesia may be helpful in some patients.

LATERAL EPICONDYLITIS

Lateral epicondylitis, also known as tennis elbow, is a painful condition involving the soft tissue over the lateral aspect of the elbow. The pain originates at or near the site of attachment of the common extensors to the lateral epicondyle and may radiate into the forearm and dorsum of the wrist. The pain usually appears after work or recreational activities involving repeated motions of wrist extension and supination against resistance. Most patients with this disorder injure themselves in activities other than tennis, such as pulling weeds, carrying suitcases or briefcases, or using a screwdriver. The injury in tennis usually occurs when hitting a backhand with the elbow flexed. Shaking hands and opening doors can reproduce the pain. Striking the lateral elbow against a solid object may also induce pain.

The treatment is usually rest along with administration of an NSAID. Ultrasound, icing, and friction massage may also help relieve pain. When pain is severe, the elbow is placed in a sling or splinted at 90° of flexion. When the pain is acute and well localized, injection of a glucocorticoid using a small-gauge needle may be effective. Following injection, the patient should be advised to rest the arm for at least 1 month and avoid activities that would aggravate the elbow. Once symptoms have subsided, the patient should begin rehabilitation to strengthen and increase flexibility of the extensor muscles before resuming physical activity involving the arm. A forearm band placed 2.5–5.0 cm (1–2 in.) below the elbow may help to reduce tension on the extensor muscles at their attachment to the lateral epicondyle. The patient should be advised to restrict activities requiring forcible extension and supination of the wrist. Improvement may take several months. The patient may continue to experience mild pain but, with care, can usually avoid the return of debilitating pain. Occasionally, surgical release of the extensor aponeurosis may be necessary.

MEDIAL EPICONDYLITIS

Medial epicondylitis is an overuse syndrome resulting in pain over the medial side of the elbow with radiation into the forearm. The cause of this syndrome is considered to be repetitive resisted motions of wrist flexion and pronation, which lead to microtears and granulation tissue at the origin of the pronator teres and forearm flexors, particularly the flexor carpi radialis. This overuse syndrome is usually seen in patients aged >35 years and is much less common than lateral epicondylitis. It occurs most often in work-related repetitive activities and also occurs with recreational activities such as swinging a golf club or throwing a baseball. On physical examination, there is tenderness just distal to the medial epicondyle over the origin of the forearm flexors. Pain can be reproduced by resisting wrist flexion and pronation with the elbow extended. Radiographs are usually normal. The differential diagnosis of patients with medial elbow symptoms includes tears of the pronator teres, acute medial collateral ligament tear, and medial collateral ligament instability. Ulnar neuritis has been found in 25–50% of patients with medial epicondylitis and is associated with tenderness over the ulnar nerve at the elbow as well as hypesthesia and paresthesia on the ulnar side of the hand.

The initial treatment of medial epicondylitis is conservative, involving rest, NSAIDs, friction massage, ultrasound, and icing. Some patients may require splinting. Injections of glucocorticoids at the painful site may also be effective. Patients should be instructed to rest for at least 1 month. Also, patients should start physical therapy once the pain has subsided. In patients with chronic debilitating medial epicondylitis that remains unresponsive after at least a year of treatment, surgical release of the flexor muscle at its origin may be necessary and is often successful.

PLANTAR FASCIITIS

Plantar fasciitis is a common cause of foot pain in adults, with the peak incidence occurring in people between the ages of 40 and 60 years. The pain originates at or near the site of the plantar fascia attachment to the medial tuberosity of the calcaneus. Several factors that increase the risk of developing plantar fasciitis include obesity, pes planus (flat foot or absence of the foot arch when standing), pes cavus (high-arched foot), limited dorsiflexion of the ankle, prolonged standing, walking on hard surfaces, and faulty shoes. In runners, excessive running and a change to a harder running surface may precipitate plantar fasciitis.

The diagnosis of plantar fasciitis can usually be made on the basis of history and physical examination alone. Patients experience severe pain with the first steps on arising in the morning or following inactivity during the day. The pain usually lessens with weight-bearing activity during the day, only to worsen with continued activity. Pain is made worse on walking barefoot or up stairs. On examination, maximal tenderness is elicited on palpation over the inferior heel corresponding to the site of attachment of the plantar fascia.

Imaging studies may be indicated when the diagnosis is not clear. Plain radiographs may show heel spurs, which are of little diagnostic significance. Ultrasonography in plantar fasciitis can demonstrate thickening of the fascia and diffuse hypoechogenicity, indicating edema at the attachment of the plantar fascia to the calcaneus. MRI is a sensitive method for detecting plantar fasciitis, but it is usually not required for establishing the diagnosis.

Resolution of symptoms occurs within 12 months in >80% of patients with plantar fasciitis. Initial treatment consists of ice, heat, massage, stretching, and eliminating activities that can exacerbate plantar fasciitis. Orthotics provide medial arch support and can be effective. Some patients may benefit from foot strapping or taping or by wearing a night splint designed to keep the ankle in a neutral position. A short course of NSAIDs can be given to patients when the benefits outweigh the risks. Local glucocorticoid injections have also been shown to be efficacious but may carry an increased risk for plantar fascia rupture. Plantar fasciotomy is reserved for those patients who have failed to improve after at least 6–12 months of conservative treatment.

FURTHER READING

Buchbinder R: Plantar fasciitis. N Engl J Med 350:2159, 2004.

Greis AC et al: Evaluation and nonsurgical management of rotator cuff calcific tendinopathy. Orthop Clin North Am 46:293, 2015.

Harrison AK, Flatow EL: Subacromial impingement syndrome. J Am Acad Orthop Surg 19:701, 2011.

Kirchgesner T et al: Drug-induced tendinopathy: From physiology to clinical applications. Joint Bone Spine 81:485, 2014.

Neviaser AS, Neviaser RJ: Adhesive capsulitis of the shoulder. J Am Acad Orthop Surg 19:536, 2011.

Section 1 Endocrinology

369 Approach to the Patient with Endocrine Disorders

J. Larry Jameson

The management of endocrine disorders requires a broad understanding of intermediary metabolism, reproductive physiology, bone metabolism, and growth. Accordingly, the practice of endocrinology is intimately linked to a conceptual framework for understanding hormone secretion, hormone action, and principles of feedback control (Chap. 370). The endocrine system is evaluated primarily by measuring hormone concentrations, arming the clinician with valuable diagnostic information. Most disorders of the endocrine system are amenable to effective treatment once the correct diagnosis is established. Endocrine deficiency disorders are treated with physiologic hormone replacement; hormone excess conditions, which usually are caused by benign glandular adenomas, are managed by removing tumors surgically or reducing hormone levels medically.

SCOPE OF ENDOCRINOLOGY

The specialty of endocrinology encompasses the study of glands and the hormones they produce. The term *endocrine* was coined by Starling to contrast the actions of hormones secreted internally (*endocrine*) with those secreted externally (*exocrine*) or into a lumen, such as the gastrointestinal tract. The term *hormone*, derived from a Greek phrase meaning "to set in motion," aptly describes the dynamic actions of hormones as they elicit cellular responses and regulate physiologic processes through feedback mechanisms.

Unlike many other specialties in medicine, it is not possible to define endocrinology strictly along anatomic lines. The classic endocrine glands—pituitary, thyroid, parathyroid, pancreatic islets, adrenals, and gonads—communicate broadly with other organs through the nervous system, hormones, cytokines, and growth factors. In addition to its traditional synaptic functions, the brain produces a vast array of peptide hormones, and this has led to the discipline of neuroendocrinology. Through the production of hypothalamic releasing factors, the central nervous system (CNS) exerts a major regulatory influence over pituitary hormone secretion (Chap. 371). The peripheral nervous system stimulates the adrenal medulla. The immune and endocrine systems are also intimately intertwined. The adrenal hormone cortisol is a powerful immunosuppressant. Cytokines and interleukins (ILs) have profound effects on the functions of the pituitary, adrenal, thyroid, and gonads. Common endocrine diseases such as autoimmune thyroid disease and type 1 diabetes mellitus are caused by dysregulation of immune surveillance and tolerance. Less common diseases such as polyglandular failure, Addison's disease, and lymphocytic hypophysitis also have an immunologic basis.

The interdigitation of endocrinology with physiologic processes in other specialties sometimes blurs the role of hormones. For example, hormones play an important role in maintenance of blood pressure, intravascular volume, and peripheral resistance in the cardiovascular system. Vasoactive substances such as catecholamines, angiotensin II, endothelin, and nitric oxide are involved in dynamic changes of vascular tone in addition to their multiple roles in other tissues. The heart is the principal source of atrial natriuretic peptide, which acts in classic endocrine fashion to induce natriuresis at a distant target organ (the kidney). Erythropoietin, a traditional circulating hormone, is made in the kidney and stimulates erythropoiesis in bone marrow (Chap. 59). The kidney is also integrally involved in the renin-angiotensin axis (Chap. 379) and is a primary target of several hormones, including parathyroid hormone (PTH), mineralocorticoids, and vasopressin. The gastrointestinal tract produces a vast array of peptide hormones, such as cholecystokinin, ghrelin, gastrin, secretin, and vasoactive intestinal peptide, among many others. Carcinoid and islet tumors can secrete excessive amounts of these hormones, leading to specific clinical syndromes (Chap. 80). Many of these gastrointestinal hormones are also produced in the CNS, where their functions are poorly understood. Adipose tissue produces leptin, which acts centrally to control appetite, along with adiponectin, resistin, and other hormones that regulate metabolism. As hormones such as inhibin, ghrelin, and leptin are discovered, they become integrated into the science and practice of medicine on the basis of their functional roles rather than their tissues of origin.

Characterization of hormone receptors frequently reveals unexpected relationships to factors in nonendocrine disciplines. The growth hormone (GH) and leptin receptors, for example, are members of the cytokine receptor family. The G protein–coupled receptors (GPCRs), which mediate the actions of many peptide hormones, are used in numerous physiologic processes, including vision, smell, and neurotransmission.

PATHOLOGIC MECHANISMS OF ENDOCRINE DISEASE

Endocrine diseases can be divided into three major types of conditions: (1) hormone excess, (2) hormone deficiency, and (3) hormone resistance (Table 369-1).

CAUSES OF HORMONE EXCESS

Syndromes of hormone excess can be caused by neoplastic growth of endocrine cells, autoimmune disorders, and excess hormone administration. Benign endocrine tumors, including parathyroid, pituitary, and adrenal adenomas, often retain the capacity to produce hormones, reflecting the fact that these tumors are relatively well differentiated. Many endocrine tumors exhibit subtle defects in their "set points" for feedback regulation. For example, in Cushing's disease, impaired feedback inhibition of adrenocorticotropic hormone (ACTH) secretion is associated with autonomous function. However, the tumor cells are not completely resistant to feedback, as evidenced by ACTH suppression by higher doses of dexamethasone (e.g., high-dose dexamethasone test) (Chap. 379). Similar set point defects are also typical of parathyroid adenomas and autonomously functioning thyroid nodules.

The molecular basis of some endocrine tumors, such as the multiple endocrine neoplasia (MEN) syndromes (MEN1, 2A, 2B), has provided important insights into tumorigenesis (Chap. 381). MEN1 is characterized primarily by the triad of parathyroid, pancreatic islet, and pituitary tumors. MEN2 predisposes to medullary thyroid carcinoma, pheochromocytoma, and hyperparathyroidism. The *MEN1* gene, located on chromosome 11q13, encodes a putative tumor-suppressor gene, menin. Analogous to the paradigm first described for retinoblastoma, the affected individual inherits a mutant copy of the *MEN1* gene, and tumorigenesis ensues after a somatic "second hit" leads to loss of function of the normal *MEN1* gene (through deletion or point mutations).

In contrast to inactivation of a tumor-suppressor gene, as occurs in MEN1 and most other inherited cancer syndromes, MEN2 is caused by activating mutations in a single allele. In this case, activating mutations of the *RET* protooncogene, which encodes a receptor tyrosine kinase, leads to thyroid C cell hyperplasia in childhood before the development of medullary thyroid carcinoma. Elucidation of this pathogenic mechanism has allowed early genetic screening for *RET* mutations in individuals at risk for MEN2, permitting identification of those who may benefit from prophylactic thyroidectomy and biochemical screening for pheochromocytoma and hyperparathyroidism.

TABLE 369-1 Causes of Endocrine Dysfunction

TYPE OF ENDOCRINE DISORDER	EXAMPLES
Hyperfunction	
Neoplastic	
Benign	Pituitary adenomas, hyperparathyroidism, autonomous thyroid or adrenal nodules, pheochromocytoma
Malignant	Adrenal cancer, medullary thyroid cancer, carcinoid
Ectopic	Ectopic ACTH, SIADH secretion
Multiple endocrine neoplasia (MEN)	MEN1, MEN2
Autoimmune	Graves' disease
Iatrogenic	Cushing's syndrome, hypoglycemia
Infectious/inflammatory	Subacute thyroiditis
Activating receptor mutations	LH, TSH, Ca^{2+}, PTH receptors, $G_s\alpha$
Hypofunction	
Autoimmune	Hashimoto's thyroiditis, type 1 diabetes mellitus, Addison's disease, polyglandular failure
Iatrogenic	Radiation-induced hypopituitarism, hypothyroidism, surgical
Infectious/inflammatory	Adrenal insufficiency, hypothalamic sarcoidosis
Hormone mutations	GH, LHβ, FSHβ, vasopressin
Enzyme defects	21-Hydroxylase deficiency
Developmental defects	Kallmann's syndrome, Turner's syndrome, transcription factors
Nutritional/vitamin deficiency	Vitamin D deficiency, iodine deficiency
Hemorrhage/infarction	Sheehan's syndrome, adrenal insufficiency
Hormone Resistance	
Receptor mutations	
Membrane	GH, vasopressin, LH, FSH, ACTH, GnRH, GHRH, PTH, leptin, Ca^{2+}
Nuclear	AR, TR, VDR, ER, GR, PPARγ
Signaling pathway mutations	Albright's hereditary osteodystrophy
Postreceptor	Type 2 diabetes mellitus, leptin resistance

Abbreviations: ACTH, adrenocorticotropic hormone; AR, androgen receptor; ER, estrogen receptor; FSH, follicle-stimulating hormone; GHRH, growth hormone–releasing hormone; GnRH, gonadotropin-releasing hormone; GR, glucocorticoid receptor; LH, luteinizing hormone; PPAR, peroxisome proliferator activated receptor; PTH, parathyroid hormone; SIADH, syndrome of inappropriate antidiuretic hormone; TR, thyroid hormone receptor; TSH, thyroid-stimulating hormone; VDR, vitamin D receptor.

Mutations that activate hormone receptor signaling have been identified in several GPCRs. For example, activating mutations of the luteinizing hormone (LH) receptor cause a dominantly transmitted form of male-limited precocious puberty, reflecting premature stimulation of testosterone synthesis in Leydig cells **(Chap. 384)**. Activating mutations in these GPCRs are located predominantly in the transmembrane domains and induce receptor coupling to $G_s\alpha$ even in the absence of hormone. Consequently, adenylate cyclase is activated, and cyclic adenosine monophosphate (AMP) levels increase in a manner that mimics hormone action. A similar phenomenon results from activating mutations in $G_s\alpha$. When these mutations occur early in development, they cause McCune-Albright syndrome. When they occur only in somatotropes, the activating $G_s\alpha$ mutations cause GH-secreting tumors and acromegaly **(Chap. 373)**.

In autoimmune Graves' disease, antibody interactions with the thyroid-stimulating hormone (TSH) receptor mimic TSH action, leading to hormone overproduction **(Chap. 375)**. Analogous to the effects of activating mutations of the TSH receptor, these stimulating autoantibodies induce conformational changes that release the receptor from a constrained state, thereby triggering receptor coupling to G proteins.

CAUSES OF HORMONE DEFICIENCY

Most examples of hormone deficiency states can be attributed to glandular destruction caused by autoimmunity, surgery, infection, inflammation, infarction, hemorrhage, or tumor infiltration (Table 369-1). Autoimmune damage to the thyroid gland (Hashimoto's thyroiditis) and pancreatic islet β cells (type 1 diabetes mellitus) is a prevalent cause of endocrine disease. Mutations in a number of hormones, hormone receptors, transcription factors, enzymes, and channels can also lead to hormone deficiencies.

HORMONE RESISTANCE

Most severe hormone resistance syndromes are due to inherited defects in membrane receptors, nuclear receptors, or the pathways that transduce receptor signals. These disorders are characterized by defective hormone action despite the presence of increased hormone levels. In complete androgen resistance, for example, mutations in the androgen receptor result in a female phenotypic appearance in genetic (XY) males, even though LH and testosterone levels are increased **(Chap. 381)**. In addition to these relatively rare genetic disorders, more common acquired forms of functional hormone resistance include insulin resistance in type 2 diabetes mellitus, leptin resistance in obesity, and GH resistance in catabolic states. The pathogenesis of functional resistance involves receptor downregulation and postreceptor desensitization of signaling pathways; functional forms of resistance are generally reversible.

CLINICAL EVALUATION OF ENDOCRINE DISORDERS

Because most glands are relatively inaccessible, the physical examination usually focuses on the manifestations of hormone excess or deficiency as well as direct examination of palpable glands, such as the thyroid and gonads. For these reasons, it is important to evaluate patients in the context of their presenting symptoms, review of systems, family and social history, and exposure to medications that may affect the endocrine system. Astute clinical skills are required to detect subtle symptoms and signs suggestive of underlying endocrine disease. For example, a patient with Cushing's syndrome may manifest specific findings, such as central fat redistribution, skin striae, and proximal muscle weakness, in addition to features seen commonly in the general population, such as obesity, plethora, hypertension, and glucose intolerance. Similarly, the insidious onset of hypothyroidism—with mental slowing, fatigue, dry skin, and other features—can be difficult to distinguish from similar, nonspecific findings in the general population. Clinical judgment that is based on knowledge of disease prevalence and pathophysiology is required to decide when to embark on more extensive evaluation of these disorders. Laboratory testing plays an essential role in endocrinology by allowing quantitative assessment of hormone levels and dynamics. Radiologic imaging tests such as

computed tomography (CT) scan, magnetic resonance imaging (MRI), thyroid scan, and ultrasound are also used for the diagnosis of endocrine disorders. However, these tests generally are employed only after a hormonal abnormality has been established by biochemical testing.

HORMONE MEASUREMENTS AND ENDOCRINE TESTING

Immunoassays are the most important diagnostic tool in endocrinology, as they allow sensitive, specific, and quantitative determination of steady-state and dynamic changes in hormone concentrations. Immunoassays use antibodies to detect specific hormones. For many peptide hormones, these measurements are now configured to use two different antibodies to increase binding affinity and specificity. There are many variations of these assays; a common format involves using one antibody to capture the antigen (hormone) onto an immobilized surface and a second antibody, coupled to a chemiluminescent (immunochemiluminescent assay [ICMA]) or radioactive (immunoradiometric assay [IRMA]) signal, to detect the antigen. These assays are sensitive enough to detect plasma hormone concentrations in the picomolar to nanomolar range, and they can readily distinguish structurally related proteins, such as PTH from PTH-related peptide (PTHrP). A variety of other techniques are used to measure specific hormones, including mass spectroscopy, various forms of chromatography, and enzymatic methods; bioassays are now used rarely.

Most hormone measurements are based on plasma or serum samples. However, urinary hormone determinations remain useful for the evaluation of some conditions. Urinary collections over 24 h provide an integrated assessment of the production of a hormone or metabolite, many of which vary during the day. It is important to ensure complete collections of 24-h urine samples; simultaneous measurement of creatinine provides an internal control for the adequacy of collection and can be used to normalize some hormone measurements. A 24-h urine-free cortisol measurement largely reflects the amount of unbound cortisol, thus providing a reasonable index of biologically available hormone. Other commonly used urine determinations include 17-hydroxycorticosteroids, 17-ketosteroids, vanillylmandelic acid, metanephrine, catecholamines, 5-hydroxyindoleacetic acid, and calcium.

The value of quantitative hormone measurements lies in their correct interpretation in a clinical context. The normal range for most hormones is relatively broad, often varying by a factor of two- to tenfold. The normal ranges for many hormones are sex- and age-specific. Thus, using the correct normative database is an essential part of interpreting hormone tests. The pulsatile nature of hormones and factors that can affect their secretion, such as sleep, meals, and medications, must also be considered. Cortisol values increase fivefold between midnight and dawn; reproductive hormone levels vary dramatically during the female menstrual cycle.

For many endocrine systems, much information can be gained from basal hormone testing, particularly when different components of an endocrine axis are assessed simultaneously. For example, low testosterone and elevated LH levels suggest a primary gonadal problem, whereas a hypothalamic-pituitary disorder is likely if both LH and testosterone are low. Because TSH is a sensitive indicator of thyroid function, it is generally recommended as a first-line test for thyroid disorders. An elevated TSH level is almost always the result of primary hypothyroidism, whereas a low TSH is most often caused by thyrotoxicosis. These predictions can be confirmed by determining the free thyroxine level. In the less common circumstance when free thyroxine and TSH are both low, it is important to consider secondary hypopituitarism caused by hypothalamic-pituitary disease. Elevated calcium and PTH levels suggest hyperparathyroidism, whereas PTH is suppressed in hypercalcemia caused by malignancy or granulomatous diseases. A suppressed ACTH in the setting of hypercortisolemia, or increased urine-free cortisol, is seen with hyperfunctioning adrenal adenomas.

It is not uncommon, however, for baseline hormone levels associated with pathologic endocrine conditions to overlap with the normal range. In this circumstance, dynamic testing is useful to separate the two groups further. There are a multitude of dynamic endocrine tests, but all are based on principles of feedback regulation, and most responses can be rationalized based on principles that govern the regulation of endocrine axes. *Suppression tests* are used in the setting of suspected endocrine hyperfunction. An example is the dexamethasone suppression test used to evaluate Cushing's syndrome **(Chaps. 373 and 379)**. *Stimulation tests* generally are used to assess endocrine hypofunction. The ACTH stimulation test, for example, is used to assess the adrenal gland response in patients with suspected adrenal insufficiency. Other stimulation tests use hypothalamic-releasing factors such as corticotropin-releasing hormone (CRH) and growth hormone–releasing hormone (GHRH) to evaluate pituitary hormone reserve **(Chap. 373)**. Insulin-induced hypoglycemia evokes pituitary ACTH and GH responses. Stimulation tests based on reduction or inhibition of endogenous hormones are now used infrequently. Examples include metyrapone inhibition of cortisol synthesis and clomiphene inhibition of estrogen feedback.

SCREENING AND ASSESSMENT OF COMMON ENDOCRINE DISORDERS

Many endocrine disorders are prevalent in the adult population **(Table 369-2)** and can be diagnosed and managed by general internists,

TABLE 369-2 Examples of Prevalent Endocrine and Metabolic Disorders in the Adult

DISORDER	APPROX. PREVALENCE IN ADULTS[a]	SCREENING/TESTING RECOMMENDATIONS[b]	CHAPTER(S)
Obesity	36% BMI ≥30 70% BMI ≥25	Calculate BMI Measure waist circumference Exclude secondary causes Consider comorbid complications	395
Type 2 diabetes mellitus	>8%	Beginning at age 45, screen every 3 years, or earlier in high-risk groups: FPG >126 mg/dL Random plasma glucose >200 mg/dL An elevated HbA_{1c} Consider comorbid complications	396
Hyperlipidemia	20–25%	Cholesterol screening at least every 5 years; more often in high-risk groups Lipoprotein analysis (LDL, HDL) for increased cholesterol, CAD, diabetes Consider secondary causes	400

(*Continued*)

TABLE 369-2 Examples of Prevalent Endocrine and Metabolic Disorders in the Adult (Continued)

DISORDER	APPROX. PREVALENCE IN ADULTS[a]	SCREENING/TESTING RECOMMENDATIONS[b]	CHAPTER(S)
Metabolic syndrome	35%	Measure waist circumference, FPG, BP, lipids	401
Hypothyroidism	5–10%, women 0.5–2%, men	TSH; confirm with free T_4 Screen women after age 35 and every 5 years thereafter	377
Graves' disease	1–3%, women 0.1%, men	TSH, free T_4	376
Thyroid nodules and neoplasia	2–5% palpable >25% by ultrasound	Physical examination of thyroid Fine-needle aspiration biopsy	378
Osteoporosis	5–10%, women 2–5%, men	Bone mineral density measurements in women >65 years or in postmenopausal women or men at risk Exclude secondary causes	404
Hyperparathyroidism	0.1–0.5%, women > men	Serum calcium PTH, if calcium is elevated Assess comorbid conditions	403
Infertility	10%, couples	Investigate both members of couple Semen analysis in male Assess ovulatory cycles in female Specific tests as indicated	384, 385
Polycystic ovarian syndrome	5–10%, women	Free testosterone, DHEAS Consider comorbid conditions	385
Hirsutism	5–10%	Free testosterone, DHEAS Exclude secondary causes Additional tests as indicated	387
Menopause	Median age, 51	FSH	388
Hyperprolactinemia	15% in women with amenorrhea or galactorrhea	PRL level MRI, if not medication-related	373
Erectile dysfunction	10–25%	Careful history, PRL, testosterone Consider secondary causes (e.g., diabetes)	390
Hypogonadism, male	1–2%	Testosterone, LH	384
Gynecomastia	15%	Often, no tests are indicated Consider Klinefelter's syndrome Consider medications, hypogonadism, liver disease	384
Klinefelter's syndrome	0.2%, men	Karyotype Testosterone	383
Vitamin D deficiency	10%	Measure serum 25-OH vitamin D Consider secondary causes	402
Turner's syndrome	0.03%, women	Karyotype Consider comorbid conditions	383

[a]The prevalence of most disorders varies among ethnic groups and with aging. Data based primarily on U.S. population. [b]See individual chapters for additional information on evaluation and treatment. Early testing is indicated in patients with signs and symptoms of disease and in those at increased risk.

Abbreviations: BMI, body mass index; BP, blood pressure; CAD, coronary artery disease; DHEAS, dehydroepiandrosterone; FPG, fasting plasma glucose; FSH, follicle-stimulating hormone; HbA_{1c}, hemoglobin A_{1c}; HDL, high-density lipoprotein; LDL, low-density lipoprotein; LH, luteinizing hormone; MRI, magnetic resonance imaging; PRL, prolactin; PTH, parathyroid hormone; TSH, thyroid-stimulating hormone.

family practitioners, or other primary health care providers. The high prevalence and clinical impact of certain endocrine diseases justifies vigilance for features of these disorders during routine physical examinations; laboratory screening is indicated in selected high-risk populations.

FURTHER READING

Golden SH et al: Prevalence and incidence of endocrine and metabolic disorders in the United States: A comprehensive review. J Clin Endocrinol Metab 94:1853, 2009.

Golden SH et al: Health disparities in endocrine disorders: Biological, clinical, and nonclinical factors—an Endocrine Society Scientific Statement. J Clin Endocrinol Metab 97:E1579, 2012.

Jameson JL, DeGroot LJ (eds): *Endocrinology: Adult and Pediatric*, 7th ed. Philadelphia, Elsevier, 2016.

Loriaux DL: *A Biographical History of Endocrinology*. Hoboken, Wiley Blackwell, 2016.

Melmed S, Polonsky KS, Larsen PR, Kronenberg HM (eds): *Williams Textbook of Endocrinology*, 13th ed. Philadelphia, Elsevier, 2016.

370 Mechanisms of Hormone Action

J. Larry Jameson

Hormones function as a communication system within the body. The endocrine system, composed of various glands and the hormones they produce, interacts with essentially all other physiologic systems to regulate growth, metabolism, homeostasis, and reproduction. Because hormones circulate and act via receptors in target tissues, they serve to integrate physiologic responses to external or internal cues. For example, the light-dark cycle, sensed through the visual system, modulates hypothalamic corticotopin-releasing hormone (CRH), which increases pituitary adrenocorticotropin hormone (ACTH) production, leading to increased adrenal cortisol production before the time of waking in the morning. Increased cortisol, in turn, circulates throughout the body, acting via the nuclear glucocorticoid receptor, to activate numerous genetic programs that influence metabolism, the cardiovascular system, behavior, and the immune system. This chapter provides an overview of the different types of hormones and how they function at the cellular level to control myriad physiologic processes.

CLASSES OF HORMONES

Hormones can be divided into five major types: (1) *amino acid derivatives* such as dopamine, catecholamine, and thyroid hormone; (2) *small neuropeptides* such as gonadotropin-releasing hormone (GnRH), thyrotropin-releasing hormone (TRH), somatostatin, and vasopressin; (3) *large proteins* such as insulin, luteinizing hormone (LH), and parathyroid hormone (PTH); (4) *steroid hormones* such as cortisol and estrogen that are synthesized from cholesterol-based precursors; and (5) *vitamin derivatives* such as retinoids (vitamin A) and vitamin D. A variety of *peptide growth factors*, most of which act locally, share actions with hormones. As a rule, amino acid derivatives and peptide hormones interact with cell-surface membrane receptors. Steroids, thyroid hormones, vitamin D, and retinoids are lipid-soluble and interact with intracellular nuclear receptors, although many also interact with membrane receptors or intracellular signaling proteins as well.

HORMONE AND RECEPTOR FAMILIES

Hormones and receptors can be grouped into families, reflecting structural similarities and evolutionary origins **(Table 370-1)**. The evolution of these families generates diverse but highly selective pathways of hormone action. Recognition of these relationships has proven useful for extrapolating information gleaned from one hormone or receptor to other family members.

The glycoprotein hormone family, consisting of thyroid-stimulating hormone (TSH), follicle-stimulating hormone (FSH), LH, and human chorionic gonadotropin (hCG), illustrates many features of evolutionarily related hormones. The glycoprotein hormones are heterodimers that share the α subunit in common; the β subunits are distinct and confer specific biologic actions. The overall three-dimensional architecture of the β subunits is similar, reflecting the locations of conserved disulfide bonds that restrain protein conformation. The cloning of the β-subunit genes from multiple species suggests that this family arose from a common ancestral gene, probably by gene duplication and subsequent divergence to evolve new biologic functions.

As hormone families enlarge and diverge, their receptors must co-evolve to derive new biologic functions. Related G protein–coupled receptors (GPCRs), for example, have evolved for each of the glycoprotein hormones. These receptors are also structurally similar, and each is coupled predominantly to the $G_s\alpha$ signaling pathway. However, there is minimal overlap of hormone binding. For example, TSH binds with high specificity to the TSH receptor but interacts minimally with the LH or FSH receptors. Nonetheless, there can be subtle physiologic consequences of hormone cross-reactivity with other receptors. Very high levels of hCG during pregnancy stimulate the TSH receptor and increase thyroid hormone levels, resulting via feedback inhibition in a compensatory decrease in TSH.

TABLE 370-1 Examples of Membrane Receptor Families and Signaling Pathways

RECEPTORS	EFFECTORS	SIGNALING PATHWAYS
G Protein–Coupled Seven-Transmembrane Receptor (GPCR)		
β-Adrenergic, LH, FSH, TSH	$G_s\alpha$, adenylate cyclase	Stimulation of cyclic AMP production, protein kinase A
Glucagon, PTH, PTHrP, ACTH, MSH, GHRH, CRH	Ca^{2+} channels	Calmodulin, Ca^{2+}-dependent kinases
α-Adrenergic, somatostatin	$G_i\alpha$	Inhibition of cyclic AMP production
		Activation of K^+, Ca^{2+} channels
TRH, GnRH	G_q, G_{11}	Phospholipase C, diacyl-glycerol, IP_3, protein kinase C, voltage-dependent Ca^{2+} channels
Receptor Tyrosine Kinase		
Insulin, IGF-I	Tyrosine kinases, IRS	MAP kinases, PI 3-kinase; AKT
EGF, NGF	Tyrosine kinases, ras	Raf, MAP kinases, RSK
Cytokine Receptor–Linked Kinase		
GH, PRL	JAK, tyrosine kinases	STAT, MAP kinase, PI 3-kinase, IRS-1
Serine Kinase		
Activin, TGF-β, MIS	Serine kinase	Smads

Abbreviations: IP_3, inositol triphosphate; IRS, insulin receptor substrates; MAP, mitogen-activated protein; MSH, melanocyte-stimulating hormone; NGF, nerve growth factor; PI, phosphatidylinositol; RSK, ribosomal S6 kinase; TGF-β, transforming growth factor β. For all other abbreviations, see text. Note that most receptors interact with multiple effectors and activate networks of signaling pathways.

Insulin and insulin-like growth factor I (IGF-I) and IGF-II have structural similarities that are most apparent when precursor forms of the proteins are compared. In contrast to the high degree of specificity seen with the glycoprotein hormones, there is moderate cross-talk among the members of the insulin/IGF family. High concentrations of an IGF-II precursor produced by certain tumors (e.g., sarcomas) can cause hypoglycemia, partly because of binding to insulin and IGF-I receptors **(Chap. 403)**. High concentrations of insulin also bind to the IGF-I receptor, perhaps accounting for some of the clinical manifestations seen in conditions with chronic hyperinsulinemia.

Another important example of receptor cross-talk is seen with PTH and parathyroid hormone–related peptide (PTHrP) **(Chap. 403)**. PTH is produced by the parathyroid glands, whereas PTHrP is expressed at high levels during development and by a variety of tumors **(Chap. 89)**. These hormones have amino acid sequence similarity, particularly in their amino-terminal regions. Both hormones bind to the PTH1R receptor that is expressed in bone and kidney. Hypercalcemia and hypophosphatemia therefore may result from excessive production of either hormone, making it difficult to distinguish hyperparathyroidism from hypercalcemia of malignancy solely on the basis of serum chemistries. However, sensitive and specific assays for PTH and PTHrP now allow these disorders to be distinguished more readily.

Based on their specificities for DNA-binding sites, the nuclear receptor family can be subdivided into type 1 receptors (glucocorticoid receptor, mineralocorticoid receptor, androgen receptor, estrogen receptor, progesterone receptor) that bind steroids and type 2 receptors (thyroid hormone receptor, vitamin D receptor, retinoic acid receptor, peroxisome proliferator activated receptor) that bind thyroid hormone, vitamin D, retinoic acid, or lipid derivatives, respectively. Certain functional domains in nuclear receptors, such as the zinc finger DNA-binding domains, are highly conserved. However, selective amino acid differences within this domain confer DNA sequence specificity. The hormone-binding domains are more variable, providing great diversity in the array of small molecules that bind to different nuclear receptors. With few

exceptions, hormone binding is highly specific for a single type of nuclear receptor. One exception involves the glucocorticoid and mineralocorticoid receptors. Because the mineralocorticoid receptor also binds glucocorticoids with high affinity, an enzyme (11β-hydroxysteroid dehydrogenase) in renal tubular cells inactivates glucocorticoids, allowing selective responses to mineralocorticoids such as aldosterone. However, when very high glucocorticoid concentrations occur, as in Cushing's syndrome, the glucocorticoid degradation pathway becomes saturated, allowing excessive cortisol levels to bind mineralocorticoid receptors leading to sodium retention and potassium wasting. This phenomenon is particularly pronounced in ectopic adrenocorticotropic hormone (ACTH) syndromes **(Chap. 379)**. Another example of relaxed nuclear receptor specificity involves the estrogen receptor, which can bind an array of compounds, some of which have little apparent structural similarity to the high-affinity ligand estradiol. This feature of the estrogen receptor makes it susceptible to activation by "environmental estrogens" such as resveratrol, octylphenol, and many other aromatic hydrocarbons. However, this lack of specificity provides an opportunity to synthesize a remarkable series of clinically useful antagonists (e.g., tamoxifen) and selective estrogen response modulators (SERMs) such as raloxifene. These compounds generate distinct conformations that alter receptor interactions with components of the transcription machinery (see below), thereby conferring their unique actions.

HORMONE SYNTHESIS AND PROCESSING

The synthesis of peptide hormones and their receptors occurs through a classic pathway of gene expression: transcription → mRNA → protein → posttranslational protein processing → intracellular sorting, followed by membrane integration or secretion.

Many hormones are embedded within larger precursor polypeptides that are proteolytically processed to yield the biologically active hormone. Examples include proopiomelanocortin (POMC) → ACTH; proglucagon → glucagon; proinsulin → insulin; and pro-PTH → PTH, among others. In many cases, such as POMC and proglucagon, these precursors generate multiple biologically active peptides. It is provocative that hormone precursors are typically inactive, presumably adding an additional level of regulatory control. Prohormone conversion occurs not only for peptide hormones but also for certain steroids (testosterone → dihydrotestosterone) and thyroid hormone ($T_4 \rightarrow T_3$).

Peptide precursor processing is intimately linked to intracellular sorting pathways that transport proteins to appropriate vesicles and enzymes, resulting in specific cleavage steps, followed by protein folding and translocation to secretory vesicles. Hormones destined for secretion are translocated across the endoplasmic reticulum under the guidance of an amino-terminal signal sequence that subsequently is cleaved. Cell-surface receptors are inserted into the membrane via short segments of hydrophobic amino acids that remain embedded within the lipid bilayer. During translocation through the Golgi and endoplasmic reticulum, hormones and receptors are subject to a variety of posttranslational modifications, such as glycosylation and phosphorylation, which can alter protein conformation, modify circulating half-life, and alter biologic activity.

Synthesis of most steroid hormones is based on modifications of the precursor, cholesterol. Multiple regulated enzymatic steps are required for the synthesis of testosterone **(Chap. 384)**, estradiol **(Chap. 385)**, cortisol **(Chap. 379)**, and vitamin D **(Chap. 402)**. This large number of synthetic steps predisposes to multiple genetic and acquired disorders of steroidogenesis.

Endocrine genes contain regulatory DNA elements similar to those found in many other genes, but their exquisite control by hormones reflects the presence of specific hormone response elements. For example, the TSH genes are repressed directly by thyroid hormones acting through the thyroid hormone receptor (TR), a member of the nuclear receptor family. Steroidogenic enzyme gene expression requires specific transcription factors, such as steroidogenic factor-1 (SF-1), acting in conjunction with signals transmitted by trophic hormones (e.g., ACTH or LH). Once activated, SF-1 functions as a master regulator, inducing a large array of genes required for steroidogenic and metabolic pathways required for steroid synthesis. For some hormones, substantial regulation occurs at the level of translational efficiency. Insulin biosynthesis, although it requires ongoing gene transcription, is regulated primarily at the translational and secretory levels in response to elevated levels of glucose or amino acids.

HORMONE SECRETION, TRANSPORT, AND DEGRADATION

The circulating level of a hormone is determined by its rate of secretion and its half-life. After protein processing, peptide hormones (e.g., GnRH, insulin, growth hormone [GH]) are stored in secretory granules. As these granules mature, they are poised beneath the plasma membrane for imminent release into the circulation. In most instances, the stimulus for hormone secretion is a releasing factor or neural signal that induces rapid changes in intracellular calcium concentrations, leading to secretory granule fusion with the plasma membrane and release of its contents into the extracellular environment and bloodstream. Steroid hormones, in contrast, diffuse into the circulation as they are synthesized. Thus, their secretory rates are closely aligned with rates of synthesis. For example, ACTH and LH induce steroidogenesis by stimulating the activity of the *st*eroidogenic *a*cute *r*egulatory (StAR) protein (transports cholesterol into the mitochondrion) along with other rate-limiting steps (e.g., cholesterol side-chain cleavage enzyme, CYP11A1) in the steroidogenic pathway.

Hormone transport and degradation dictate the rapidity with which a hormonal signal decays. Some hormone signals are evanescent (e.g., somatostatin), whereas others are longer-lived (e.g., TSH). Because somatostatin exerts effects in virtually every tissue, a short half-life allows its concentrations and actions to be controlled locally. Structural modifications that impair somatostatin degradation have been useful for generating long-acting therapeutic analogues such as octreotide **(Chap. 373)**. In contrast, the actions of TSH are highly specific for the thyroid gland. Its prolonged half-life accounts for relatively constant serum levels even though TSH is secreted in discrete pulses.

An understanding of circulating hormone half-life is important for achieving physiologic hormone replacement, as the frequency of dosing and the time required to reach steady state are intimately linked to rates of hormone decay. T_4, for example, has a circulating half-life of 7 days. Consequently, >1 month is required to reach a new steady state, and single daily doses are sufficient to achieve constant hormone levels. T_3, in contrast, has a half-life of 1 day. Its administration is associated with more dynamic serum levels, and it must be administered two to three times per day. Similarly, synthetic glucocorticoids vary widely in their half-lives; those with longer half-lives (e.g., dexamethasone) are associated with greater suppression of the hypothalamic-pituitary-adrenal (HPA) axis. Most protein hormones (e.g., ACTH, GH, prolactin [PRL], PTH, LH) have relatively short half-lives (<20 min), leading to sharp peaks of secretion and decay. The only accurate way to profile the pulse frequency and amplitude of these hormones is to measure levels in frequently sampled blood (every 10 min or less) over long durations (8–24 h). Because this is not practical in a clinical setting, an alternative strategy is to pool three to four samples drawn at about 30-min intervals, or interpret the results in the context of a relatively wide normal range. Rapid hormone decay is useful in certain clinical settings. For example, the short half-life of PTH allows the use of intraoperative PTH determinations to confirm successful removal of a parathyroid adenoma. This is particularly valuable diagnostically when there is a possibility of multicentric disease or parathyroid hyperplasia, as occurs with multiple endocrine neoplasia (MEN) or renal insufficiency.

Many hormones circulate in association with serum-binding proteins. Examples include (1) T_4 and T_3 binding to thyroxine-binding globulin (TBG), albumin, and thyroxine-binding prealbumin (TBPA); (2) cortisol binding to cortisol-binding globulin (CBG); (3) androgen and estrogen binding to sex hormone–binding globulin (SHBG); (4) IGF-I and II binding to multiple IGF-binding proteins (IGFBPs); (5) GH interactions with GH-binding protein (GHBP), a circulating fragment of the GH receptor extracellular domain; and (6) activin binding to follistatin. These interactions provide a hormonal reservoir, prevent otherwise rapid degradation of unbound hormones,

restrict hormone access to certain sites (e.g., IGFBPs), and modulate the unbound, or "free," hormone concentrations. Although a variety of binding protein abnormalities have been identified, most have little clinical consequence aside from creating diagnostic problems. For example, TBG deficiency can reduce total thyroid hormone levels greatly but the free concentrations of T_4 and T_3 remain normal. Liver disease and certain medications can also influence binding protein levels (e.g., estrogen increases TBG) or cause displacement of hormones from binding proteins (e.g., salsalate displaces T_4 from TBG). In general, only unbound hormone is available to interact with receptors and thus elicit a biologic response. Short-term perturbations in binding proteins change the free hormone concentration, which in turn induces compensatory adaptations through feedback loops. SHBG changes in women are an exception to this self-correcting mechanism. When SHBG decreases because of insulin resistance or androgen excess, the unbound testosterone concentration is increased, potentially contributing to hirsutism in women with polycystic ovary syndrome (PCOS) **(Chap. 387)**. The increased unbound testosterone level does not result in an adequate compensatory feedback correction because estrogen, not testosterone, is the primary regulator of the reproductive axis.

An additional exception to the unbound hormone hypothesis involves megalin, a member of the low-density lipoprotein (LDL) receptor family that serves as an endocytotic receptor for thyroglobulin, carrier-bound vitamins A and D and SHBG-bound androgens and estrogens. After internalization, the carrier proteins are degraded in lysosomes and release their bound ligands within the cells. Membrane transporters have also been identified for thyroid hormones.

Hormone degradation can be an important mechanism for regulating concentrations locally. As noted above, 11β-hydroxysteroid dehydrogenase inactivates glucocorticoids in renal tubular cells, preventing actions through the mineralocorticoid receptor. Thyroid hormone deiodinases convert T_4 to T_3 and can inactivate T_3. During development, degradation of retinoic acid by Cyp26b1 prevents primordial germ cells in the male from entering meiosis, as occurs in the female ovary.

HORMONE ACTION THROUGH RECEPTORS

Receptors for hormones are divided into two major classes: membrane and nuclear. *Membrane receptors* primarily bind peptide hormones and catecholamines. *Nuclear receptors* bind small molecules that can diffuse across the cell membrane, such as steroids and vitamin D. Certain general principles apply to hormone-receptor interactions regardless of the class of receptor. Hormones bind to receptors with specificity and an affinity that generally coincides with the dynamic range of circulating hormone concentrations. Low concentrations of free hormone (usually 10^{-12} to 10^{-9} *M*) rapidly associate and dissociate from receptors in a bimolecular reaction such that the occupancy of the receptor at any given moment is a function of hormone concentration and the receptor's affinity for the hormone. Receptor numbers vary greatly in different target tissues, providing one of the major determinants of specific tissue responses to circulating hormones. For example, ACTH receptors are located almost exclusively in the adrenal cortex, and FSH receptors are found predominantly in the gonads. In contrast, insulin and TRs are widely distributed, reflecting the need for metabolic responses in all tissues.

MEMBRANE RECEPTORS

Membrane receptors for hormones can be divided into several major groups: (1) seven transmembrane GPCRs, (2) tyrosine kinase receptors, (3) cytokine receptors, and (4) serine kinase receptors **(Fig. 370-1)**. The *seven transmembrane GPCR family* binds a remarkable array of hormones, including large proteins (e.g., LH, PTH), small peptides (e.g., TRH, somatostatin), catecholamines (epinephrine, dopamine), and even minerals (e.g., calcium). The extracellular domains of GPCRs vary widely in size and are the major binding site for large hormones. The transmembrane-spanning regions are composed of hydrophobic α-helical domains that traverse the lipid bilayer. Like some channels, these domains are thought to circularize and form a hydrophobic pocket into which certain small ligands fit. Hormone binding induces conformational changes in these domains, transducing structural changes to the intracellular domain, which is a docking site for G proteins.

The large family of *G proteins*, so named because they bind guanine nucleotides (guanosine triphosphate [GTP], guanosine diphosphate [GDP]), provides great diversity for coupling receptors to different signaling pathways. G proteins form a heterotrimeric complex that is composed of various α and βγ subunits **(Fig. 370-2)**. The α subunit contains the guanine nucleotide–binding site and an intrinsic GTPase that hydrolyzes GTP → GDP. The βγ subunits are tightly associated and modulate the activity of the α subunit as well as mediating their own effector signaling pathways. G protein activity is regulated by a cycle that involves GTP hydrolysis and dynamic interactions between the α and βγ subunits. Hormone binding to the receptor induces GDP dissociation, allowing Gα to bind GTP and dissociate from the βγ complex. Under these conditions, the Gα subunit is activated and mediates signal transduction through various enzymes, such as adenylate cyclase and phospholipase C. GTP hydrolysis to GDP allows reassociation with the βγ subunits and restores the inactive state. G proteins interact with other cellular proteins, including kinases, channels, G protein–coupled receptor kinases (GRKs), and arrestins, that mediate signaling as well as receptor desensitization and recycling.

A variety of endocrinopathies result from mutations in GPCR receptors that alter their interactions with G proteins **(Table 370-2)**. Loss-of-function mutations are generally recessive and inactivate the relevant hormone signaling pathway. Because many of these receptors are important for development as well as signaling, patient presentations often resemble glandular failure syndromes (e.g., mutations in LH-R, FSH-R, TSH-R). Gain-of-function mutations involve a more complex mechanism. Selected mutations induce conformational changes in the GPCR that mimic the activated state normally induced by hormone binding. These mutations result in a constitutively active state in which G protein coupling stimulates cell signaling pathways, most commonly via cyclic adenosine 5′-monophosphate (cAMP) and protein kinase A. When mutations occur in the germline, the conditions are heritable and present in early life (e.g., LH-R, TSH-R). Somatic mutations can also occur and result in clonal expansion of hyperfunctioning cells.

Mutations in the TSH-R illustrate the range of possible clinical consequences of GPCR mutations. Recessive inactivating mutations in the TSH-R cause congenital hypothyroidism with thyroid gland hypoplasia and resistance to TSH. Clinically, the hormone profile resembles

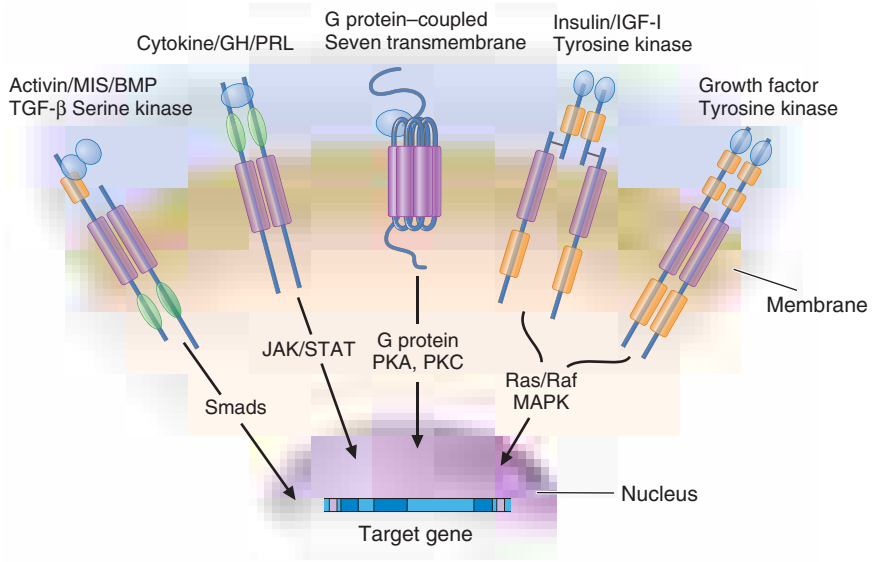


FIGURE 370-1 Membrane receptor signaling. MAPK, mitogen-activated protein kinase; PKA, C, protein kinase A, C; TGF, transforming growth factor. For other abbreviations, see text.

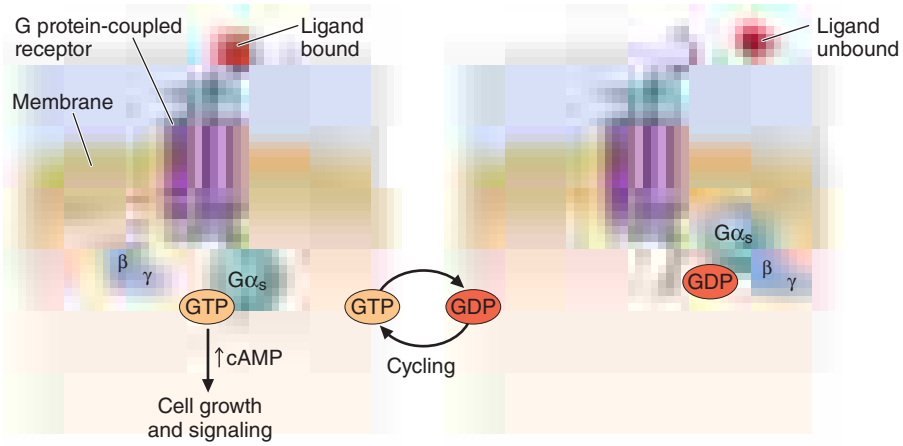


FIGURE 370-2 G protein signaling. G protein-coupled receptors signal via the family of G proteins, so-named because they bind guanylyl nucleotides. In the example shown, a G-protein-coupled receptor (GPCR) bound to a ligand induces GDP dissociation, allowing Gsα to bind GTP and dissociate from the βγ complex. GTP-bound Gsα increases cAMP production by adenylyl cyclase and activates the protein kinase A pathway. Not shown are separate signaling pathways activated by the βγ complex. When GTP is converted to GDP by an intrinsic GTPase, the βγ subunits reassociate with GDP-bound Gsα and the complex returns to an inactive state. As noted in the text, mutations in Gsα that eliminate GTPase activity result in constitutive activation of receptor signaling pathways because GTP-bound Gsα cannot be converted to its GDP-bound inactive state. cAMP, cyclic adenosine 5′-monophosphate; GDP, guanosine diphosphate; Gsα, G protein α; GTP, guanosine triphosphate.

primary hypothyroidism with low T4 and high TSH. On the other hand, germline activating mutations cause congenital hyperthyroidism. The disorder is autosomal dominant because an activating mutation of one TSH-R allele is sufficient to induce cellular hyperfunction and disease. Because the TSH-R is activated in every cell of the thyroid, there is hyperplastic growth and hyperfunction that resembles the pathology seen in Graves' disease. This unusual disorder presents in infancy and must be distinguished from the more common clinical circumstance in which maternal antibodies in women with active or previously treated Graves' disease cross the placenta and stimulate the thyroid gland of the fetus. If an activating TSH-R mutation occurs later in life, in the somatic tissue, there is clonal expansion of the thyrocyte harboring the mutation, ultimately leading to an autonomous hyperfunctioning thyroid nodule. Of note, a similar condition can be caused by somatic mutations in Gsα. In this case, the Gsα GTPase is inactivated and GTP cannot be converted to GDP. Consequently, the Gsα signaling pathway in this particular cell is constitutively active, mimicking chronic TSH stimulation and again leading to clonal expansion and an autonomous hyperfunctioning thyroid nodule. About one-third of hyperfunctioning "hot" thyroid nodules harbor mutations in either the TSH-R or Gsα (TSH-R mutations are more common).

Gsα mutations in tissues other than the thyroid can also cause endocrine disease. For example, Gsα mutations in pituitary somatotropes mimic activation of the growth hormone–releasing hormone (GHRH) pathway and lead to GH-producing adenomas and acromegaly. Rarely, mutations in other components of the protein kinase A pathway in somatotropes can also cause GH-producing adenomas. Gsα mutations that occur early in development (typically mosaic) cause McCune-Albright syndrome **(Chap. 405)** and the clinical features are manifest because the activated G protein pathway mimics the actions of various hormones (PTH, melanocyte stimulating hormone [MSH], TSH, GHRH) in different tissues. Germline inactivating Gsα mutations cause a range of disorders that are transmitted and expressed in a complex manner because the locus is imprinted **(Chap. 403)**. These conditions include Albright's hereditary osteodystrophy (AHO), pseudopseudohypoparathyroidism (PPHP), and pseudohypoparathyroidism types 1b, 1c, and 2.

The *tyrosine kinase receptors* transduce signals for insulin and a variety of growth factors, such as IGF-I, epidermal growth factor (EGF), nerve growth factor, platelet-derived growth factor, and fibroblast growth factor. The cysteine-rich extracellular domains contain binding sites for the growth factors. After ligand binding, this class of receptors undergoes autophosphorylation, inducing interactions with intracellular adaptor proteins such as Shc and insulin receptor substrates (IRS). In the case of the insulin receptor, multiple kinases are activated, including the Raf-Ras-MAPK and the Akt/protein kinase B pathways. The tyrosine kinase receptors play a prominent role in cell growth and differentiation as well as in intermediary metabolism.

The GH and PRL receptors belong to the *cytokine receptor* family. Analogous to the tyrosine kinase receptors, ligand binding induces receptor interaction with intracellular kinases—the Janus kinases (JAKs), which phosphorylate members of the signal transduction and activators of transcription (STAT)

TABLE 370-2 Genetic Causes of G protein Receptor Disorders

RECEPTOR	DISORDER	GENETICS
LH	Leydig cell hypoplasia (male)	AR, inactivating
	Primary amenorrhea, resistance to LH (female)	AR, inactivating
	Familial male precocious puberty (male)	AD, activating
	Leydig cell adenoma, precocious puberty (male)	Sporadic, activating
FSH	Hypergonadotropic ovarian failure (female)	AR, inactivating
	Hypospermia (male)	AR, inactivating
	Ovarian hyperstimulation (female)	AD, activating
TSH	Congenital hypothyroidism, TSH resistance	AR, AD, inactivating
	Nonautoimmune familial hyperthyroidism	AD, activating
	Hyperfunctioning thyroid adenoma	Sporadic, activating
GnRH	Hypogonadotropic hypogonadism	AR, inactivating
Kisspeptin	Hypogonadotropic hypogonadism	AR, inactivating
	Precocious puberty	AD, activating
TRH	Central hypothyroidism	AR, inactivating
GHRH	GH deficiency	AR, inactivating
PTH	Blomstrand chondrodysplasia	AR, inactivating
	Jansen metaphyseal chondrodysplasia	AD, activating
Calcium sensing receptor	Familial hypocalciuric hypercalcemia	AD, inactivating
	Neonatal severe hyperparathyroidism	AR, inactivating
	Familial hypocalcemic hypercalciura	AD, activating
Arginine vasopressin receptor 2	Nephrogenic diabetes insipidus	XL, inactivating
	Nephrogenic SIADH	XL, activating
ACTH	Familial ACTH resistance	AR, inactivating
Melanocortin 4	Severe obesity	Codominant, inactivating

Abbreviations: ACTH, adrenocorticotropin hormone; AD, autosomal dominant; AR, autosomal recessive; FSH, follicle-stimulating hormone; GH, growth hormone; GHRH, growth hormone–releasing hormone; GnRH, gonadotropin-releasing hormone; LH, luteinizing hormone; PTH, parathyroid hormone; SIADH, syndrome of inappropriate antidiuretic hormone secretion; TRH, thyrotropin-releasing hormone; TSH, thyroid-stimulating hormone; XL, X-linked.

family—as well as with other signaling pathways (Ras, PI3-K, MAPK). The activated STAT proteins translocate to the nucleus and stimulate expression of target genes.

The *serine kinase receptors* mediate the actions of activins, transforming growth factor β, müllerian-inhibiting substance (MIS, also known as anti-müllerian hormone, AMH), and bone morphogenic proteins (BMPs). This family of receptors (consisting of type I and II subunits) signals through proteins termed *smads* (fusion of terms for *Caenorhabditis elegans* sma + mammalian mad). Like the STAT proteins, the smads serve a dual role of transducing the receptor signal and acting as transcription factors. The pleomorphic actions of these growth factors dictate that they act primarily in a local (paracrine or autocrine) manner. Binding proteins such as follistatin (which binds activin and other members of this family) function to inactivate the growth factors and restrict their distribution.

Disease-causing mutations also occur in each of these classes of receptors. For example, insulin receptor mutations cause an extreme form of insulin resistance. GH receptor mutations cause Laron-type dwarfism, characterized by low IGF-1 and high GH. AMH receptor mutations cause persistent Müllerian duct syndrome. These hormone resistance syndromes are autosomal recessive and relatively uncommon. Unlike the GPCRs, activating mutations are unusual, although they do occur for the RET tyrosine kinase receptor, which causes the autosomal dominant disorder multiple endocrine neoplasia type 2 (MEN-2) **(Chap. 381)**.

NUCLEAR RECEPTORS

The family of nuclear receptors has grown to nearly 100 members, many of which are still classified as orphan receptors because their ligands, if they exist, have not been identified **(Fig. 370-3)**. Otherwise, most nuclear receptors are classified on the basis of their ligands. Although all nuclear receptors ultimately act to increase or decrease gene transcription, some (e.g., glucocorticoid receptor) reside primarily in the cytoplasm, whereas others (e.g., TR) are located in the nucleus. After ligand binding, the cytoplasmically localized receptors translocate to the nucleus. There is growing evidence that certain nuclear receptors (e.g., glucocorticoid, estrogen) can also act at the membrane or in the cytoplasm to activate or repress signal transduction pathways, providing a mechanism for cross-talk between membrane and nuclear receptors.

The structures of nuclear receptors have been studied extensively, including by x-ray crystallography. The DNA-binding domain, consisting of two zinc fingers, contacts-specific DNA recognition sequences in target genes. Most nuclear receptors bind to DNA as dimers. Consequently, each monomer recognizes an individual DNA motif, referred to as a "half-site." The steroid receptors, including the glucocorticoid, estrogen, progesterone, and androgen receptors, bind to DNA as homodimers. Consistent with this twofold symmetry, their DNA recognition half-sites are palindromic. The thyroid, retinoid, peroxisome proliferator activated, and vitamin D receptors bind to DNA preferentially as heterodimers in combination with retinoid X receptors (RXRs). Their DNA half-sites are typically arranged as direct repeats.

The carboxy-terminal hormone-binding domain mediates transcriptional control. For type II receptors such as TR and retinoic acid receptor (RAR), co-repressor proteins bind to the receptor in the absence of ligand and silence gene transcription. Hormone binding induces conformational changes, triggering the release of co-repressors and inducing the recruitment of coactivators that stimulate transcription. Thus, these receptors are capable of mediating dramatic changes in the level of gene activity. Disease states can be associated with defective regulation of these events. For example, in promyelocytic leukemia, fusion of RARα to other nuclear proteins causes aberrant gene silencing that prevents normal cellular differentiation. Treatment with retinoic acid reverses this repression and allows cellular differentiation and apoptosis to occur. Most type 1 steroid receptors interact weakly with co-repressors, but ligand binding still induces interactions with an array of coactivators. X-ray crystallography shows that various SERMs induce distinct estrogen receptor conformations. The tissue-specific responses caused by these agents in breast, bone, and uterus appear to reflect distinct interactions with coactivators. The receptor-coactivator complex stimulates gene transcription by several pathways, including (1) recruitment of enzymes (histone acetyl transferases) that modify chromatin structure, (2) interactions with additional transcription factors on the target gene, and (3) direct interactions with components of the general transcription apparatus to enhance the rate of RNA polymerase II–mediated transcription. Studies of nuclear receptor-mediated transcription show that these are dynamic events that involve relatively rapid (e.g., 30–60 min) cycling of transcription complexes on any specific target gene.

Nuclear receptor mutations are an important cause of endocrine disease. Androgen receptor mutations cause androgen insensitivity syndrome (AIS) **(Chap. 383)**. Because the androgen receptor is located on the X-chromosome, mutations are more commonly manifest than with other nuclear receptor disorders. Affected individuals with AIS are XY phenotypic females with retained testes and male-range testosterone levels. Tissue insensitivity to androgens varies based on the severity of the mutation. Müllerian structures are absent because Sertoli cells of the testis produce AMH during development. Female carriers of androgen receptor mutations are phenotypically normal. Recessive mutations of the estrogen, glucocorticoid, and vitamin D receptors are rare.

Thyroid hormone receptor β (TRβ) mutations have an unusual pathophysiology. They are autosomal dominant and function via a "dominant negative" mechanism to cause resistance to thyroid hormone (RTH) **(Chap. 375)**. The mutations occur in selected regions of the TRβ hormone-binding domain and preserve the ability of the mutant receptor to heterodimerize with RXR and bind to DNA regulatory sites. The mutant receptors function as antagonists of receptors from the normal copy of the TRβ gene. Affected patients have high T_4 and T_3 and inappropriately elevated (unsuppressed) TSH, reflecting impaired feedback regulation of the hypothalamic-pituitary-thyroid axis. Organ systems are variably resistant to thyroid hormones based upon the relative expression of TRβ and

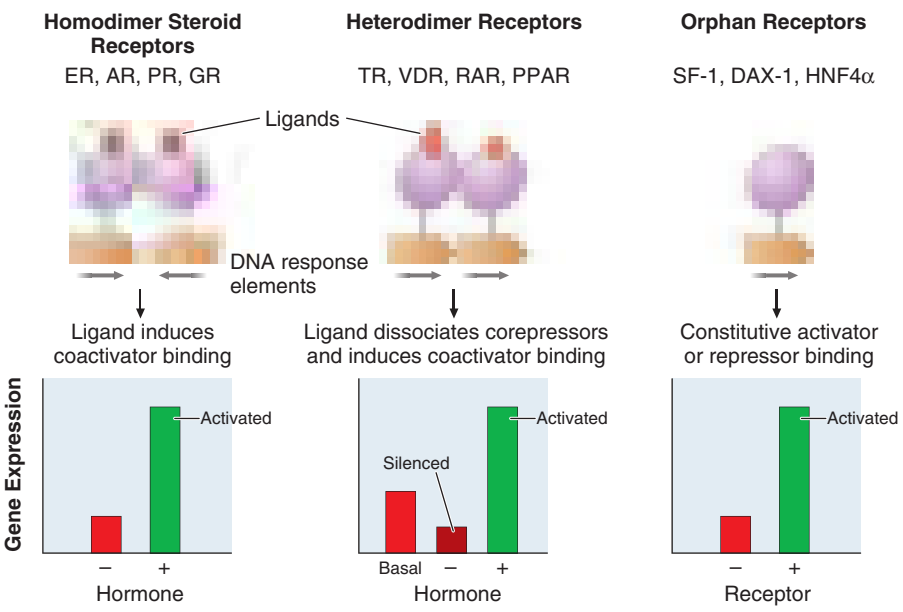


FIGURE 370-3 Nuclear receptor signaling. AR, androgen receptor; DAX, *d*osage-sensitive sex-reversal, adrenal hypoplasia congenita, *X*-chromosome; ER, estrogen receptor; GR, glucocorticoid receptor; HNF4α, hepatic nuclear factor 4α; PPAR, peroxisome proliferator activated receptor; PR, progesterone receptor; RAR, retinoic acid receptor; SF-1, steroidogenic factor-1; TR, thyroid hormone receptor; VDR, vitamin D receptor.

TRα proteins. Mutations in the genes encoding TRα and PPARγ can also cause disease by functioning in an analogous dominant negative manner.

FUNCTIONS OF HORMONES

The functions of individual hormones are described in detail in subsequent chapters. Nevertheless, it is useful to illustrate how most biologic responses require integration of several different hormone pathways. The physiologic functions of hormones can be divided into three general areas: (1) growth and differentiation, (2) maintenance of homeostasis, and (3) reproduction.

GROWTH

Multiple hormones and nutritional factors mediate the complex phenomenon of growth **(Chap. 371)**. Short stature may be caused by GH deficiency, hypothyroidism, Cushing's syndrome, precocious puberty, malnutrition, chronic illness, or genetic abnormalities that affect the epiphyseal growth plates (e.g., *FGFR3* and *SHOX* mutations). Many factors (GH, IGF-I, thyroid hormones) stimulate growth, whereas others (sex steroids) lead to epiphyseal closure. Understanding these hormonal interactions is important in the diagnosis and management of growth disorders. For example, delaying exposure to high levels of sex steroids may enhance the efficacy of GH treatment.

MAINTENANCE OF HOMEOSTASIS

Although virtually all hormones affect homeostasis, the most important among them are the following:

1. Thyroid hormone—controls about 25% of basal metabolism in most tissues.
2. Cortisol—exerts a permissive action for many hormones in addition to its own direct effects.
3. PTH—regulates calcium and phosphorus levels.
4. Vasopressin—regulates serum osmolality by controlling renal free-water clearance.
5. Mineralocorticoids—control vascular volume and serum electrolyte (Na^+, K^+) concentrations.
6. Insulin—maintains euglycemia in the fed and fasted states.

The defense against hypoglycemia is an impressive example of integrated hormone action **(Chap. 399)**. In response to the fasting state and falling blood glucose, insulin secretion is suppressed, resulting in decreased glucose uptake and enhanced glycogenolysis, lipolysis, proteolysis, and gluconeogenesis to mobilize fuel sources. If hypoglycemia develops (usually from insulin administration or sulfonylureas), an orchestrated counterregulatory response occurs—glucagon and epinephrine rapidly stimulate glycogenolysis and gluconeogenesis, whereas GH and cortisol act over several hours to raise glucose levels and antagonize insulin action.

Although free-water clearance is controlled primarily by vasopressin, cortisol and thyroid hormone are also important for facilitating renal tubular responses to vasopressin **(Chap. 374)**. PTH and vitamin D function in an interdependent manner to control calcium metabolism **(Chap. 402)**. PTH stimulates renal synthesis of 1,25-dihydroxyvitamin D, which increases calcium absorption in the gastrointestinal tract and enhances PTH action in bone. Increased calcium, along with vitamin D, feeds back to suppress PTH, thus maintaining calcium balance.

Depending on the severity of a specific stress and whether it is acute or chronic, multiple endocrine and cytokine pathways are activated to mount an appropriate physiologic response. In severe acute stress such as trauma or shock, the sympathetic nervous system is activated and catecholamines are released, leading to increased cardiac output and a primed musculoskeletal system. Catecholamines also increase mean blood pressure and stimulate glucose production. Multiple stress-induced pathways converge on the hypothalamus, stimulating several hormones, including vasopressin and corticotropin-releasing hormone (CRH). These hormones, in addition to cytokines (tumor necrosis factor α, interleukin [IL] 2, IL-6) increase ACTH and GH production. ACTH stimulates the adrenal gland, increasing cortisol, which in turn helps sustain blood pressure and dampen the inflammatory response. Increased vasopressin acts to conserve free water.

REPRODUCTION

The stages of reproduction include (1) sex determination during fetal development **(Chap. 383)**; (2) sexual maturation during puberty **(Chaps. 384 and 385)**; (3) conception, pregnancy, lactation, and child rearing **(Chap. 385)**; and (4) cessation of reproductive capability at menopause **(Chap. 388)**. Each of these stages involves an orchestrated interplay of multiple hormones, a phenomenon well illustrated by the dynamic hormonal changes that occur during each 28-day menstrual cycle. In the early follicular phase, pulsatile secretion of LH and FSH stimulates the progressive maturation of the ovarian follicle. This results in gradually increasing estrogen and progesterone levels, leading to enhanced pituitary sensitivity to GnRH, which, when combined with accelerated GnRH secretion, triggers the LH surge and rupture of the mature follicle. Inhibin, a protein produced by the granulosa cells, enhances follicular growth and feeds back to the pituitary to selectively suppress FSH without affecting LH. Growth factors such as EGF and IGF-I modulate follicular responsiveness to gonadotropins. Vascular endothelial growth factor and prostaglandins play a role in follicle vascularization and rupture.

During pregnancy, the increased production of prolactin, in combination with placentally derived steroids (e.g., estrogen and progesterone), prepares the breast for lactation. Estrogens induce the production of progesterone receptors, allowing for increased responsiveness to progesterone. In addition to these and other hormones involved in lactation, the nervous system and oxytocin mediate the suckling response and milk release.

HORMONAL FEEDBACK REGULATORY SYSTEMS

Feedback control, both negative and positive, is a fundamental feature of endocrine systems. Each of the major hypothalamic-pituitary-hormone axes is governed by negative feedback, a process that maintains hormone levels within a relatively narrow range **(Chap. 371)**. Examples of hypothalamic-pituitary negative feedback include (1) thyroid hormones on the TRH-TSH axis, (2) cortisol on the CRH-ACTH axis, (3) gonadal steroids on the GnRH-LH/FSH axis, and (4) IGF-I on the GHRH-GH axis **(Fig. 370-4)**. These regulatory loops include both positive (e.g., TRH, TSH) and negative (e.g., T_4, T_3) components, allowing for exquisite control of hormone levels. As an example, a small

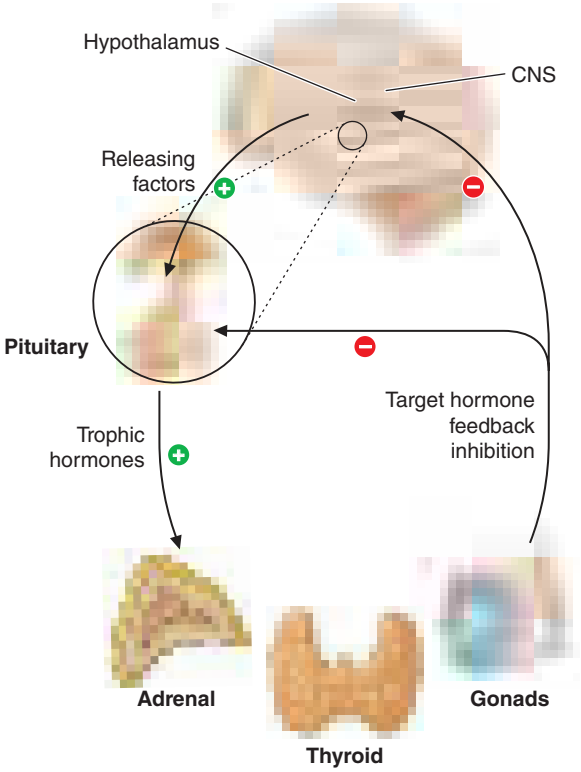


FIGURE 370-4 Feedback regulation of endocrine axes. CNS, central nervous system.

reduction of thyroid hormone triggers a rapid increase of TRH and TSH secretion, resulting in thyroid gland stimulation and increased thyroid hormone production. When thyroid hormone reaches a normal level, it feeds back to suppress TRH and TSH, and a new steady state is attained. Feedback regulation also occurs for endocrine systems that do not involve the pituitary gland, such as calcium feedback on PTH, glucose inhibition of insulin secretion, and leptin feedback on the hypothalamus. An understanding of feedback regulation provides important insights into endocrine testing paradigms (see below).

Positive feedback control also occurs but is not well understood. The primary example is estrogen-mediated stimulation of the midcycle LH surge. Although chronic low levels of estrogen are inhibitory, gradually rising estrogen levels stimulate LH secretion. This effect, which is illustrative of an endocrine rhythm (see below), involves activation of the hypothalamic GnRH pulse generator. In addition, estrogen-primed gonadotropes are extraordinarily sensitive to GnRH, leading to amplification of LH release.

PARACRINE AND AUTOCRINE CONTROL

The previously mentioned examples of feedback control involve classic endocrine pathways in which hormones are released by one gland and act on a distant target gland. However, local regulatory systems, often involving growth factors, are increasingly recognized. *Paracrine regulation* refers to factors released by one cell that act on an adjacent cell in the same tissue. For example, somatostatin secretion by pancreatic islet δ cells inhibits insulin secretion from nearby β cells. *Autocrine regulation* describes the action of a factor on the same cell from which it is produced. IGF-I acts on many cells that produce it, including chondrocytes, breast epithelium, and gonadal cells. Unlike endocrine actions, paracrine and autocrine control are difficult to document because local growth factor concentrations cannot be measured readily.

Anatomic relationships of glandular systems also greatly influence hormonal exposure: the physical organization of islet cells enhances their intercellular communication; the portal vasculature of the hypothalamic-pituitary system exposes the pituitary to high concentrations of hypothalamic releasing factors; testicular seminiferous tubules gain exposure to high testosterone levels produced by the interdigitated Leydig cells; the pancreas receives nutrient information and local exposure to peptide hormones (incretins) from the gastrointestinal tract; and the liver is the proximal target of insulin action because of portal drainage from the pancreas.

HORMONAL RHYTHMS

The feedback regulatory systems described above are superimposed on hormonal rhythms that are used for adaptation to the environment. Seasonal changes, the daily occurrence of the light-dark cycle, sleep, meals, and stress are examples of the many environmental events that affect hormonal rhythms. The *menstrual cycle* is repeated on average every 28 days, reflecting the time required to follicular maturation and ovulation **(Chap. 385)**. Essentially all pituitary hormone rhythms are entrained to sleep and to the *circadian cycle*, generating reproducible patterns that are repeated approximately every 24 h. The HPA axis, for example, exhibits characteristic peaks of ACTH and cortisol production in the early morning, with a nadir during the night. Recognition of these rhythms is important for endocrine testing and treatment. Patients with Cushing's syndrome characteristically exhibit increased midnight cortisol levels compared with normal individuals **(Chap. 379)**. In contrast, morning cortisol levels are similar in these groups, as cortisol is normally high at this time of day in normal individuals. The HPA axis is more susceptible to suppression by glucocorticoids administered at night as they blunt the early-morning rise of ACTH. Understanding these rhythms allows glucocorticoid replacement that mimics diurnal production by administering larger doses in the morning than in the afternoon. Disrupted sleep rhythms can alter hormonal regulation. For example, sleep deprivation causes mild insulin resistance, food craving, and hypertension, which are reversible, at least in the short term. Emerging evidence indicates that circadian clock pathways not only regulate sleep-wake cycles but also play important roles in virtually every cell type. For example, tissue-specific deletion of clock genes alters rhythms and levels of gene expression, as well as metabolic responses in liver, adipose, and other tissues.

Other endocrine rhythms occur on a more rapid time scale. Many peptide hormones are secreted in discrete bursts every few hours. LH and FSH secretion are exquisitely sensitive to GnRH pulse frequency. Intermittent pulses of GnRH are required to maintain pituitary sensitivity, whereas continuous exposure to GnRH causes pituitary gonadotrope desensitization. This feature of the hypothalamic-pituitary-gonadotrope axis forms the basis for using long-acting GnRH agonists to treat central precocious puberty or to decrease testosterone levels in the management of prostate cancer. It is important to be aware of the pulsatile nature of hormone secretion and the rhythmic patterns of hormone production in relating serum hormone measurements to normal values. For some hormones, integrated markers have been developed to circumvent hormonal fluctuations. Examples include 24-h urine collections for cortisol, IGF-I as a biologic marker of GH action, and HbA_{1c} as an index of long-term (weeks to months) blood glucose control.

Often, one must interpret endocrine data only in the context of other hormones. For example, PTH levels typically are assessed in combination with serum calcium concentrations. A high serum calcium level in association with elevated PTH is suggestive of hyperparathyroidism, whereas a suppressed PTH in this situation is more likely to be caused by hypercalcemia of malignancy or other causes of hypercalcemia. Similarly, TSH should be elevated when T_4 and T_3 concentrations are low, reflecting reduced feedback inhibition. When this is not the case, it is important to consider secondary hypothyroidism, which is caused by a defect at the level of the pituitary.

FURTHER READING

Evans RM, Mangelsdorf DJ: Nuclear receptors, RXR, and the big bang. Cell 157:255, 2014.

Jameson JL, DeGroot LJ (eds): *Endocrinology: Adult and Pediatric*, 7th ed. Philadelphia, Elsevier, 2016.

Melmed S et al (eds): *Williams Textbook of Endocrinology*, 13th ed. Philadelphia, Elsevier, 2016.

Moran C, Chatterjee K: Resistance to thyroid hormone α-emerging definition of a disorder of the thyroid hormone action. J Clin Endocrinol Metab 101:2317, 2016.

Vassart G, Costagliola S: G protein-coupled receptors: Mutations and endocrine diseases. Nature Rev Endocrinol 7:362, 2011.

371 Physiology of Anterior Pituitary Hormones

Shlomo Melmed, J. Larry Jameson

The anterior pituitary often is referred to as the "master gland" because, together with the hypothalamus, it orchestrates the complex regulatory functions of many other endocrine glands. The anterior pituitary gland produces six major hormones: (1) prolactin (PRL), (2) growth hormone (GH), (3) adrenocorticotropic hormone (ACTH), (4) luteinizing hormone (LH), (5) follicle-stimulating hormone (FSH), and (6) thyroid-stimulating hormone (TSH) **(Table 371-1)**. Pituitary hormones are secreted in a pulsatile manner, reflecting regulation by an array of specific hypothalamic releasing factors. Each of these pituitary hormones elicits specific trophic responses in peripheral target tissues. The hormonal products of those peripheral glands, in turn, exert feedback control at the level of the hypothalamus and pituitary to modulate pituitary function **(Fig. 371-1)**. Pituitary tumors cause characteristic hormone excess syndromes. Hormone deficiency may be inherited or acquired. Fortunately, there are efficacious treatments for many pituitary hormone excess and deficiency syndromes. Nonetheless, these diagnoses are often elusive; this emphasizes the importance of recognizing subtle clinical manifestations and performing the correct

TABLE 371-1 Anterior Pituitary Hormone Expression and Regulation

CELL	CORTICOTROPE	SOMATOTROPE	LACTOTROPE	THYROTROPE	GONADOTROPE
Tissue-specific transcription factor	T-Pit	Prop-1, Pit-1	Prop-1, Pit-1	Prop-1, Pit-1, TEF	SF-1, DAX-1
Fetal appearance	6 weeks	8 weeks	12 weeks	12 weeks	12 weeks
Hormone	POMC	GH	PRL	TSH	FSH, LH
Protein	Polypeptide	Polypeptide	Polypeptide	Glycoprotein α, β subunits	Glycoprotein α, β subunits
Amino acids	266 (ACTH 1–39)	191	198	211	210, 204
Stimulators	CRH, AVP, gp-130 cytokines	GHRH, ghrelin	Estrogen, TRH, VIP	TRH	GnRH, activins, estrogen
Inhibitors	Glucocorticoids	Somatostatin, IGF-I	Dopamine	T_3, T_4, dopamine, somatostatin, glucocorticoids	Sex steroids, inhibin
Target gland	Adrenal	Liver, bone, other tissues	Breast, other tissues	Thyroid	Ovary, testis
Trophic effect	Steroid production	IGF-I production, growth induction, insulin antagonism	Milk production	T_4 synthesis and secretion	Sex steroid production, follicle growth, germ cell maturation
Normal range	ACTH, 4–22 pg/L	<0.5 μg/L[a]	M <15 μg/L; F <20 μg/L	0.1–5 mU/L	M, 5–20 IU/L; F (basal), 5–20 IU/L

[a]Hormone secretion integrated over 24 h.

Abbreviations: M, male; F, female. For other abbreviations, see text.

Source: Adapted from S Melmed: Hypothalmic-pituitary regulation, in P Conn (ed): Conn's Translational Neuroscience. San Diego, CA, Elsevier, 2017.

laboratory diagnostic tests. **For discussion of disorders of the posterior pituitary, or neurohypophysis, see Chap. 374.**

ANATOMY AND DEVELOPMENT

ANATOMY

The pituitary gland weighs ~600 mg and is located within the sella turcica ventral to the diaphragma sella; it consists of anatomically and functionally distinct anterior and posterior lobes. The bony sella is contiguous to vascular and neurologic structures, including the cavernous sinuses, cranial nerves, and optic chiasm. Thus, expanding intrasellar pathologic processes may have significant central mass effects in addition to their endocrinologic impact.

Hypothalamic neural cells synthesize specific releasing and inhibiting hormones that are secreted directly into the portal vessels of the pituitary stalk. Blood supply of the pituitary gland comes from the superior and inferior hypophyseal arteries **(Fig. 371-2)**. The hypothalamic-pituitary portal plexus provides the major blood source for the anterior pituitary, allowing reliable transmission of hypothalamic peptide pulses without significant systemic dilution; consequently, pituitary cells are exposed to releasing or inhibiting factors and in turn release their respective hormones as discrete pulses into the systemic circulation **(Fig. 371-3)**.

The posterior pituitary is supplied by the inferior hypophyseal arteries. In contrast to the anterior pituitary, the posterior lobe is directly innervated by hypothalamic neurons (supraopticohypophyseal and tuberohypophyseal nerve tracts) via the pituitary stalk **(Chap. 374)**. Thus, posterior pituitary production of vasopressin (antidiuretic hormone [ADH]) and oxytocin is particularly sensitive to neuronal damage by lesions that affect the pituitary stalk or hypothalamus.

PITUITARY DEVELOPMENT

The embryonic differentiation and maturation of anterior pituitary cells have been elucidated in considerable detail. Pituitary development from Rathke's pouch involves a complex interplay of lineage-specific transcription factors expressed in pluripotent precursor cells and gradients of locally produced growth factors (Table 371-1). The transcription factor Prop-1 induces pituitary development of Pit-1-specific lineages as well as gonadotropes. The transcription factor Pit-1 determines cell-specific expression of GH, PRL, and TSH in somatotropes, lactotropes, and thyrotropes. Expression of high levels of estrogen receptors in cells that contain Pit-1 favors PRL expression, whereas thyrotrope embryonic factor (TEF) induces TSH expression. Pit-1 binds to GH, PRL, and TSH gene regulatory elements as well as to recognition sites on its own promoter, providing a mechanism for determining specific pituitary hormone phenotypic stability. Gonadotrope cell development is further defined by the cell-specific expression of the nuclear receptors steroidogenic factor (SF-1) and *d*osage-sensitive sex reversal, *a*drenal hypoplasia critical region, on chromosome *X*, gene *1* (DAX-1). Development of corticotrope cells, which express the proopiomelanocortin (*POMC*) gene, requires the T-Pit transcription factor. Abnormalities of pituitary development caused by mutations of Pit-1, Prop-1, SF-1, DAX-1, and T-Pit result in rare, selective or combined pituitary hormone deficit syndromes.

ANTERIOR PITUITARY HORMONES

Each anterior pituitary hormone is under unique control, and each exhibits highly specific normal and dysregulated secretory characteristics.

PROLACTIN

Synthesis PRL consists of 198 amino acids and has a molecular mass of 21,500 kDa; it is weakly homologous to GH and human placental lactogen (hPL), reflecting the duplication and divergence of a common GH-PRL-hPL precursor gene. PRL is synthesized in lactotropes, which constitute about 20% of anterior pituitary cells. Lactotropes and somatotropes are derived from a common precursor cell that may give rise to a tumor that secretes both PRL and GH. Marked lactotrope cell hyperplasia develops during pregnancy and the first few months of lactation. These transient functional changes in the lactotrope population are induced by estrogen.

Secretion Normal adult serum PRL levels are about 10–25 μg/L in women and 10–20 μg/L in men. PRL secretion is pulsatile, with the highest secretory peaks occurring during rapid eye movement sleep. Peak serum PRL levels (up to 30 μg/L) occur between 4:00 and 6:00 A.M. The circulating half-life of PRL is about 50 min.

PRL is unique among the pituitary hormones in that the predominant central control mechanism is inhibitory, reflecting tonic dopamine-mediated suppression of PRL release. This regulatory pathway accounts for the spontaneous PRL hypersecretion that occurs with pituitary stalk section, often a consequence of head trauma or compressive mass lesions at the skull base. Pituitary dopamine type 2 (D_2) receptors mediate inhibition of PRL synthesis and secretion. Targeted disruption (gene knockout) of the murine D_2 receptor in mice results in hyperprolactinemia and lactotrope proliferation. As discussed below, dopamine agonists play a central role in the management of hyperprolactinemic disorders.

Thyrotropin-releasing hormone (TRH) (pyro Glu-His-Pro-NH_2) is a hypothalamic tripeptide that elicits PRL release within 15–30 min

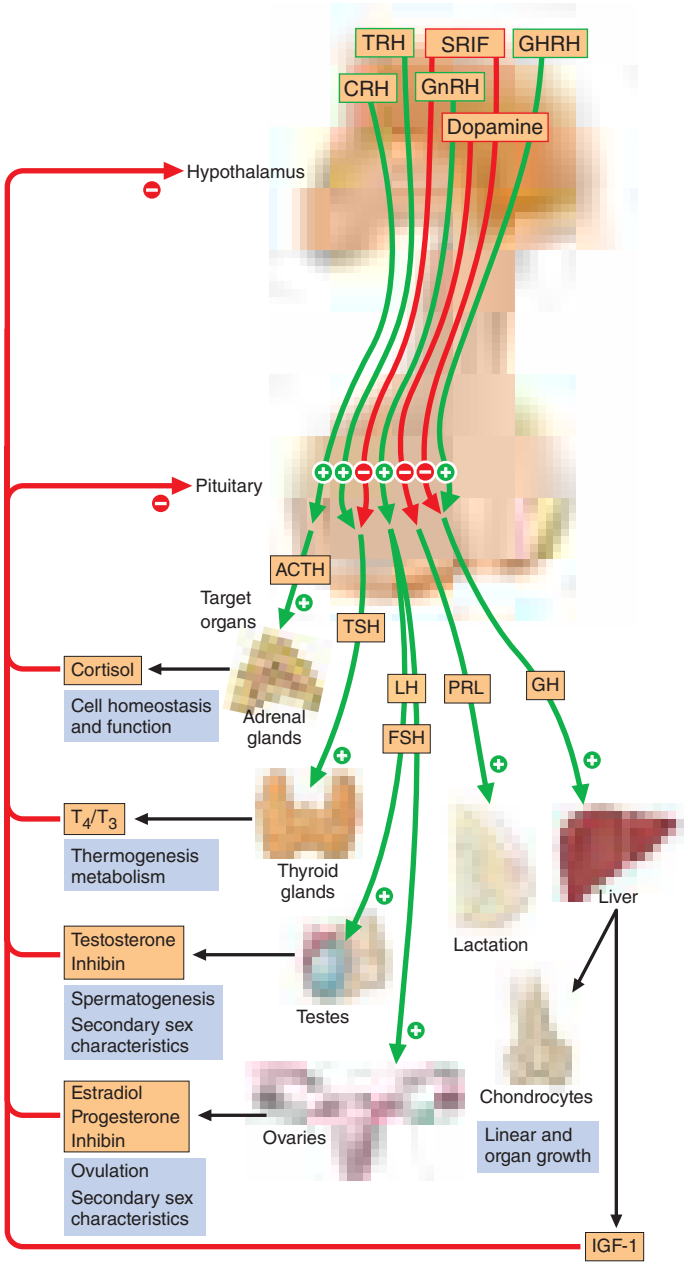


FIGURE 371-1 Diagram of pituitary axes. Hypothalamic hormones regulate anterior pituitary trophic hormones that in turn determine target gland secretion. Peripheral hormones feed back to regulate hypothalamic and pituitary hormones. For abbreviations, see text.

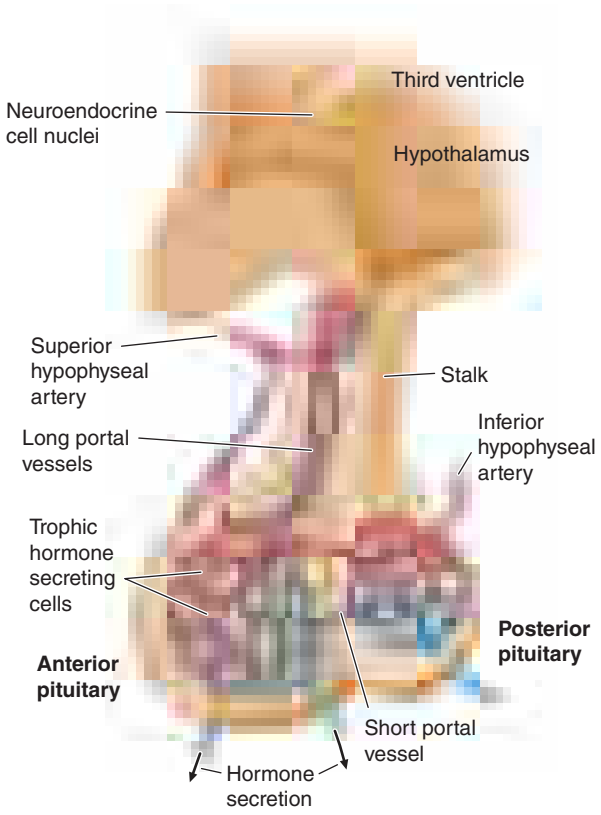


FIGURE 371-2 Diagram of hypothalamic-pituitary vasculature. The hypothalamic nuclei produce hormones that traverse the portal system and impinge on anterior pituitary cells to regulate pituitary hormone secretion. Posterior pituitary hormones are derived from direct neural extensions.

after intravenous injection. TRH primarily regulates TSH, and the physiologic relevance of TRH for PRL regulation is unclear **(Chap. 375)**. *Vasoactive intestinal peptide* (VIP) also induces PRL release, whereas glucocorticoids and thyroid hormone weakly suppress PRL secretion.

Serum PRL levels rise transiently after exercise, meals, sexual intercourse, minor surgical procedures, general anesthesia, chest wall injury, acute myocardial infarction, and other forms of acute stress. PRL levels increase markedly (about tenfold) during pregnancy and decline rapidly within 2 weeks of parturition. If breast-feeding is initiated, basal PRL levels remain elevated; suckling stimulates transient reflex increases in PRL levels that last for about 30–45 min. Breast suckling activates afferent neural pathways in the hypothalamus that induce PRL release. With time, suckling-induced responses diminish and interfeeding PRL levels return to normal.

Action The PRL receptor is a member of the type I cytokine receptor family that also includes GH and interleukin (IL) 6 receptors. Ligand binding induces receptor dimerization and intracellular signaling by Janus kinase (JAK), which stimulates translocation of the signal transduction and activators of transcription (STAT) family to activate target genes. Heterozygous mutations of the PRL receptor result in PRL insensitivity, hyperprolactinemia, and oligomenorrhea. In the breast, the lobuloalveolar epithelium proliferates in response to PRL, placental lactogens, estrogen, progesterone, and local paracrine growth factors, including insulin-like growth factor I (IGF-I).

PRL acts to induce and maintain lactation, decrease reproductive function, and suppress sexual drive. These functions are geared toward ensuring that maternal lactation is sustained and not interrupted by pregnancy. PRL inhibits reproductive function by suppressing hypothalamic gonadotropin-releasing hormone (GnRH) and pituitary gonadotropin secretion and by impairing gonadal steroidogenesis in both women and men. In the ovary, PRL blocks folliculogenesis and inhibits granulosa cell aromatase activity, leading to hypoestrogenism and anovulation. PRL also has a luteolytic effect, generating a shortened, or inadequate, luteal phase of the menstrual cycle. In men, attenuated LH secretion leads to low testosterone levels and decreased spermatogenesis. These hormonal changes decrease libido and reduce fertility in patients with hyperprolactinemia.

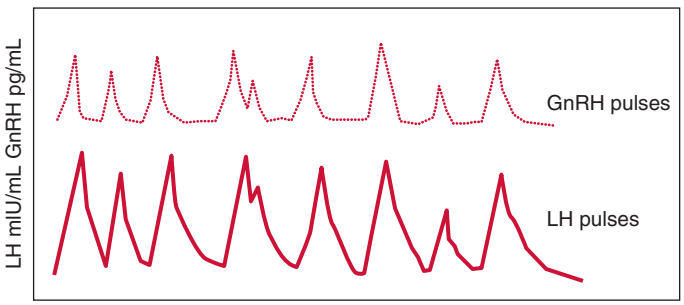


FIGURE 371-3 Hypothalamic gonadotropin-releasing hormone (GnRH) pulses induce secretory pulses of luteinizing hormone (LH).

Synthesis GH is the most abundant anterior pituitary hormone, and GH-secreting somatotrope cells constitute up to 50% of the total anterior pituitary cell population. Mammosomatotrope cells, which coexpress PRL with GH, can be identified by using double immunostaining techniques. Somatotrope development and GH transcription are determined by expression of the cell-specific Pit-1 nuclear transcription factor. Five distinct genes encode GH and related proteins. The pituitary GH gene (*hGH-N*) produces two alternatively spliced products that give rise to 22-kDa GH (191 amino acids) and a less abundant 20-kDa GH molecule with similar biologic activity. Placental syncytiotrophoblast cells express a GH variant (*hGH-V*) gene; the related hormone human chorionic somatotropin (HCS) is expressed by distinct members of the gene cluster.

Secretion GH secretion is controlled by complex hypothalamic and peripheral factors. *GH-releasing hormone* (GHRH) is a 44-amino-acid hypothalamic peptide that stimulates GH synthesis and release. Ghrelin, an octanoylated gastric-derived peptide, and synthetic agonists of the *GHS-R* induce GHRH and also directly stimulate GH release. *Somatostatin* (somatotropin-release inhibiting factor [SRIF]) is synthesized in the medial preoptic area of the hypothalamus and inhibits GH secretion. GHRH is secreted in discrete spikes that elicit GH pulses, whereas SRIF sets basal GH secretory tone. SRIF also is expressed in many extrahypothalamic tissues, including the central nervous system (CNS), gastrointestinal tract, and pancreas, where it also acts to inhibit islet hormone secretion. *IGF-I*, the peripheral target hormone for GH, feeds back to inhibit GH; estrogen induces GH, whereas chronic glucocorticoid excess suppresses GH release.

Surface receptors on the somatotrope regulate GH synthesis and secretion. The GHRH receptor is a G protein–coupled receptor (GPCR) that signals through the intracellular cyclic AMP pathway to stimulate somatotrope cell proliferation as well as GH production. Inactivating mutations of the GHRH receptor cause profound dwarfism. A distinct surface receptor for ghrelin, the gastric-derived GH secretagogue, is expressed in both the hypothalamus and pituitary. Somatostatin binds to five distinct receptor subtypes (SST1 to SST5); SST2 and SST5 subtypes preferentially suppress GH (and TSH) secretion and SST5 signals to suppress ACTH secretion.

GH secretion is pulsatile, with highest peak levels occurring at night, generally correlating with sleep onset. GH secretory rates decline markedly with age so that hormone levels in middle age are about 15% of pubertal levels. These changes are paralleled by an age-related decline in lean muscle mass. GH secretion is also reduced in obese individuals, although IGF-I levels may not be suppressed, suggesting a change in the setpoint for feedback control. Elevated GH levels occur within an hour of deep sleep onset as well as after exercise, physical stress, and trauma and during sepsis. Integrated 24-h GH secretion is higher in women and is also enhanced by estrogen replacement, likely reflective of increased peripheral GH resistance. Using standard assays, random GH measurements are undetectable in ~50% of daytime samples obtained from healthy subjects and are also undetectable in most obese and elderly subjects. Thus, single random GH measurements do not distinguish patients with adult GH deficiency from normal persons.

GH secretion is profoundly influenced by nutritional factors. Using newer ultrasensitive GH assays with a sensitivity of 0.002 μg/L, a glucose load suppresses GH to <0.7 μg/L in women and to <0.07 μg/L in men. Increased GH pulse frequency and peak amplitudes occur with chronic malnutrition or prolonged fasting. GH is stimulated by intravenous L-arginine, dopamine, and apomorphine (a dopamine receptor agonist), as well as by α-adrenergic pathways. β-Adrenergic blockade induces basal GH and enhances GHRH- and insulin-evoked GH release.

Action The pattern of GH secretion may affect tissue responses. The higher GH pulsatility observed in men compared with the relatively continuous basal GH secretion in women may be an important biologic determinant of linear growth patterns and liver enzyme induction.

The 70-kDa peripheral GH receptor protein has structural homology with the cytokine/hematopoietic superfamily. A fragment of the receptor extracellular domain generates a soluble GH binding protein (GHBP) that interacts with GH in the circulation. The liver and cartilage express the greatest number of GH receptors. GH binding to preformed receptor dimers is followed by internal rotation and subsequent signaling through the JAK/STAT pathway. Activated STAT proteins translocate to the nucleus, where they modulate expression of GH-regulated target genes. GH analogues that bind to the receptor but are incapable of mediating receptor signaling are potent antagonists of GH action. A GH receptor antagonist (pegvisomant) is approved for treatment of acromegaly.

GH induces protein synthesis and nitrogen retention and also impairs glucose tolerance by antagonizing insulin action. GH also stimulates lipolysis, leading to increased circulating fatty acid levels, reduced omental fat mass, and enhanced lean body mass. GH promotes sodium, potassium, and water retention and elevates serum levels of inorganic phosphate. Linear bone growth occurs as a result of complex hormonal and growth factor actions, including those of IGF-I. GH stimulates epiphyseal prechondrocyte differentiation. These precursor cells produce IGF-I locally, and their proliferation is also responsive to the growth factor.

Insulin-Like Growth Factors Although GH exerts direct effects in target tissues, many of its physiologic effects are mediated indirectly through IGF-I, a potent growth and differentiation factor. The liver is the major source of circulating IGF-I. In peripheral tissues, IGF-I also exerts local paracrine actions that appear to be both dependent on and independent of GH. Thus, GH administration induces circulating IGF-I as well as stimulating local IGF-I production in multiple tissues.

Both IGF-I and IGF-II are bound to high-affinity circulating IGF-binding proteins (IGFBPs) that regulate IGF availability and bioactivity. Levels of IGFBP3 are GH-dependent, and it serves as the major carrier protein for circulating IGF-I. GH deficiency and malnutrition usually are associated with low IGFBP3 levels. IGFBP1 and IGFBP2 regulate local tissue IGF action but do not bind appreciable amounts of circulating IGF-I.

Serum IGF-I concentrations are profoundly affected by physiologic factors. Levels increase during puberty, peak at 16 years, and subsequently decline by >80% during the aging process. IGF-I concentrations are higher in women than in men. Because GH is the major determinant of hepatic IGF-I synthesis, abnormalities of GH synthesis or action (e.g., pituitary failure, GHRH receptor defect, GH receptor defect or pharmacologic GH receptor blockade) reduce IGF-I levels. Hypocaloric states are associated with GH resistance; IGF-I levels are therefore low with cachexia, malnutrition, and sepsis. In acromegaly, IGF-I levels are invariably high and reflect a log-linear relationship with circulating GH concentrations.

IGF-I PHYSIOLOGY Injected IGF-I (100 μg/kg) induces hypoglycemia, and lower doses improve insulin sensitivity in patients with severe insulin resistance and diabetes. In cachectic subjects, IGF-I infusion (12 μg/kg per h) enhances nitrogen retention and lowers cholesterol levels. Longer-term subcutaneous IGF-I injections enhance protein synthesis and are anabolic. Although bone formation markers are induced, bone turnover also may be stimulated by IGF-I. IGF-I is approved for use in patients with GH-resistance syndromes.

IGF-I side effects are dose-dependent, and overdose may result in hypoglycemia, hypotension, fluid retention, temporomandibular jaw pain, and increased intracranial pressure, all of which are reversible. Avascular femoral head necrosis has been reported. Chronic excess IGF-I administration presumably would result in features of acromegaly.

■ ADRENOCORTICOTROPIC HORMONE

(See also Chap. 379)

Synthesis ACTH-secreting corticotrope cells constitute about 20% of the pituitary cell population. ACTH (39 amino acids) is derived from the POMC precursor protein (266 amino acids) that also generates

several other peptides, including β-lipotropin, β-endorphin, met-enkephalin, α-melanocyte-stimulating hormone (α-MSH), and corticotropin-like intermediate lobe protein (CLIP). The *POMC* gene is potently suppressed by glucocorticoids and induced by corticotropin-releasing hormone (CRH), arginine vasopressin (AVP), and proinflammatory cytokines, including IL-6, as well as leukemia inhibitory factor.

CRH, a 41-amino-acid hypothalamic peptide synthesized in the paraventricular nucleus as well as in higher brain centers, is the predominant stimulator of ACTH synthesis and release. The CRH receptor is a GPCR that is expressed on the corticotrope and signals to induce *POMC* transcription.

Secretion ACTH secretion is pulsatile and exhibits a characteristic circadian rhythm, peaking at about 6:00 A.M. and reaching a nadir about midnight. Adrenal glucocorticoid secretion, which is driven by ACTH, follows a parallel diurnal pattern. ACTH circadian rhythmicity is determined by variations in secretory pulse amplitude rather than changes in pulse frequency. Superimposed on this endogenous rhythm, ACTH levels are increased by physical and psychological stress, exercise, acute illness, and insulin-induced hypoglycemia.

Glucocorticoid-mediated negative regulation of the hypothalamic-pituitary-adrenal (HPA) axis occurs as a consequence of both hypothalamic CRH suppression and direct attenuation of pituitary *POMC* gene expression and ACTH release. In contrast, loss of cortisol feedback inhibition, as occurs in primary adrenal failure, results in extremely high ACTH levels.

Acute inflammatory or septic insults activate the HPA axis through the integrated actions of proinflammatory cytokines, bacterial toxins, and neural signals. The overlapping cascade of ACTH-inducing cytokines (tumor necrosis factor [TNF]; IL-1, -2, and -6; and leukemia inhibitory factor) activates hypothalamic CRH and AVP secretion, pituitary *POMC* gene expression, and local pituitary paracrine cytokine networks. The resulting cortisol elevation restrains the inflammatory response and enables host protection. Concomitantly, cytokine-mediated central glucocorticoid receptor resistance impairs glucocorticoid suppression of the HPA. Thus, the neuroendocrine stress response reflects the net result of highly integrated hypothalamic, intrapituitary, and peripheral hormone and cytokine signals acting to regulate cortisol secretion.

Action The major function of the HPA axis is to maintain metabolic homeostasis and mediate the neuroendocrine stress response. ACTH induces adrenocortical steroidogenesis by sustaining adrenal cell proliferation and function. The receptor for ACTH, designated *melanocortin-2 receptor*, is a GPCR that induces steroidogenesis by stimulating a cascade of steroidogenic enzymes **(Chap. 379)**.

GONADOTROPINS: FSH AND LH

Synthesis and Secretion Gonadotrope cells constitute about 10% of anterior pituitary cells and produce two gonadotropin hormones—LH and FSH. Like TSH and hCG, LH and FSH are glycoprotein hormones that comprise α and β subunits. The α subunit is common to these glycoprotein hormones; specificity of hormone function is conferred by the β subunits, which are expressed by separate genes.

Gonadotropin synthesis and release are dynamically regulated. This is particularly true in women, in whom rapidly fluctuating gonadal steroid levels vary throughout the menstrual cycle. Hypothalamic GnRH, a 10-amino-acid peptide, regulates the synthesis and secretion of both LH and FSH. Brain kisspeptin, a product of the *KISSI* gene, regulates hypothalamic GnRH release. GnRH is secreted in discrete pulses every 60–120 min, and the pulses in turn elicit LH and FSH pulses (Fig. 371-3). The pulsatile mode of GnRH input is essential to its action; pulses prime gonadotrope responsiveness, whereas continuous GnRH exposure induces desensitization. Based on this phenomenon, long-acting GnRH agonists are used to suppress gonadotropin levels in children with precocious puberty and in men with prostate cancer **(Chap. 83)** and are used in some ovulation-induction protocols to reduce levels of endogenous gonadotropins **(Chap. 385)**. Estrogens act at both the hypothalamus and the pituitary to modulate gonadotropin secretion. Chronic estrogen exposure is inhibitory, whereas rising estrogen levels, as occur during the preovulatory surge, exert positive feedback to increase gonadotropin pulse frequency and amplitude. Progesterone slows GnRH pulse frequency but enhances gonadotropin responses to GnRH. Testosterone feedback in men also occurs at the hypothalamic and pituitary levels and is mediated in part by its conversion to estrogens.

Although GnRH is the main regulator of LH and FSH secretion, FSH synthesis is also under separate control by the gonadal peptides inhibin and activin, which are members of the transforming growth factor β (TGF-β) family. Inhibin selectively suppresses FSH, whereas activin stimulates FSH synthesis **(Chap. 385)**.

Action The gonadotropin hormones interact with their respective GPCRs expressed in the ovary and testis, evoking germ cell development and maturation and steroid hormone biosynthesis. In women, FSH regulates ovarian follicle development and stimulates ovarian estrogen production. LH mediates ovulation and maintenance of the corpus luteum. In men, LH induces Leydig cell testosterone synthesis and secretion, and FSH stimulates seminiferous tubule development and regulates spermatogenesis.

THYROID-STIMULATING HORMONE

Synthesis and Secretion TSH-secreting thyrotrope cells constitute 5% of the anterior pituitary cell population. TSH shares a common α subunit with LH and FSH but contains a specific TSH β subunit. TRH is a hypothalamic tripeptide (pyroglutamyl histidylprolinamide) that acts through a pituitary GPCR to stimulate TSH synthesis and secretion; it also stimulates the lactotrope cell to secrete PRL. TSH secretion is stimulated by TRH, whereas thyroid hormones, dopamine, somatostatin, and glucocorticoids suppress TSH by overriding TRH induction.

Thyrotrope cell proliferation and TSH secretion are both induced when negative feedback inhibition by thyroid hormones is removed. Thus, thyroid damage (including surgical thyroidectomy), radiation-induced hypothyroidism, chronic thyroiditis, and prolonged goitrogen exposure are associated with increased TSH levels. Long-standing untreated hypothyroidism can lead to elevated TSH levels, which may be associated with thyrotrope hyperplasia and pituitary enlargement and may sometimes be evident on magnetic resonance imaging.

Action TSH is secreted in pulses, although the excursions are modest in comparison to other pituitary hormones because of the low amplitude of the pulses and the relatively long half-life of TSH. Consequently, single determinations of TSH suffice to precisely assess its circulating levels. TSH binds to a GPCR on thyroid follicular cells to stimulate thyroid hormone synthesis and release **(Chap. 375)**.

FURTHER READING

BERNARD V et al: New insights in prolactin: Pathological implications. Nat Rev Endocrinol 11:265, 2015.

BROOKS AJ, WATERS MJ: The growth hormone receptor: Mechanism of activation and clinical implications. Nat Rev Endocrinol 6:515, 2010.

HO K: The year in pituitary 2014. J Clin Endocrinol Metab 99:4449, 2014.

KELBERMAN D et al: Genetic regulation of pituitary gland development in human and mouse. Endocr Rev 30:790, 2009.

LANGLAIS D et al: Adult pituitary cell maintenance: Lineage-specific contribution of self-duplication. Mol Endocrinol 27:1103, 2013.

MURRAY PG, HIGHAM CE, CLAYTON PE: 60 years of neuroendocrinology: The hypothalamo-GH axis: The past 60 years. J Endocrinol 226:T123, 2015.

NEWEY PJ et al: Mutant prolactin receptor and familial hyperprolactinemia. N Engl J Med 369:2012, 2013.

372 Hypopituitarism

Shlomo Melmed, J. Larry Jameson

Inadequate production of anterior pituitary hormones leads to features of hypopituitarism. Impaired production of one or more of the anterior pituitary trophic hormones can result from inherited disorders; more commonly, adult hypopituitarism is acquired and reflects the compressive mass effects of tumors or the consequences of local pituitary or hypothalamic traumatic, inflammatory, or vascular damage. These processes also may impair synthesis or secretion of hypothalamic hormones, with resultant pituitary failure (Table 372-1).

DEVELOPMENTAL AND GENETIC CAUSES OF HYPOPITUITARISM

Pituitary Dysplasia Pituitary dysplasia may result in aplastic, hypoplastic, or ectopic pituitary gland development. Because pituitary development follows midline cell migration from the nasopharyngeal Rathke's pouch, midline craniofacial disorders may be associated with pituitary dysplasia. Acquired pituitary failure in the newborn also can be caused by birth trauma, including cranial hemorrhage, asphyxia, and breech delivery.

TABLE 372-1 Etiology of Hypopituitarism[a]

Development/structural
Transcription factor defect
Pituitary dysplasia/aplasia
Congenital central nervous system mass, encephalocele
Primary empty sella
Congenital hypothalamic disorders (septo-optic dysplasia, Prader-Willi syndrome, Bardet-Biedl syndrome, Kallmann syndrome)
Traumatic
Surgical resection
Radiation damage
Head injuries
Neoplastic
Pituitary adenoma
Parasellar mass (germinoma, ependymoma, glioma)
Rathke's cyst
Craniopharyngioma
Hypothalamic hamartoma, gangliocytoma
Pituitary metastases (breast, lung, colon carcinoma)
Lymphoma and leukemia
Meningioma
Infiltrative/inflammatory
Lymphocytic hypophysitis
Hemochromatosis
Sarcoidosis
Histiocytosis X
Granulomatous hypophysitis
Transcription factor antibodies
Immunotherapy
Vascular
Pituitary apoplexy
Pregnancy-related (infarction with diabetes; postpartum necrosis)
Sickle cell disease
Arteritis
Infections
Fungal (histoplasmosis)
Parasitic (toxoplasmosis)
Tuberculosis
Pneumocystis jirovecii

[a]Trophic hormone failure associated with pituitary compression or destruction usually occurs sequentially: growth hormone > follicle-stimulating hormone > luteinizing hormone > thyroid-stimulating hormone > adrenocorticotropic hormone. During childhood, growth retardation is often the presenting feature, and in adults, hypogonadism is the earliest symptom.

SEPTO-OPTIC DYSPLASIA Hypothalamic dysfunction and hypopituitarism may result from dysgenesis of the septum pellucidum or corpus callosum. Affected children have mutations in the *HESX1* gene, which is involved in early development of the ventral prosencephalon. These children exhibit variable combinations of cleft palate, syndactyly, ear deformities, hypertelorism, optic nerve hypoplasia, micropenis, and anosmia. Pituitary dysfunction leads to diabetes insipidus, growth hormone (GH) deficiency and short stature, and, occasionally, thyroid-stimulating hormone (TSH) deficiency.

Tissue-Specific Factor Mutations Several pituitary cell–specific transcription factors, such as Pit-1 and Prop-1, are critical for determining the development and committed function of differentiated anterior pituitary cell lineages. Autosomal dominant or recessive Pit-1 mutations cause combined GH, prolactin (PRL), and TSH deficiencies. These patients usually present with growth failure and varying degrees of hypothyroidism. The pituitary may appear hypoplastic on magnetic resonance imaging (MRI).

Prop-1 is expressed early in pituitary development and appears to be required for Pit-1 function. Familial and sporadic *PROP1* mutations result in combined GH, PRL, TSH, and gonadotropin deficiency. Over 80% of these patients have growth retardation; by adulthood, all are deficient in TSH and gonadotropins, and a small minority later develop adrenocorticotropic hormone (ACTH) deficiency. Because of gonadotropin deficiency, these individuals do not enter puberty spontaneously. In some cases, the pituitary gland appears enlarged on MRI. *TPIT* mutations result in ACTH deficiency associated with hypocortisolism.

Developmental Hypothalamic Dysfunction • KALLMANN SYNDROME Kallmann syndrome results from defective hypothalamic gonadotropin-releasing hormone (GnRH) synthesis and is associated with anosmia or hyposmia due to olfactory bulb agenesis or hypoplasia (Chap. 384). Classically, the syndrome may also be associated with color blindness, optic atrophy, nerve deafness, cleft palate, renal abnormalities, cryptorchidism, and neurologic abnormalities such as mirror movements. The initial genetic cause was identified in the X-linked *KAL* gene, mutations of which impair embryonic migration of GnRH neurons from the hypothalamic olfactory placode to the hypothalamus. Since then, at least a dozen additional genetic abnormalities, in addition to *KAL* mutations, have been found to cause isolated GnRH deficiency. Autosomal recessive (i.e., *GPR54, KISS1*) and dominant (i.e., *FGFR1*) modes of transmission have been described, and there is a growing list of genes associated with GnRH deficiency (including *GNRH1, PROK2, PROKR2, CHD7, PCSK1, FGF8, NELF, WDR11, TAC3, TACR3,* and *SEMA3E*). Some patients have oligogenic mutations. Associated clinical features, in addition to GnRH deficiency, vary depending on the genetic cause. GnRH deficiency prevents progression through puberty. Males present with delayed puberty and pronounced hypogonadal features, including micropenis, probably the result of low testosterone levels during infancy. Females present with primary amenorrhea and failure of secondary sexual development.

Kallmann syndrome and other causes of congenital GnRH deficiency are characterized by low luteinizing hormone (LH) and follicle-stimulating hormone (FSH) levels and low concentrations of sex steroids (testosterone or estradiol). In sporadic cases of isolated gonadotropin deficiency, the diagnosis is often one of exclusion after other known causes of hypothalamic-pituitary dysfunction have been eliminated. Repetitive GnRH administration restores normal pituitary gonadotropin responses, pointing to a hypothalamic defect in these patients.

Long-term treatment of males with human chorionic gonadotropin (hCG) or testosterone restores pubertal development and secondary sex characteristics; women can be treated with cyclic estrogen and progestin. Fertility may be restored by the administration of

gonadotropins or by using a portable infusion pump to deliver subcutaneous, pulsatile GnRH.

BARDET-BIEDL SYNDROME This very rare genetically heterogeneous disorder is characterized by mental retardation, renal abnormalities, obesity, and hexadactyly, brachydactyly, or syndactyly. Central diabetes insipidus may or may not be associated. GnRH deficiency occurs in 75% of males and half of affected females. Retinal degeneration begins in early childhood, and most patients are blind by age 30. Numerous subtypes of Bardet-Biedl syndrome (BBS) have been identified, with genetic linkage to at least nine different loci. Several of the loci encode genes involved in basal body cilia function, and this may account for the diverse clinical manifestations.

LEPTIN AND LEPTIN RECEPTOR MUTATIONS Deficiencies of leptin or its receptor cause a broad spectrum of hypothalamic abnormalities, including hyperphagia, obesity, and central hypogonadism **(Chap. 394)**. Decreased GnRH production in these patients results in attenuated pituitary FSH and LH synthesis and release.

PRADER-WILLI SYNDROME This is a contiguous gene syndrome that results from deletion of the paternal copies of the imprinted *SNRPN* gene, the *NECDIN* gene, and possibly other genes on chromosome 15q. Prader-Willi syndrome is associated with hypogonadotropic hypogonadism, hyperphagia-obesity, chronic muscle hypotonia, mental retardation, and adult-onset diabetes mellitus. Multiple somatic defects also involve the skull, eyes, ears, hands, and feet. Diminished hypothalamic oxytocin- and vasopressin-producing nuclei have been reported. Deficient GnRH synthesis is suggested by the observation that chronic GnRH treatment restores pituitary LH and FSH release.

ACQUIRED HYPOPITUITARISM

Hypopituitarism may be caused by accidental or neurosurgical trauma; vascular events such as apoplexy; pituitary or hypothalamic neoplasms, craniopharyngioma, lymphoma, or metastatic tumors; inflammatory disease such as lymphocytic hypophysitis; autoimmune hypophysitis associated with checkpoint inhibitor cancer immunotherapy; infiltrative disorders such as sarcoidosis, hemochromatosis **(Chap. 407)**, and tuberculosis; or irradiation.

Increasing evidence suggests that patients with brain injury, including contact sports trauma, subarachnoid hemorrhage, and irradiation, have transient hypopituitarism and require intermittent long-term endocrine follow-up, because permanent hypothalamic or pituitary dysfunction will develop in 25–40% of these patients.

Hypothalamic Infiltration Disorders These disorders—including sarcoidosis, histiocytosis X, amyloidosis, and hemochromatosis—frequently involve both hypothalamic and pituitary neuronal and neurochemical tracts. Consequently, diabetes insipidus occurs in half of patients with these disorders. Growth retardation is seen if attenuated GH secretion occurs before puberty. Hypogonadotropic hypogonadism and hyperprolactinemia are also common.

Inflammatory Lesions Pituitary damage and subsequent secretory dysfunction can be seen with chronic site infections such as tuberculosis, with opportunistic fungal infections associated with AIDS, and in tertiary syphilis. Other inflammatory processes, such as granulomas and sarcoidosis, may mimic the features of a pituitary adenoma. These lesions may cause extensive hypothalamic and pituitary damage, leading to trophic hormone deficiencies.

Cranial Irradiation Cranial irradiation may result in long-term hypothalamic and pituitary dysfunction, especially in children and adolescents, as they are more susceptible to damage after whole-brain or head and neck therapeutic irradiation. The development of hormonal abnormalities correlates strongly with irradiation dosage and the time interval after completion of radiotherapy. Up to two-thirds of patients ultimately develop hormone insufficiency after a median dose of 50 Gy (5000 rad) directed at the skull base. The development of hypopituitarism occurs over 5–15 years and usually reflects hypothalamic damage rather than primary destruction of pituitary cells. Although the pattern of hormone loss is variable, GH deficiency is most common, followed by gonadotropin and ACTH deficiency. When deficiency of one or more hormones is documented, the possibility of diminished reserve of other hormones is likely. Accordingly, anterior pituitary function should be continually evaluated over the long term in previously irradiated patients, and replacement therapy instituted when appropriate (see below).

Lymphocytic Hypophysitis This occurs most often in postpartum women; it usually presents with hyperprolactinemia and MRI evidence of a prominent pituitary mass that often resembles an adenoma, with mildly elevated PRL levels. Pituitary failure caused by diffuse lymphocytic infiltration may be transient or permanent but requires immediate evaluation and treatment. Rarely, isolated pituitary hormone deficiencies have been described, suggesting a selective autoimmune process targeted to specific cell types. Most patients manifest symptoms of progressive mass effects with headache and visual disturbance. The erythrocyte sedimentation rate often is elevated. Because the MRI image may be indistinguishable from that of a pituitary adenoma, hypophysitis should be considered in a postpartum woman with a newly diagnosed pituitary mass before an unnecessary surgical intervention is undertaken. The inflammatory process often resolves after several months of glucocorticoid treatment, and pituitary function may be restored, depending on the extent of damage.

Immunotherapy and Hypophysitis Pituitary cells express cytotoxic T lymphocyte antigen-4 (CTLA-4) and up to 20% of patients receiving cancer immunotherapy with CTLA-4 blockers (e.g., ipilimumab) may develop hypophysitis with associated thyroid adrenal and gonadal failure. Pituitary hormone replacement, with or without high-dose glucocorticoids, may be safely tolerated with continued immunotherapy.

Pituitary Apoplexy Acute intrapituitary hemorrhagic vascular events can cause substantial damage to the pituitary and surrounding sellar structures. Pituitary apoplexy may occur spontaneously in a preexisting adenoma; postpartum (Sheehan's syndrome); or in association with diabetes, hypertension, sickle cell anemia, or acute shock. The hyperplastic enlargement of the pituitary, which occurs normally during pregnancy, increases the risk for hemorrhage and infarction. Apoplexy is an endocrine emergency that may result in severe hypoglycemia, hypotension and shock, central nervous system (CNS) hemorrhage, and death. Acute symptoms may include severe headache with signs of meningeal irritation, bilateral visual changes, ophthalmoplegia, and, in severe cases, cardiovascular collapse and loss of consciousness. Pituitary computed tomography (CT) or MRI may reveal signs of intratumoral or sellar hemorrhage, with pituitary stalk deviation and compression of pituitary tissue.

Patients with no evident visual loss or impaired consciousness can be observed and managed conservatively with high-dose glucocorticoids. Those with significant or progressive visual loss, cranial nerve palsy, or loss of consciousness require urgent surgical decompression. Visual recovery after sellar surgery is inversely correlated with the length of time after the acute event. Therefore, severe ophthalmoplegia or visual deficits are indications for early surgery. Hypopituitarism is common after apoplexy.

Empty Sella A partial or apparently totally empty sella is often an incidental MRI finding, and may be associated with intracranial hypertension. These patients usually have normal pituitary function, implying that the surrounding rim of pituitary tissue is fully functional. Hypopituitarism, however, may develop insidiously. Pituitary masses also may undergo clinically silent infarction and involution with development of a partial or totally empty sella by cerebrospinal fluid (CSF) filling the dural herniation. Rarely, small but functional pituitary adenomas may arise within the rim of normal pituitary tissue, and they are not always visible on MRI.

PRESENTATION AND DIAGNOSIS

The clinical manifestations of hypopituitarism depend on which hormones are lost and the extent of the hormone deficiency. GH deficiency causes growth disorders in children and leads to abnormal body

composition in adults (see below). Gonadotropin deficiency causes menstrual disorders and infertility in women and decreased sexual function, infertility, and loss of secondary sexual characteristics in men. TSH and ACTH deficiencies usually develop later in the course of pituitary failure. TSH deficiency causes growth retardation in children and features of hypothyroidism in children and adults. The secondary form of adrenal insufficiency caused by ACTH deficiency leads to hypocortisolism with relative preservation of mineralocorticoid production. PRL deficiency causes failure of lactation. When lesions involve the posterior pituitary, polyuria and polydipsia reflect loss of vasopressin secretion. In patients with long-standing pituitary damage, epidemiologic studies document an increased mortality rate, primarily from increased cardiovascular and cerebrovascular disease. Previous head or neck irradiation is also a determinant of increased mortality rates in patients with hypopituitarism, especially from cerebrovascular disease.

LABORATORY INVESTIGATION

Biochemical diagnosis of pituitary insufficiency is made by demonstrating low levels of respective pituitary trophic hormones in the setting of low levels of target hormones. For example, low free thyroxine in the setting of a low or inappropriately normal TSH level suggests secondary hypothyroidism. Similarly, a low testosterone level without elevation of gonadotropins suggests hypogonadotropic hypogonadism. Provocative tests may be required to assess pituitary reserve **(Table 372-2)**. GH responses to insulin-induced hypoglycemia, arginine, L-dopa, growth hormone–releasing hormone (GHRH), or growth hormone–releasing peptides (GHRPs) can be used to assess GH reserve. Corticotropin-releasing hormone (CRH) administration induces ACTH release, and administration of synthetic ACTH (cosyntropin) evokes adrenal cortisol release as an indirect indicator of pituitary ACTH reserve **(Chap. 379)**. ACTH reserve is most reliably assessed by measuring ACTH and cortisol levels during insulin-induced hypoglycemia. However, this test should be performed cautiously in patients with suspected adrenal insufficiency because of enhanced susceptibility to hypoglycemia and hypotension. Administering insulin to induce hypoglycemia is contraindicated in patients with active coronary artery disease or known seizure disorders.

TREATMENT

Hypopituitarism

Hormone replacement therapy, including glucocorticoids, thyroid hormone, sex steroids, GH, and vasopressin, is usually safe and free of complications. Treatment regimens that mimic physiologic hormone production allow for maintenance of satisfactory clinical homeostasis. Effective dosage schedules are outlined in **Table 372-3**. Patients in need of glucocorticoid replacement require careful dose adjustments during stressful events such as acute illness, dental procedures, trauma, and acute hospitalization.

TABLE 372-2 Tests of Pituitary Sufficiency

HORMONE	TEST	BLOOD SAMPLES	INTERPRETATION
Growth hormone (GH)	Insulin tolerance test: Regular insulin (0.05–0.15 U/kg IV)	–30, 0, 30, 60, 120 min for glucose and GH	Glucose <40 mg/dL; GH should be >3 μg/L
	GHRH test: 1 μg/kg IV	0, 15, 30, 45, 60, 120 min for GH	Normal response is GH >3 μg/L
	L-Arginine test: 30 g IV over 30 min	0, 30, 60, 120 min for GH	Normal response is GH >3 μg/L
	L-Dopa test: 500 mg PO	0, 30, 60, 120 min for GH	Normal response is GH >3 μg/L
Prolactin	TRH test: 200–500 μg IV	0, 20, and 60 min for TSH and PRL	Normal prolactin is >2 μg/L and increase >200% of baseline
ACTH	Insulin tolerance test: regular insulin (0.05–0.15 U/kg IV)	–30, 0, 30, 60, 90 min for glucose and cortisol	Glucose <40 mg/dL Cortisol should increase by >7 μg/dL or to >20 μg/dL
	CRH test: 1 μg/kg ovine CRH IV at 8 A.M.	0, 15, 30, 60, 90, 120 min for ACTH and cortisol	Basal ACTH increases 2- to 4-fold and peaks at 20–100 pg/mL Cortisol levels >20–25 μg/dL
	Metyrapone test: Metyrapone (30 mg/kg) at midnight	Plasma 11-deoxycortisol and cortisol at 8 A.M.; ACTH can also be measured	Plasma cortisol should be <4 g/dL to assure an adequate response Normal response is 11-deoxycortisol >7.5 μg/dL or ACTH >75 pg/mL
	Standard ACTH stimulation test: ACTH 1-24 (cosyntropin), 0.25 mg IM or IV	0, 30, 60 min for cortisol and aldosterone	Normal response is cortisol >21 g/dL and aldosterone response of >4 ng/dL above baseline
	Low-dose ACTH test: ACTH 1-24 (cosyntropin), 1 μg IV	0, 30, 60 min for cortisol	Cortisol should be >21 g/dL
	3-day ACTH stimulation test consists of 0.25 mg ACTH 1-24 given IV over 8 h each day		Cortisol >21 g/dL
TSH	Basal thyroid function tests: T_4, T_3, TSH	Basal measurements	Low free thyroid hormone levels in the setting of TSH levels that are not appropriately increased indicate pituitary insufficiency
	TRH test: 200–500 μg IV	0, 20, 60 min for TSH and PRL[a]	TSH should increase by >5 mU/L unless thyroid hormone levels are increased
LH, FSH	LH, FSH, testosterone, estrogen	Basal measurements	Basal LH and FSH should be increased in postmenopausal women Low testosterone levels in the setting of low LH and FSH indicate pituitary insufficiency
	GnRH test: GnRH (100 μg) IV	0, 30, 60 min for LH and FSH	In most adults, LH should increase by 10 IU/L and FSH by 2 IU/L Normal responses are variable
Multiple hormones	Combined anterior pituitary test: GHRH (1 g/kg), CRH (1 μg/kg), GnRH (100 g), TRH (200 μg) are given IV	–30, 0, 15, 30, 60, 90, 120 min for GH, ACTH, cortisol, LH, FSH, and TSH	Combined or individual releasing hormone responses must be elevated in the context of basal target gland hormone values and may not be uniformly diagnostic (see text)

[a]Evoked PRL response indicates lactotrope integrity.

Abbreviations: T_3, triiodothyronine; T_4, thyroxine; TRH, thyrotropin-releasing hormone. For other abbreviations, see text.

TABLE 372-3 Hormone Replacement Therapy for Adult Hypopituitarism[a]

TROPHIC HORMONE DEFICIT	HORMONE REPLACEMENT
ACTH	Hydrocortisone (10–20 mg/d in divided doses) Cortisone acetate (15–25 mg/d in divided doses) Prednisone (5 mg A.M.)
TSH	L-Thyroxine (0.075–0.15 mg daily)
FSH/LH	Males Testosterone gel (5–10 g/d) Testosterone skin patch (5 mg/d) Testosterone enanthate (200 mg IM every 2 weeks) Females Conjugated estrogen (0.65–1.25 mg qd for 25 days) Progesterone (5–10 mg qd) on days 16–25 Estradiol skin patch (0.025–0.1 mg every week), adding progesterone on days 16–25 if uterus intact For fertility: menopausal gonadotropins, human chorionic gonadotropins
GH	Adults: Somatotropin (0.1–1.25 mg SC qd) Children: Somatotropin (0.02–0.05 mg/kg per day)
Vasopressin	Intranasal desmopressin (5–20 g twice daily) Oral 300–600 μg qd

[a]All doses shown should be individualized for specific patients and should be reassessed during stress, surgery, or pregnancy. Male and female fertility requirements should be managed as discussed in **Chaps. 384 and 385**.

Note: For abbreviations, see text.

DISORDERS OF GROWTH AND DEVELOPMENT

Skeletal Maturation and Somatic Growth The growth plate is dependent on a variety of hormonal stimuli, including GH, insulin-like growth factor (IGF)-I, sex steroids, thyroid hormones, paracrine growth factors, and cytokines. The growth-promoting process also requires caloric energy, amino acids, vitamins, and trace metals and consumes about 10% of normal energy production. Malnutrition impairs chondrocyte activity, increases GH resistance, and reduces circulating IGF-I and IGF binding protein (IGBP)-3 levels.

Linear bone growth rates are very high in infancy and are pituitary-dependent. Mean growth velocity is ~6 cm/year in later childhood and usually is maintained within a given range on a standardized percentile chart. Peak growth rates occur during midpuberty when bone age is 12 (girls) or 13 (boys). Secondary sexual development is associated with elevated sex steroids that cause progressive epiphyseal growth plate closure. *Bone age* is delayed in patients with all forms of true GH deficiency or GH receptor defects that result in attenuated GH action.

Short stature may occur as a result of constitutive intrinsic growth defects or because of acquired extrinsic factors that impair growth. In general, delayed bone age in a child with short stature is suggestive of a hormonal or systemic disorder, whereas normal bone age in a short child is more likely to be caused by a genetic cartilage dysplasia or growth plate disorder **(Chap. 406)**.

GH Deficiency in Children • **GH DEFICIENCY** Isolated GH deficiency is characterized by short stature, micropenis, increased fat, high-pitched voice, and a propensity to hypoglycemia due to relatively unopposed insulin action. Familial modes of inheritance are seen in at least one-third of these individuals and may be autosomal dominant, recessive, or X-linked. About 10% of children with GH deficiency have mutations in the *GH-N* gene, including gene deletions and a wide range of point mutations. Mutations in transcription factors Pit-1 and Prop-1, which control somatotrope development, result in GH deficiency in combination with other pituitary hormone deficiencies, which may become manifest only in adulthood. The diagnosis of *idiopathic GH deficiency* (IGHD) should be made only after known molecular defects have been rigorously excluded.

GHRH RECEPTOR MUTATIONS Recessive mutations of the GHRH receptor gene in subjects with severe proportionate dwarfism are associated with low basal GH levels that cannot be stimulated by exogenous GHRH, GHRP, or insulin-induced hypoglycemia, as well as anterior pituitary hypoplasia. The syndrome exemplifies the importance of the GHRH receptor for somatotrope cell proliferation and hormonal responsiveness.

GH INSENSITIVITY This is caused by defects of GH receptor structure or signaling. Homozygous or heterozygous mutations of the GH receptor are associated with partial or complete GH insensitivity and growth failure (*Laron syndrome*). The diagnosis is based on normal or high GH levels, with decreased circulating GH-binding protein (GHBP), and low IGF-I levels. Very rarely, defective IGF-I, IGF-I receptor, or IGF-I signaling defects are also encountered. *STAT5B* mutations result in both immunodeficiency as well as abrogated GH signaling, leading to short stature with normal or elevated GH levels and low IGF-I levels. Circulating GH receptor antibodies may rarely cause peripheral GH insensitivity.

NUTRITIONAL SHORT STATURE Caloric deprivation and malnutrition, uncontrolled diabetes, and chronic renal failure represent secondary causes of abrogated GH receptor function. These conditions also stimulate production of proinflammatory cytokines, which act to exacerbate the block of GH-mediated signal transduction. Children with these conditions typically exhibit features of acquired short stature with normal or elevated GH and low IGF-I levels.

PSYCHOSOCIAL SHORT STATURE Emotional and social deprivation lead to growth retardation accompanied by delayed speech, discordant hyperphagia, and an attenuated response to administered GH. A nurturing environment restores growth rates.

PRESENTATION AND DIAGNOSIS

Short stature is commonly encountered in clinical practice, and the decision to evaluate these children requires clinical judgment in association with auxologic data and family history. Short stature should be evaluated comprehensively if a patient's height is >3 standard deviations (SD) below the mean for age or if the growth rate has decelerated. Skeletal maturation is best evaluated by measuring a radiologic bone age, which is based mainly on the degree of wrist bone growth plate fusion. Final height can be predicted using standardized scales (Bayley-Pinneau or Tanner-Whitehouse) or estimated by adding 6.5 cm (boys) or subtracting 6.5 cm (girls) from the midparental height.

LABORATORY INVESTIGATION

Because GH secretion is pulsatile, GH deficiency is best assessed by examining the response to provocative stimuli, including exercise, insulin-induced hypoglycemia, and other pharmacologic tests that normally increase GH to >7 μg/L in children. Random GH measurements do not distinguish normal children from those with true GH deficiency. Adequate adrenal and thyroid hormone replacement should be assured before testing. Age- and sex-matched IGF-I levels are not sufficiently sensitive or specific to make the diagnosis but can be useful to confirm GH deficiency. Pituitary MRI may reveal pituitary mass lesions or structural defects. Molecular analyses for known mutations should be undertaken when the cause of short stature remains cryptic, or when additional clinical features suggest a genetic cause.

TREATMENT

Disorders of Growth and Development

Replacement therapy with recombinant GH (0.02–0.05 mg/kg per day SC) restores growth velocity in GH-deficient children to ~10 cm/year. If pituitary insufficiency is documented, other associated hormone deficits should be corrected, especially adrenal steroids. GH treatment is also moderately effective for accelerating growth rates in children with Turner syndrome and chronic renal failure.

In patients with GH insensitivity and growth retardation due to mutations of the GH receptor, treatment with IGF-I bypasses the dysfunctional GH receptor.

ADULT GH DEFICIENCY (AGHD)

This disorder usually is caused by acquired hypothalamic or pituitary somatotrope damage. Acquired pituitary hormone deficiency follows a typical pattern in which loss of adequate GH reserve foreshadows subsequent hormone deficits. The sequential order of hormone loss is usually GH → FSH/LH → TSH → ACTH. Patients previously diagnosed with childhood-onset GH deficiency should be retested as adults to affirm the diagnosis.

PRESENTATION AND DIAGNOSIS

The clinical features of AGHD include changes in body composition, lipid metabolism, and quality of life and cardiovascular dysfunction **(Table 372-4)**. Body composition changes are common and include reduced lean body mass, increased fat mass with selective deposition of intraabdominal visceral fat, and increased waist-to-hip ratio. Hyperlipidemia, left ventricular dysfunction, hypertension, and increased plasma fibrinogen levels also may be present. Bone mineral content is reduced, with resultant increased fracture rates. Patients may experience social isolation, depression, and difficulty maintaining gainful employment. Adult hypopituitarism is associated with a threefold increase in cardiovascular mortality rates in comparison to age- and sex-matched controls, and this may be due to GH deficiency, as patients in these studies were replaced with other deficient pituitary hormones.

LABORATORY INVESTIGATION

AGHD is rare, and in light of the nonspecific nature of associated clinical symptoms, patients appropriate for testing should be selected carefully on the basis of well-defined criteria. With few exceptions, testing should be restricted to patients with the following predisposing factors: (1) pituitary surgery, (2) pituitary or hypothalamic tumor or granulomas, (3) history of cranial irradiation, (4) radiologic evidence of a pituitary lesion, and (5) childhood requirement for GH replacement therapy. The transition of a GH-deficient adolescent to adulthood requires retesting to document subsequent AGHD. Up to 20% of patients previously treated for childhood-onset GH deficiency are found to be GH-sufficient on repeat testing as adults.

A significant proportion (~25%) of truly GH-deficient adults have low-normal IGF-I levels. Thus, as in the evaluation of GH deficiency in children, valid age- and sex-matched IGF-I measurements provide a useful index of therapeutic responses but are not sufficiently sensitive for diagnostic purposes. The most validated test to distinguish pituitary-sufficient patients from those with AGHD is insulin-induced (0.05–0.1 U/kg) hypoglycemia. After glucose reduction to ~40 mg/dL, most individuals experience neuroglycopenic symptoms **(Chap. 399)**, and peak GH release occurs at 60 min and remains elevated for up to 2 h. About 90% of healthy adults exhibit GH responses >5 μg/L; AGHD is defined by a peak GH response to hypoglycemia of <3 μg/L. Although insulin-induced hypoglycemia is safe when performed under appropriate supervision, it is contraindicated in patients with diabetes, ischemic heart disease, cerebrovascular disease, or epilepsy and in elderly patients. Alternative stimulatory tests include intravenous arginine (30 g), GHRH (1 μg/kg), GHRP-6 (90 μg), and glucagon (1 mg). Combinations of these tests may evoke GH secretion in subjects who are not responsive to a single test.

TABLE 372-4 Features of Adult Growth Hormone Deficiency

Clinical
Impaired quality of life
Decreased energy and drive
Poor concentration
Low self-esteem
Social isolation
Body composition changes
Increased body fat mass
Central fat deposition
Increased waist-to-hip ratio
Decreased lean body mass
Reduced exercise capacity
Reduced maximum O_2 uptake
Impaired cardiac function
Reduced muscle mass
Cardiovascular risk factors
Impaired cardiac structure and function
Abnormal lipid profile
Decreased fibrinolytic activity
Atherosclerosis
Omental obesity
Imaging
Pituitary: mass or structural damage
Bone: reduced bone mineral density
Abdomen: excess omental adiposity
Laboratory
Evoked GH <3 ng/mL
IGF-I and IGFBP3 low or normal
Increased LDL cholesterol
Concomitant gonadotropin, TSH, and/or ACTH reserve deficits may be present

Abbreviation: LDL, low-density lipoprotein. For other abbreviations, see text.

TREATMENT

Adult GH Deficiency

Once the diagnosis of AGHD is unequivocally established, replacement of GH may be indicated. Contraindications to therapy include the presence of an active neoplasm, intracranial hypertension, and uncontrolled diabetes and retinopathy. The starting adult dose of 0.1–0.2 mg/d should be titrated (up to a maximum of 1.25 mg/d) to maintain IGF-I levels in the mid-normal range for age- and sex-matched controls **(Fig. 372-1)**. Women require higher doses than men, and elderly patients require less GH. Long-term GH maintenance sustains normal IGF-I levels and is associated with persistent body composition changes (e.g., enhanced lean body mass and lower body fat). High-density lipoprotein cholesterol increases, but total cholesterol and insulin levels may not change significantly. Lumbar spine bone mineral density increases, but this response is gradual (>1 year). Many patients note significant improvement in quality of life when evaluated by standardized questionnaires. The effect of GH replacement on mortality rates in GH-deficient patients is currently the subject of long-term prospective investigation.

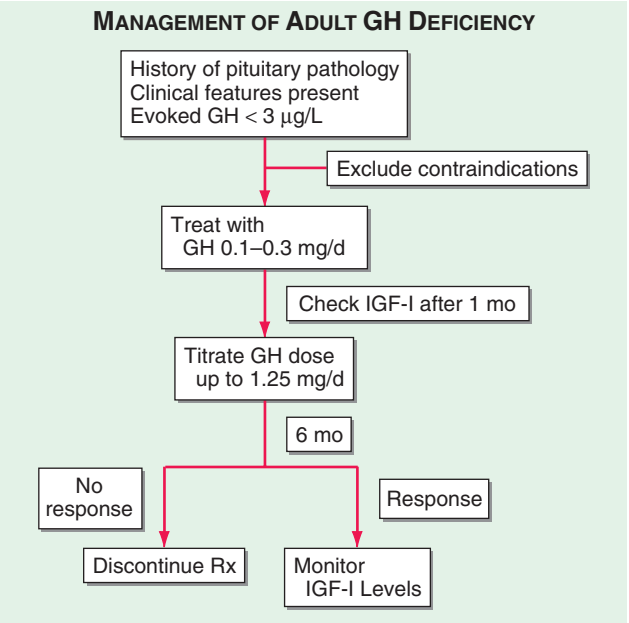


FIGURE 372-1 Management of adult growth hormone (GH) deficiency. IGF, insulin-like growth factor; Rx, Treatment.

About 30% of patients exhibit reversible dose-related fluid retention, joint pain, and carpal tunnel syndrome, and up to 40% exhibit myalgias and paresthesia. Patients receiving insulin require careful monitoring for dosing adjustments, as GH is a potent counterregulatory hormone for insulin action. Patients with type 2 diabetes mellitus may initially develop further insulin resistance. However, glycemic control usually improves with the sustained loss of abdominal fat associated with long-term GH replacement. Headache, increased intracranial pressure, hypertension, and tinnitus occur rarely. Pituitary tumor regrowth and progression of skin lesions or other tumors have not been encountered in long-term surveillance programs with appropriate replacement doses.

ACTH DEFICIENCY

PRESENTATION AND DIAGNOSIS

Secondary adrenal insufficiency occurs as a result of pituitary ACTH deficiency. It is characterized by fatigue, weakness, anorexia, nausea, vomiting, and, occasionally, hypoglycemia. In contrast to primary adrenal failure, hypocortisolism associated with pituitary failure usually is not accompanied by hyperpigmentation or mineralocorticoid deficiency.

ACTH deficiency is commonly due to glucocorticoid withdrawal after treatment-associated suppression of the hypothalamic-pituitary-adrenal (HPA) axis. Isolated ACTH deficiency may occur after surgical resection of an ACTH-secreting pituitary adenoma that has suppressed the HPA axis; this phenomenon is in fact suggestive of a surgical cure. The mass effects of other pituitary adenomas or sellar lesions may lead to ACTH deficiency, usually in combination with other pituitary hormone deficiencies. Partial ACTH deficiency may be unmasked in the presence of an acute medical or surgical illness, when clinically significant hypocortisolism reflects diminished ACTH reserve. Rarely, *TPIT* or *POMC* mutations result in primary ACTH deficiency.

LABORATORY DIAGNOSIS

Inappropriately low ACTH levels in the setting of low cortisol levels are characteristic of diminished ACTH reserve. Low basal serum cortisol levels are associated with blunted cortisol responses to ACTH stimulation and impaired cortisol response to insulin-induced hypoglycemia, or testing with metyrapone or CRH. **For a description of provocative ACTH tests, see Chap. 379.**

TREATMENT

ACTH Deficiency

Glucocorticoid replacement therapy improves most features of ACTH deficiency. The total daily dose of hydrocortisone replacement preferably should generally not exceed 20 mg daily, divided into two or three doses. Prednisone (5 mg each morning) is longer acting and has fewer mineralocorticoid effects than hydrocortisone. Some authorities advocate lower maintenance doses in an effort to avoid cushingoid side effects. Doses should be increased severalfold during periods of acute illness or stress. Patients should wear medi-alert bracelets and/or carry identification cards with information about their glucocorticoid requirements.

GONADOTROPIN DEFICIENCY

Hypogonadism is the most common presenting feature of adult hypopituitarism even when other pituitary hormones are also deficient. It is often a harbinger of hypothalamic or pituitary lesions that impair GnRH production or delivery through the pituitary stalk. As noted below, hypogonadotropic hypogonadism is a common presenting feature of hyperprolactinemia.

A variety of inherited and acquired disorders are associated with *isolated hypogonadotropic hypogonadism* (IHH) **(Chap. 384)**. Hypothalamic defects associated with GnRH deficiency include Kallmann syndrome and mutations in more than a dozen genes that regulate GnRH neuron migration, development, and function (see above). Mutations in *GPR54, DAX1,* kisspeptin, the GnRH receptor, and the LHβ or FSHβ subunit genes also cause pituitary gonadotropin deficiency. Acquired forms of GnRH deficiency leading to hypogonadotropism are seen in association with anorexia nervosa, stress, starvation, and extreme exercise but also may be idiopathic. Hypogonadotropic hypogonadism in these disorders is reversed by removal of the stressful stimulus or by caloric replenishment.

PRESENTATION AND DIAGNOSIS

In premenopausal women, hypogonadotropic hypogonadism presents as diminished ovarian function leading to oligomenorrhea or amenorrhea, infertility, decreased vaginal secretions, decreased libido, and breast atrophy. In hypogonadal adult men, secondary testicular failure is associated with decreased libido and potency, infertility, decreased muscle mass with weakness, reduced beard and body hair growth, soft testes, and characteristic fine facial wrinkles. Osteoporosis occurs in both untreated hypogonadal women and men.

LABORATORY INVESTIGATION

Central hypogonadism is associated with low or inappropriately normal serum gonadotropin levels in the setting of low sex hormone concentrations (testosterone in men, estradiol in women). Because gonadotropin secretion is pulsatile, valid assessments may require repeated measurements or the use of pooled serum samples. Men have reduced sperm counts.

Intravenous GnRH (100 μg) stimulates gonadotropes to secrete LH (which peaks within 30 min) and FSH (which plateaus during the ensuing 60 min). Normal responses vary according to menstrual cycle stage, age, and sex of the patient. Generally, LH levels increase about threefold, whereas FSH responses are less pronounced. In the setting of gonadotropin deficiency, a normal gonadotropin response to GnRH indicates intact pituitary gonadotrope function and suggests a hypothalamic abnormality. An absent response, however, does not reliably distinguish pituitary from hypothalamic causes of hypogonadism. For this reason, GnRH testing usually adds little to the information gained from baseline evaluation of the hypothalamic-pituitary-gonadotrope axis except in cases of isolated GnRH deficiency (e.g., Kallmann syndrome).

MRI examination of the sellar region and assessment of other pituitary functions usually are indicated in patients with documented central hypogonadism.

TREATMENT

Gonadotropin Deficiency

In males, testosterone replacement is necessary to achieve and maintain normal growth and development of the external genitalia, secondary sex characteristics, male sexual behavior, and androgenic anabolic effects, including maintenance of muscle function and bone mass. Testosterone may be administered by intramuscular injections every 1–4 weeks or by using skin patches or testosterone gels **(Chap. 384)**. Gonadotropin injections (hCG or human menopausal gonadotropin [hMG]) over 12–18 months are used to restore fertility. Pulsatile GnRH therapy (25–150 ng/kg every 2 h), administered by a subcutaneous infusion pump, is also effective for treatment of hypothalamic hypogonadism when fertility is desired.

In premenopausal women, cyclical replacement of estrogen and progesterone maintains secondary sexual characteristics and integrity of genitourinary tract mucosa and prevents premature osteoporosis **(Chap. 385)**. Gonadotropin therapy is used for ovulation induction. Follicular growth and maturation are initiated using hMG or recombinant FSH; hCG or human luteinizing hormone (hLH) is subsequently injected to induce ovulation. As in men, pulsatile GnRH therapy can be used to treat hypothalamic causes of gonadotropin deficiency.

DIABETES INSIPIDUS

See Chap. 374 for diagnosis and treatment of diabetes insipidus.

FURTHER READING

ALATZOGLOU KS, DATTANI MT: Genetic causes and treatment of isolated growth hormone deficiency—An update. Nat Rev Endocrinol 6:562, 2010.

CLEMMONS DR: The diagnosis and treatment of growth hormone deficiency in adults. Curr Opin Endocrinol Diabetes Obes 17:377, 2010.

DEKKERS OM et al: Multisystem morbidity and mortality in Cushing's syndrome: A cohort study. J Clin Endocrinol Metab 98:2277, 2013.

FLESERIU M et al: Hormonal replacement in hypopituitarism in adults: An Endocrine Society clinical practice guideline. J Clin Endocrinol Metab 101:3888, 2016.

HONEGGER J et al: Diagnosis of primary hypophysitis in Germany. J Clin Endocrinol Metab 100:3841, 2015.

ROMIJN JA: The chronic syndromes after previous treatment of pituitary tumours. Nat Rev Endocrinol 12:547, 2016.

SHERLOCK M et al: Mortality in patients with pituitary disease. Endocr Rev 31:301, 2010.

TANRIVERDI F et al: Pituitary dysfunction after traumatic brain injury: A clinical and pathophysiological approach. Endocr Rev 36:305, 2015.

373 Pituitary Tumor Syndromes

Shlomo Melmed, J. Larry Jameson

HYPOTHALAMIC, PITUITARY, AND OTHER SELLAR MASSES

EVALUATION OF SELLAR MASSES

Local Mass Effects Clinical manifestations of sellar lesions vary, depending on the anatomic location of the mass and the direction of its extension (Table 373-1). The dorsal sellar diaphragm presents the least resistance to soft tissue expansion from the sella; consequently, pituitary adenomas frequently extend in a suprasellar direction. Bony invasion may occur as well.

Headaches are common features of small intrasellar tumors, even with no demonstrable suprasellar extension. Because of the confined nature of the pituitary, small changes in intrasellar pressure stretch the dural plate; however, headache severity correlates poorly with adenoma size or extension.

Suprasellar extension can lead to visual loss by several mechanisms, the most common being compression of the optic chiasm, but rarely, direct invasion of the optic nerves or obstruction of cerebrospinal fluid (CSF) flow leading to secondary visual disturbances can occur. Pituitary stalk compression by a hormonally active or inactive intrasellar mass may compress the portal vessels, disrupting pituitary access to hypothalamic hormones and dopamine; this results in early hyperprolactinemia and later concurrent loss of other pituitary hormones. This "stalk section" phenomenon may also be caused by trauma, whiplash injury with posterior clinoid stalk compression, or skull base fractures. Lateral mass invasion may impinge on the cavernous sinus and compress its neural contents, leading to cranial nerve III, IV, and VI palsies as well as effects on the ophthalmic and maxillary branches of the fifth cranial nerve (Chap. 433). Patients may present with diplopia, ptosis, ophthalmoplegia, and decreased facial sensation, depending on the extent of neural damage. Extension into the sphenoid sinus indicates that the pituitary mass has eroded through the sellar floor. Aggressive tumors rarely invade the palate roof and cause nasopharyngeal obstruction, infection, and CSF leakage. Temporal and frontal lobe involvement may rarely lead to uncinate seizures, personality disorders, and anosmia. Direct hypothalamic encroachment by an invasive pituitary mass may cause important metabolic sequelae, including precocious puberty or hypogonadism, diabetes insipidus, sleep disturbances, dysthermia, and appetite disorders.

TABLE 373-1 Features of Sellar Mass Lesions[a]

IMPACTED STRUCTURE	CLINICAL IMPACT
Pituitary	Hypogonadism
	Hypothyroidism
	Growth failure and adult hyposomatotropism
	Hypoadrenalism
Optic chiasm	Loss of red perception
	Bitemporal hemianopia
	Superior or bitemporal field defect
	Scotoma
	Blindness
Hypothalamus	Temperature dysregulation
	Appetite and thirst disorders
	Obesity
	Diabetes insipidus
	Sleep disorders
	Behavioral dysfunction
	Autonomic dysfunction
Cavernous sinus	Ophthalmoplegia with or without ptosis or diplopia
	Facial numbness
Frontal lobe	Personality disorder
	Anosmia
Brain	Headache
	Hydrocephalus
	Psychosis
	Dementia
	Laughing seizures

[a]As the intrasellar mass expands, it first compresses intrasellar pituitary tissue, then usually invades dorsally through the dura to lift the optic chiasm or laterally to the cavernous sinuses. Bony erosion is rare, as is direct brain compression. Microadenomas may present with headache.

Magnetic Resonance Imaging Sagittal and coronal T1-weighted magnetic resonance imaging (MRI) before and after administration of gadolinium allows precise visualization of the pituitary gland with clear delineation of the hypothalamus, pituitary stalk, pituitary tissue and surrounding suprasellar cisterns, cavernous sinuses, sphenoid sinus, and optic chiasm. Pituitary gland height ranges from 6 mm in children to 8 mm in adults; during pregnancy and puberty, the height may reach 10–12 mm. The upper aspect of the adult pituitary is flat or slightly concave, but in adolescent and pregnant individuals, this surface may be convex, reflecting physiologic pituitary enlargement. The stalk should be midline and vertical. Computed tomography (CT) scan is reserved to define the extent of bony erosion or the presence of calcification.

Anterior pituitary gland soft tissue consistency is slightly heterogeneous on MRI, and signal intensity resembles that of brain matter on T1-weighted imaging (Fig. 373-1). Adenoma density is usually lower than that of surrounding normal tissue on T1-weighted imaging, and the signal intensity increases with T2-weighted images. The high phospholipid content of the posterior pituitary results in a "pituitary bright spot."

Sellar masses are encountered commonly as incidental findings on MRI, and most of them are pituitary adenomas (incidentalomas). In the absence of hormone hypersecretion, these small intrasellar lesions can be monitored safely with MRI, which is performed annually and then less often if there is no evidence of further growth. Resection should be considered for incidentally discovered larger macroadenomas, because about one-third become invasive or cause local pressure effects. If

FIGURE 373-1 Pituitary adenoma. Coronal T1-weighted postcontrast magnetic resonance image shows a homogeneously enhancing mass (*arrowheads*) in the sella turcica and suprasellar region compatible with a pituitary adenoma; the *small arrows* outline the carotid arteries.

hormone hypersecretion is evident, specific therapies are indicated as described below. When larger masses (>1 cm) are encountered, they should also be distinguished from nonadenomatous lesions. Meningiomas often are associated with bony hyperostosis; craniopharyngiomas may be calcified and are usually hypodense, whereas gliomas are hyperdense on T2-weighted images.

Ophthalmologic Evaluation Because optic tracts may be contiguous to an expanding pituitary mass, reproducible visual field assessment using perimetry techniques should be performed on all patients with sellar mass lesions that impinge the optic chiasm **(Chap. 28)**. Bitemporal hemianopia, often more pronounced superiorly, is observed classically. It occurs because nasal ganglion cell fibers, which cross in the optic chiasm, are especially vulnerable to compression of the ventral optic chiasm. Occasionally, homonymous hemianopia occurs from postchiasmal compression or monocular temporal field loss from prechiasmal compression. Invasion of the cavernous sinus can produce diplopia from ocular motor nerve palsy. Early diagnosis reduces the risk of optic atrophy, vision loss, or eye misalignment.

Laboratory Investigation The presenting clinical features of functional pituitary adenomas (e.g., acromegaly, prolactinomas, or Cushing syndrome) should guide the laboratory studies **(Table 373-2)**. However, for a sellar mass with no obvious clinical features of hormone excess, laboratory studies are geared toward determining the nature of the tumor and assessing the possible presence of hypopituitarism. When a pituitary adenoma is suspected based on MRI, initial hormonal evaluation usually includes (1) basal prolactin (PRL); (2) insulin-like growth factor (IGF)-I; (3) 24-h urinary free cortisol (UFC) and/or overnight oral dexamethasone (1 mg) suppression test; (4) α subunit, follicle-stimulating hormone (FSH), and luteinizing hormone (LH); and (5) thyroid function tests. Additional hormonal evaluation may be indicated based on the results of these tests. Pending more detailed assessment of hypopituitarism, a menstrual history, measurement of testosterone and 8 A.M. cortisol levels, and thyroid function tests usually identify patients with pituitary hormone deficiencies that require hormone replacement before further testing or surgery.

Histologic Evaluation Immunohistochemical staining of pituitary tumor specimens obtained at transsphenoidal surgery confirms clinical and laboratory studies and provides a histologic diagnosis when hormone studies are equivocal and in cases of clinically nonfunctioning tumors.

TABLE 373-2 Screening Tests for Functional Pituitary Adenomas

	TEST	COMMENTS
Acromegaly	Serum IGF-I	Interpret IGF-I relative to age- and sex-matched controls
	Oral glucose tolerance test with GH obtained at 0, 30, and 60 min	Normal subjects should suppress growth hormone to <1 μ/L
Prolactinoma	Serum PRL	Exclude medications
		MRI of the sella should be ordered if PRL is elevated
Cushing's disease	24-h urinary free cortisol	Ensure urine collection is total and accurate
	Dexamethasone (1 mg) at 11 P.M. and fasting plasma cortisol measured at 8 A.M.	Normal subjects suppress to <5 μ/dL
	ACTH assay	Distinguishes adrenal adenoma (ACTH suppressed) from ectopic ACTH or Cushing's disease (ACTH normal or elevated)

Abbreviations: ACTH, adrenocorticotropin hormone; GH, growth hormone; IGF-I, insulin-like growth factor I; MRI, magnetic resonance imaging; PRL, prolactin.

TREATMENT

Hypothalamic, Pituitary, and Other Sellar Masses

OVERVIEW

Successful management of sellar masses requires accurate diagnosis as well as selection of optimal therapeutic modalities. Most pituitary tumors are benign and slow-growing. Clinical features result from local mass effects and hormonal hyper- or hyposecretion syndromes caused directly by the adenoma or occurring as a consequence of treatment. Thus, lifelong management and follow-up are necessary for these patients.

MRI with gadolinium enhancement for pituitary visualization, new advances in transsphenoidal surgery and in stereotactic radiotherapy (including gamma-knife radiotherapy), and novel therapeutic agents have improved pituitary tumor management. The goals of pituitary tumor treatment include normalization of excess pituitary secretion, amelioration of symptoms and signs of hormonal hypersecretion syndromes, and shrinkage or ablation of large tumor masses with relief of adjacent structure compression. Residual anterior pituitary function should be preserved during treatment and sometimes can be restored by removing the tumor mass. Ideally, adenoma recurrence should be prevented.

TRANSSPHENOIDAL SURGERY

Transsphenoidal rather than transfrontal resection is the desired surgical approach for pituitary tumors, except for the rare invasive suprasellar mass surrounding the frontal or middle fossa or the optic nerves or invading posteriorly behind the clivus. Intraoperative microscopy facilitates visual distinction between adenomatous and normal pituitary tissue as well as microdissection of small tumors that may not be visible by MRI **(Fig. 373-2)**. Transsphenoidal surgery also avoids the cranial invasion and manipulation of brain tissue required by subfrontal surgical approaches. Endoscopic techniques with three-dimensional intraoperative localization have improved visualization and access to tumor tissue. Individual surgical experience is a major determinant of outcome efficacy with these techniques.

In addition to correction of hormonal hypersecretion, pituitary surgery is indicated for mass lesions that impinge on surrounding structures. Surgical decompression and resection are required for an expanding pituitary mass, which may be asymptomatic or accompanied by persistent headache, progressive visual field defects, cranial nerve palsies, hydrocephalus, and, occasionally, intrapituitary hemorrhage and apoplexy. Transsphenoidal surgery sometimes is used for pituitary tissue biopsy to establish a histologic diagnosis.

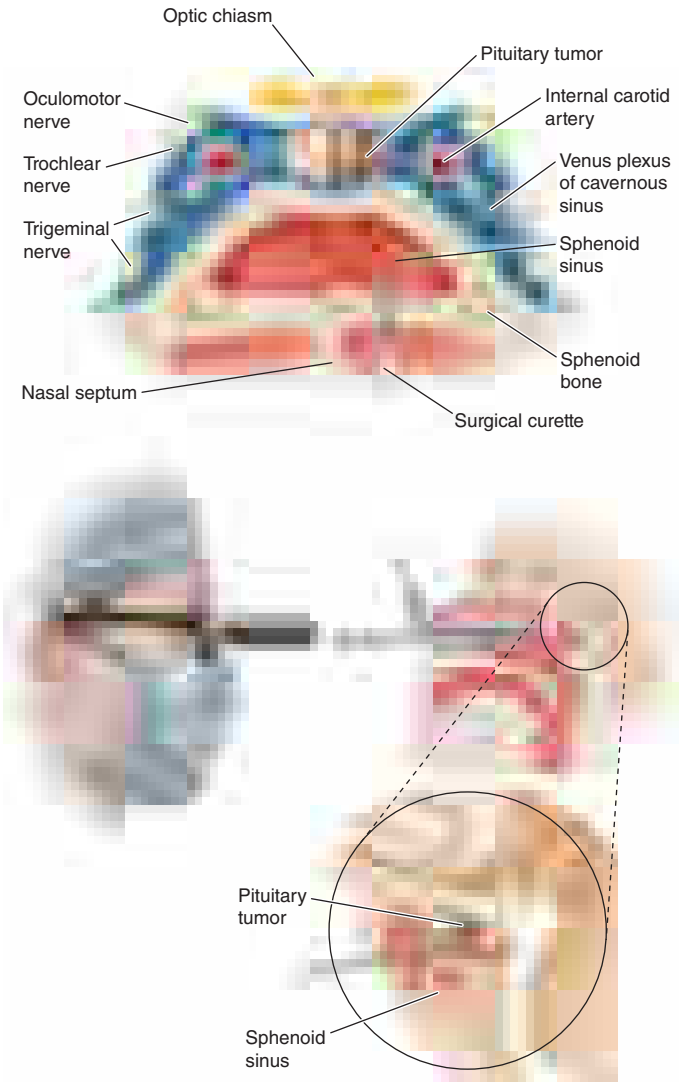


FIGURE 373-2 Transsphenoidal resection of pituitary mass via the endonasal approach. *(Adapted from R Fahlbusch: Endocrinol Metab Clin 21:669, 1992.)*

Whenever possible, the pituitary mass lesion should be selectively excised; normal pituitary tissue should be manipulated or resected only when critical for effective mass dissection. Nonselective hemihypophysectomy or total hypophysectomy may be indicated if no hypersecreting mass lesion is clearly discernible, multifocal lesions are present, or the remaining nontumorous pituitary tissue is obviously necrotic. This strategy, however, increases the likelihood of postoperative hypopituitarism and the need for lifelong hormone replacement.

Preoperative mass effects, including visual field defects and compromised pituitary function, may be reversed by surgery, particularly when the deficits are not long-standing. For large and invasive tumors, it is necessary to determine the optimal balance between maximal tumor resection and preservation of anterior pituitary function, especially for preserving growth and reproductive function in younger patients. Similarly, tumor invasion outside the sella is rarely amenable to surgical cure; the surgeon must judge the risk-versus-benefit ratio of extensive tumor resection.

Side Effects Tumor size, the degree of invasiveness, and experience of the surgeon largely determine the incidence of surgical complications. Operative mortality rate is ~1%. Transient diabetes insipidus and hypopituitarism occur in up to 20% of patients. Permanent diabetes insipidus, cranial nerve damage, nasal septal perforation, or visual disturbances may be encountered in up to 10% of patients. CSF leaks occur in 4% of patients. Less common complications include carotid artery injury, loss of vision, hypothalamic damage, and meningitis. Permanent side effects are rare after surgery for microadenomas.

RADIATION

Radiation is used either as a primary therapy for pituitary or parasellar masses or, more commonly, as an adjunct to surgery or medical therapy. Focused megavoltage irradiation is achieved by precise MRI localization, using a high-voltage linear accelerator and accurate isocentric rotational arcing. A major determinant of accurate irradiation is reproduction of the patient's head position during multiple visits and maintenance of absolute head immobility. A total of <50 Gy (5000 rad) is given as 180-cGy (180-rad) fractions divided over ~6 weeks. Stereotactic radiosurgery delivers a large single high-energy dose from a cobalt-60 source (gamma knife), linear accelerator, or cyclotron. Long-term effects of gamma-knife surgery are unclear but appear to be similar to those encountered with conventional radiation. Proton beam therapy is available in some centers and provides concentrated radiation doses within a localized region.

The role of radiation therapy in pituitary tumor management depends on multiple factors, including the nature of the tumor, the age of the patient, and the availability of surgical and radiation expertise. Because of its relatively slow onset of action, radiation therapy is usually reserved for postsurgical management. As an adjuvant to surgery, radiation is used to treat residual tumor and in an attempt to prevent regrowth. Irradiation offers the only means for potentially ablating significant postoperative residual nonfunctioning tumor tissue. In contrast, PRL-, growth hormone (GH)–, and adrenocorticotropin hormone (ACTH)–secreting residual tumor tissues are amenable to medical therapy.

Side Effects In the short term, radiation may cause transient nausea and weakness. Alopecia and loss of taste and smell may be more long-lasting. Failure of pituitary hormone synthesis is common in patients who have undergone head and neck or pituitary-directed irradiation. More than 50% of patients develop loss of GH, ACTH, thyroid-stimulating hormone (TSH), and/or gonadotropin secretion within 10 years, usually due to hypothalamic damage. Lifelong follow-up with testing of anterior pituitary hormone reserve is therefore required after radiation treatment. Optic nerve damage with impaired vision due to optic neuritis is reported in ~2% of patients who undergo pituitary irradiation. Cranial nerve damage is uncommon now that radiation doses are <2 Gy (200 rad) at any one treatment session and the maximum dose is <50 Gy (5000 rad). The use of stereotactic radiotherapy may reduce damage to adjacent structures. Radiotherapy for pituitary tumors has been associated with adverse mortality rates, mainly from cerebrovascular disease. The cumulative risk of developing a secondary tumor after conventional radiation is 1.3% after 10 years and 1.9% after 20 years.

MEDICAL

Medical therapy for pituitary tumors is highly specific and depends on tumor type. For prolactinomas, dopamine agonists are the treatment of choice. For acromegaly, somatostatin analogues and a GH receptor antagonist are indicated. For TSH-secreting tumors, somatostatin analogues and occasionally dopamine agonists are indicated. ACTH-secreting tumors may respond to somatostatin analogues, and adrenal-directed therapy may also be of benefit. Nonfunctioning tumors are generally not responsive to medications and require surgery and/or irradiation.

SELLAR MASSES

Sellar masses other than pituitary adenomas may arise from brain, hypothalamic, or pituitary tissues. Each exhibit features related to the lesion location but also unique to the specific etiology.

Hypothalamic Lesions Lesions involving the anterior and preoptic hypothalamic regions cause paradoxical vasoconstriction, tachycardia, and hyperthermia. Acute hyperthermia usually is due to a

hemorrhagic insult, but poikilothermia may also occur. Central disorders of thermoregulation result from posterior hypothalamic damage. The *periodic hypothermia syndrome* is characterized by episodic attacks of rectal temperatures <30°C (86°F), sweating, vasodilation, vomiting, and bradycardia **(Chap. 454)**. Damage to the ventromedial hypothalamic nuclei by craniopharyngiomas, hypothalamic trauma, or inflammatory disorders may be associated with *hyperphagia* and *obesity*. This region appears to contain an energy-satiety center where melanocortin receptors are influenced by leptin, insulin, pro-opiomelanocortin (POMC) products, and gastrointestinal peptides **(Chap. 394)**. Polydipsia and hypodipsia are associated with damage to central osmoreceptors located in preoptic nuclei **(Chap. 374)**. Slow-growing hypothalamic lesions can cause increased somnolence and disturbed sleep cycles as well as obesity, hypothermia, and emotional outbursts. Lesions of the central hypothalamus may stimulate sympathetic neurons, leading to elevated serum catecholamine and cortisol levels. These patients are predisposed to cardiac arrhythmias, hypertension, and gastric erosions.

Craniopharyngiomas are benign, suprasellar cystic masses that present with headaches, visual field deficits, and variable degrees of hypopituitarism. They are derived from Rathke's pouch and arise near the pituitary stalk, commonly extending into the suprasellar cistern. Craniopharyngiomas are often large, cystic, and locally invasive. Many are partially calcified, exhibiting a characteristic appearance on skull x-ray and CT images. More than half of all patients present before age 20, usually with signs of increased intracranial pressure, including headache, vomiting, papilledema, and hydrocephalus. Associated symptoms include visual field abnormalities, personality changes and cognitive deterioration, cranial nerve damage, sleep difficulties, and weight gain. Hypopituitarism can be documented in ~90%, and diabetes insipidus occurs in ~10% of patients. About half of affected children present with growth retardation. MRI is generally superior to CT for evaluating cystic structure and tissue components of craniopharyngiomas. CT is useful to define calcifications and evaluate invasion into surrounding bony structures and sinuses.

Treatment usually involves transcranial or transsphenoidal surgical resection followed by postoperative radiation of residual tumor. Surgery alone is curative in less than half of patients because of recurrences due to adherence to vital structures or because of small tumor deposits in the hypothalamus or brain parenchyma. The goal of surgery is to remove as much tumor as possible without risking complications associated with efforts to remove firmly adherent or inaccessible tissue. In the absence of radiotherapy, ~75% of craniopharyngiomas recur, and 10-year survival is <50%. In patients with incomplete resection, radiotherapy improves 10-year survival to 70–90% but is associated with increased risk of secondary malignancies. Most patients require lifelong pituitary hormone replacement.

Developmental failure of Rathke's pouch obliteration may lead to *Rathke's cysts*, which are small (<5 mm) cysts entrapped by squamous epithelium and are found in ~20% of individuals at autopsy. Although Rathke's cleft cysts do not usually grow and are often diagnosed incidentally, about a third present in adulthood with compressive symptoms, diabetes insipidus, and hyperprolactinemia due to stalk compression. Rarely, hydrocephalus develops. The diagnosis is suggested preoperatively by visualizing the cyst wall on MRI, which distinguishes these lesions from craniopharyngiomas. Cyst contents range from CSF-like fluid to mucoid material. *Arachnoid cysts* are rare and generate an MRI image that is isointense with CSF.

Sella chordomas usually present with bony clival erosion, local invasiveness, and, on occasion, calcification. Normal pituitary tissue may be visible on MRI, distinguishing chordomas from aggressive pituitary adenomas. Mucinous material may be obtained by fine-needle aspiration.

Meningiomas arising in the sellar region may be difficult to distinguish from nonfunctioning pituitary adenomas. Meningiomas typically enhance on MRI and may show evidence of calcification or bony erosion. Meningiomas may cause compressive symptoms.

Histiocytosis X includes a variety of syndromes associated with foci of eosinophilic granulomas. Diabetes insipidus, exophthalmos, and punched-out lytic bone lesions (*Hand-Schüller-Christian disease*) are associated with granulomatous lesions visible on MRI, as well as a characteristic axillary skin rash. Rarely, the pituitary stalk may be involved.

Pituitary metastases occur in ~3% of cancer patients. Bloodborne metastatic deposits are found almost exclusively in the posterior pituitary. Accordingly, diabetes insipidus can be a presenting feature of lung, gastrointestinal, breast, and other pituitary metastases. About half of pituitary metastases originate from breast cancer; ~25% of patients with metastatic breast cancer have such deposits. Rarely, pituitary stalk involvement results in anterior pituitary insufficiency. The MRI diagnosis of a metastatic lesion may be difficult to distinguish from an aggressive pituitary adenoma; the diagnosis may require histologic examination of excised tumor tissue. Primary or metastatic lymphoma, leukemias, and plasmacytomas also occur within the sella.

Hypothalamic hamartomas and *gangliocytomas* may arise from astrocytes, oligodendrocytes, and neurons with varying degrees of differentiation. These tumors may overexpress hypothalamic neuropeptides, including gonadotropin-releasing hormone (GnRH), growth hormone–releasing hormone (GHRH), and corticotropin-releasing hormone (CRH). With GnRH-producing tumors, children present with precocious puberty, psychomotor delay, and laughing-associated seizures. Medical treatment of GnRH-producing hamartomas with long-acting GnRH analogues effectively suppresses gonadotropin secretion and controls premature pubertal development. Rarely, hamartomas also are associated with craniofacial abnormalities; imperforate anus; cardiac, renal, and lung disorders; and pituitary failure as features of *Pallister-Hall syndrome*, which is caused by mutations in the carboxy terminus of the *GLI3* gene. Hypothalamic hamartomas are often contiguous with the pituitary, and preoperative MRI diagnosis may not be possible. Histologic evidence of hypothalamic neurons in tissue resected at transsphenoidal surgery may be the first indication of a primary hypothalamic lesion.

Hypothalamic gliomas and *optic gliomas* occur mainly in childhood and usually present with visual loss. Adults have more aggressive tumors; about a third are associated with neurofibromatosis.

Brain germ cell tumors may arise within the sellar region. They include *dysgerminomas*, which frequently are associated with diabetes insipidus and visual loss. They rarely metastasize. *Germinomas, embryonal carcinomas*, *teratomas*, and *choriocarcinomas* may arise in the parasellar region and produce hCG. These germ cell tumors present with precocious puberty, diabetes insipidus, visual field defects, and thirst disorders. Many patients are GH-deficient with short stature.

PITUITARY ADENOMAS AND HYPERSECRETION SYNDROMES

Pituitary adenomas are the most common cause of pituitary hormone hypersecretion and hyposecretion syndromes in adults. They account for ~15% of all intracranial neoplasms and have been identified with a population prevalence of ~80/100,000. At autopsy, up to one-quarter of all pituitary glands harbor an unsuspected microadenoma (<10 mm diameter). Similarly, pituitary imaging detects small clinically inapparent pituitary lesions in at least 10% of individuals.

Pathogenesis Pituitary adenomas are benign neoplasms that arise from one of the five anterior pituitary cell types. The clinical and biochemical phenotypes of pituitary adenomas depend on the cell type from which they are derived. Thus, tumors arising from lactotrope (PRL), somatotrope (GH), corticotrope (ACTH), thyrotrope (TSH), or gonadotrope (LH, FSH) cells hypersecrete their respective hormones **(Table 373-3)**. Plurihormonal tumors express various combinations of GH, PRL, TSH, ACTH, or the glycoprotein hormone α or β subunits. They may be diagnosed by careful immunocytochemistry or may manifest as clinical syndromes that combine features of these hormonal hypersecretory syndromes. Morphologically, these tumors may arise from a single polysecreting cell type or include cells with mixed function within the same tumor.

Hormonally active tumors are characterized by autonomous hormone secretion with diminished feedback responsiveness to physiologic inhibitory pathways. Hormone production does not always

TABLE 373-3 Classification of Pituitary Adenomas[a]

ADENOMA CELL ORIGIN	HORMONE PRODUCT	CLINICAL SYNDROME
Lactotrope	PRL	Hypogonadism, galactorrhea
Gonadotrope	FSH, LH, subunits	Silent or hypogonadism
Somatotrope	GH	Acromegaly/gigantism
Corticotrope	ACTH/none	Cushing's disease or silent
Mixed growth hormone and prolactin cell	GH, PRL	Acromegaly, hypogonadism, galactorrhea
Other plurihormonal cell	Any	Mixed
Acidophil stem cell	PRL, GH	Hypogonadism, galactorrhea, acromegaly
Mammosomatotrope	PRL, GH	Hypogonadism, galactorrhea, acromegaly
Thyrotrope	TSH	Thyrotoxicosis
Null cell	None	Pituitary failure/none
Oncocytoma	None	Pituitary failure/none

[a]Hormone-secreting tumors are listed in decreasing order of frequency. All tumors may cause local pressure effects, including visual disturbances, cranial nerve palsy, and headache.

Note: For abbreviations, see text.

Source: Adapted from S Melmed: Nat Rev Endocrinol 7:257, 2011.

correlate with tumor size. Small hormone-secreting adenomas may cause significant clinical perturbations, whereas larger adenomas that produce less hormone may be clinically silent and remain undiagnosed (if no central compressive effects occur). About one-third of all adenomas are clinically nonfunctioning and produce no distinct clinical hypersecretory syndrome. Most of them arise from gonadotrope cells and may secrete small amounts of α- and β-glycoprotein hormone subunits or, very rarely, intact circulating gonadotropins. True pituitary carcinomas with documented extracranial metastases are exceedingly rare.

Almost all pituitary adenomas are monoclonal in origin, implying the acquisition of one or more somatic mutations that confer a selective growth advantage. Consistent with their clonal origin, complete surgical resection of small pituitary adenomas usually cures hormone hypersecretion. Nevertheless, hypothalamic hormones such as GHRH and CRH also enhance mitotic activity of their respective pituitary target cells in addition to their role in pituitary hormone regulation. Thus, patients who harbor rare abdominal or chest tumors that elaborate ectopic GHRH or CRH may present with somatotrope or corticotrope hyperplasia with GH or ACTH hypersecretion.

Several etiologic genetic events have been implicated in the development of pituitary tumors. The pathogenesis of sporadic forms of acromegaly has been particularly informative as a model of tumorigenesis. GHRH, after binding to its G protein–coupled somatotrope receptor, uses cyclic adenosine monophosphate (AMP) as a second messenger to stimulate GH secretion and somatotrope proliferation. A subset (~35%) of GH-secreting pituitary tumors contains sporadic mutations in Gsα (Arg 201 → Cys or His; Gln 227 → Arg). These mutations attenuate intrinsic GTPase activity, resulting in constitutive elevation of cyclic AMP, Pit-1 induction, and activation of cyclic AMP response element binding protein (CREB), thereby promoting somatotrope cell proliferation and GH secretion.

Characteristic loss of heterozygosity (LOH) in various chromosomes has been documented in large or invasive macroadenomas, suggesting the presence of putative tumor suppressor genes at these loci in up to 20% of sporadic pituitary tumors, including GH-, PRL-, and ACTH-producing adenomas and some nonfunctioning tumors. Lineage-specific cell cycle disruptions with elevated levels of CDK inhibitors are present in most pituitary adenomas.

Compelling evidence also favors growth factor promotion of pituitary tumor proliferation. Basic fibroblast growth factor (bFGF) is abundant in the pituitary and stimulates pituitary cell mitogenesis, whereas epithelial growth factor receptor (EGFR) signaling induces both hormone synthesis and cell proliferation. Mutations of *USP8* may result in overexpressed EGFR in a subset of ACTH-secreting tumors. Other factors involved in initiation and promotion of pituitary tumors include loss of negative-feedback inhibition (as seen with primary hypothyroidism or hypogonadism) and estrogen-mediated or paracrine angiogenesis. Growth characteristics and neoplastic behavior also may be influenced by several activated oncogenes, including *RAS* and pituitary tumor transforming gene (*PTTG*), or inactivation of growth suppressor genes, including *MEG3*.

Genetic Syndromes Associated with Pituitary Tumors

Several familial syndromes are associated with pituitary tumors, and the genetic mechanisms for some of them have been unraveled **(Table 373-4)**.

Multiple endocrine neoplasia (MEN) 1 is an autosomal dominant syndrome characterized primarily by a genetic predisposition to parathyroid, pancreatic islet, and pituitary adenomas **(Chap. 381)**. MEN1 is caused by inactivating germline mutations in *MENIN*, a constitutively expressed tumor-suppressor gene located on chromosome 11q13. Loss of heterozygosity or a somatic mutation of the remaining normal *MENIN* allele leads to tumorigenesis. About half of affected patients develop prolactinomas; acromegaly and Cushing syndrome are less commonly encountered.

Carney complex is characterized by spotty skin pigmentation, myxomas, and endocrine tumors, including testicular, adrenal, and pituitary adenomas. Acromegaly occurs in ~20% of these patients. A subset of patients have mutations in the R1α regulatory subunit of protein kinase A (*PRKAR1A*).

McCune-Albright syndrome consists of polyostotic fibrous dysplasia, pigmented skin patches, and a variety of endocrine disorders, including acromegaly, adrenal adenomas, and autonomous ovarian function **(Chap. 405)**. Hormonal hypersecretion results from constitutive cyclic AMP production caused by inactivation of the GTPase activity of Gsα. The Gsα mutations occur postzygotically, leading to a mosaic pattern of mutant expression.

Familial acromegaly is a rare disorder in which family members may manifest either acromegaly or gigantism. A subset of families with a predisposition for familial pituitary tumors, especially acromegaly, have been found to harbor germline mutations in the *AIP* gene, which encodes the aryl hydrocarbon receptor interacting protein.

HYPERPROLACTINEMIA

Etiology Hyperprolactinemia is the most common pituitary hormone hypersecretion syndrome in both men and women. PRL-secreting

TABLE 373-4 Familial Pituitary Tumor Syndromes

	GENE MUTATED	CLINICAL FEATURES
Multiple endocrine neoplasia 1 (MEN 1)	*MEN1* (11q13)	Hyperparathyroidism Pancreatic neuroendocrine tumors Foregut carcinoids Adrenal adenomas Skin lesions Pituitary adenomas (40%)
Multiple endocrine neoplasia 4 (MEN 4)	*CDKNIB* (12p13)	Hyperparathyroidism Pituitary adenomas Other tumors
Carney complex	*PRKAR1A* (17q23-24)	Pituitary hyperplasia and adenomas (10%) Atrial myxomas Schwannomas Adrenal hyperplasia Lentigines
Familial pituitary adenomas	*AIP* (11q.13.2)	Acromegaly/gigantism (~15% of afflicted families)

pituitary adenomas (prolactinomas) are the most common cause of PRL levels >200 µg/L (see below). Less pronounced PRL elevation can also be seen with microprolactinomas but is more commonly caused by drugs, pituitary stalk compression, hypothyroidism, or renal failure (Table 373-5).

TABLE 373-5 Etiology of Hyperprolactinemia

I. Physiologic hypersecretion
Pregnancy
Lactation
Chest wall stimulation
Sleep
Stress
II. Hypothalamic–pituitary stalk damage
Tumors
Craniopharyngioma
Suprasellar pituitary mass
Meningioma
Dysgerminoma
Metastases
Empty sella
Lymphocytic hypophysitis
Adenoma with stalk
Compression
Granulomas
Rathke cyst
Irradiation
Trauma
Pituitary stalk section
Suprasellar surgery
III. Pituitary hypersecretion
Prolactinoma
Acromegaly
IV. Systemic disorders
Chronic renal failure
Hypothyroidism
Cirrhosis
Pseudocyesis
Epileptic seizures
V. Drug-induced hypersecretion
Dopamine receptor blockers
Atypical antipsychotics: risperidone
Phenothiazines: chlorpromazine, perphenazine
Butyrophenones: haloperidol
Thioxanthenes
Metoclopramide
Dopamine synthesis inhibitors
α-Methyldopa
Catecholamine depletors
Reserpine
Opiates
H_2 antagonists
Cimetidine, ranitidine
Imipramines
Amitriptyline, amoxapine
Serotonin reuptake inhibitors
Fluoxetine
Calcium channel blockers
Verapamil
Estrogens
Thyrotropin-releasing hormone

Note: Hyperprolactinemia >200 µg/L almost invariably is indicative of a prolactin-secreting pituitary adenoma. Physiologic causes, hypothyroidism, and drug-induced hyperprolactinemia should be excluded before extensive evaluation.

Pregnancy and lactation are the important physiologic causes of hyperprolactinemia. Sleep-associated hyperprolactinemia reverts to normal within an hour of awakening. Nipple stimulation and sexual orgasm also may increase PRL. Chest wall stimulation or trauma (including chest surgery and herpes zoster) invoke the reflex suckling arc with resultant hyperprolactinemia. Chronic renal failure elevates PRL by decreasing peripheral clearance. Primary hypothyroidism is associated with mild hyperprolactinemia, probably because of compensatory TRH secretion. Mutation of the PRL receptor is a rare cause of hyperprolactinemia.

Lesions of the hypothalamic-pituitary region that disrupt hypothalamic dopamine synthesis, portal vessel delivery, or lactotrope responses are associated with hyperprolactinemia. Thus, hypothalamic tumors, cysts, infiltrative disorders, and radiation-induced damage cause elevated PRL levels, usually in the range of 30–100 µg/L. Plurihormonal adenomas (including GH and ACTH tumors) may hypersecrete PRL directly. Pituitary masses, including clinically nonfunctioning pituitary tumors, may compress the pituitary stalk to cause hyperprolactinemia.

Drug-induced inhibition or disruption of dopaminergic receptor function is a common cause of hyperprolactinemia (Table 373-5). Thus, antipsychotics and antidepressants are a relatively common cause of mild hyperprolactinemia. Most patients receiving risperidone have elevated prolactin levels, sometimes exceeding 200 µg/L. Methyldopa inhibits dopamine synthesis, and verapamil blocks dopamine release, also leading to hyperprolactinemia. Hormonal agents that induce PRL include estrogens and thyrotropin-releasing hormone (TRH).

Presentation and Diagnosis Amenorrhea, galactorrhea, and infertility are the hallmarks of hyperprolactinemia in women. If hyperprolactinemia develops before menarche, primary amenorrhea results. More commonly, hyperprolactinemia develops later in life and leads to oligomenorrhea and ultimately to amenorrhea. If hyperprolactinemia is sustained, vertebral bone mineral density can be reduced compared with age-matched controls, particularly when it is associated with pronounced hypoestrogenemia. Galactorrhea is present in up to 80% of hyperprolactinemic women. Although usually bilateral and spontaneous, it may be unilateral or expressed only manually. Patients also may complain of decreased libido, weight gain, and mild hirsutism.

In men with hyperprolactinemia, diminished libido, infertility, and visual loss (from optic nerve compression) are the usual presenting symptoms. Gonadotropin suppression leads to reduced testosterone, impotence, and oligospermia. True galactorrhea is uncommon in men with hyperprolactinemia. If the disorder is long-standing, secondary effects of hypogonadism are evident, including osteopenia, reduced muscle mass, and decreased beard growth.

The diagnosis of idiopathic hyperprolactinemia is made by exclusion of known causes of hyperprolactinemia in the setting of a normal pituitary MRI. Some of these patients may harbor small microadenomas below visible MRI sensitivity (~2 mm).

GALACTORRHEA

Galactorrhea, the inappropriate discharge of milk-containing fluid from the breast, is considered abnormal if it persists longer than 6 months after childbirth or discontinuation of breast-feeding. Postpartum galactorrhea associated with amenorrhea is a self-limiting disorder usually associated with moderately elevated PRL levels. Galactorrhea may occur spontaneously, or it may be elicited by nipple pressure. In both men and women, galactorrhea may vary in color and consistency (transparent, milky, or bloody) and arise either unilaterally or bilaterally. Mammography or ultrasound is indicated for bloody discharges (particularly from a single nipple), which may be caused by breast cancer. Galactorrhea is commonly associated with hyperprolactinemia caused by any of the conditions listed in Table 373-5. Acromegaly is associated with galactorrhea in about one-third of patients. Treatment of galactorrhea usually involves managing the underlying disorder (e.g., replacing T_4 for hypothyroidism, discontinuing a medication, treating prolactinoma).

Laboratory Investigation Basal, fasting morning PRL levels (normally <20 μg/L) should be measured to assess hypersecretion. Both false-positive and false-negative results may be encountered. In patients with markedly elevated PRL levels (>1000 μg/L), reported results may be falsely lowered because of assay artifacts; sample dilution is required to measure these high values accurately. Falsely elevated values may be caused by aggregated forms of circulating PRL, which are usually biologically inactive (macroprolactinemia). Hypothyroidism should be excluded by measuring TSH and T_4 levels.

TREATMENT

Hyperprolactinemia

Treatment of hyperprolactinemia depends on the cause of elevated PRL levels. Regardless of the etiology, however, treatment should be aimed at normalizing PRL levels to alleviate suppressive effects on gonadal function, halt galactorrhea, and preserve bone mineral density. Dopamine agonists are effective for most causes of hyperprolactinemia (see the treatment section for prolactinoma, below) regardless of the underlying cause.

If the patient is taking a medication known to cause hyperprolactinemia, the drug should be withdrawn, if possible. For psychiatric patients who require neuroleptic agents, supervised dose titration or the addition of a dopamine agonist can help restore normoprolactinemia and alleviate reproductive symptoms. However, dopamine agonists may worsen the underlying psychiatric condition, especially at high doses. Hyperprolactinemia usually resolves after adequate thyroid hormone replacement in hypothyroid patients or after renal transplantation in patients undergoing dialysis. Resection of hypothalamic or sellar mass lesions can reverse hyperprolactinemia caused by stalk compression and reduced dopamine tone. Granulomatous infiltrates occasionally respond to glucocorticoid administration. In patients with irreversible hypothalamic damage, no treatment may be warranted. In up to 30% of patients with hyperprolactinemia—usually without a visible pituitary microadenoma—the condition may resolve spontaneously.

PROLACTINOMA

Etiology and Prevalence Tumors arising from lactotrope cells account for about half of all functioning pituitary tumors, with a population prevalence of ~10/100,000 in men and ~30/100,000 in women. Mixed tumors that secrete combinations of GH and PRL, ACTH and PRL, and rarely TSH and PRL are also seen. These plurihormonal tumors are usually recognized by immunohistochemistry, sometimes without apparent clinical manifestations from the production of additional hormones. Microadenomas are classified as <1 cm in diameter and usually do not invade the parasellar region. Macroadenomas are >1 cm in diameter and may be locally invasive and impinge on adjacent structures. The female-to-male ratio for microprolactinomas is 20:1, whereas the sex ratio is near 1:1 for macroadenomas. Tumor size generally correlates directly with PRL concentrations; values >250 μg/L usually are associated with macroadenomas. Men tend to present with larger tumors than women, possibly because the features of male hypogonadism are less readily evident. PRL levels remain stable in most patients, reflecting the slow growth of these tumors. About 5% of microadenomas progress in the long term to macroadenomas.

Presentation and Diagnosis Women usually present with amenorrhea, infertility, and galactorrhea. If the tumor extends outside the sella, visual field defects or other mass effects may be seen. Men often present with impotence, loss of libido, infertility, or signs of central nervous system (CNS) compression, including headaches and visual defects. Assuming that physiologic and medication-induced causes of hyperprolactinemia are excluded (Table 373-5), the diagnosis of prolactinoma is likely with a PRL level >200 μg/L. PRL levels <100 μg/L may be caused by microadenomas, other sellar lesions that decrease dopamine inhibition, or nonneoplastic causes of hyperprolactinemia. For this reason, an MRI should be performed in all patients with hyperprolactinemia. It is important to remember that hyperprolactinemia caused secondarily by the mass effects of nonlactotrope lesions is also corrected by treatment with dopamine agonists despite failure to shrink the underlying mass. Consequently, PRL suppression by dopamine agonists does not necessarily indicate that the underlying lesion is a prolactinoma.

TREATMENT

Prolactinoma

Because microadenomas rarely progress to become macroadenomas, no treatment may be needed if patients are asymptomatic and fertility is not desired; these patients should be monitored by regular serial PRL measurements and MRI scans. For symptomatic microadenomas, therapeutic goals include control of hyperprolactinemia, reduction of tumor size, restoration of menses and fertility, and resolution of galactorrhea. Dopamine agonist doses should be titrated to achieve maximal PRL suppression and restoration of reproductive function **(Fig. 373-3)**. A normalized PRL level does not ensure reduced tumor size. However, tumor shrinkage usually is not seen in those who do not respond with lowered PRL levels. For macroadenomas, formal visual field testing should be performed before initiating dopamine agonists. MRI and visual fields should be assessed at 6- to 12-month intervals until the mass shrinks and annually thereafter until maximum size reduction has occurred.

MEDICAL

Oral dopamine agonists (cabergoline and bromocriptine) are the mainstay of therapy for patients with micro- or macroprolactinomas. Dopamine agonists suppress PRL secretion and synthesis as well as lactotrope cell proliferation. In patients with microadenomas who have achieved normoprolactinemia and significant reduction of tumor mass, the dopamine agonist may be withdrawn after 2 years. These patients should be monitored carefully for evidence of prolactinoma recurrence. About 20% of patients (especially males) are resistant to dopaminergic treatment; these adenomas may exhibit decreased D_2 dopamine receptor numbers or a postreceptor defect. D_2 receptor gene mutations in the pituitary have not been reported.

Cabergoline An ergoline derivative, cabergoline is a long-acting dopamine agonist with high D_2 receptor affinity. The drug effectively suppresses PRL for >14 days after a single oral dose and induces prolactinoma shrinkage in most patients. Cabergoline (0.5–1.0 mg twice weekly) achieves normoprolactinemia and resumption of normal gonadal function in ~80% of patients with microadenomas; galactorrhea improves or resolves in 90% of patients. Cabergoline normalizes PRL and shrinks ~70% of macroprolactinomas. Mass effect symptoms, including headaches and visual disorders, usually improve dramatically within days after cabergoline initiation; improvement of sexual function requires several weeks of treatment but may occur before complete normalization of PRL levels. After initial control of PRL levels has been achieved, cabergoline should be reduced to the lowest effective maintenance dose. In ~5% of treated patients harboring a microadenoma, hyperprolactinemia may resolve and not recur when dopamine agonists are discontinued after long-term treatment. Cabergoline also may be effective in patients resistant to bromocriptine. Adverse effects and drug intolerance are encountered less commonly than with bromocriptine.

Bromocriptine The ergot alkaloid bromocriptine mesylate is a dopamine receptor agonist that suppresses PRL secretion. Because it is short-acting, the drug is preferred when pregnancy is desired. In microadenomas, bromocriptine rapidly lowers serum PRL levels to normal in up to 70% of patients, decreases tumor size, and restores gonadal function. In patients with macroadenomas, PRL levels are also normalized in 70% of patients, and tumor mass shrinkage (≥50%) is achieved in most patients.

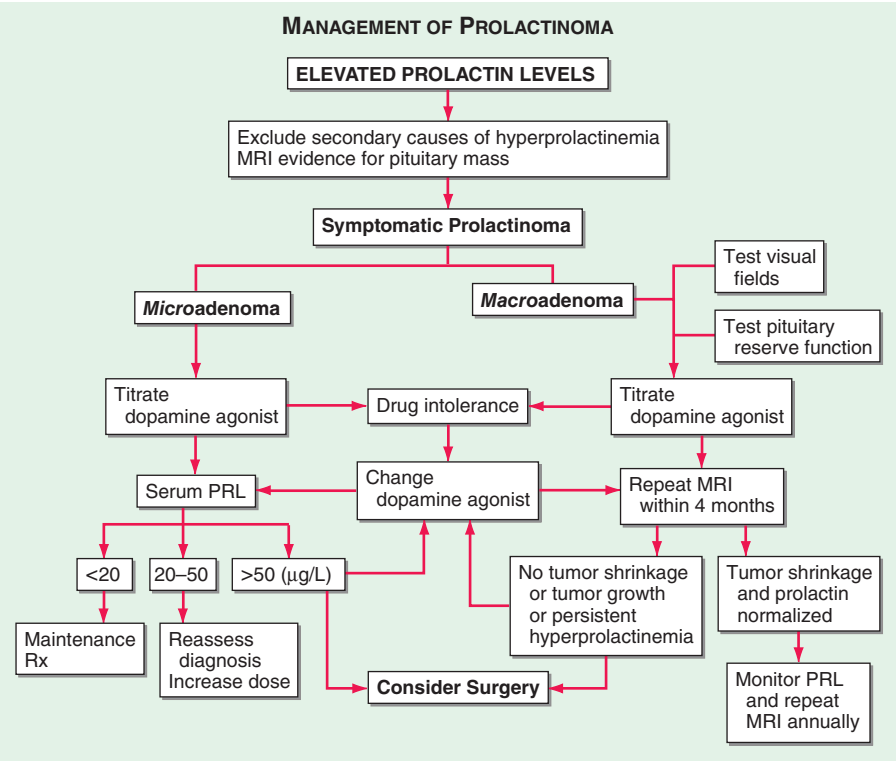


FIGURE 373-3 Management of prolactinoma. MRI, magnetic resonance imaging; PRL, prolactin.

Therapy is initiated by administering a low bromocriptine dose (0.625–1.25 mg) at bedtime with a snack, followed by gradually increasing the dose. Most patients are controlled with a daily dose of <7.5 mg (2.5 mg tid).

SIDE EFFECTS

Side effects of dopamine agonists include constipation, nasal stuffiness, dry mouth, nightmares, insomnia, and vertigo; decreasing the dose usually alleviates these problems. Nausea, vomiting, and postural hypotension with faintness may occur in ~25% of patients after the initial dose. These symptoms may persist in some patients. In general, fewer side effects are reported with cabergoline. For the ~15% of patients who are intolerant of oral bromocriptine, cabergoline may be better tolerated. Intravaginal administration of bromocriptine is often efficacious in patients with intractable gastrointestinal side effects. Auditory hallucinations, delusions, and mood swings have been reported in up to 5% of patients and may be due to the dopamine agonist properties or to the lysergic acid derivative of the compounds. Rare reports of leukopenia, thrombocytopenia, pleural fibrosis, cardiac arrhythmias, and hepatitis have been described. Patients with Parkinson disease who receive at least 3 mg of cabergoline daily have been reported to be at risk for development of cardiac valve regurgitation. Studies analyzing >500 prolactinoma patients receiving recommended doses of cabergoline (up to 2 mg weekly) have shown no evidence for an increased incidence of valvular disorders. Nevertheless, because no controlled prospective studies in pituitary tumor patients are available, it is prudent to perform echocardiograms before initiating standard-dose cabergoline therapy.

Surgery Indications for surgical adenoma debulking include dopamine resistance or intolerance and the presence of an invasive macroadenoma with compromised vision that fails to improve after drug treatment. Initial PRL normalization is achieved in ~70% of microprolactinomas after surgical resection, but only 30% of macroadenomas can be resected successfully. Follow-up studies have shown that hyperprolactinemia recurs in up to 20% of patients within the first year after surgery; long-term recurrence rates exceed 50% for macroadenomas. Radiotherapy for prolactinomas is reserved for patients with aggressive tumors that do not respond to maximally tolerated dopamine agonists and/or surgery.

PREGNANCY

The pituitary increases in size during pregnancy, reflecting the stimulatory effects of estrogen and perhaps other growth factors on pituitary vascularity and lactotrope cell hyperplasia. About 5% of microadenomas significantly increase in size, but 15–30% of macroadenomas grow during pregnancy. Bromocriptine has been used for >30 years to restore fertility in women with hyperprolactinemia, without evidence of teratogenic effects. Nonetheless, most authorities recommend strategies to minimize fetal exposure to the drug. For women taking bromocriptine who desire pregnancy, mechanical contraception should be used through three regular menstrual cycles to allow for conception timing. When pregnancy is confirmed, bromocriptine should be discontinued and PRL levels followed serially, especially if headaches or visual symptoms occur. For women harboring macroadenomas, regular visual field testing is recommended, and the drug should be reinstituted if tumor growth is apparent. Although pituitary MRI may be safe during pregnancy, this procedure should be reserved for symptomatic patients with severe headache and/or visual field defects. Surgical decompression may be indicated if vision is threatened. Although comprehensive data support the efficacy and relative safety of bromocriptine-facilitated fertility, patients should be advised of potential unknown deleterious effects and the risk of tumor growth during pregnancy. Because cabergoline is long-acting with a high D_2-receptor affinity, it is not recommended for use in women when fertility is desired.

■ ACROMEGALY

Etiology GH hypersecretion is usually the result of a somatotrope adenoma but may rarely be caused by extrapituitary lesions (Table 373-6). In addition to the more common GH-secreting somatotrope adenomas, mixed mammosomatotrope tumors and acidophilic stem-cell adenomas secrete both GH and PRL. In patients with acidophilic stem-cell adenomas, features of hyperprolactinemia

TABLE 373-6 Causes of Acromegaly

	PREVALENCE, %
Excess Growth Hormone Secretion	
Pituitary	98
Densely or sparsely granulated GH cell adenoma	60
Mixed GH cell and PRL cell adenoma	25
Mammosomatotrope cell adenoma	10
Plurihormonal adenoma	
GH cell carcinoma or metastases	
Multiple endocrine neoplasia 1 (GH cell adenoma)	
McCune-Albright syndrome	
Ectopic sphenoid or parapharyngeal sinus pituitary adenoma	
Extrapituitary tumor	<1
Pancreatic islet cell tumor	
Lymphoma	
Excess Growth Hormone–Releasing Hormone Secretion	
Central	<1
Hypothalamic hamartoma, choristoma, ganglioneuroma	
Peripheral	<1
Bronchial carcinoid, pancreatic islet cell tumor, small cell lung cancer, adrenal adenoma, medullary thyroid carcinoma, pheochromocytoma	

Abbreviations: GH, growth hormone; PRL, prolactin.

Source: Adapted from S Melmed: N Engl J Med 355:2558, 2006.

(hypogonadism and galactorrhea) predominate over the less clinically evident signs of acromegaly. Occasionally, mixed plurihormonal tumors are encountered that also secrete ACTH, the glycoprotein hormone α subunit, or TSH in addition to GH. Patients with partially empty sellae may present with GH hypersecretion due to a small GH-secreting adenoma within the compressed rim of pituitary tissue; some of these may reflect the spontaneous necrosis of tumors that were previously larger. GH-secreting tumors rarely arise from ectopic pituitary tissue remnants in the nasopharynx or midline sinuses.

There are case reports of ectopic GH secretion by tumors of pancreatic, ovarian, lung, or hematopoietic origin. Rarely, excess GHRH production may cause acromegaly because of chronic stimulation of somatotropes. These patients present with classic features of acromegaly, elevated GH levels, pituitary enlargement on MRI, and pathologic characteristics of pituitary hyperplasia. The most common cause of GHRH-mediated acromegaly is a chest or abdominal carcinoid tumor. Although these tumors usually express positive GHRH immunoreactivity, clinical features of acromegaly are evident in only a minority of patients with carcinoid disease. Excessive GHRH also may be elaborated by hypothalamic tumors, usually choristomas or neuromas.

Presentation and Diagnosis Protean manifestations of GH and IGF-I hypersecretion are indolent and often are not clinically diagnosed for 10 years or more. Acral bony overgrowth results in frontal bossing, increased hand and foot size, mandibular enlargement with prognathism, and widened space between the lower incisor teeth. In children and adolescents, initiation of GH hypersecretion before epiphyseal long bone closure is associated with development of pituitary gigantism **(Fig. 373-4)**. Soft tissue swelling results in increased heel pad thickness, increased shoe or glove size, ring tightening, characteristic coarse facial features, and a large fleshy nose. Other commonly encountered clinical features include hyperhidrosis, a deep and hollow-sounding voice, oily skin, arthropathy, kyphosis, carpal tunnel syndrome, proximal muscle weakness and fatigue, acanthosis nigricans, and skin tags. Generalized visceromegaly occurs, including cardiomegaly, macroglossia, and thyroid gland enlargement.

The most significant clinical impact of GH excess occurs with respect to the cardiovascular system. Cardiomyopathy with arrhythmias, left ventricular hypertrophy, decreased diastolic function, and hypertension ultimately occur in most patients if untreated. Upper airway obstruction with sleep apnea occurs in >60% of patients and is associated with both soft tissue laryngeal airway obstruction and central sleep dysfunction. Diabetes mellitus develops in 25% of patients with

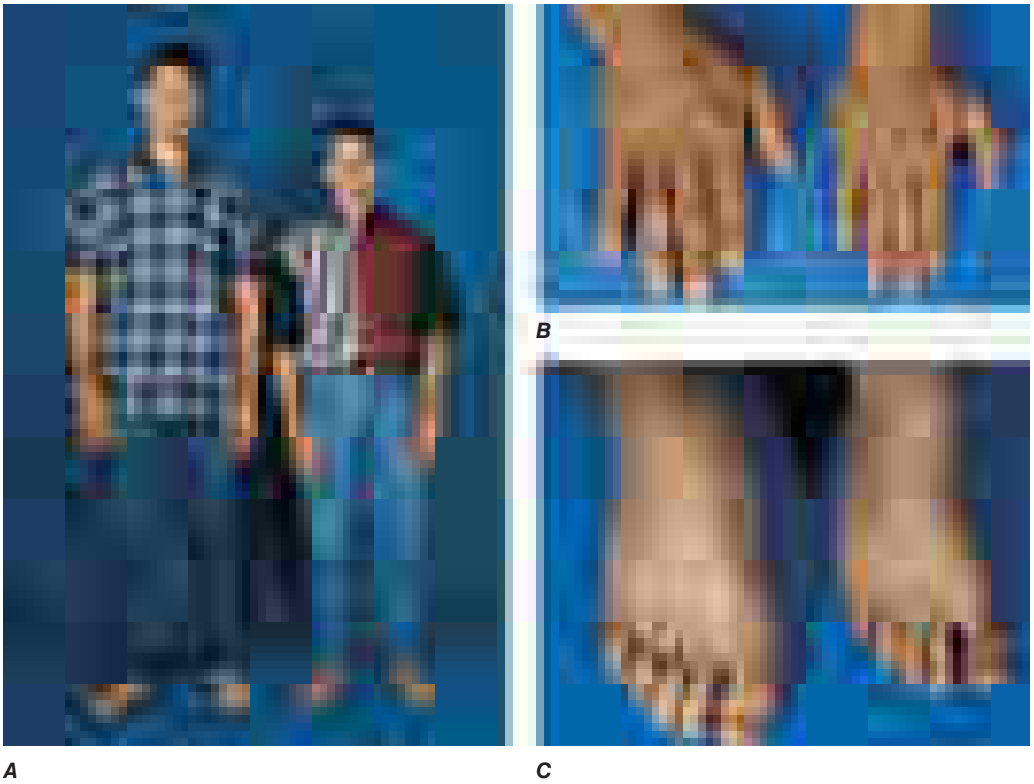


FIGURE 373-4 Features of acromegaly/gigantism. A 22-year-old man with gigantism due to excess growth hormone is shown to the left of his identical twin. The increased height and prognathism (***A***) and enlarged hand (***B***) and foot (***C***) of the affected twin are apparent. Their clinical features began to diverge at the age of ~13 years. *(Reproduced from R Gagel, IE McCutcheon: N Engl J Med 324:524, 1999; with permission.)*

acromegaly, and most patients are intolerant of a glucose load (as GH counteracts the action of insulin). Acromegaly is associated with an increased risk of colon polyps and mortality from colonic malignancy; polyps are diagnosed in up to one-third of patients. Overall mortality is increased about threefold and is due primarily to cardiovascular and cerebrovascular disorders and respiratory disease. Unless GH levels are controlled, survival is reduced by an average of 10 years compared with an age-matched control population.

Laboratory Investigation Age-matched serum IGF-I levels are elevated in acromegaly. Consequently, an IGF-I level provides a useful laboratory screening measure when clinical features raise the possibility of acromegaly. Owing to the pulsatility of GH secretion, measurement of a single random GH level is not useful for the diagnosis or exclusion of acromegaly and does not correlate with disease severity. The diagnosis of acromegaly is confirmed by demonstrating the failure of GH suppression to <0.4 μg/L within 1–2 h of an oral glucose load (75 g). When newer ultrasensitive GH assays are used, normal nadir GH levels are even lower (<0.05 μg/L). About 20% of patients exhibit a paradoxical GH rise after glucose. PRL should be measured, as it is elevated in ~25% of patients with acromegaly. Thyroid function, gonadotropins, and sex steroids may be attenuated because of tumor mass effects. Because most patients will undergo surgery with glucocorticoid coverage, tests of ACTH reserve in asymptomatic patients are more efficiently deferred until after surgery.

TREATMENT

Acromegaly

The goal of treatment is to control GH and IGF-I hypersecretion, ablate or arrest tumor growth, ameliorate comorbidities, restore mortality rates to normal, and preserve pituitary function.

Surgical resection of GH-secreting adenomas is the initial treatment for most patients **(Fig. 373-5)**. Somatostatin analogues are used as adjuvant treatment for preoperative shrinkage of large invasive macroadenomas, immediate relief of debilitating symptoms, and reduction of GH hypersecretion; in frail patients experiencing morbidity; and in patients who decline surgery or, when surgery fails, to achieve biochemical control. Irradiation or repeat surgery may be required for patients who cannot tolerate or do not respond to adjunctive medical therapy. The high rate of late hypopituitarism and the slow rate (5–15 years) of biochemical response are the main disadvantages of radiotherapy. Irradiation is also relatively ineffective in normalizing IGF-I levels. Stereotactic ablation of GH-secreting adenomas by gamma-knife radiotherapy is promising, but long-term results and side effects appear similar to those observed with conventional radiation. Somatostatin analogues may be required while awaiting the full benefits of radiotherapy. Systemic comorbid sequelae of acromegaly, including cardiovascular disease, diabetes, and arthritis, should be managed aggressively. Mandibular surgical repair may be indicated.

SURGERY

Transsphenoidal surgical resection by an experienced surgeon is the preferred primary treatment for both microadenomas (remission rate ~70%) and macroadenomas (<50% in remission). Soft tissue swelling improves immediately after tumor resection. GH levels return to normal within an hour, and IGF-I levels are normalized within 3–4 days. In ~10% of patients, acromegaly may recur several years after apparently successful surgery; hypopituitarism develops in up to 15% of patients after surgery.

SOMATOSTATIN ANALOGUES

Somatostatin analogues exert their therapeutic effects through SSTR2 and SSTR5 receptors, both of which are expressed by GH-secreting tumors. Octreotide acetate is an eight-amino-acid synthetic somatostatin analogue. In contrast to native somatostatin, the analogue is relatively resistant to plasma degradation. It has a 2-h serum half-life and possesses 40-fold greater potency than native somatostatin to suppress GH. Octreotide is administered by subcutaneous injection, beginning with 50 μg tid; the dose can be increased gradually up to 1500 μg/d. Octreotide suppresses integrated GH levels and normalizes IGF-I levels in ~60% of treated patients.

The long-acting somatostatin depot formulations, octreotide and lanreotide, are the preferred medical treatment for patients with acromegaly. *Octreotide LAR* is a sustained-release, long-acting formulation of octreotide incorporated into microspheres that sustain

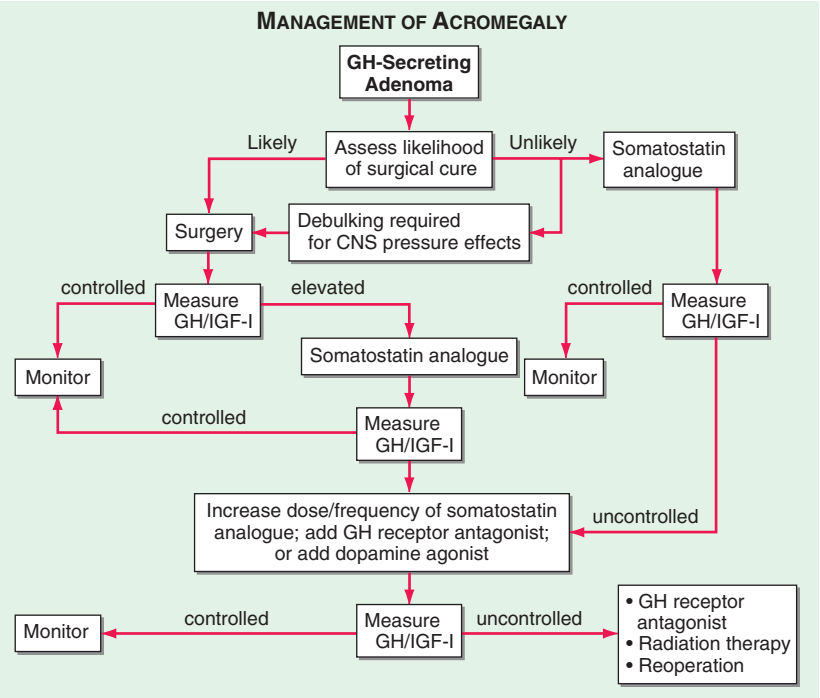


FIGURE 373-5 Management of acromegaly. CNS, central nervous system; IGF, insulin-like growth factor; GH, growth hormone. *(Adapted from S Melmed et al: J Clin Endocrinol Metab 94:1509–1517, 2009; © The Endocrine Society.)*

drug levels for several weeks after intramuscular injection. GH suppression occurs for as long as 6 weeks after a 30-mg intramuscular injection; long-term monthly treatment sustains GH and IGF-I suppression and also reduces pituitary tumor size in ~50% of patients. *Lanreotide* Autogel, a slow-release depot somatostatin preparation, is a cyclic somatostatin octapeptide analogue that suppresses GH and IGF-I hypersecretion after a 60-mg subcutaneous injection. Long-term (every 4–6 weeks) administration controls GH hypersecretion in about two-thirds of treated patients and improves patient compliance because of the long interval required between drug injections. Rapid relief of headache and soft tissue swelling occurs in ~75% of patients within days to weeks of somatostatin analogue initiation. Most patients report symptomatic improvement, including amelioration of headache, perspiration, obstructive apnea, and cardiac failure. For those resistant to octreotide, pasireotide, with preferential SST5 binding, has been shown to exhibit efficacy.

Side Effects Somatostatin analogues are well tolerated in most patients. Adverse effects are short-lived and mostly relate to drug-induced suppression of gastrointestinal motility and secretion. Transient nausea, abdominal discomfort, fat malabsorption, diarrhea, and flatulence occur in one-third of patients, and these symptoms usually remit within 2 weeks. Octreotide suppresses postprandial gallbladder contractility and delays gallbladder emptying; up to 30% of patients develop long-term echogenic sludge or asymptomatic cholesterol gallstones. Other side effects include mild glucose intolerance due to transient insulin suppression, asymptomatic bradycardia, hypothyroxinemia, and local injection site discomfort. Pasireotide is associated with a higher prevalence of glucose intolerance or new-onset diabetes mellitus.

GH RECEPTOR ANTAGONIST

Pegvisomant antagonizes endogenous GH action by blocking peripheral GH binding to its receptor. Consequently, serum IGF-I levels are suppressed, reducing the deleterious effects of excess endogenous GH. Pegvisomant is administered by daily subcutaneous injection (10–20 mg) and normalizes IGF-I in ~70% of patients. GH levels, however, remain elevated as the drug does not target the pituitary adenoma. Side effects include reversible liver enzyme elevation, lipodystrophy, and injection site pain. Tumor size should be monitored by MRI.

Combined treatment with monthly somatostatin analogues and weekly or biweekly pegvisomant injections has been used effectively in resistant patients.

DOPAMINE AGONISTS

Bromocriptine and cabergoline may modestly suppress GH secretion in some patients. Very high doses of bromocriptine (≥20 mg/d) or cabergoline (0.5 mg/d) are usually required to achieve modest GH therapeutic efficacy. Combined treatment with octreotide and cabergoline may induce additive biochemical control compared with either drug alone.

RADIATION

External radiation therapy or high-energy stereotactic techniques are used as adjuvant therapy for acromegaly. An advantage of radiation is that patient compliance with long-term treatment is not required. Tumor mass is reduced, and GH levels are attenuated over time. However, 50% of patients require at least 8 years for GH levels to be suppressed to <5 μg/L; this level of GH reduction is achieved in ~90% of patients after 18 years but represents suboptimal GH suppression. Patients may require interim medical therapy for several years before attaining maximal radiation benefits. Most patients also experience hypothalamic-pituitary damage, leading to gonadotropin, ACTH, and/or TSH deficiency within 10 years of therapy.

In summary, surgery is the preferred primary treatment for GH-secreting microadenomas (Fig. 373-5). The high frequency of GH hypersecretion after macroadenoma resection usually necessitates adjuvant or primary medical therapy for these larger tumors. Patients unable to receive or respond to unimodal medical treatment may benefit from combined treatments, or they can be offered radiation.

CUSHING'S DISEASE (ACTH-PRODUCING ADENOMA)

(See also Chap. 379)

Etiology and Prevalence Pituitary corticotrope adenomas (Cushing's disease) account for 70% of patients with endogenous causes of Cushing's syndrome. However, it should be emphasized that iatrogenic hypercortisolism is the most common cause of cushingoid features. Ectopic tumor ACTH production, cortisol-producing adrenal adenomas, adrenal carcinoma, and adrenal hyperplasia account for the other causes; rarely, ectopic tumor CRH production is encountered.

ACTH-producing adenomas account for ~10–15% of all pituitary tumors. Because the clinical features of Cushing's syndrome often lead to early diagnosis, most ACTH-producing pituitary tumors are relatively small microadenomas. However, macroadenomas also are seen and some ACTH-expressing adenomas are clinically silent. Cushing's disease is 5–10 times more common in women than in men. These pituitary adenomas exhibit unrestrained ACTH secretion, with resultant hypercortisolemia. However, they retain partial suppressibility in the presence of high doses of administered glucocorticoids, providing the basis for dynamic testing to distinguish pituitary from nonpituitary causes of Cushing's syndrome.

Presentation and Diagnosis The diagnosis of Cushing's syndrome presents two great challenges: (1) to distinguish patients with pathologic cortisol excess from those with physiologic or other disturbances of cortisol production and (2) to determine the etiology of pathologic cortisol excess.

Typical features of chronic cortisol excess include thin skin, central obesity, hypertension, plethoric moon facies, purple striae and easy bruisability, glucose intolerance or diabetes mellitus, gonadal dysfunction, osteoporosis, proximal muscle weakness, signs of hyperandrogenism (acne, hirsutism), and psychological disturbances (depression, mania, and psychoses) **(Table 373-7)**. Hematopoietic features of hypercortisolism include leukocytosis, lymphopenia, and eosinopenia. Immune suppression includes delayed hypersensitivity and infection propensity. These protean yet commonly encountered manifestations of hypercortisolism make it challenging to decide which patients mandate formal laboratory evaluation. Certain features make pathologic

TABLE 373-7 Clinical Features of Cushing's Syndrome (All Ages)

SYMPTOMS/SIGNS	FREQUENCY, %
Obesity or weight gain (>115% ideal body weight)	80
Thin skin	80
Moon facies	75
Hypertension	75
Purple skin striae	65
Hirsutism	65
Menstrual disorders (usually amenorrhea)	60
Plethora	60
Abnormal glucose tolerance	55
Impotence	55
Proximal muscle weakness	50
Truncal obesity	50
Acne	45
Bruising	45
Mental changes	45
Osteoporosis	40
Edema of lower extremities	30
Hyperpigmentation	20
Hypokalemic alkalosis	15
Diabetes mellitus	15

Source: Adapted from MA Magiokou et al, in ME Wierman (ed): Diseases of the Pituitary. Totowa, NJ, Humana, 1997.

causes of hypercortisolism more likely; they include characteristic central redistribution of fat, thin skin with striae and bruising, and proximal muscle weakness. In children and young females, early osteoporosis may be particularly prominent. The primary cause of death is cardiovascular disease, but life-threatening infections and risk of suicide are also increased.

Rapid development of features of hypercortisolism associated with skin hyperpigmentation and severe myopathy suggests an ectopic tumor source of ACTH. Hypertension, hypokalemic alkalosis, glucose intolerance, and edema are also more pronounced in these patients. Serum potassium levels <3.3 mmol/L are evident in ~70% of patients with ectopic ACTH secretion but are seen in <10% of patients with pituitary-dependent Cushing's syndrome.

Laboratory Investigation The diagnosis of Cushing's disease is based on laboratory documentation of endogenous hypercortisolism. Measurement of 24-h urine free cortisol (UFC) is a precise and cost-effective screening test. Alternatively, the failure to suppress plasma cortisol after an overnight 1-mg dexamethasone suppression test can be used to identify patients with hypercortisolism. As nadir levels of cortisol occur at night, elevated midnight serum or salivary samples of cortisol are suggestive of Cushing's disease. Basal plasma ACTH levels often distinguish patients with ACTH-independent (adrenal or exogenous glucocorticoid) from those with ACTH-dependent (pituitary, ectopic ACTH) Cushing's syndrome. Mean basal ACTH levels are about eightfold higher in patients with ectopic ACTH secretion than in those with pituitary ACTH-secreting adenomas. However, extensive overlap of ACTH levels in these two disorders precludes using ACTH measurements to make the distinction. Preferably, dynamic testing based on differential sensitivity to glucocorticoid feedback or ACTH stimulation in response to CRH or cortisol reduction is used to distinguish ectopic from pituitary sources of excess ACTH **(Table 373-8)**. Very rarely, circulating CRH levels are elevated, reflecting ectopic tumor-derived secretion of CRH and often ACTH. **For further discussion of dynamic testing for Cushing's syndrome, see Chap. 379.**

Most ACTH-secreting pituitary tumors are <5 mm in diameter, and about half are undetectable by sensitive MRI. The high prevalence of incidental pituitary microadenomas diminishes the ability to distinguish ACTH-secreting pituitary tumors accurately from nonsecreting incidentalomas.

Inferior Petrosal Venous Sampling Because pituitary MRI with gadolinium enhancement is insufficiently sensitive to detect small (<2 mm) pituitary ACTH-secreting adenomas, bilateral inferior petrosal sinus ACTH sampling before and after CRH administration may be required to distinguish these lesions from ectopic ACTH-secreting tumors that may have similar clinical and biochemical characteristics. Simultaneous assessment of ACTH in each inferior petrosal vein and in the diagnosis of peripheral circulation provides a strategy for confirming and localizing pituitary ACTH production. Sampling is performed at baseline and 2, 5, and 10 min after intravenous bovine CRH (1 μg/kg) injection. An increased ratio (>2) of inferior petrosal:peripheral vein ACTH confirms pituitary Cushing's syndrome. After CRH injection, peak petrosal:peripheral ACTH ratios ≥3 confirm the presence of a pituitary ACTH-secreting tumor. The sensitivity of this test is >95%, with very rare false-positive results. False-negative results may be encountered in patients with aberrant venous drainage. Petrosal sinus catheterizations are technically difficult, and ~0.05% of patients develop neurovascular complications. The procedure should not be performed in patients with hypertension, in patients with known cerebrovascular disease, or in the presence of a well-visualized pituitary adenoma on MRI.

TABLE 373-8 Differential Diagnosis of ACTH-Dependent Cushing's Syndrome[a]

	ACTH-SECRETING PITUITARY TUMOR	ECTOPIC ACTH SECRETION
Etiology	Pituitary corticotrope adenoma Plurihormonal adenoma	Bronchial, abdominal carcinoid Small cell lung cancer Thymoma
Sex	F > M	M > F
Clinical features	Slow onset	Rapid onset Pigmentation Severe myopathy
Serum potassium <3.3 μg/L	<10%	75%
24-h UFC	High	High
Basal ACTH level	Inappropriately high	Very high
Dexamethasone suppression 1 mg overnight Low-dose(0.5 mg q6h) High-dose (2 mg q6h)	 Cortisol >5 μg/dL Cortisol <5 μg/dL	 Cortisol >5 μg/dL Cortisol >5 μg/dL
UFC >80% suppressed	Microadenomas: 90% Macroadenomas: 50%	10%
Inferior petrosal sinus sampling (IPSS) Basal IPSS: peripheral CRH-induced IPSS: peripheral	 >2 >3	 <2 <3

[a]ACTH-independent causes of Cushing's syndrome are diagnosed by suppressed ACTH levels and an adrenal mass in the setting of hypercortisolism. Iatrogenic Cushing's syndrome is excluded by history.

Abbreviations: ACTH, adrenocorticotropic hormone; CRH, corticotropin-releasing hormone; F, female; M, male; IPSS, inferior petrosal sinus sampling; UFC, urinary free cortisol.

TREATMENT

Cushing's Disease

Selective transsphenoidal resection is the treatment of choice for Cushing's disease **(Fig. 373-6)**. The remission rate for this procedure is ~80% for microadenomas but <50% for macroadenomas. However, surgery is rarely successful when the adenoma is not visible

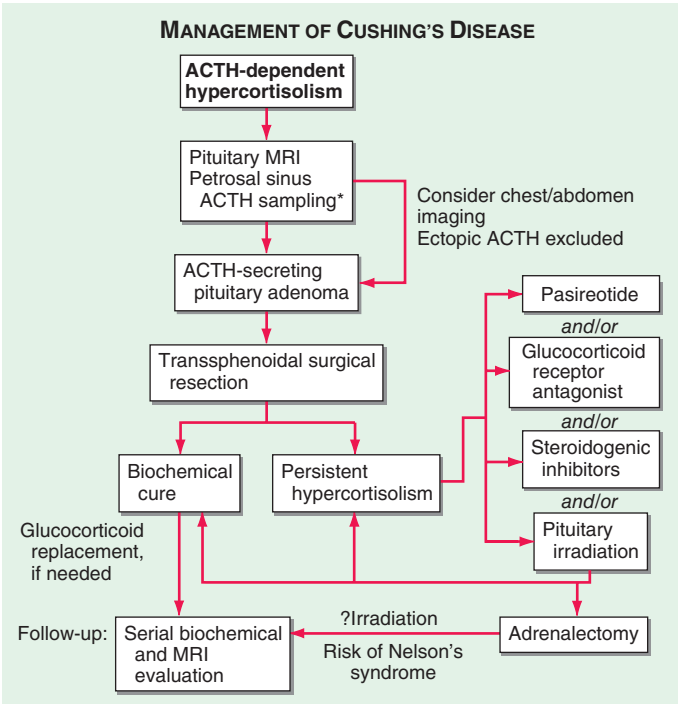


FIGURE 373-6 Management of Cushing's disease. ACTH, adrenocorticotropin hormone; MRI, magnetic resonance imaging; *, Not usually required.

on MRI. After successful tumor resection, most patients experience a postoperative period of symptomatic ACTH deficiency that may last up to 12 months. This usually requires low-dose cortisol replacement, as patients experience both steroid withdrawal symptoms and have a suppressed hypothalamic-pituitary-adrenal axis. Biochemical recurrence occurs in ~5% of patients in whom surgery was initially successful.

When initial surgery is unsuccessful, repeat surgery is sometimes indicated, particularly when a pituitary source for ACTH is well documented. In older patients, in whom issues of growth and fertility are less important, hemi- or total hypophysectomy may be necessary if a discrete pituitary adenoma is not recognized. Pituitary irradiation may be used after unsuccessful surgery, but it cures only ~15% of patients. Because the effects of radiation are slow and only partially effective in adults, steroidogenic inhibitors are used in combination with pituitary irradiation to block adrenal effects of persistently high ACTH levels.

Pasireotide (600 or 900 μg/d subcutaneously), a somatostatin analogue with high affinity for SST5 > SST2 receptors, may control hypercortisolemia in a subset of patients with ACTH-secreting pituitary tumors when surgery is not an option or has not been successful. In clinical trials, the drug lowered plasma ACTH levels and normalized 24-h UFC levels in ~20% of patients, and resulted in up to 40% mean pituitary tumor shrinkage. Side effects include development of hyperglycemia and diabetes in up to 70% of patients, likely due to suppressed pancreatic secretion of insulin and incretins. Because patients with hypercortisolism are insulin-resistant, hyperglycemia should be rigorously managed. Other side effects are similar to those encountered for other somatostatin analogues and include transient abdominal discomfort, diarrhea, nausea, and gallstones (20% of patients). The drug requires consistent long-term administration.

Ketoconazole, an imidazole derivative antimycotic agent, inhibits several P450 enzymes and effectively lowers cortisol in most patients with Cushing's disease when administered twice daily (600–1200 mg/d). Elevated hepatic transaminases, gynecomastia, impotence, gastrointestinal upset, and edema are common side effects.

Mifepristone (300–1200 mg/d), a glucocorticoid receptor antagonist, blocks peripheral cortisol action and is approved to treat hyperglycemia in Cushing's disease. Because the drug does not target the pituitary tumor, both ACTH and cortisol levels remain elevated, thus obviating a reliable circulating biomarker. Side effects are largely due to general antagonism of other steroid hormones and include hypokalemia, endometrial hyperplasia, hypoadrenalism, and hypertension.

Metyrapone (2–4 g/d) inhibits 11β-hydroxylase activity and normalizes plasma cortisol in up to 75% of patients. Side effects include nausea and vomiting, rash, and exacerbation of acne or hirsutism. *Mitotane* (*o,p'*-DDD; 3–6 g/d orally in four divided doses) suppresses cortisol hypersecretion by inhibiting 11β-hydroxylase and cholesterol side-chain cleavage enzymes and by destroying adrenocortical cells. Side effects of mitotane include gastrointestinal symptoms, dizziness, gynecomastia, hyperlipidemia, skin rash, and hepatic enzyme elevation. It also may lead to hypoaldosteronism. Other agents include *aminoglutethimide* (250 mg tid), *trilostane* (200–1000 mg/d), *cyproheptadine* (24 mg/d), and IV *etomidate* (0.3 mg/kg/h). Glucocorticoid insufficiency is a potential side effect of agents used to block steroidogenesis.

The use of steroidogenic inhibitors has decreased the need for bilateral adrenalectomy. Surgical removal of both adrenal glands corrects hypercortisolism but may be associated with significant morbidity rates and necessitates permanent glucocorticoid and mineralocorticoid replacement. Adrenalectomy in the setting of residual corticotrope adenoma tissue predisposes to the development of *Nelson's syndrome*, a disorder characterized by rapid pituitary tumor enlargement and increased pigmentation secondary to high ACTH levels. Prophylactic radiation therapy may be indicated to prevent the development of Nelson's syndrome after adrenalectomy.

NONFUNCTIONING AND GONADOTROPIN-PRODUCING PITUITARY ADENOMAS

Etiology and Prevalence Nonfunctioning pituitary adenomas include those that secrete little or no pituitary hormones as well as tumors that produce too little hormone to result in recognizable clinical features. They are the most common type of pituitary adenoma and are usually macroadenomas at the time of diagnosis because clinical features are not apparent until tumor mass effects occur. Based on immunohistochemistry, most clinically nonfunctioning adenomas can be shown to originate from gonadotrope cells. These tumors typically produce small amounts of intact gonadotropins (usually FSH) as well as uncombined α, LH β, and FSH β subunits. Tumor secretion may lead to elevated α and FSH β subunits and, rarely, to increased LH β subunit levels. Some adenomas express α subunits without FSH or LH. TRH administration often induces an atypical increase of tumor-derived gonadotropins or subunits.

Presentation and Diagnosis Clinically nonfunctioning tumors often present with optic chiasm pressure and other symptoms of local expansion or may be incidentally discovered on an MRI performed for another indication (incidentaloma). Rarely, menstrual disturbances or ovarian hyperstimulation occur in women with large tumors that produce FSH and LH. More commonly, adenoma compression of the pituitary stalk or surrounding pituitary tissue leads to attenuated LH and features of hypogonadism. PRL levels are usually slightly increased, also because of stalk compression. It is important to distinguish this circumstance from true prolactinomas, as nonfunctioning tumors do not shrink in response to treatment with dopamine agonists.

Laboratory Investigation The goal of laboratory testing in clinically nonfunctioning tumors is to classify the type of the tumor, identify hormonal markers of tumor activity, and detect possible hypopituitarism. Free α subunit levels may be elevated in 10–15% of patients with nonfunctioning tumors. In female patients, peri- or postmenopausal basal FSH concentrations are difficult to distinguish from tumor-derived FSH elevation. Premenopausal women have cycling FSH levels, also preventing clear-cut diagnostic distinction from tumor-derived FSH. In men, gonadotropin-secreting tumors may be diagnosed because of slightly increased gonadotropins (FSH > LH) in the setting of a pituitary mass. Testosterone levels are usually low despite the normal or increased LH level, perhaps reflecting reduced LH bioactivity or the loss of normal LH pulsatility. Because this pattern of hormone test results is also seen in primary gonadal failure and, to some extent, with aging (Chap. 384), the finding of increased gonadotropins alone is insufficient for the diagnosis of a gonadotropin-secreting tumor. In the majority of patients with gonadotrope adenomas, TRH administration stimulates LH β subunit secretion; this response is not seen in normal individuals. GnRH testing, however, is not helpful for making the diagnosis. For nonfunctioning and gonadotropin-secreting tumors, the diagnosis usually rests on immunohistochemical analyses of surgically resected tumor tissue, as the mass effects of these tumors usually necessitate resection.

Although acromegaly or Cushing's syndrome usually presents with unique clinical features, clinically inapparent (silent) somatotrope or corticotrope adenomas may only be diagnosed by immunostaining of resected tumor tissue. If PRL levels are <100 μg/L in a patient harboring a pituitary mass, a nonfunctioning adenoma causing pituitary stalk compression should be considered.

TREATMENT

Nonfunctioning and Gonadotropin-Producing Pituitary Adenomas

Asymptomatic small nonfunctioning microadenomas with no threat to vision may be followed with regular MRI and visual field testing without immediate intervention. However, for macroadenomas, transsphenoidal surgery is indicated to reduce tumor size and

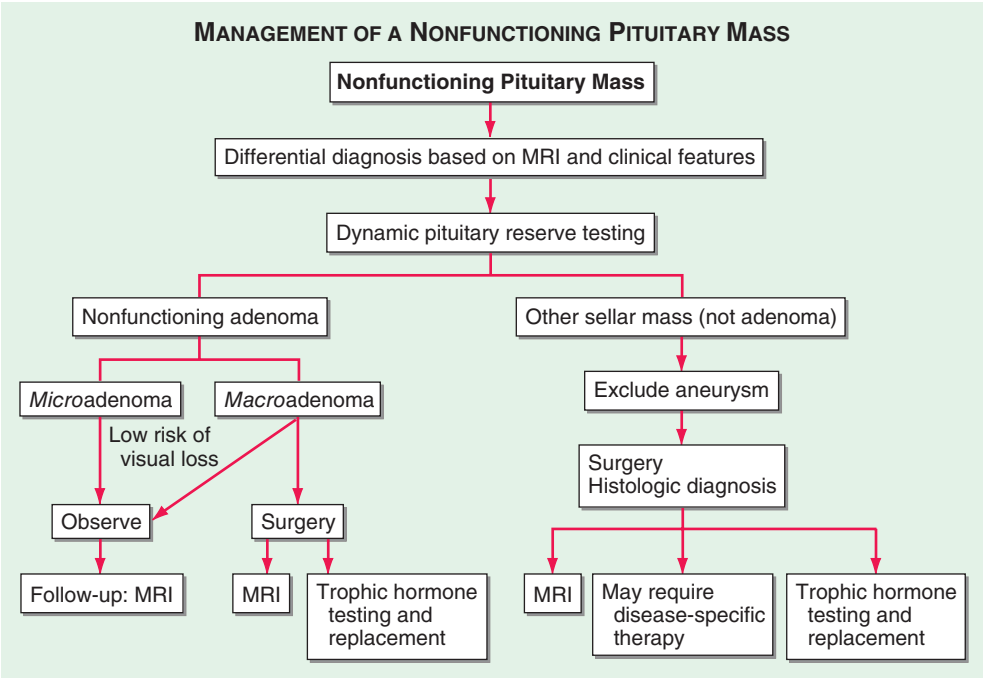


FIGURE 373-7 Management of a nonfunctioning pituitary mass. MRI, magnetic resonance imaging.

relieve mass effects (Fig. 373-7). Although it is not usually possible to remove all adenoma tissue surgically, vision improves in 70% of patients with preoperative visual field defects. Preexisting hypopituitarism that results from tumor mass effects may improve or resolve completely. Beginning ~6 months postoperatively, MRI scans should be performed yearly to detect tumor regrowth. Within 5–6 years after successful surgical resection, ~15% of nonfunctioning tumors recur. When substantial tumor remains after transsphenoidal surgery, adjuvant radiotherapy may be indicated to prevent tumor regrowth. Radiotherapy may be deferred if no postoperative residual mass is evident. Nonfunctioning pituitary tumors respond poorly to dopamine agonist treatment and somatostatin analogues are largely ineffective for shrinking these tumors. The selective GnRH antagonist Nal-Glu GnRH suppresses FSH hypersecretion but has no effect on adenoma size.

TSH-SECRETING ADENOMAS

TSH-producing macroadenomas are very rare but are often large and locally invasive when they occur. Patients usually present with thyroid goiter and hyperthyroidism, reflecting overproduction of TSH. Diagnosis is based on demonstrating elevated serum-free T_4 levels, inappropriately normal or high TSH secretion, and MRI evidence of a pituitary adenoma. Elevated uncombined α subunits are seen in many patients.

It is important to exclude other causes of inappropriate TSH secretion, such as resistance to thyroid hormone, an autosomal dominant disorder caused by mutations in the thyroid hormone β receptor (Chap. 375). The presence of a pituitary mass and elevated β subunit levels are suggestive of a TSH-secreting tumor. Dysalbuminemic hyperthyroxinemia syndromes, caused by mutations in serum thyroid hormone binding proteins, are also characterized by elevated thyroid hormone levels, but with normal rather than suppressed TSH levels. Moreover, free thyroid hormone levels are normal in these disorders, most of which are familial.

TREATMENT

TSH-Secreting Adenomas

The initial therapeutic approach is to remove or debulk the tumor mass surgically, usually using a transsphenoidal approach. Total resection is not often achieved as most of these adenomas are large and locally invasive. Normal circulating thyroid hormone levels are achieved in about two-thirds of patients after surgery. Thyroid ablation or antithyroid drugs (methimazole and propylthiouracil) can be used to reduce thyroid hormone levels. Somatostatin analogue treatment effectively normalizes TSH and α subunit hypersecretion, shrinks the tumor mass in 50% of patients, and improves visual fields in 75% of patients; euthyroidism is restored in most patients. Because somatostatin analogues markedly suppress TSH, biochemical hypothyroidism often requires concomitant thyroid hormone replacement, which may also further control tumor growth.

FURTHER READING

Di Ieva A et al: Aggressive pituitary adenomas—diagnosis and emerging treatments. Nat Rev Endocrinol 10:423, 2014.

Famini P, Maya MM, Melmed S: Pituitary magnetic resonance imaging for sellar and parasellar masses: Ten-year experience in 2598 patients. J Clin Endocrinol Metab 96:1633, 2011.

Freda PU et al: Pituitary incidentaloma: An Endocrine Society clinical practice guideline. J Clin Endocrinol Metab 96:894, 2011.

Gittleman H et al: Descriptive epidemiology of pituitary tumors in the United States, 2004–2009. J Neurosurg 121:527, 2014.

Hannah-Shmouni F, Trivellin G, Stratakis CA: Genetics of gigantism and acromegaly. Growth Horm IGF Res 30–31:37–41, 2016.

Melmed S: Medical progress: acromegaly. N Engl J Med 355:2558, 2006.

Melmed S: Pathogenesis of pituitary tumors. Nat Rev Endocrinol 7:257, 2011.

Melmed S et al: Diagnosis and treatment of hyperprolactinemia: An Endocrine Society clinical practice guideline. J Clin Endocrinol Metab 96:273, 2011.

Nieman LK: Cushing's syndrome: Update on signs, symptoms and biochemical screening. Eur J Endocrinol 173:M33, 2015.

Ntali G et al: Clinical review: Functioning gonadotroph adenomas. J Clin Endocrinol Metab 99:4423, 2014.

374 Disorders of the Neurohypophysis

Gary L. Robertson

The neurohypophysis, or posterior pituitary, is formed by axons that originate in large cell bodies in the supraoptic and paraventricular nuclei of the hypothalamus. It produces two hormones: (1) arginine vasopressin (AVP), also known as antidiuretic hormone, and (2) oxytocin. AVP acts on the renal tubules to reduce water loss by concentrating the urine. Oxytocin stimulates postpartum milk letdown in response to suckling. A deficiency of AVP secretion or action causes diabetes insipidus (DI), a syndrome characterized by the production of large amounts of dilute urine. Excessive or inappropriate AVP production impairs urinary water excretion and predisposes to hyponatremia if water intake is not reduced in parallel with urine output.

VASOPRESSIN

SYNTHESIS AND SECRETION

AVP is a nonapeptide composed of a six-member disulfide ring and a tripeptide tail **(Fig. 374-1)**. It is synthesized via a polypeptide precursor that includes AVP, neurophysin, and copeptin, all encoded by a single gene on chromosome 20. After preliminary processing and folding, the precursor is packaged in neurosecretory vesicles, where it is transported down the axon; further processed to AVP, neurophysin, and copeptin; and stored in neurosecretory vesicles until released by exocytosis into peripheral blood.

AVP secretion is regulated primarily by the "effective" osmotic pressure of body fluids. This control is mediated by specialized hypothalamic cells known as *osmoreceptors*, which are extremely sensitive to small changes in the plasma concentration of sodium and its anions but normally are insensitive to other solutes such as urea and glucose. The osmoreceptors appear to include inhibitory as well as stimulatory components that function in concert to form a threshold, or set point, control system. Below this threshold, plasma AVP is suppressed to levels that permit the development of a maximum water diuresis. Above it, plasma AVP rises steeply in direct proportion to plasma osmolarity, quickly reaching levels sufficient to effect a maximum antidiuresis. The absolute levels of plasma osmolarity/sodium at which minimally and maximally effective levels of plasma AVP occur, vary appreciably from person to person, apparently due to genetic influences on the set and sensitivity of the system. However, the average threshold, or set point, for AVP release corresponds to a plasma osmolarity or sodium of about 280 mosmol/L or 135 meq/L, respectively; levels only 2–4% higher normally result in maximum antidiuresis.

Although it is relatively stable in a healthy adult, the set point of the osmoregulatory system can be lowered by pregnancy, the menstrual cycle, estrogen, and relatively large, acute reductions in blood pressure or volume. Those reductions are mediated largely by neuronal afferents that originate in transmural pressure receptors of the heart and large arteries and project via the vagus and glossopharyngeal nerves to the brainstem, from which postsynaptic projections ascend to the hypothalamus. These pathways maintain a tonic inhibitory tone that decreases when blood volume or pressure falls by >10–20%. This

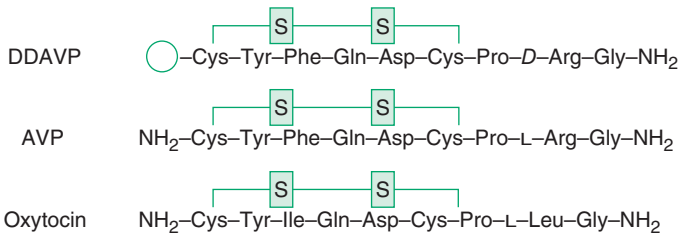


FIGURE 374-1 Primary structures of arginine vasopressin (AVP), oxytocin, and desmopressin (DDAVP).

baroregulatory system is probably of minor importance in the physiology of AVP secretion because the hemodynamic changes required to affect it usually do not occur during normal activities. However, the baroregulatory system undoubtedly plays an important role in AVP secretion in patients with disorders that produce large, acute disturbances of hemodynamic function. AVP secretion also can be stimulated by nausea, acute hypoglycemia, glucocorticoid deficiency, smoking, and, possibly, hyperangiotensinemia. The emetic stimuli are extremely potent since they typically elicit immediate, 50- to 100-fold increases in plasma AVP even when the nausea is transient and is not associated with vomiting or other symptoms. They appear to act via the emetic center in the medulla and can be blocked completely by treatment with antiemetics such as fluphenazine. There is no evidence that pain or other noxious stresses have any effect on AVP unless they elicit a vasovagal reaction with its associated nausea and hypotension.

ACTION

The most important, if not the only, physiologic action of AVP is to reduce water excretion by promoting concentration of urine. This antidiuretic effect is achieved by increasing the hydroosmotic permeability of cells that line the distal tubule and medullary collecting ducts of the kidney **(Fig. 374-2)**. In the absence of AVP, these cells are impermeable to water and reabsorb little, if any, of the relatively large volume of dilute filtrate that enters from the proximal nephron. The lack of reabsorption results in the excretion of very large volumes (as much as 0.2 mL/kg per min) of maximally dilute urine (specific gravity and osmolarity ~1.000 and 50 mosmol/L, respectively), a condition known as *water diuresis*. In the presence of AVP, these cells become selectively permeable to water, allowing the water to diffuse back down the osmotic gradient created by the hypertonic renal medulla. As a result, the dilute fluid passing through the tubules is concentrated and the rate of urine flow decreases. The magnitude of this effect varies in direct proportion to the plasma AVP concentration and the rate of solute excretion. At maximum levels of AVP and normal rates of solute excretion, it approximates a urine flow rate as low as 0.35 mL/min and a urine osmolarity as high as 1200 mosmol/L. This effect is reduced by a solute diuresis such as glucosuria in diabetes mellitus.

At high concentrations, AVP also causes contraction of smooth muscle in blood vessels in the skin and gastrointestinal tract, induces glycogenolysis in the liver, and potentiates adrenocorticotropic hormone (ACTH) release by corticotropin-releasing factor. These effects are mediated by V_{1a} or V_{1b} receptors that are coupled to phospholipase C. Their role, if any, in human physiology/pathophysiology is uncertain.

METABOLISM

AVP distributes rapidly into a space roughly equal to the extracellular fluid volume. It is cleared irreversibly with a half-life ($t_{1/2}$) of 10–30 min. Most AVP clearance is due to degradation in the liver and kidneys. During pregnancy, the metabolic clearance of AVP is increased three- to fourfold due to placental production of an N-terminal peptidase.

THIRST

Because AVP cannot reduce water loss below a certain minimum level obligated by urinary solute load and evaporation from skin and lungs, a mechanism for ensuring adequate intake is essential for preventing dehydration. This vital function is performed by the thirst mechanism. Like AVP, thirst is regulated primarily by an osmostat that is situated in the anteromedial hypothalamus and is able to detect very small changes in the plasma concentration of sodium and its anions. The thirst osmostat appears to be "set" about 3% higher than the AVP osmostat. This arrangement ensures that thirst, polydipsia, and dilution of body fluids do not occur until plasma osmolarity/sodium starts to exceed the defensive capacity of the antidiuretic mechanism.

OXYTOCIN

Oxytocin is also a nonapeptide that differs from AVP only at positions 3 and 8 (Fig. 374-1). However, it has relatively little antidiuretic effect and seems to act mainly on mammary ducts to facilitate milk letdown during nursing. It also may help initiate or facilitate labor by stimulating

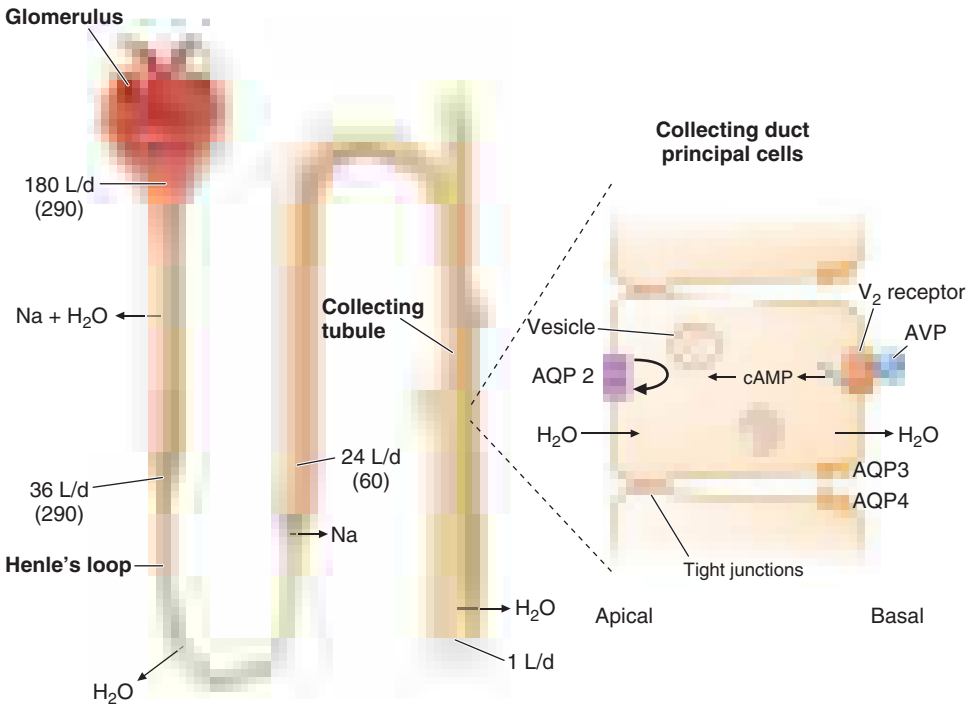


FIGURE 374-2 Antidiuretic effect of arginine vasopressin (AVP) in the regulation of urine volume. In a typical 70-kg adult, the kidney filters ~180 L/d of plasma. Of this, ~144 L (80%) is reabsorbed isosmotically in the proximal tubule and another 8 L (4–5%) is reabsorbed without solute in the descending limb of Henle's loop. The remainder is diluted to an osmolarity of ~60 mmol/kg by selective reabsorption of sodium and chloride in the ascending limb. In the absence of AVP, the urine issuing from the loop passes largely unmodified through the distal tubules and collecting ducts, resulting in a maximum water diuresis. In the presence of AVP, solute-free water is reabsorbed osmotically through the principal cells of the collecting ducts, resulting in the excretion of a much smaller volume of concentrated urine. This antidiuretic effect is mediated via a G protein–coupled V_2 receptor that increases intracellular cyclic AMP, thereby inducing translocation of aquaporin 2 (AQP 2) water channels into the apical membrane. The resultant increase in permeability permits an influx of water that diffuses out of the cell through AQP 3 and AQP 4 water channels on the basal-lateral surface. The net rate of flux across the cell is determined by the number of AQP 2 water channels in the apical membrane and the strength of the osmotic gradient between tubular fluid and the renal medulla. Tight junctions on the lateral surface of the cells serve to prevent unregulated water flow. The V2 receptors and aquaporin 2 are encoded by genes on chromosome Xq28 and 12q13, respectively.

contraction of uterine smooth muscle, but it is not clear if this action is physiologic or necessary for normal delivery.

DEFICIENCIES OF AVP SECRETION AND ACTION

DIABETES INSIPIDUS

Clinical Characteristics A decrease of 75% or more in the secretion or action of AVP usually results in DI, a syndrome characterized by the production of abnormally large volumes of dilute urine. The 24-h urine volume exceeds 40 mL/kg body weight, and the osmolarity is <300 mosmol/L. The polyuria produces symptoms of urinary frequency, enuresis, and/or nocturia, which may disturb sleep and cause mild daytime fatigue or somnolence. It also results in a slight rise in plasma osmolarity that stimulates thirst and a commensurate increase in fluid intake (polydipsia). Overt clinical signs of dehydration are uncommon unless thirst and/or the compensatory increase of fluid intake are also impaired.

Etiology A primary deficiency of AVP secretion usually results from agenesis or irreversible destruction of the neurohypophysis. It is referred to variously as *neurohypophyseal DI*, *neurogenic DI*, *pituitary DI*, *cranial DI*, or *central DI*. It can be caused by a variety of congenital, acquired, or genetic disorders, but in about one-half of all adult patients, it is idiopathic (Table 374-1). Pituitary DI caused by surgery in or around the neurohypophysis usually appears within 24 h. After a few days, it may transition to a 2- to 3-week period of inappropriate antidiuresis, after which the DI may or may not recur permanently. Five genetic forms of pituitary DI are now known. By far, the most common is transmitted in an autosomal dominant mode and is caused by diverse mutations in the coding region of one allele of the AVP–neurophysin II (or *AVP-NPII*) gene. All the mutations alter one or more amino acids known to be critical for correct processing and/or folding of the prohormone, thus interfering with its trafficking through the endoplasmic reticulum. The misfolded mutant precursor accumulates and interferes with production of AVP by the normal allele, eventually destroying the magnocellular neurons in which it is produced. The AVP deficiency and DI are usually not present at birth but develop gradually over a period of several months to years, progressing from partial to severe at different rates depending on the mutation. Once established, the deficiency of AVP is permanent, but for unknown reasons, the DI occasionally improves or remits spontaneously in late middle age. The parvocellular neurons that make AVP and the magnocellular neurons that make oxytocin appear to be unaffected. There are also rare autosomal recessive forms of pituitary DI. One is due to an inactivating mutation in the AVP portion of the gene; another is due to a large deletion involving the majority of the AVP gene and regulatory sequences in the intergenic region. A third form is caused by mutations of the *WFS 1* gene responsible for Wolfram's syndrome (DI, diabetes mellitus, optic atrophy, and neural deafness [DIDMOAD]). An X-linked recessive form linked to a region on Xq28 has also been described but the causative gene has not yet been identified.

A primary deficiency of plasma AVP also can result from increased metabolism by an N-terminal aminopeptidase produced by the placenta. It is referred to as *gestational DI* because the signs and symptoms manifest during pregnancy and usually remit several weeks after delivery.

Secondary deficiencies of AVP secretion result from inhibition by excessive intake of fluids. They are referred to as *primary polydipsia* and can be divided into three subcategories. One of them, *dipsogenic DI*, is characterized by inappropriate thirst caused by a reduction in the set of the osmoregulatory mechanism. It sometimes occurs in association with multifocal diseases of the brain such as neurosarcoid, tuberculous meningitis, and multiple sclerosis but is often idiopathic. The second subtype, *psychogenic polydipsia*, is not associated with thirst, and the polydipsia seems to be a feature of psychosis or obsessive compulsive disorder. The third subtype, *iatrogenic polydipsia*, results from recommendations to increase fluid intake for its presumed health benefits.

Primary deficiencies in the antidiuretic action of AVP result in *nephrogenic DI*. The causes can be genetic, acquired, or drug induced (Table 374-1). The most common genetic form is transmitted in a semirecessive X-linked manner. It is caused by mutations in the coding region of the V_2 receptor gene that impair trafficking and/or ligand binding of the mutant receptor. There are also autosomal recessive or dominant forms of nephrogenic DI. They are caused by *AQP2* gene mutations that result in complete or partial defects in trafficking and function of the water channels that mediate antidiuresis in the distal and collecting tubules of the kidney.

Secondary deficiencies in the antidiuretic response to AVP result from polyuria per se. They are caused by washout of the medullary concentration gradient and/or suppression of aquaporin function. They usually resolve 24–48 h after the polyuria is corrected but can

TABLE 374-1 Causes of Diabetes Insipidus

Pituitary diabetes insipidus
Acquired
Head trauma (closed and penetrating) including pituitary surgery
Neoplasms
Primary
Craniopharyngioma
Pituitary adenoma (suprasellar)
Dysgerminoma
Meningioma
Metastatic (lung, breast)
Hematologic (lymphoma, leukemia)
Granulomas
Sarcoidosis
Histiocytosis
Xanthoma disseminatum
Infectious
Chronic meningitis
Viral encephalitis
Toxoplasmosis
Inflammatory
Lymphocytic infundibuloneurohypophysitis
Granulomatosis with polyangiitis (Wegener's)
Lupus erythematosus
Scleroderma
Chemical toxins
Tetrodotoxin
Snake venom
Vascular
Sheehan's syndrome
Aneurysm (internal carotid)
Aortocoronary bypass
Hypoxic encephalopathy
Idiopathic
Congenital malformations
Septo-optic dysplasia
Midline craniofacial defects
Holoprosencephaly
Hypogenesis, ectopia of pituitary
Genetic
Autosomal dominant (*AVP-neurophysin* gene)
Autosomal recessive
Type A (*AVP-neurophysin* gene)
Type B (*AVP-neurophysin* gene)
Type C (*Wolfram's [4p-WFS 1]* gene)
X-linked recessive (Xq28)
Gestational diabetes insipidus
Pregnancy (second and third trimesters)
Nephrogenic diabetes insipidus
Acquired
Drugs
Lithium
Demeclocycline
Methoxyflurane
Amphotericin B
Aminoglycosides
Cisplatin
Rifampin
Foscarnet
Metabolic
Hypercalcemia, hypercalciuria
Hypokalemia
Obstruction (ureter or urethra)
Vascular
Sickle cell disease and trait
Ischemia (acute tubular necrosis)
Granulomas
Sarcoidosis
Neoplasms
Sarcoma
Infiltration
Amyloidosis
Idiopathic
Genetic
X-linked recessive (*AVP receptor-2* gene)
Autosomal recessive (*AQP2* gene)
Autosomal dominant (*AQP2* gene)
Primary polydipsia
Acquired
Psychogenic
Schizophrenia
Obsessive compulsive disorder
Dipsogenic (abnormal thirst)
Granulomas (sarcoidosis)
Infectious (tuberculous meningitis)
Head trauma (closed and penetrating)
Demyelination (multiple sclerosis)
Drugs
Idiopathic
Iatrogenic

complicate interpretation of some acute tests used for differential diagnosis.

Pathophysiology In pituitary, gestational, or nephrogenic DI, the polyuria results in a small (1–2%) decrease in body water and a commensurate increase in plasma osmolarity and sodium that stimulates thirst and a compensatory increase in water intake. As a result, *hypernatremia and other overt physical or laboratory signs of dehydration do not develop unless the patient also has a defect in thirst or fails to increase fluid intake for some other reason.*

In pituitary and nephrogenic DI, the severity of the defect in AVP secretion or action varies significantly from patient to patient. In some, the defect is so severe that it cannot be overcome by even an intense stimulus such as nausea or severe dehydration. In others, the defect in AVP secretion or action is incomplete, and a modest stimulus such as a few hours of fluid deprivation, smoking, or a vasovagal reaction can raise urine osmolarity as high as 800 mosmol/L. However, even when the defects are partial, the relation of urine osmolarity to plasma AVP in patients with nephrogenic DI **(Fig. 374-3*A*)** or of plasma AVP to plasma osmolarity and sodium in patients with pituitary DI **(Fig. 374-3*B*)** is subnormal.

In primary polydipsia, the pathogenesis of the polydipsia and polyuria is the reverse of that in pituitary, nephrogenic, and gestational DI. In primary polydipsia, an abnormality in cognition or thirst causes excessive intake of fluids and an increase in body water that reduces plasma osmolarity/sodium, AVP secretion, and urinary concentration. Dilution of the urine, in turn, results in a compensatory increase in urinary free-water excretion that usually offsets the increase in intake and stabilizes plasma osmolarity/sodium at a level only 1–2% below basal. Thus, hyponatremia or clinically appreciable overhydration is uncommon unless the polydipsia is very severe or the compensatory water diuresis is impaired by a drug or disease that stimulates or mimics the antidiuretic effect of endogenous AVP. A rise in plasma osmolarity and sodium produced by fluid deprivation or hypertonic saline infusion increases plasma AVP normally. However, the resultant increase in urine concentration is often subnormal because polyuria per se temporarily reduces the capacity of the kidney to concentrate the urine. Thus, the maximum level of urine osmolarity achieved during fluid deprivation is often indistinguishable from that in patients with partial pituitary or partial nephrogenic DI.

Differential Diagnosis When symptoms of urinary frequency, enuresis, nocturia, and/or persistent thirst are present in the absence of glucosuria, the possibility of DI should be evaluated by collecting a 24-h urine on ad libitum fluid intake. If the osmolarity is <300 mosmol/L and the volume >50 mL/kg per day, the patient has DI and should be evaluated further to determine the type and select appropriate therapy. If the volume and osmolarity are not concordant, the possibility of inaccurate collection can be evaluated by determining if total urinary creatinine is normal for the size of the patient (20–30 mg/kg/day).

The type of DI can sometimes be inferred from the clinical setting or medical history. Often, however, such information is lacking, ambiguous, or misleading, and other approaches to differential diagnosis are needed. If basal plasma osmolarity and sodium are within normal limits, the traditional approach is to determine the effect of fluid deprivation and injection of antidiuretic hormone on urine osmolarity. This approach suffices for differential diagnosis *if* fluid deprivation raises plasma osmolarity and sodium above the normal range *without* inducing concentration of the urine. In that event, primary polydipsia and partial defects in AVP secretion and action are excluded, and the effect on urine osmolarity of injecting 2 μg of the AVP analogue, desmopressin indicates whether the patient has severe pituitary DI or severe nephrogenic DI. However, this approach is of little or no diagnostic value if fluid deprivation results in concentration of the urine because the increases in urine osmolarity achieved both before and after the injection of desmopressin are similar in patients with *partial* pituitary DI, *partial* nephrogenic DI, and primary polydipsia. These disorders can be differentiated by measuring plasma AVP during fluid deprivation and relating it to the concurrent level of plasma and urine osmolarity (Fig. 374-3). However, this approach does not always differentiate clearly between partial pituitary DI and primary polydipsia unless the measurement is made when plasma osmolarity and sodium are at or above the normal range. This level is difficult to achieve by fluid deprivation alone once urinary concentration occurs. Therefore, it is usually necessary to give a short infusion of 3% saline condition (0.1 mL/kg body weight per minute for 60–90 min) and repeat the measurement of plasma AVP. This approach is highly reliable for differential diagnosis but it is often stressful for the patient and requires special facilities and staff to perform safely and accurately.

A simpler, and less stressful, but equally reliable way to differentiate between pituitary DI, nephrogenic DI, and primary polydipsia is

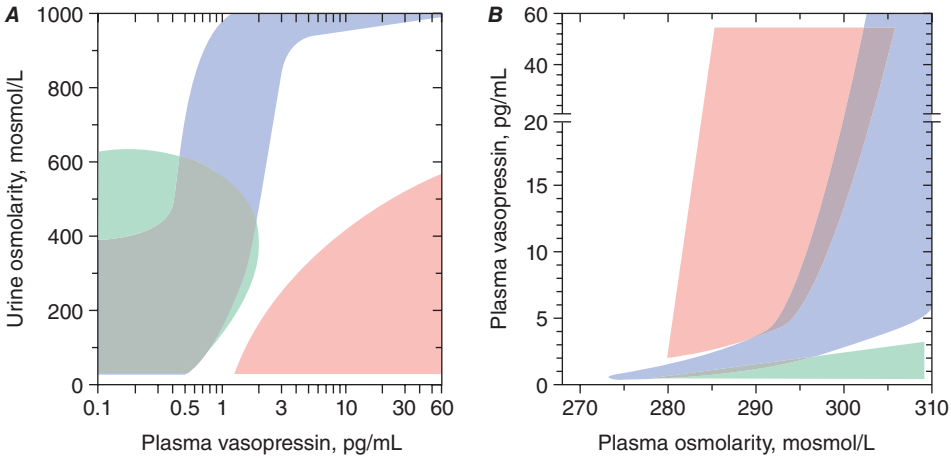


FIGURE 374-3 Relationship of plasma AVP to urine osmolarity (*A*) and plasma osmolarity (*B*) before and during fluid deprivation–hypertonic saline infusion test in patients who are normal or have primary polydipsia (blue zones), pituitary diabetes insipidus (green zones), or nephrogenic diabetes insipidus (pink zones).

to start by measuring basal plasma AVP and urine osmolarity under conditions of unrestricted fluid intake (Fig. 374-4). If AVP is normal or elevated (>1 pg/mL) and the concurrent urine osmolarity is low (<300 mosm/L), the patient has nephrogenic DI and the only additional evaluation required is to determine the cause. If, however, basal plasma AVP is low or undetectable (<1 pg/mL), nephrogenic DI is very unlikely and MRI of the brain can be performed to differentiate pituitary DI from primary polydipsia. In most healthy adults and children, the posterior pituitary emits a hyperintense signal visible in T1-weighted midsagittal images. This "bright spot" is almost always present in patients with primary polydipsia but is always absent or abnormally small in patients with pituitary DI, even if their AVP deficiency is partial. The MRI is also useful in searching for pathology responsible for pituitary DI or the dipsogenic form of primary polydipsia (Fig. 374-2). The principal caveat is that MRI is not reliable for differential diagnosis of DI in patients with empty sella because they typically lack a bright spot even when their AVP secretion and action are normal. MRI also cannot be used to differentiate pituitary from nephrogenic DI because many patients with nephrogenic DI also lack a posterior pituitary bright spot, probably because they have an abnormally high rate of AVP secretion and turnover.

If MRI and/or AVP assays with the requisite sensitivity and specificity are unavailable and a fluid deprivation test is impractical or undesirable, a third way to differentiate between pituitary DI, nephrogenic DI, and primary polydipsia is a trial of desmopressin therapy. Such a trial should be conducted with very close monitoring of serum sodium as well as urine output, preferably in hospital, because desmopressin will produce hyponatremia in 8–24 h if the patient has primary polydipsia.

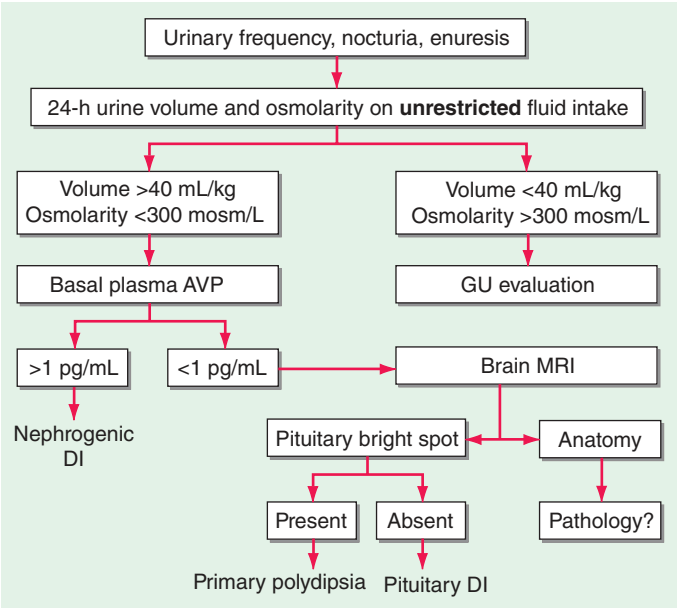


FIGURE 374-4 Simplified approach to the differential diagnosis of diabetes insipidus. When symptoms suggest diabetes insipidus (DI), the syndrome should be differentiated from a genitourinary (GU) abnormality by measuring the 24-h urine volume and osmolarity on unrestricted fluid intake. If DI is confirmed, basal plasma arginine vasopressin (AVP) should be measured on unrestricted fluid intake. If AVP is normal or elevated (>1 pg/mL), the patient probably has nephrogenic DI. However, if plasma AVP is low or undetectable, the patient has either pituitary DI or primary polydipsia. In that case, magnetic resonance imaging (MRI) of the brain can be performed to differentiate between these two conditions by determining whether or not the normal posterior pituitary bright spot is visible on T1-weighted midsagittal images. In addition, the MRI anatomy of the pituitary hypothalamic area can be examined to look for evidence of pathology that sometimes causes pituitary DI or the dipsogenic form of primary polydipsia. MRI is not reliable for differential diagnosis unless nephrogenic DI has been excluded because the bright spot is also absent, small, or faint in this condition.

TREATMENT

Diabetes Insipidus

The signs and symptoms of uncomplicated pituitary DI can be eliminated by treatment with desmopressin (DDAVP), a synthetic analogue of AVP (Fig. 374-1). DDAVP acts selectively at V2 receptors to increase urine concentration and decrease urine flow in a dose-dependent manner. It is also more resistant to degradation than is AVP and has a three- to fourfold longer duration of action. DDAVP can be given by IV or SC injection, nasal inhalation, or orally by means of a tablet or melt. The doses required to control pituitary DI vary widely, depending on the patient and the route of administration. However, among adults, they usually range from 1–2 μg qd or bid by injection, 10–20 μg bid or tid by nasal spray, or 100–400 μg bid or tid orally. The onset of antidiuresis is rapid, ranging from as little as 15 min after injection to 60 min after oral administration. When given in a dose that normalizes 24-h urinary osmolarity (400–800 mosmol/L) and volume (15–30 mL/kg body weight), DDAVP produces a slight (1–3%) increase in total body water and a decrease in plasma osmolarity/sodium that rapidly eliminates thirst and polydipsia (Fig. 374-5). Consequently, water balance is maintained within the normal range. Hyponatremia rarely develops unless urine volume is reduced too far (to <10 mL/kg per day) or fluid intake is excessive due to an associated abnormality in thirst or cognition. Fortunately, thirst abnormalities are rare, and if the patient is taught to drink only when truly thirsty, DDAVP can be given safely in doses sufficient to normalize urine output without the need for allowing intermittent escape to prevent water intoxication.

Primary polydipsia cannot be treated safely with DDAVP or any other antidiuretic drug because eliminating the polyuria does

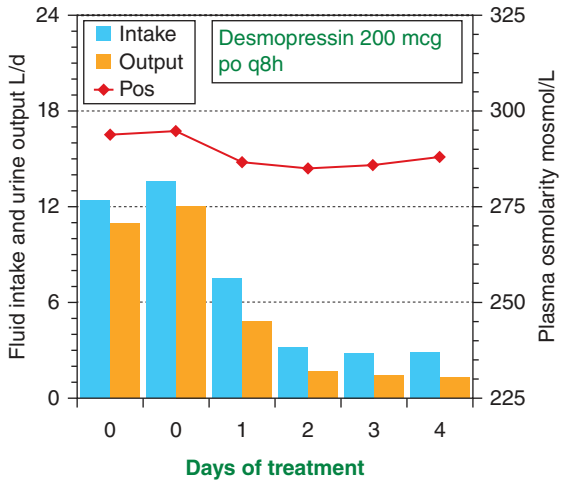


FIGURE 374-5 Effect of desmopressin therapy on fluid intake (*blue bars*), urine output (*orange bars*), and plasma osmolarity (*red line*) in a patient with uncomplicated pituitary diabetes insipidus. Note that treatment rapidly reduces fluid intake and urine output to normal, with only a slight increase in body water as evidenced by the slight decrease in plasma osmolarity.

not eliminate the urge to drink. Therefore, it invariably produces hyponatremia and/or other signs of water intoxication, usually within 8–24 h if urine output is normalized completely. There is no consistently effective way to correct dipsogenic or psychogenic polydipsia, but the iatrogenic form may respond to patient education. To minimize the risk of water intoxication, all patients should be warned about the use of other drugs such as thiazide diuretics or carbamazepine (Tegretol) that can impair urinary free-water excretion directly or indirectly.

The polyuria and polydipsia of nephrogenic DI are not affected by treatment with standard doses of DDAVP. If resistance is partial, it may be overcome by tenfold higher doses, but this treatment is too expensive and inconvenient for long-term use. However, treatment with conventional doses of a thiazide diuretic and/or amiloride in conjunction with a low-sodium diet and a prostaglandin synthesis inhibitor (e.g., indomethacin) usually reduces the polyuria and polydipsia by 30–70% and may eliminate them completely in some patients. Side effects such as hypokalemia and gastric irritation can be minimized by the use of amiloride or potassium supplements and by taking medications with meals.

HYPODIPSIC HYPERNATREMIA

An increase in plasma osmolarity/sodium above the normal range (hypertonic hypernatremia) can be caused by either a decrease in total body water or an increase in total body sodium. The former results from a failure to drink enough to replace normal or increased urinary and insensible water loss. The deficient intake can be due either to water deprivation or a lack of thirst (hypodipsia). The most common cause of an increase in total body sodium is primary hyperaldosteronism (Chap. 379). Rarely, it can also result from ingestion of hypertonic saline in the form of sea water or incorrectly prepared infant formula. However, even in these forms of hypernatremia, inadequate intake of water also contributes. This chapter focuses on hypodipsic hypernatremia, the form of hypernatremia due to a primary defect in the thirst mechanism.

Clinical Characteristics Hypodipsic hypernatremia is a syndrome characterized by chronic or recurrent hypertonic dehydration. The hypernatremia varies widely in severity and usually is associated with signs of hypovolemia such as tachycardia, postural hypotension, azotemia, hyperuricemia, and hypokalemia due to secondary hyperaldosteronism. Muscle weakness, pain, rhabdomyolysis, hyperglycemia, hyperlipidemia, and acute renal failure may also occur. Obtundation or coma may be present but are often absent. Despite inappropriately low levels of plasma AVP, DI usually is not evident at presentation but may develop during rehydration as blood volume, blood pressure, and plasma osmolarity/sodium return toward normal, further reducing plasma AVP.

Etiology Hypodipsia is usually due to hypogenesis or destruction of the osmoreceptors in the anterior hypothalamus that regulate thirst. These defects can result from various congenital malformations of midline brain structures or may be acquired due to diseases such as occlusions of the anterior communicating artery, primary, or metastatic tumors in the hypothalamus, head trauma, surgery, granulomatous diseases such as sarcoidosis and histiocytosis, AIDS, and cytomegalovirus encephalitis. Because of their proximity, the osmoreceptors that regulate AVP secretion also are usually impaired. Thus, AVP secretion responds poorly or not at all to hyperosmotic stimulation (Fig. 374-6) but, in most cases, increases normally to nonosmotic stimuli such as nausea or large reductions in blood volume or blood pressure, indicating that the neurohypophysis is intact.

Pathophysiology Hypodipsia results in a failure to drink enough water to replenish obligatory renal and extrarenal losses. Consequently, plasma osmolarity and sodium rise often to extremely high levels before the disorder is recognized. In most cases, urinary loss of water contributes little, if any, to the dehydration because AVP continues to be secreted in the small amounts necessary to concentrate the urine. In some patients this appears to be due to hypovolemic stimulation and/or incomplete destruction of AVP osmoreceptors because plasma AVP

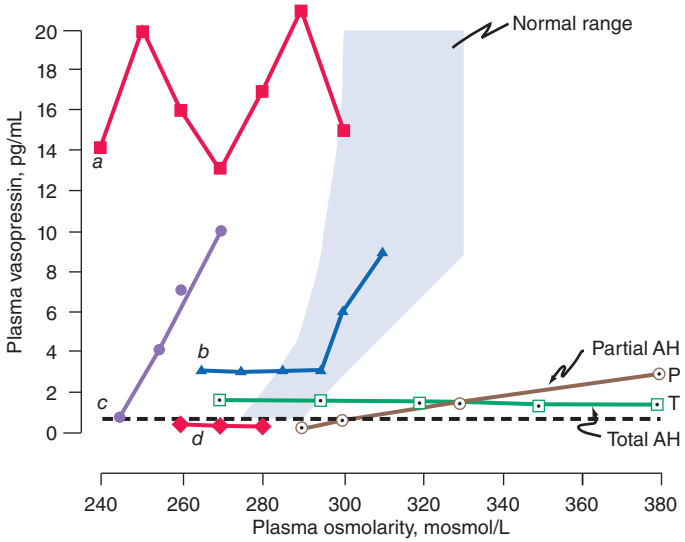


FIGURE 374-6 Heterogeneity of osmoregulatory dysfunction in adipsic hypernatremia (AH) and the syndrome of inappropriate antidiuresis (SIAD). Each line depicts schematically the relationship of plasma arginine vasopressin (AVP) to plasma osmolarity during water loading and/or infusion of 3% saline in a patient with either AH (*open symbols*) or SIAD (*closed symbols*). The shaded area indicates the normal range of the relationship. The horizontal broken line indicates the plasma AVP level below which the hormone is undetectable and urinary concentration usually does not occur. Lines P and T represent patients with a selective deficiency in the osmoregulation of thirst and AVP that is either partial (⊙) or total (⊡). In the latter, plasma AVP does not change in response to increases or decreases in plasma osmolarity but remains within a range sufficient to concentrate the urine even if overhydration produces hypotonic hyponatremia. In contrast, if the osmoregulatory deficiency is partial (⊙), rehydration of the patient suppresses plasma AVP to levels that result in urinary dilution and polyuria before plasma osmolarity and sodium are reduced to normal. Lines *a–d* represent different defects in the osmoregulation of plasma AVP observed in patients with SIADH or SIAD. In *a* (■), plasma AVP is markedly elevated and fluctuates widely without relation to changes in plasma osmolarity, indicating complete loss of osmoregulation. In *b* (▲), plasma AVP remains fixed at a slightly elevated level until plasma osmolarity reaches the normal range, at which point it begins to rise appropriately, indicating a selective defect in the inhibitory component of the osmoregulatory mechanism. In *c* (●), plasma AVP rises in close correlation with plasma osmolarity before the latter reaches the normal range, indicating downward resetting of the osmostat. In *d* (◆), plasma AVP appears to be osmoregulated normally, suggesting that the inappropriate antidiuresis is caused by some other abnormality.

declines and DI develops during rehydration (Fig. 374-6). In others, however, plasma AVP does not decline during rehydration even if they are overhydrated. Consequently, they develop a hyponatremic syndrome indistinguishable from inappropriate antidiuresis. This suggests that the AVP osmoreceptors normally provide inhibitory and stimulatory input to the neurohypophysis and the patients can no longer osmotically stimulate or suppress tonic secretion of the hormone because both inputs have been totally eliminated by the same pathology that destroyed the osmoregulation of thirst. In a few patients, the neurohypophysis is also destroyed, resulting in a combination of chronic pituitary DI and hypodipsia that is particularly difficult to manage.

Differential Diagnosis Hypodipsic hypernatremia usually can be distinguished from other causes of inadequate fluid intake (e.g., coma, paralysis, restraints, absence of fresh water) by the clinical history and setting. Previous episodes and/or denial of thirst and failure to drink spontaneously when the patient is conscious, unrestrained, and hypernatremic are virtually diagnostic. The hypernatremia caused by excessive retention or intake of sodium can be distinguished by the presence of thirst as well as the physical and laboratory signs of hypervolemia rather than hypovolemia.

TREATMENT

Hypodipsic Hypernatremia

Hypodipsic hypernatremia should be treated by administering water orally if the patient is alert and cooperative or by infusing hypotonic fluids (0.45% saline or 5% dextrose and water) if the patient is not. The amount of free water in liters required to correct the deficit (ΔFW) can be estimated from body weight in kg (BW) and the serum sodium concentration in mmol/L (S_{Na}) by the formula $\Delta FW = 0.5BW \times ([S_{Na} - 140]/140)$. If serum glucose ($S_{Glu}$) is elevated, the measured S_{Na} should be corrected (S_{Na}^*) by the formula $S_{Na}^* = S_{Na} + ([S_{Glu} - 90]/36)$. This amount plus an allowance for continuing insensible and urinary losses should be given over a 24- to 48-h period. Close monitoring of serum sodium as well as fluid intake and urinary output is essential because, depending on the extent of osmoreceptor deficiency, some patients will develop AVP-deficient DI, requiring DDAVP therapy to complete rehydration; others will develop hyponatremia and a syndrome of inappropriate antidiuresis (SIAD)-like picture if overhydrated. If hyperglycemia and/or hypokalemia are present, insulin and/or potassium supplements should be given with the expectation that both can be discontinued soon after rehydration is complete. Plasma urea/creatinine should be monitored closely for signs of acute renal failure caused by rhabdomyolysis, hypovolemia, and hypotension.

Once the patient has been rehydrated, an MRI of the brain and tests of anterior pituitary function should be performed to look for the cause and collateral defects in other hypothalamic functions. A long-term management plan to prevent or minimize recurrence of the fluid and electrolyte imbalance also should be developed. This should include a practical method to regulate fluid intake in accordance with variations in water balance as indicated by changes in body weight or serum sodium determined by home monitoring analyzers. Prescribing a constant fluid intake is ineffective and potentially dangerous because it does not take into account the large, uncontrolled variations in insensible loss that inevitably result from changes in ambient temperature and physical activity.

HYPONATREMIA DUE TO INAPPROPRIATE ANTIDIURESIS

A decrease in plasma osmolarity/sodium below the normal range (hypotonic hyponatremia) can be due to any of three different types of salt and water imbalance: (1) an increase in total body water that exceeds the increase in total body sodium (hypervolemic hyponatremia); (2) a decrease in body sodium greater than the decrease in body water (hypovolemic hyponatremia); or (3) an increase in body water with little or no change in body sodium (euvolemic hyponatremia) (Chap. 49). All three forms are associated with a failure to fully dilute the urine and mount a water diuresis in the face of hypotonic hyponatremia. However, the disorders with which they are associated and the types of salt and water imbalance that result differ. The hypervolemic form typically occurs in disorders like severe congestive heart failure or cirrhosis in which water is retained in excessive of sodium. The hypovolemic form typically occurs in disorders such as severe diarrhea, diuretic abuse, or mineralocorticoid deficiency in which sodium is lost in excess of water. Euvolemic hyponatremia, however, is due mainly to expansion of total body water caused by excessive intake in the face of a failure to dilute the urine in response to excessive water intake. The impaired dilution is usually caused by a defect in the osmotic suppression of AVP that can have either of two causes. One is a nonhemodynamic stimulus such as nausea or a cortisol deficiency, which can be corrected quickly by treatment with antiemetics or cortisol. The other is a primary defect in osmoregulation caused by another disorder such as malignancy, stroke, or pneumonia that cannot be easily or quickly corrected. The latter is commonly known as the syndrome of inappropriate antidiuretic hormone (SIADH). Much less often, euvolemic hyponatremia can also result from AVP-independent activation of renal V_2 receptors, a variant known as nephrogenic inappropriate antidiuresis or NSIAD. Both of the latter will be discussed in this chapter.

Clinical Characteristics Antidiuresis of any cause decreases the volume and increases the concentration of urine. If not accompanied by a commensurate reduction in fluid intake or an increase in insensible loss, the reduction in urine output results in excess water retention which expands and dilutes body fluids. If the hyponatremia develops gradually or has been present for more than a few days, it may be largely asymptomatic. However, if it develops acutely, it is usually accompanied by symptoms and signs of water intoxication that may include mild headache, confusion, anorexia, nausea, vomiting, coma, and convulsions. Severe acute hyponatremia may be lethal. Other clinical signs and symptoms vary greatly, depending on the type of hyponatremia. The hypervolemic form is characterized by generalized edema and other signs of marked volume expansion. The opposite is evident in the hypovolemic form. However, overt signs of volume expansion or contraction are absent in SIADH, SIAD, NSIAD, and other forms of euvolemic hyponatremia.

Etiology In SIADH, the inappropriate secretion of AVP can have many different causes. They include ectopic production of AVP by lung cancer or other neoplasms; eutopic release induced by various diseases or drugs; and exogenous administration of AVP, DDAVP, or large doses of oxytocin (Table 374-2). The ectopic forms result from abnormal expression of the *AVP-NPII* gene by primary or metastatic malignancies. The eutopic forms occur most often in patients with acute infections or strokes but have also been associated with many other neurologic diseases and injuries. The mechanisms by which these diseases interfere with osmotic suppression of AVP are not known. The defect in osmoregulation can take any of four distinct forms (Fig. 374-6). In one of the most common (reset osmostat), AVP secretion remains fully responsive to changes in plasma osmolarity/sodium, but the threshold, or set point, of the osmoregulatory system is abnormally low. These patients differ from those with the other types of SIADH in that they are able to maximally suppress plasma AVP and dilute their urine if their fluid intake is high enough to reduce their plasma osmolarity and/or sodium to the lower set point. In most patients, SIADH is self-limited and remits spontaneously within 2–3 weeks, but about 10% of cases are chronic. Another, smaller subgroup (~10% of the total) has inappropriate antidiuresis without a demonstrable defect in the osmoregulation of plasma AVP (Fig. 374-6). In some of them, all young boys, the inappropriate antidiuresis has been traced to a constitutively activating mutation of the V_2 receptor gene. This unusual variant may be referred to as familial nephrogenic SIAD (NSIAD) to distinguish it from other possible causes of the syndrome. The inappropriate antidiuresis in these patients appears to be permanent, although the hyponatremia is variable owing presumably to individual differences in fluid intake.

TABLE 374-2 Causes of Syndrome of Inappropriate Antidiuretic Hormone (SIADH)

Neoplasms	Neurologic
Carcinomas	Guillain-Barré syndrome
Lung	Multiple sclerosis
Duodenum	Delirium tremens
Pancreas	Amyotrophic lateral sclerosis
Ovary	Hydrocephalus
Bladder, ureter	Psychosis
Other neoplasms	Peripheral neuropathy
Thymoma	Congenital malformations
Mesothelioma	Agenesis corpus callosum
Bronchial adenoma	Cleft lip/palate
Carcinoid	Other midline defects
Gangliocytoma	Metabolic
Ewing's sarcoma	Acute intermittent porphyria
Head trauma (closed and penetrating)	Pulmonary
Infections	Asthma
Pneumonia, bacterial or viral	Pneumothorax
Abscess, lung or brain	Positive-pressure respiration
Cavitation (aspergillosis)	Drugs
Tuberculosis, lung or brain	Vasopressin or desmopressin
Meningitis, bacterial or viral	Serotonin reuptake inhibitors
Encephalitis	Oxytocin, high dose
AIDS	Vincristine
Vascular	Carbamazepine
Cerebrovascular occlusions, hemorrhage	Nicotine
Cavernous sinus thrombosis	Phenothiazines
	Cyclophosphamide
	Tricyclic antidepressants
	Monoamine oxidase inhibitors

Pathophysiology Impaired osmotic suppression of antidiuresis results in excessive retention of water and dilution of body fluids only if water intake exceeds insensible and urinary losses. The excess intake is sometimes due to an associated defect in the osmoregulation of thirst (dipsogenic) but can also be psychogenic or iatrogenic, including excessive IV administration of hypotonic fluids. In SIADH and other forms of euvolemic hyponatremia, the decrease in plasma osmolarity/sodium and the increase in extracellular and intracellular volume are proportional to the amount of water retained. Thus, an increase in body water of 10% (~4 L in a 70-kg adult) reduces plasma osmolarity and sodium by ~10% (~28 mosmol/L or 14 meq/L). An increase in body water of this magnitude is rarely detectable on physical examination but will be reflected in a weight gain of about 4 kg. It also increases glomerular filtration and atrial natriuretic hormone and suppresses plasma renin activity, thereby increasing urinary sodium excretion. The resultant reduction in total body sodium decreases the expansion of extracellular volume but aggravates the hyponatremia and further expands intracellular volume. The latter increases brain swelling and intracranial pressure, which probably produces most of the symptoms of acute water intoxication. Within a few days, this swelling may be counteracted by inactivation or elimination of intracellular solutes, resulting in the remission of symptoms even though the hyponatremia persists.

In type I (hypervolemic) or type II (hypovolemic) hyponatremia, osmotic suppression of AVP secretion appears to be counteracted by a hemodynamic stimulus resulting from a large reduction in cardiac output and/or effective blood volume. The resultant antidiuresis is enhanced by decreased distal delivery of glomerular filtrate that results from increased reabsorption of sodium in proximal nephron. If the reduction in urine output is not associated with a commensurate reduction in water intake or an increase in insensible loss, body fluids are expanded and diluted, resulting in hyponatremia despite an increase in body sodium. Unlike SIADH and other forms of euvolemic hyponatremia, however, glomerular filtration is reduced and plasma renin activity and aldosterone are elevated. Thus, the rate of urinary sodium excretion is low (unless sodium reabsorption is impaired by a diuretic), and the hyponatremia is usually accompanied by edema, hypokalemia, azotemia, and hyperuricemia. In type II (hypovolemic) hyponatremia, sodium and water are also retained as an appropriate compensatory response to the severe depletion.

Differential Diagnosis SIADH is a diagnosis of exclusion that usually can be made from the history, physical examination, and basic laboratory data. If hyperglycemia is present, its contribution to the reduction in plasma sodium can be estimated either by measuring plasma osmolarity for a more accurate estimate of the true "effective" tonicity of body fluids or by correcting the measured plasma sodium for the reduction caused by the hyperglycemia using the simplified formula

$$\text{corrected } P_{na} = \text{measured } P_{na} + (P_{glu} - 90)/36$$

where P_{na} = plasma sodium in meq/L and P_{glu} = plasma glucose in mg/dL.

If the plasma osmolarity and/or corrected plasma sodium are below normal limits, hypotonic hyponatremia is present and further evaluation to determine the type should be undertaken in order to administer safe and effective treatment. This differentiation is usually possible by evaluating standard clinical indicators of the extracellular fluid volume **(Table 374-3)**. If these findings are ambiguous or contradictory, measuring plasma renin activity or the *rate* of urinary sodium excretion may be helpful *provided* that the hyponatremia is not in the recovery phase or is due to a primary defect in renal conservation of sodium, diuretic abuse, or hyporeninemic hypoaldosteronism. The latter may be suspected if serum potassium is elevated instead of low, as it usually is in types I and II hyponatremia. Measurements of plasma AVP are currently of no value in differentiating SIADH from the other types of hyponatremia since the plasma levels are elevated similarly in all. In patients who fulfill the clinical criteria for type III (euvolemic) hyponatremia, morning plasma cortisol should also be measured to exclude secondary adrenal insufficiency. If it is normal and there is no history of nausea/vomiting, the diagnosis of SIADH is confirmed, and a careful search for occult lung cancer or other common causes of the syndrome (Table 374-2) should be undertaken.

SIAD due to an activating mutation of the V_2 receptor gene should be suspected if the hyponatremia occurs in a child or several members of the family or is refractory to treatment with a vaptan (see below). In that case, plasma AVP should be measured to confirm that it is appropriately suppressed while the hyponatremia and antidiuresis are present, and the V_2 receptor gene should be sequenced, if possible.

TREATMENT

Hyponatremia

The management of hyponatremia differs depending on the type and the severity and duration of symptoms. In acute symptomatic SIADH, the aim should be to raise plasma osmolarity and/or plasma sodium at a rate ~1% an hour until they reach levels of ~270 mosmol/L or 130 meq/L, respectively. This can be accomplished in either of two ways. One is to infuse hypertonic (3%) saline at a rate of about 0.05 mL/kg body weight per min. This treatment often produces a solute diuresis that serves to remove some of the excess water. The other treatment for acute, symptomatic SIADH is to reduce body water by giving an AVP receptor-2 antagonist (vaptan) to block the antidiuretic effect of AVP and increase urine output **(Fig. 374-7)**. One of the vaptans, a combined V_2/V_{1a} antagonist (Conivaptan), has been approved for short-term, in-hospital IV treatment of SIADH. It should be given as a loading dose of 20 mg IV over 30 min followed by a continuous infusion of 20 mg over 24 h. Another vaptan (Tolvaptan) can be given orally starting at a dose of 15 mg po and increasing to 30 mg or 60 mg at 24 h intervals depending on the effect. With either approach, fluid intake should be restricted to less than urine output. Because the

TABLE 374-3 Differential Diagnosis of Hyponatremia Based on Clinical Assessment of Extracellular Fluid Volume (ECFV)

CLINICAL FINDINGS	TYPE I, HYPERVOLEMIC	TYPE II, HYPOVOLEMIC	TYPE III, EUVOLEMIC	SIADH AND SIAD EUVOLEMIC
History				
CHF, cirrhosis, or nephrosis	Yes	No	No	No
Salt and water loss	No	Yes	No	No
ACTH–cortisol deficiency and/or nausea and vomiting	No	No	Yes	No
Physical examination				
Generalized edema, ascites	Yes	No	No	No
Postural hypotension	Maybe	Maybe	Maybe[a]	No
Laboratory				
BUN, creatinine	High-normal	High-normal	Low-normal	Low-normal
Uric acid	High-normal	High-normal	Low-normal	Low-normal
Serum potassium	Low-normal	Low-normal[b]	Normal[c]	Normal
Serum urate	High	High	Low	Low
Serum albumin	Low-normal	High-normal	Normal	Normal
Serum cortisol	Normal-high	Normal-high[d]	Low[e]	Normal
Plasma renin activity	High	High	Low[f]	Low
Urinary sodium (meq per unit of time)[g]	Low	Low[h]	High[i]	High[i]

[a]Postural hypotension may occur in secondary (ACTH-dependent) adrenal insufficiency even though extracellular fluid volume and aldosterone are usually normal. [b]Serum potassium may be high if hypovolemia is due to aldosterone deficiency. [c]Serum potassium may be low if vomiting causes alkalosis. [d]Serum cortisol is low if hypovolemia is due to primary adrenal insufficiency (Addison's disease). [e]Serum cortisol will be normal or high if the cause is nausea and vomiting rather than secondary (ACTH-dependent) adrenal insufficiency. [f]Plasma renin activity may be high if the cause is secondary (ACTH) adrenal insufficiency. [g]Urinary sodium should be expressed as the *rate of excretion* rather than the concentration. In a hyponatremic adult, an excretion rate >25 meq/d (or 25 μeq/mg of creatinine) could be considered high. [h]The rate of urinary sodium excretion may be high if the hypovolemia is due to diuretic abuse, primary adrenal insufficiency, or other causes of renal sodium wasting. [i]The rate of urinary sodium excretion may be low if intake is curtailed by symptoms or treatment.

Abbreviations: ACTH, adrenocorticotropic hormone; BUN, blood urea nitrogen; CHF, congestive heart failure; SIAD, syndrome of inappropriate antidiuresis.

aquaretic effect of the vaptans varies in magnitude from patient to patient, the rate of rise in serum sodium also varies if fluid intake is fixed at a constant rate. This variability in effect can be reduced or eliminated by continuously monitoring the rate of urine output and adjusting the rate of IV or oral fluid intake so as to reduce body water at a constant rate. Regulating fluid intake so that it under replaces urine output by 5mL/kg body weight/h will raise serum sodium at a rate of about 1% an hour. In any event, serum sodium should be checked every 2–4 h to ensure it is not raised faster than 1mEq/L per hour or above the lower limit of normal. Doing so may result in central pontine myelinolysis, an acute, potentially fatal neurologic syndrome characterized by quadriparesis, ataxia, and abnormal extraocular movements.

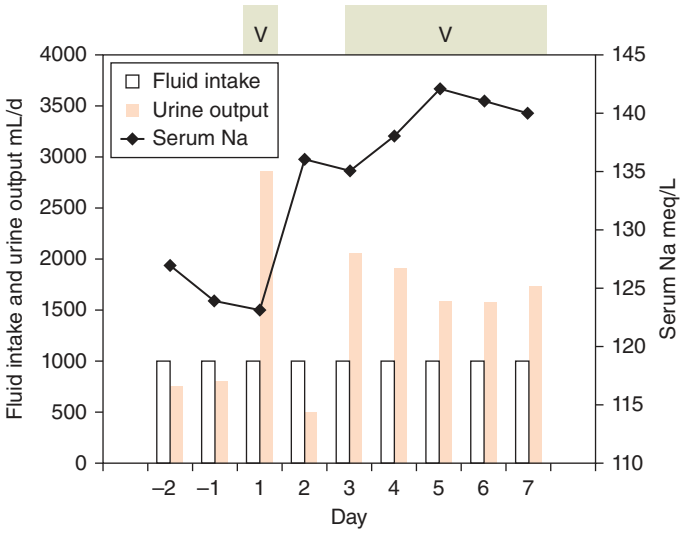


FIGURE 374-7 The effect of vaptan therapy on water balance in a patient with chronic syndrome of inappropriate antidiuretic hormone (SIADH). The periods of vaptan (V) therapy are indicated by the green shaded boxes at the top. Urine output is indicated by orange bars. Fluid intake is shown by the open bars. Intake was restricted to 1 L/d throughout. Serum sodium is indicated by the black line. Note that sodium increased progressively when vaptan increased urine output to levels that clearly exceeded fluid intake.

In chronic and/or minimally symptomatic SIADH, the hyponatremia can and should be corrected more gradually. This can be achieved by restricting total fluid intake to less than the sum of urinary and insensible losses. Because the water derived from food (300–700 mL/d) usually approximates basal insensible losses in adults, the aim should be to reduce total discretionary intake (all liquids) to ~500 mL less than urinary output. Adherence to this regimen is often problematic and, even if achieved, usually reduces body water and increases serum sodium by only about 1–2% per day. Therefore, it is often necessary to add a treatment that increases urinary water excretion. The oral AVP2 antagonist, tolvaptan, is best suited for this purpose. The best approach for treatment of chronic SIADH is the administration of an oral vaptan, tolvaptan, a selective V_2 antagonist that also increases urinary water excretion by blocking the antidiuretic effect of AVP. Some restriction of fluid intake may also be necessary to achieve satisfactory control of the hyponatremia. It is approved for treatment of nonemergent SIADH with initial in-hospital dosing. Other approaches include demeclocycline, 150–300 mg PO tid or qid, which induces a reversible form of nephrogenic DI in 1–2 weeks, or fludrocortisone, 0.05–0.2 mg PO bid. The effect of the demeclocycline manifests in 7–14 days and is due to induction of a reversible form of nephrogenic DI. Fludrocortisone, 0.05–0.2 mg po bid, also raises serum sodium gradually over 1–2 weeks. Its mechanism of action is unclear but probably involves increased retention of sodium. It also increases urinary potassium excretion, which may require replacement through dietary adjustments or supplements and may induce hypertension, occasionally necessitating discontinuation of the treatment.

In the type of euvolemic hyponatremia caused by protracted nausea and vomiting or isolated glucocorticoid deficiency (type III), all abnormalities can be corrected quickly and completely by giving an antiemetic or stress doses of hydrocortisone (for glucocorticoid deficiency). As with other treatments, care must be taken to ensure that serum sodium does not rise too quickly or too far.

In SIAD due to an activating mutation of the V_2 receptor, the V_2 antagonists may not block the antidiuresis or raise plasma

osmolarity/sodium. In that condition, use of an osmotic diuretic such as urea is reported to be effective in preventing or correcting hyponatremia. However, some vaptans may be effective in patients with a different type of activating mutation so the response to this therapy may be neither predictable nor diagnostic.

In hypervolemic hyponatremia, fluid restriction is also appropriate and somewhat effective if it can be maintained. The infusion of hypertonic saline is contraindicated because it further increases total body sodium and edema and may precipitate cardiovascular decompensation. However, as in SIADH, the V_2 receptor antagonists are also safe and effective in the treatment of hypervolemic hyponatremia caused by congestive heart failure. Tolvaptan is approved by the Food and Drug Administration for this indication with the caveat that treatment should be initiated or reinitiated in hospital. Its use should also be limited to 30 days at a time because of reports that longer periods may be associated with abnormal liver chemistries.

In hypovolemic hyponatremia, the imbalance can be corrected easily and quickly by stopping the loss of sodium and water and/or replacing the deficits by mouth or IV infusion of normal or hypertonic saline. As with the treatment of other forms of hyponatremia, care must be taken to ensure that plasma sodium does not increase too rapidly or too far. Fluid restriction and administration of AVP antagonists are contraindicated in type II hyponatremia because they would only aggravate the underlying volume depletion and could result in hemodynamic collapse.

GLOBAL PERSPECTIVES

The incidence, clinical characteristics, etiology, pathophysiology, differential diagnosis, and treatments of fluid and electrolyte disorders in tropical and nonindustrialized countries differ in some respects from those in the United States and other industrialized parts of the world. Hyponatremia, for example, appears to be more common and is more likely to be due to infectious diseases such as cholera, shigellosis, and other diarrheal disorders. In these circumstances, hyponatremia is probably due to gastrointestinal losses of salt and water (hypovolemia type II), but other abnormalities, including undefined infectious toxins, also may contribute. The causes of DI are similar worldwide except that malaria and venoms from snake or insect bites are much more common in some tropical climates.

FURTHER READING

ROBERTSON GL: Vaptans for the treatment of hyponatremia. Nature Rev Endocrin 7:151, 2011.

ROBERTSON GL: Diabetes insipidus: Differential diagnosis and management. Best Prac Res Clin Endocrin Metabol 30:205, 2016.

375 Thyroid Gland Physiology and Testing

J. Larry Jameson, Susan J. Mandel, Anthony P. Weetman

The thyroid gland produces two related hormones, thyroxine (T_4) and triiodothyronine (T_3) **(Fig. 375-1)**. Acting through thyroid hormone receptors α and β, these hormones play a critical role in cell differentiation and organogenesis during development and help maintain thermogenic and metabolic homeostasis in the adult. Autoimmune disorders of the thyroid gland can stimulate overproduction of thyroid hormones (*thyrotoxicosis*) or cause glandular destruction and hormone deficiency (*hypothyroidism*). Benign nodules and various forms of thyroid cancer are relatively common and amenable to detection by physical examination or various imaging techniques.

ANATOMY AND DEVELOPMENT

The thyroid (Greek *thyreos*, shield, plus *eidos*, form) consists of two lobes connected by an isthmus. It is located anterior to the trachea between the cricoid cartilage and the suprasternal notch. The normal thyroid is 12–20 g in size, highly vascular, and soft in consistency. Four parathyroid glands, which produce parathyroid hormone **(Chap. 403)**, are located posterior to each pole of the thyroid. The recurrent laryngeal nerves traverse the lateral borders of the thyroid gland and must be identified during thyroid surgery to avoid injury and vocal cord paralysis.

The thyroid gland develops from the floor of the primitive pharynx during the third week of gestation. The developing gland migrates along the thyroglossal duct to reach its final location in the neck. This feature accounts for the rare ectopic location of thyroid tissue at the base of the tongue (lingual thyroid) as well as the occurrence of thyroglossal duct cysts along this developmental tract. Thyroid hormone synthesis normally begins at about 11 weeks' gestation.

Neural crest derivatives from the ultimobranchial body give rise to thyroid medullary C cells that produce calcitonin, a calcium-lowering hormone. The C cells are interspersed throughout the thyroid gland, although their density is greatest in the juncture of the upper one-third and lower two-thirds of the gland. Calcitonin plays a minimal role in calcium homeostasis in humans but the C-cells are important because of their involvement in medullary thyroid cancer.

Thyroid gland development is orchestrated by the coordinated expression of several developmental transcription factors. Thyroid transcription factor (TTF)-1, TTF-2, NKX2-1, and paired homeobox-8 (PAX-8) are expressed selectively, but not exclusively, in the thyroid gland. In combination, they dictate thyroid cell development and the induction of thyroid-specific genes such as thyroglobulin (Tg), thyroid peroxidase (TPO), the sodium iodide symporter (Na^+/I^-, NIS), and the thyroid-stimulating hormone receptor (TSH-R). Mutations in these developmental transcription factors or their downstream target genes are rare causes of thyroid agenesis or dyshormonogenesis, although the causes of most forms of congenital hypothyroidism remain unknown **(see Chap. 376, Table 376-1)**. Because congenital hypothyroidism occurs in ~1 in 4000 newborns, neonatal screening is now performed in most industrialized countries. Transplacental passage of maternal thyroid hormone occurs before the fetal thyroid gland begins to function and provides significant hormone support to a fetus with congenital hypothyroidism. Early thyroid hormone replacement in newborns with congenital hypothyroidism prevents potentially severe developmental abnormalities.

The thyroid gland consists of numerous spherical follicles composed of thyroid follicular cells that surround secreted colloid, a proteinaceous fluid containing large amounts of thyroglobulin, the protein precursor of thyroid hormones **(Fig. 375-2)**. The thyroid follicular cells are polarized—the basolateral surface is apposed to the bloodstream and an apical surface faces the follicular lumen. Increased demand for thyroid hormone is regulated by TSH, which binds to its receptor on the basolateral surface of the follicular cells. This binding leads to Tg reabsorption from the follicular lumen and proteolysis within the cytoplasm, yielding thyroid hormones for secretion into the bloodstream.

REGULATION OF THE THYROID AXIS

TSH, secreted by the thyrotrope cells of the anterior pituitary, plays a pivotal role in control of the thyroid axis and serves as the most useful physiologic marker of thyroid hormone action. TSH is a 31-kDa hormone composed of α and β subunits; the α subunit is common to the other glycoprotein hormones (luteinizing hormone, follicle-stimulating hormone, human chorionic gonadotropin [hCG]), whereas the TSH β subunit is unique to TSH. The extent and nature of carbohydrate modification are modulated by thyrotropin-releasing hormone (TRH) stimulation and influence the biologic activity of the hormone.

The thyroid axis is a classic example of an endocrine feedback loop **(Chap. 370)**. Hypothalamic TRH stimulates pituitary production of TSH, which, in turn, stimulates thyroid hormone synthesis and secretion. Thyroid hormones act via negative feedback predominantly through thyroid hormone receptor β2 (TRβ2) to inhibit TRH and TSH production (Fig. 375-2). The "set-point" in this axis is established by

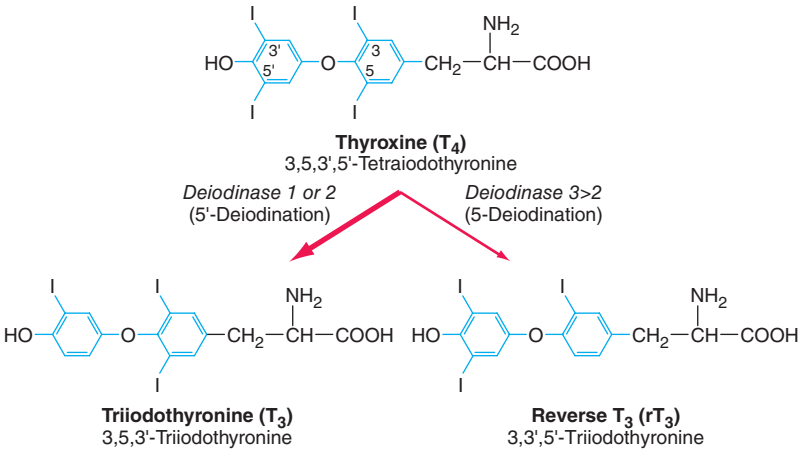


FIGURE 375-1 Structures of thyroid hormones. Thyroxine (T_4) contains four iodine atoms. Deiodination leads to production of the potent hormone triiodothyronine (T_3) or the inactive hormone reverse T_3.

TSH. TRH is the major positive regulator of TSH synthesis and secretion. Peak TSH secretion occurs ~15 min after administration of exogenous TRH. Dopamine, glucocorticoids, and somatostatin suppress TSH but are not of major physiologic importance except when these agents are administered in pharmacologic doses. Reduced levels of thyroid hormone increase basal TSH production and enhance TRH-mediated stimulation of TSH. High thyroid hormone levels rapidly and directly suppress TSH gene expression secretion and inhibit TRH stimulation of TSH, indicating that thyroid hormones are the dominant regulator of TSH production. Like other pituitary hormones, TSH is released in a pulsatile manner and exhibits a diurnal rhythm; its highest levels occur at night. However, these TSH excursions are modest in comparison to those of other pituitary hormones, in part, because TSH has a relatively long plasma half-life (50 min). Consequently, single measurements of TSH are adequate for assessing its circulating level. TSH is measured using immunoradiometric assays that are highly sensitive and specific. These assays readily distinguish between normal and suppressed TSH values; thus, TSH can be used for the diagnosis of primary hyperthyroidism (low TSH) or primary hypothyroidism (high TSH).

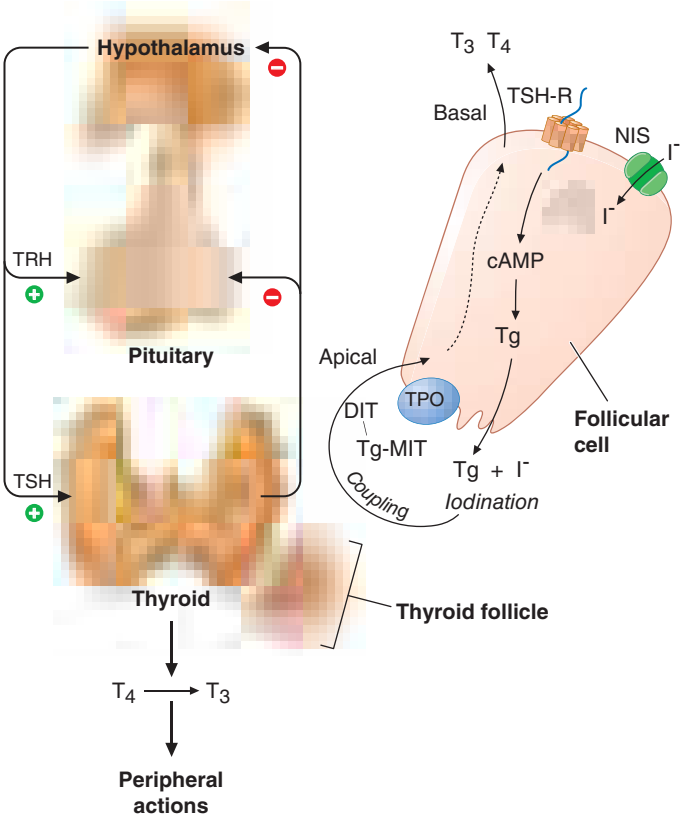


FIGURE 375-2 Regulation of thyroid hormone synthesis. ***Left.*** Thyroid hormones T_4 and T_3 feedback to inhibit hypothalamic production of thyrotropin-releasing hormone (TRH) and pituitary production of thyroid-stimulating hormone (TSH). TSH stimulates thyroid gland production of T_4 and T_3. ***Right.*** Thyroid follicles are formed by thyroid epithelial cells surrounding proteinaceous colloid, which contains thyroglobulin. Follicular cells, which are polarized, synthesize thyroglobulin and carry out thyroid hormone biosynthesis (see text for details). DIT, diiodotyrosine; MIT, monoiodotyrosine; NIS, sodium iodide symporter; Tg, thyroglobulin; TPO, thyroid peroxidase; TSH-R, thyroid-stimulating hormone receptor.

THYROID HORMONE SYNTHESIS, METABOLISM, AND ACTION

THYROID HORMONE SYNTHESIS

Thyroid hormones are derived from Tg, a large iodinated glycoprotein. After secretion into the thyroid follicle, Tg is iodinated on tyrosine residues that are subsequently coupled via an ether linkage. Reuptake of Tg into the thyroid follicular cell allows proteolysis and the release of newly synthesized T_4 and T_3.

Iodine Metabolism and Transport Iodide uptake is a critical first step in thyroid hormone synthesis. Ingested iodine is bound to serum proteins, particularly albumin. Unbound iodine is excreted in the urine. The thyroid gland extracts iodine from the circulation in a highly efficient manner. For example, 10–25% of radioactive tracer (e.g., ^{123}I) is taken up by the normal thyroid gland over 24 h in an iodine-replete state; this value can rise to 70–90% in Graves' disease. Iodide uptake is mediated by NIS, which is expressed at the basolateral membrane of thyroid follicular cells. NIS is most highly expressed in the thyroid gland, but low levels are present in the salivary glands, lactating breast, and placenta. The iodide transport mechanism is highly regulated, allowing adaptation to variations in dietary supply. Low iodine levels increase the amount of NIS and stimulate uptake, whereas high iodine levels suppress NIS expression and uptake. The selective expression of NIS in the thyroid allows isotopic scanning, treatment of hyperthyroidism, and ablation of thyroid cancer with radioisotopes of iodine, without significant effects on other organs. Mutation of the *NIS* gene is a rare cause of congenital hypothyroidism, underscoring its importance in thyroid hormone synthesis. Another iodine transporter, pendrin, is located on the apical surface of thyroid cells and mediates iodine efflux into the lumen. Mutation of the *pendrin* gene causes *Pendred syndrome*, a disorder characterized by defective organification of iodine, goiter, and sensorineural deafness.

Iodine deficiency is prevalent in many mountainous regions and in central Africa, central South America, and northern Asia (Fig. 375-3). Europe remains mildly iodine-deficient, and health surveys indicate that iodine intake has been falling in the United States and Australia. The World Health Organization (WHO) estimates that about 2 billion people are iodine-deficient, based on urinary excretion data. In areas of relative iodine deficiency, there is an increased prevalence of goiter and, when deficiency is severe, hypothyroidism and cretinism. *Cretinism* is characterized by mental and growth retardation and occurs when children who live in iodine-deficient regions are not treated with iodine or thyroid hormone to restore normal thyroid hormone levels during early life. These children are often born to mothers with iodine deficiency, and it is likely that maternal thyroid hormone deficiency worsens the condition. Concomitant selenium deficiency may also contribute to the neurologic manifestations of cretinism. Iodine supplementation of salt, bread, and other food substances has markedly reduced the prevalence of cretinism. Unfortunately, however, iodine deficiency remains the most common cause of preventable mental deficiency, often because of societal resistance to food additives or the cost of supplementation. In addition to overt cretinism, mild iodine deficiency can lead to subtle reduction of IQ. Oversupply of iodine, through supplements or foods enriched in iodine (e.g., shellfish, kelp), is associated with an increased incidence of

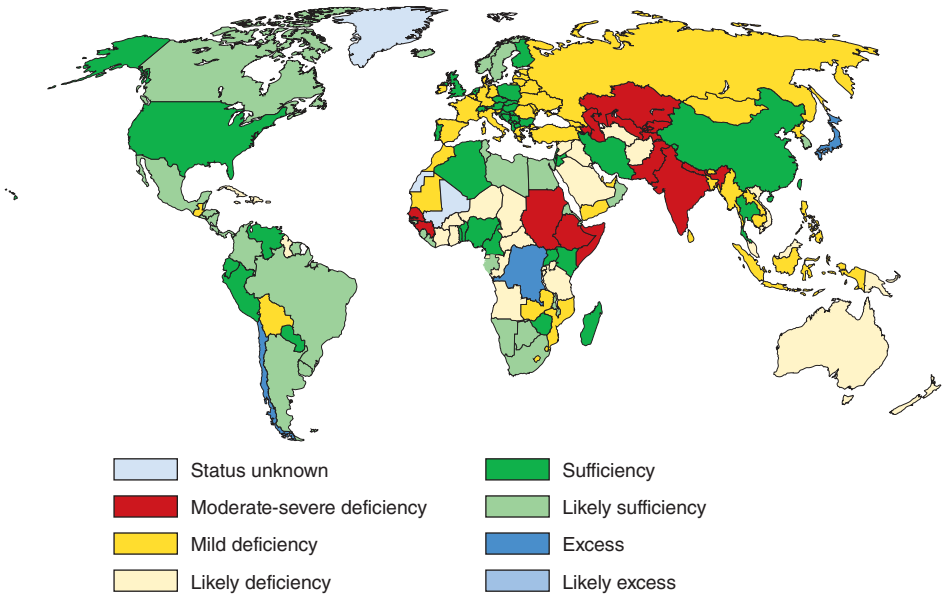


FIGURE 375-3 Worldwide iodine nutrition. Data are from the World Health Organization and the International Council for the Control of Iodine Deficiency Disorders *(http://indorgs.virginia.edu/iccidd/mi/cidds.html)*.

autoimmune thyroid disease. The RDA is 220 µg iodine per day for pregnant women and 290 µg iodine per day for breastfeeding women. Because the effects of iodine deficiency are most severe in pregnant women and their babies, the American Thyroid Association has recommended that all pregnant and breastfeeding women in the United States and Canada take a prenatal multivitamin containing 150 µg iodine per day. Urinary iodine is >10 µg/dL in iodine-sufficient populations.

Organification, Coupling, Storage, and Release After iodide enters the thyroid, it is trapped and transported to the apical membrane of thyroid follicular cells, where it is oxidized in an organification reaction that involves TPO and hydrogen peroxide produced by dual oxidase (DUOX) and DUOX maturation factor (DUOXA). The reactive iodine atom is added to specific tyrosyl residues within Tg, a large (660 kDa) dimeric protein that consists of 2769 amino acids. The iodotyrosines in Tg are then coupled via an ether linkage in a reaction that is also catalyzed by TPO. Either T_4 or T_3 can be produced by this reaction, depending on the number of iodine atoms present in the iodotyrosines. After coupling, Tg is taken back into the thyroid cell, where it is processed in lysosomes to release T_4 and T_3. Uncoupled mono- and diiodotyrosines (MIT, DIT) are deiodinated by the enzyme dehalogenase, thereby recycling any iodide that is not converted into thyroid hormones.

Disorders of thyroid hormone synthesis are rare causes of congenital hypothyroidism **(Chap. 376)**. The vast majority of these disorders are due to recessive mutations in TPO or Tg, but defects have also been identified in the TSH-R, NIS, pendrin, hydrogen peroxide generation, and dehalogenase, as well as genes involved in thyroid gland development. Because of the biosynthetic defect, the gland is incapable of synthesizing adequate amounts of hormone, leading to increased TSH and a large goiter.

TSH Action TSH regulates thyroid gland function through the TSH-R, a seven-transmembrane G protein–coupled receptor (GPCR). The TSH-R is coupled to the α subunit of stimulatory G protein ($G_{s\alpha}$), which activates adenylyl cyclase, leading to increased production of cyclic adenosine monophosphate (AMP). TSH also stimulates phosphatidylinositol turnover by activating phospholipase C. The functional role of the TSH-R is exemplified by the consequences of naturally occurring mutations. Recessive loss-of-function mutations cause thyroid hypoplasia and congenital hypothyroidism. Dominant gain-of-function mutations cause sporadic or familial hyperthyroidism that is characterized by goiter, thyroid cell hyperplasia, and autonomous function **(Chap. 377)**. Most of these activating mutations occur in the transmembrane domain of the receptor. They mimic the conformational changes induced by TSH binding or the interactions of thyroid-stimulating immunoglobulins (TSI) in Graves' disease. Activating TSH-R mutations also occur as somatic events, leading to clonal selection and expansion of the affected thyroid follicular cell and autonomously functioning thyroid nodules.

Other Factors That Influence Hormone Synthesis and Release Although TSH is the dominant hormonal regulator of thyroid gland growth and function, a variety of growth factors, most produced locally in the thyroid gland, also influence thyroid hormone synthesis. These include insulin-like growth factor I (IGF-I), epidermal growth factor, transforming growth factor β (TGF-β), endothelins, and various cytokines. The quantitative roles of these factors are not well understood, but they are important in selected disease states. In acromegaly, for example, increased levels of growth hormone and IGF-I are associated with goiter and predisposition to multinodular goiter (MNG). Certain cytokines and interleukins (ILs) produced in association with autoimmune thyroid disease induce thyroid growth, whereas others lead to apoptosis. Iodine deficiency increases thyroid blood flow and upregulates the NIS, stimulating more efficient iodine uptake. Excess iodide transiently inhibits thyroid iodide organification, a phenomenon known as the *Wolff-Chaikoff effect*. In individuals with a normal thyroid, the gland escapes from this inhibitory effect and iodide organification resumes; the suppressive action of high iodide may persist, however, in patients with underlying autoimmune thyroid disease.

THYROID FUNCTION IN PREGNANCY

Five factors alter thyroid function in pregnancy: (1) the transient increase in hCG during the first trimester, which weakly stimulates the TSH-R; (2) the estrogen-induced rise in TBG during the first trimester, which is sustained during pregnancy; (3) alterations in the immune system, leading to the onset, exacerbation, or amelioration of an underlying autoimmune thyroid disease; (4) increased thyroid hormone metabolism by the placenta; and (5) increased urinary iodide excretion, which can cause impaired thyroid hormone production in areas of marginal iodine sufficiency. Women with a precarious iodine intake (<50 µg/d) are most at risk of developing a goiter during pregnancy or giving birth to an infant with a goiter and hypothyroidism. The World Health Organization recommends a daily iodine intake of 250 µg during pregnancy and lactation and prenatal vitamins should contain 150 µg per tablet.

The rise in circulating hCG levels during the first trimester is accompanied by a reciprocal fall in TSH that persists into the middle of pregnancy. This reflects the weak binding of hCG, which is present at very high levels, to the TSH-R. Rare individuals have variant TSH-R sequences that enhance hCG binding and TSH-R activation. hCG-induced changes in thyroid function can result in transient gestational hyperthyroidism that may be associated with *hyperemesis gravidarum*, a condition characterized by severe nausea and vomiting and risk of volume depletion. However, since the hyperthyroidism is not causal, antithyroid drugs are not indicated unless concomitant Graves' disease is suspected. Parenteral fluid replacement usually suffices until the condition resolves.

Normative values for most thyroid function tests differ during pregnancy and, if available, trimester specific reference ranges should be used when diagnosing thyroid dysfunction during pregnancy. TSH levels decrease during the first trimester and then rise as gestation progresses. Total T_4 and T_3 levels are about 1.5× higher throughout pregnancy but the free T_4 progressively decreases so that third trimester values in healthy pregnancies are often below the nonpregnant lower reference cutoff.

During pregnancy, subclinical hypothyroidism occurs in 2% of women, but overt hypothyroidism is present in only 1 in 500. Prospective randomized controlled trials have not shown a benefit for

TABLE 375-1 Characteristics of Circulating T_4 and T_3

HORMONE PROPERTY	T_4	T_3
Serum concentrations		
Total hormone	8 μg/dL	0.14 μg/dL
Fraction of total hormone in the unbound form	0.02%	0.3%
Unbound (free) hormone	$21 \times 10^{-12} M$	$6 \times 10^{-12} M$
Serum half-life	7 d	2 d
Fraction directly from the thyroid	100%	20%
Production rate, including peripheral conversion	90 μg/d	32 μg/d
Intracellular hormone fraction	~20%	~70%
Relative metabolic potency	0.3	1
Receptor binding	$10^{-10} M$	$10^{-11} M$

universal thyroid disease screening in pregnancy. Targeted TSH testing for hypothyroidism is recommended for women planning a pregnancy if they have a strong family history of autoimmune thyroid disease, other autoimmune disorders (e.g., type 1 diabetes), infertility, prior preterm delivery or recurrent miscarriage, signs or symptoms of thyroid disease, or are older than 30 years. Thyroid hormone requirements are increased by up to 45% during pregnancy in levothyroxine-treated hypothyroid women.

THYROID HORMONE TRANSPORT AND METABOLISM

Serum-Binding Proteins T_4 is secreted from the thyroid gland in about twentyfold excess over T_3 (Table 375-1). Both hormones are bound to plasma proteins, including thyroxine-binding globulin (TBG), transthyretin (TTR, formerly known as thyroxine-binding prealbumin, or TBPA), and albumin. The plasma-binding proteins increase the pool of circulating hormone, delay hormone clearance, and may modulate hormone delivery to selected tissue sites. The concentration of TBG is relatively low (1–2 mg/dL), but because of its high affinity for thyroid hormones ($T_4 > T_3$), it carries about 80% of the bound hormones. Albumin has relatively low affinity for thyroid hormones but has a high plasma concentration (~3.5 g/dL), and it binds up to 10% of T_4 and 30% of T_3. TTR carries about 10% of T_4 but little T_3.

When the effects of the various binding proteins are combined, ~99.98% of T_4 and 99.7% of T_3 are protein-bound. Because T_3 is less tightly bound than T_4, the fraction of unbound T_3 is greater than unbound T_4, but there is less unbound T_3 in the circulation because it is produced in smaller amounts and cleared more rapidly than T_4. The unbound or "free" concentrations of the hormones are ~2×10^{-11} *M* for T_4 and ~6×10^{-12} *M* for T_3, which roughly correspond to the thyroid hormone receptor-binding constants for these hormones (see below). The unbound hormone is thought to be biologically available to tissues. Nonetheless, the homeostatic mechanisms that regulate the thyroid axis are directed toward maintenance of normal concentrations of unbound hormones.

Abnormalities of Thyroid Hormone-Binding Proteins A number of inherited and acquired abnormalities affect thyroid hormone-binding proteins. X-linked TBG deficiency is associated with very low levels of total T_4 and T_3. However, because unbound hormone levels are normal, patients are euthyroid and TSH levels are normal. It is important to recognize this disorder to avoid efforts to normalize total T_4 levels, because this leads to thyrotoxicosis and is futile because of rapid hormone clearance in the absence of TBG. TBG levels are elevated by estrogen, which increases sialylation and delays TBG clearance. Consequently, in women who are pregnant or taking estrogen-containing contraceptives, elevated TBG increases total T_4 and T_3 levels; however, unbound T_4 and T_3 levels are normal. These features are part of the explanation for why women with hypothyroidism require increased amounts of L-thyroxine replacement as TBG levels are increased by pregnancy or estrogen treatment. Mutations in TBG, TTR, and albumin may increase the binding affinity for T_4 and/or T_3 and cause disorders known as *euthyroid hyperthyroxinemia* or *familial dysalbuminemic hyperthyroxinemia* (FDH) (Table 375-2). These disorders result in increased total T_4 and/or T_3, but unbound hormone levels are normal. The familial nature of the disorders, and the fact that TSH levels are normal rather than suppressed, should suggest this diagnosis. Unbound hormone levels (ideally measured by dialysis) are normal in FDH. The diagnosis can be confirmed by using tests that measure the affinities of radiolabeled hormone binding to specific transport proteins or by performing DNA sequence analyses of the abnormal transport protein genes.

Certain medications, such as salicylates and salsalate, can displace thyroid hormones from circulating binding proteins. Although these drugs transiently perturb the thyroid axis by increasing free thyroid hormone levels, TSH is suppressed until a new steady state is reached, thereby restoring euthyroidism. Circulating factors associated with acute illness may also displace thyroid hormone from binding proteins (Chap. 377).

Deiodinases T_4 may be thought of as a precursor for the more potent T_3. T_4 is converted to T_3 by the deiodinase enzymes (Fig. 375-1). Type I deiodinase, which is located primarily in thyroid, liver, and kidneys, has a relatively low affinity for T_4. Type II deiodinase has a higher affinity for T_4 and is found primarily in the pituitary gland, brain, brown fat, and thyroid gland. Expression of type II deiodinase allows it to regulate T_3 concentrations locally, a property that may be important in the context of levothyroxine (T_4) replacement. Type II deiodinase is also regulated by thyroid hormone; hypothyroidism induces the enzyme, resulting in enhanced $T_4 \rightarrow T_3$ conversion in tissues such as brain and pituitary. $T_4 \rightarrow T_3$ conversion is impaired by fasting, systemic illness or acute trauma, oral contrast agents, and a

TABLE 375-2 Conditions Associated with Euthyroid Hyperthyroxinemia

DISORDER	CAUSE	TRANSMISSION	CHARACTERISTICS
Familial dysalbuminemic hyperthyroxinemia (FDH)	Albumin mutations, usually R218H	AD	Increased T_4 Normal unbound T_4 Rarely increased T_3
TBG			
Familial excess	Increased TBG production	XL	Increased total T_4, T_3 Normal unbound T_4, T_3
Acquired excess	Medications (estrogen), pregnancy, cirrhosis, hepatitis	Acquired	Increased total T_4, T_3 Normal unbound T_4, T_3
Transthyretin[a]			
Excess	Islet tumors	Acquired	Usually normal T_4, T_3
Mutations	Increased affinity for T_4 or T_3	AD	Increased total T_4, T_3 Normal unbound T_4, T_3
Medications: propranolol, ipodate, iopanoic acid, amiodarone	Decreased $T_4 \rightarrow T_3$ conversion	Acquired	Increased T_4 Decreased T_3 Normal or increased TSH
Resistance to thyroid hormone (RTH)	Thyroid hormone receptor β mutations	AD	Increased unbound T_4, T_3 Normal or increased TSH Some patients clinically thyrotoxic

[a]Also known as thyroxine-binding prealbumin (TBPA).

Abbreviations: AD, autosomal dominant; TBG, thyroxine-binding globulin; TSH, thyroid-stimulating hormone; XL, X-linked.

variety of medications (e.g., propylthiouracil, propranolol, amiodarone, glucocorticoids). Type III deiodinase inactivates T_4 and T_3 and is the most important source of reverse T_3 (rT_3), including in the sick euthyroid syndrome. This enzyme is expressed in the human placenta but is not active in healthy individuals. In the sick euthyroid syndrome, especially with hypoperfusion, the type III deiodinase is activated in muscle and liver. Massive hemangiomas that express type III deiodinase are a rare cause of consumptive hypothyroidism in infants.

THYROID HORMONE ACTION

Thyroid Hormone Transport Circulating thyroid hormones enter cells by passive diffusion and via specific transporters such as the monocarboxylate 8 transporter (MCT8), MCT10, and organic anion-transporting polypeptide 1C1. Mutations in the *MCT8* gene have been identified in patients with X-linked psychomotor retardation and thyroid function abnormalities (low T_4, high T_3, and high TSH). After entering cells, thyroid hormones act primarily through nuclear receptors, although they also have nongenomic actions through stimulating mitochondrial enzymatic responses and may act directly on blood vessels and the heart through integrin receptors.

Nuclear Thyroid Hormone Receptors Thyroid hormones bind with high affinity to nuclear *thyroid hormone receptors* (TRs) α and β. Both TRα and TRβ are expressed in most tissues, but their relative expression levels vary among organs; TRα is particularly abundant in brain, kidneys, gonads, muscle, and heart, whereas TRβ expression is relatively high in the pituitary and liver. Both receptors are variably spliced to form unique isoforms. The TRβ2 isoform, which has a unique amino terminus, is selectively expressed in the hypothalamus and pituitary, where it plays a role in feedback control of the thyroid axis (see above). The TRα2 isoform contains a unique carboxy terminus that precludes thyroid hormone binding; it may function to inhibit the action of other TR isoforms.

The TRs contain a central DNA-binding domain and a C-terminal ligand-binding domain. They bind to specific DNA sequences, termed *thyroid response elements* (TREs), in the promoter regions of target genes **(Fig. 375-4)**. The receptors bind as homodimers or, more commonly, as heterodimers with retinoic acid X receptors (RXRs) **(Chap. 370)**. The activated receptor can either stimulate gene transcription (e.g., myosin heavy chain α) or inhibit transcription (e.g., TSH β-subunit gene), depending on the nature of the regulatory elements in the target gene.

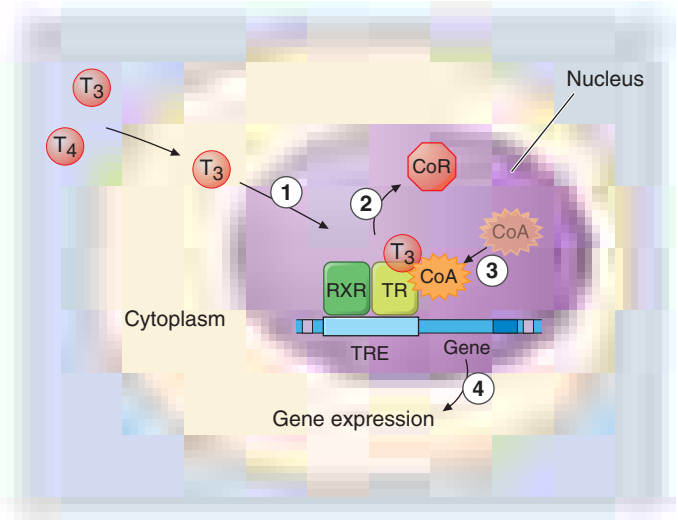


FIGURE 375-4 Mechanism of thyroid hormone receptor action. The thyroid hormone receptor (TR) and retinoid X receptor (RXR) form heterodimers that bind specifically to thyroid hormone response elements (TRE) in the promoter regions of target genes. In the absence of hormone, TR binds co-repressor (CoR) proteins that silence gene expression. The numbers refer to a series of ordered reactions that occur in response to thyroid hormone: (1) T_4 or T_3 enters the nucleus; (2) T_3 binding dissociates CoR from TR; (3) co-activators (CoA) are recruited to the T_3-bound receptor; and (4) gene expression is altered.

Thyroid hormones (T_3 and T_4) bind with similar affinities to TRα and TRβ. However, structural differences in the ligand-binding domains provide the potential for developing receptor-selective agonists or antagonists, and these are under investigation. T_3 is bound with 10–15 times greater affinity than T_4, which explains its increased hormonal potency. Although T_4 is produced in excess of T_3, receptors are occupied mainly by T_3, reflecting $T_4 \rightarrow T_3$ conversion by peripheral tissues, T_3 bioavailability in the plasma, and the greater affinity of receptors for T_3. After binding to TRs, thyroid hormone induces conformational changes in the receptors that modify its interactions with accessory transcription factors. Importantly, in the absence of thyroid hormone binding, the aporeceptors bind to co-repressor proteins that inhibit gene transcription. Hormone binding dissociates the co-repressors and allows the recruitment of co-activators that enhance transcription. The discovery of TR interactions with co-repressors explains the fact that TR silences gene expression in the absence of hormone binding. Consequently, hormone deficiency has a profound effect on gene expression because it causes gene repression as well as loss of hormone-induced stimulation. This concept has been corroborated by the finding that targeted deletion of the TR genes in mice has a less pronounced phenotypic effect than hormone deficiency.

Thyroid Hormone Resistance Resistance to thyroid hormone (RTH) is an autosomal dominant disorder characterized by elevated thyroid hormone levels and inappropriately normal or elevated TSH. Individuals with RTH do not, in general, exhibit signs and symptoms that are typical of hypothyroidism because hormone resistance is partial and is compensated by increased levels of thyroid hormone. The clinical features of RTH can include goiter, attention deficit disorder, mild reduction in IQ, delayed skeletal maturation, tachycardia, and impaired metabolic responses to thyroid hormone.

Classical forms of RTH are caused by mutations in the TRβ gene. These mutations, located in restricted regions of the ligand-binding domain, cause loss of receptor function. However, because the mutant receptors retain the capacity to dimerize with RXRs, bind to DNA, and recruit co-repressor proteins, they function as antagonists of the remaining normal TRβ and TRα receptors. This property, referred to as "dominant negative" activity, explains the autosomal dominant mode of transmission. The diagnosis is suspected when unbound thyroid hormone levels are increased without suppression of TSH. Similar hormonal abnormalities are found in other affected family members, although the TRβ mutation arises de novo in about 20% of patients. DNA sequence analysis of the TRβ gene provides a definitive diagnosis. RTH must be distinguished from other causes of euthyroid hyperthyroxinemia (e.g., FDH) and inappropriate secretion of TSH by TSH-secreting pituitary adenomas **(Chap. 373)**. In most patients, no treatment is indicated; the importance of making the diagnosis is to avoid inappropriate treatment of mistaken hyperthyroidism and to provide genetic counseling.

A distinct form of RTH is caused by mutations in the TRα gene. Affected patients have many clinical features of congenital hypothyroidism including growth retardation, skeletal dysplasia, and severe constipation. In contrast to RTH caused by mutations in TRβ, thyroid function tests include normal TSH, low or normal T_4, and normal or elevated T_3 levels. These distinct clinical and laboratory features underscore the different tissue distribution and functional roles of TRβ and TRα. Thyroxine treatment appears to alleviate some of the clinical manifestations of patients with RTH caused by TRα mutations.

PHYSICAL EXAMINATION

In addition to the examination of the thyroid itself, the physical examination should include a search for signs of abnormal thyroid function and the extrathyroidal features of ophthalmopathy and dermopathy **(Chap. 377)**. Examination of the neck begins by inspecting the seated patient from the front and side and noting any surgical scars, obvious masses, or distended veins. The thyroid can be palpated with both hands from behind or while facing the patient, using the thumbs to palpate each lobe. It is best to use a combination of these methods,

especially when nodules are small. The patient's neck should be slightly flexed to relax the neck muscles. After locating the cricoid cartilage, the isthmus, which is attached to the lower one-third of the thyroid lobes, can be identified and then followed laterally to locate either lobe (normally, the right lobe is slightly larger than the left). By asking the patient to swallow sips of water, thyroid consistency can be better appreciated as the gland moves beneath the examiner's fingers.

Features to be noted include thyroid size, consistency, nodularity, and any tenderness or fixation. An estimate of thyroid size (normally 12–20 g) should be made, and a drawing is often the best way to record findings. Ultrasound imaging provides the most accurate measurement of thyroid volume and nodularity and is useful for assessment of goiter prevalence in iodine deficient regions. However, ultrasound is not indicated if the thyroid physical examination is normal. The size, location, and consistency of any nodules should also be defined. A bruit or thrill over the gland, located over the insertion of the superior and inferior thyroid arteries (supero- or inferolaterally), indicates increased vascularity, associated with turbulent rather than laminar blood flow, as occurs in hyperthyroidism. If the lower borders of the thyroid lobes are not clearly felt, a goiter may be retrosternal. Large retrosternal goiters can cause venous distention over the neck and difficulty breathing, especially when the arms are raised (Pemberton's sign). With any central mass above the thyroid, the tongue should be extended, as thyroglossal cysts then move upward. The thyroid examination is not complete without assessment for lymphadenopathy in the supraclavicular and cervical regions of the neck.

LABORATORY EVALUATION

Measurement of Thyroid Hormones The enhanced sensitivity and specificity of *TSH assays* have greatly improved laboratory assessment of thyroid function. Because TSH levels change dynamically in response to alterations of T_4 and T_3, a logical approach to thyroid testing is to first determine whether TSH is suppressed, normal, or elevated. With rare exceptions (see below), a normal TSH level excludes a primary abnormality of thyroid function. This strategy depends on the use of immunochemiluminometric assays (ICMAs) for TSH that are sensitive enough to discriminate between the lower limit of the reference interval and the suppressed values that occur with thyrotoxicosis. Extremely sensitive assays can detect TSH levels ≤0.004 mIU/L, but, for practical purposes, assays sensitive to ≤0.1 mIU/L are sufficient. The widespread availability of the TSH ICMA has rendered the TRH stimulation test obsolete, because the failure of TSH to rise after an intravenous bolus of 200–400 μg TRH has the same implications as a suppressed basal TSH measured by ICMA.

The finding of an abnormal TSH level must be followed by measurements of circulating thyroid hormone levels to confirm the diagnosis of hyperthyroidism (suppressed TSH) or hypothyroidism (elevated TSH). Automated immunoassays are widely available for serum *total* T_4 and *total* T_3. T_4 and T_3 are highly protein-bound, and numerous factors (illness, medications, genetic factors) can influence protein binding. It is useful, therefore, to measure the free, or unbound, hormone levels, which correspond to the biologically available hormone pool. Two direct methods are used to measure *unbound thyroid hormones*: (1) unbound thyroid hormone competition with radiolabeled T_4 (or an analogue) for binding to a solid-phase antibody, and (2) physical separation of the unbound hormone fraction by ultracentrifugation or equilibrium dialysis. Although early unbound hormone immunoassays suffered from artifacts, newer assays correlate well with the results of the more technically demanding and expensive physical separation methods. An indirect method that is now less commonly used to estimate unbound thyroid hormone levels is to calculate the free T_3 or free T_4 index from the total T_4 or T_3 concentration and the *thyroid hormone binding ratio* (THBR). The latter is derived from the *T_3-resin uptake test*, which determines the distribution of radiolabeled T_3 between an absorbent resin and the unoccupied thyroid hormone binding proteins in the sample. The binding of the labeled T_3 to the resin is increased when there is reduced unoccupied protein binding sites (e.g., TBG deficiency) or increased total thyroid hormone in the sample; it is decreased under the opposite circumstances. The product of THBR and total T_3 or T_4 provides the *free T_3 or T_4 index*. In effect, the index corrects for anomalous total hormone values caused by variations in hormone-protein binding.

Total thyroid hormone levels are *elevated* when TBG is increased due to estrogens (pregnancy, oral contraceptives, hormone therapy, tamoxifen, selective estrogen receptor modulators, inflammatory liver disease) and *decreased* when TBG binding is reduced (androgens, nephrotic syndrome). Genetic disorders and acute illness can also cause abnormalities in thyroid hormone-binding proteins, and various drugs (phenytoin, carbamazepine, salicylates, and nonsteroidal anti-inflammatory drugs [NSAIDs]) can interfere with thyroid hormone binding. Because unbound thyroid hormone levels are normal and the patient is euthyroid in all of these circumstances, assays that measure unbound hormone are preferable to those for total thyroid hormones.

For most purposes, the unbound T_4 level is sufficient to confirm thyrotoxicosis, but 2–5% of patients have only an elevated T_3 level (T_3 toxicosis). Thus, unbound T_3 levels should be measured in patients with a suppressed TSH but normal unbound T_4 levels.

There are several clinical conditions in which the use of TSH as a screening test may be misleading, particularly without simultaneous unbound T_4 determinations. Any severe nonthyroidal illness can cause abnormal TSH levels. Although hypothyroidism is the most common cause of an elevated TSH level, rare causes include a TSH-secreting pituitary tumor **(Chap. 373)**, thyroid hormone resistance, and assay artifact. Conversely, a suppressed TSH level, particularly <0.01 mIU/L, usually indicates thyrotoxicosis. However, subnormal TSH levels between 0.01 and 0.1 mIU/L may be seen during the first trimester of pregnancy (due to hCG secretion), after treatment of hyperthyroidism (because TSH can remain suppressed for several months), and in response to certain medications (e.g., high doses of glucocorticoids or dopamine). TSH levels measured by immunoassay may also be suppressed in patients ingesting biotin supplements <18 hours prior to a blood draw because the TSH capture antibodies are biotinylated and the exogenous biotin can interfere with the subsequent streptavidin capture. Importantly, secondary hypothyroidism, caused by hypothalamic-pituitary disease, is associated with a variable (low to high-normal) TSH level, which is inappropriate for the low T_4 level. Thus, *TSH should not be used as an isolated laboratory test to assess thyroid function in patients with suspected or known pituitary disease*.

Tests for the end-organ effects of thyroid hormone excess or depletion, such as estimation of basal metabolic rate, tendon reflex relaxation rates, or serum cholesterol, are relatively insensitive and are not useful as clinical determinants of thyroid function.

Tests to Determine the Etiology of Thyroid Dysfunction Autoimmune thyroid disease is detected most easily by measuring circulating antibodies against TPO and Tg. Because antibodies to Tg alone are uncommon, it is reasonable to measure only TPO antibodies. About 5–15% of euthyroid women and up to 2% of euthyroid men have thyroid antibodies; such individuals are at increased risk of developing thyroid dysfunction. Almost all patients with autoimmune hypothyroidism, and up to 80% of those with Graves' disease, have TPO antibodies, usually at high levels.

TSIs are antibodies that stimulate the TSH-R in Graves' disease. They are most commonly measured by commercially available tracer displacement assays called TRAb (TSH receptor antibody) with the assumption that elevated levels in the setting of clinical hyperthyroidism reflect stimulatory effects on the TSH receptor. A bioassay is less commonly used. Remission rates in patients with Graves' disease after antithyroid drug cessation are higher with disappearance rather than persistence of TRAb. Furthermore, the TRAb assay is used to predict both fetal and neonatal thyrotoxicosis caused by transplacental passage of high maternal levels of TRAb or TSI (>3× upper limit of normal) in the last trimester of pregnancy.

Serum Tg levels are increased in all types of thyrotoxicosis except *thyrotoxicosis factitia* caused by self-administration of thyroid hormone. Tg levels are particularly increased in thyroiditis, reflecting thyroid tissue destruction and release of Tg. The main role for Tg measurement, however, is in the follow-up of thyroid cancer patients. After total

thyroidectomy and radioablation, Tg levels should be undetectable; in the absence of anti-Tg antibodies, measurable levels indicate incomplete ablation or recurrent cancer.

Radioiodine Uptake and Thyroid Scanning The thyroid gland selectively transports radioisotopes of iodine (^{123}I, ^{125}I, ^{131}I) and ^{99m}Tc pertechnetate, allowing thyroid imaging and quantitation of radioactive tracer fractional uptake.

Nuclear imaging of Graves' disease is characterized by an enlarged gland and increased tracer uptake that is distributed homogeneously. Toxic adenomas appear as focal areas of increased uptake, with suppressed tracer uptake in the remainder of the gland. In toxic MNG, the gland is enlarged—often with distorted architecture—and there are multiple areas of relatively increased (functioning nodules) or decreased tracer uptake (suppressed thyroid parenchyma or nonfunctioning nodules). Subacute, viral, and postpartum thyroiditis are associated with very low uptake because of follicular cell damage and TSH suppression. Thyrotoxicosis factitia is also associated with low uptake. In addition, if there is excessive circulating exogenous iodine (e.g., from dietary sources of iodinated contrast dye), the radionuclide uptake is low even in the presence of increased thyroid hormone production.

Thyroid scintigraphy is not used in the routine evaluation of patients with thyroid nodules, but should be performed if the serum TSH level is subnormal to determine if functioning thyroid nodules are present. Functioning or "hot" nodules are almost never malignant, and fine-needle aspiration (FNA) biopsy is not indicated. The vast majority of thyroid nodules do not produce thyroid hormone ("cold" nodules), and these are more likely to be malignant (~5–10%). Whole-body and thyroid scanning is also used in the treatment and surveillance of thyroid cancer. After thyroidectomy for thyroid cancer, the TSH level is raised by either using a thyroid hormone withdrawal protocol or recombinant human TSH injection **(Chap. 378)**. Administration of either ^{131}I or ^{123}I (in higher activities than used to image the thyroid gland alone) allows whole-body scanning (WBS) to confirm remnant ablation and to detect any functioning metastases. In addition, WBS may be helpful in surveillance of patients at risk for recurrence.

Thyroid Ultrasound Ultrasonography is valuable for the diagnosis and evaluation of patients with nodular thyroid disease **(Chap. 378)**. Evidence-based guidelines recommend thyroid ultrasonography for all patients suspected of having thyroid nodules by either physical examination or another imaging study. Using 10- to 12-MHz linear transducers, resolution and image quality are excellent, allowing the characterization of nodules and cysts >3 mm. Sonographic patterns that combine suspicious sonographic features are highly suggestive of malignancy (e.g., hypoechoic solid nodules with infiltrative borders and microcalcifications, >90% cancer risk), whereas other patterns correlate with a lower likelihood of cancer (isoechoic solid nodules, 5–10% cancer risk). Some patterns suggest benignity (e.g., spongiform nodules, defined as those with multiple small internal cystic areas, or simple cysts <3% cancer risk) **(see Chap. 378, Fig. 378-1)**. In addition to evaluating thyroid nodules, ultrasound is useful for monitoring nodule size and for the aspiration of nodules or cystic lesions. Ultrasound-guided FNA biopsy of thyroid lesions lowers the rate of inadequate sampling and decreases sample error, thereby reducing both the nondiagnostic and false-negative rates of FNA cytology. Ultrasonography of the central and lateral cervical lymph node compartments is indispensable in the evaluation thyroid cancer patients, preoperatively and during follow-up. In addition, the American College of Radiology recommends a survey of the cervical lymph nodes as part of every diagnostic thyroid sonographic examination.

FURTHER READING

Alexander EK et al: 2017 Guidelines of the American Thyroid Association for the diagnosis and management of thyroid disease during pregnancy and postpartum. Thyroid 27:315, 2017.

Ortiga-Cavalho TM, Sidhaye AR, Wondisford FE: Thyroid hormone receptors and resistance to thyroid hormone disorders. Nat Rev Endocrinol 10:582, 2014.

Rugge JB, Bougatsos C, Chou R: Screening and treatment of thyroid dysfunction: An evidence review for the U. S. Preventive Services Task Force. Ann Intern Med 162:35, 2015.

Stoupa A et al: Update of thyroid developmental genes. Endocrinol Metab Clin North Am 45:243, 2016.

Zimmermann MB, Boelaert K: Iodine deficiency and thyroid disorders. Lancet Diabetes Endocrinol 3: 286, 2015.

376 Hypothyroidism

J. Larry Jameson, Susan J. Mandel, Anthony P. Weetman

HYPOTHYROIDISM

Iodine deficiency remains a common cause of hypothyroidism worldwide. In areas of iodine sufficiency, autoimmune disease (Hashimoto's thyroiditis) and iatrogenic causes (treatment of hyperthyroidism) are most common **(Table 376-1)**.

CONGENITAL HYPOTHYROIDISM

Prevalence Hypothyroidism occurs in about 1 in 4000 newborns and neonatal screening is performed in most industrialized countries. It may be transient, especially if the mother has TSH-R blocking antibodies or has received antithyroid drugs, but permanent hypothyroidism occurs in the majority. Neonatal hypothyroidism is due to thyroid gland dysgenesis in 80–85%, to inborn errors of thyroid hormone synthesis in 10–15%, and is TSH-R antibody-mediated in 5% of affected newborns. The developmental abnormalities are twice as common in girls. Mutations that cause congenital hypothyroidism are being increasingly identified, but most remain idiopathic **(Table 376-2)**. Transplacental passage of maternal thyroid hormone occurs before the fetal thyroid gland begins to function and provides partial hormone support to a fetus with congenital hypothyroidism.

Clinical Manifestations The majority of infants appear normal at birth, and with the use of biochemical screening, few cases are now diagnosed based on clinical features, which include prolonged

TABLE 376-1 Causes of Hypothyroidism

Primary
Autoimmune hypothyroidism: Hashimoto's thyroiditis, atrophic thyroiditis
Iatrogenic: ^{131}I treatment, subtotal or total thyroidectomy, external irradiation of neck for lymphoma or cancer
Drugs: iodine excess (including iodine-containing contrast media and amiodarone), lithium, antithyroid drugs, *p*-aminosalicylic acid, interferon α and other cytokines, aminoglutethimide, tyrosine kinase inhibitors (e.g., sunitinib)
Congenital hypothyroidism: absent or ectopic thyroid gland, dyshormonogenesis, TSH-R mutation
Iodine deficiency
Infiltrative disorders: amyloidosis, sarcoidosis, hemochromatosis, scleroderma, cystinosis, Riedel's thyroiditis
Overexpression of type 3 deiodinase in infantile hemangioma and other tumors
Transient
Silent thyroiditis, including postpartum thyroiditis
Subacute thyroiditis
Withdrawal of supraphysiologic thyroxine treatment in individuals with an intact thyroid
After ^{131}I treatment or subtotal thyroidectomy for Graves' disease
Secondary
Hypopituitarism: tumors, pituitary surgery or irradiation, infiltrative disorders, Sheehan's syndrome, trauma, genetic forms of combined pituitary hormone deficiencies
Isolated TSH deficiency or inactivity
Bexarotene treatment
Hypothalamic disease: tumors, trauma, infiltrative disorders, idiopathic

Abbreviations: TSH, thyroid-stimulating hormone; TSH-R, TSH receptor.

TABLE 376-2 Genetic Causes of Congenital Hypothyroidism

DEFECTIVE GENE PROTEIN	INHERITANCE	CONSEQUENCES
PROP-1	Autosomal recessive	Combined pituitary hormone deficiencies with preservation of adrenocorticotropic hormone
PIT-1	Autosomal recessive Autosomal dominant	Combined deficiencies of growth hormone, prolactin, thyroid-stimulating hormone (TSH)
TSHβ	Autosomal recessive	TSH deficiency
TTF-1 (TITF-1)	Autosomal dominant	Variable thyroid hypoplasia, choreoathetosis, pulmonary problems
TTF-2 (FOXE-1)	Autosomal recessive	Thyroid agenesis, choanal atresia, spiky hair
PAX-8	Autosomal dominant	Thyroid dysgenesis, kidney abnormalities
NKX2-1	Autosomal dominant	Thyroid dysgenesis, brain, lung abnormalities
NKX2-5	Autosomal dominant	Thyroid dysgenesis, heart abnormalities
TSH-receptor	Autosomal recessive	Resistance to TSH
$G_{s\alpha}$ (Albright hereditary osteodystrophy)	Autosomal dominant	Resistance to TSH
Na^+/I^- symporter (SLC5A5)	Autosomal recessive	Inability to transport iodide
DUOX2 (THOX2)	Autosomal dominant	Organification defect
DUOXA2	Autosomal recessive	Organification defect
Thyroid peroxidase	Autosomal recessive	Defective organification of iodide
Thyroglobulin	Autosomal recessive	Defective synthesis of thyroid hormone
Pendrin (SLC26A4)	Autosomal recessive	Pendred syndrome: sensorineural deafness and partial organification defect in thyroid
Dehalogenase 1 (IYD)	Autosomal recessive	Loss of iodide reutilization

TABLE 376-3 Signs and Symptoms of Hypothyroidism (Descending Order of Frequency)

SYMPTOMS	SIGNS
Tiredness, weakness	Dry coarse skin; cool peripheral extremities
Dry skin	Puffy face, hands, and feet (myxedema)
Feeling cold	Diffuse alopecia
Hair loss	Bradycardia
Difficulty concentrating and poor memory	Peripheral edema
Constipation	Delayed tendon reflex relaxation
Weight gain with poor appetite	Carpal tunnel syndrome
Dyspnea	Serous cavity effusions
Hoarse voice	
Menorrhagia (later oligomenorrhea or amenorrhea)	
Paresthesia	
Impaired hearing	

jaundice, feeding problems, hypotonia, enlarged tongue, delayed bone maturation, and umbilical hernia. Importantly, permanent neurologic damage results if treatment is delayed. Typical features of adult hypothyroidism may also be present (Table 376-3). Other congenital malformations, especially cardiac, are four times more common in congenital hypothyroidism.

Diagnosis and Treatment Because of the severe neurologic consequences of untreated congenital hypothyroidism, neonatal screening programs have been established. These are generally based on measurement of TSH or T_4 levels in heel-prick blood specimens. When the diagnosis is confirmed, T_4 is instituted at a dose of 10–15 μg/kg per day, and the dose is adjusted by close monitoring of TSH levels. T_4 requirements are relatively great during the first year of life, and a high circulating T_4 level is usually needed to normalize TSH. Early treatment with T_4 results in normal IQ levels, but subtle neurodevelopmental abnormalities may occur in those with the most severe hypothyroidism at diagnosis or when treatment is delayed or suboptimal. If transient hypothyroidism is suspected, or the diagnosis is unclear, treatment can be stopped safely after the age of 3 years followed by further evaluation.

AUTOIMMUNE HYPOTHYROIDISM

Classification Autoimmune hypothyroidism may be associated with a goiter (Hashimoto's, or *goitrous thyroiditis*) or, at the later stages of the disease, minimal residual thyroid tissue (*atrophic thyroiditis*). Because the autoimmune process gradually reduces thyroid function, there is a phase of compensation when normal thyroid hormone levels are maintained by a rise in TSH. Although some patients may have minor symptoms, this state is called *subclinical hypothyroidism*. Later, unbound T_4 levels fall and TSH levels rise further; symptoms become more readily apparent at this stage (usually TSH >10 mIU/L), which is referred to as *clinical hypothyroidism* or *overt hypothyroidism*.

Prevalence The mean annual incidence rate of autoimmune hypothyroidism is up to 4 per 1000 women and 1 per 1000 men. It is more common in certain populations, such as the Japanese, probably because of genetic factors and chronic exposure to a high-iodine diet. The mean age at diagnosis is 60 years, and the prevalence of overt hypothyroidism increases with age. Subclinical hypothyroidism is found in 6–8% of women (10% over the age of 60) and 3% of men. The annual risk of developing clinical hypothyroidism is about 4% when subclinical hypothyroidism is associated with positive thyroid peroxidase (TPO) antibodies.

Pathogenesis In Hashimoto's thyroiditis, there is a marked lymphocytic infiltration of the thyroid with germinal center formation, atrophy of the thyroid follicles accompanied by oxyphil metaplasia, absence of colloid, and mild to moderate fibrosis. In atrophic thyroiditis, the fibrosis is much more extensive, lymphocyte infiltration is less pronounced, and thyroid follicles are almost completely absent. Atrophic thyroiditis usually represents the end stage of Hashimoto's thyroiditis rather than a separate disorder, although a distinct form of marked fibrosis occurs in which the gland is infiltrated with IgG4-positive plasma cells.

As with most autoimmune disorders, susceptibility to autoimmune hypothyroidism is determined by a combination of genetic and environmental factors, and the risk of either autoimmune hypothyroidism or Graves' disease is increased among siblings. HLA-DR polymorphisms are the best documented genetic risk factors for autoimmune hypothyroidism, especially HLA-DR3, DR4, and DR5 in Caucasians. A weak association also exists between polymorphisms in *CTLA-4*, a T cell–regulatory gene, and autoimmune hypothyroidism. Both of these genetic associations are shared by other autoimmune diseases, which may explain the relationship between autoimmune hypothyroidism and other autoimmune diseases, especially type 1 diabetes mellitus, Addison's disease, pernicious anemia, and vitiligo. HLA-DR and *CTLA-4* polymorphisms account for approximately half of the genetic susceptibility to autoimmune hypothyroidism and the role of other contributory loci remains to be clarified. A gene on chromosome 21 may be responsible for the association between autoimmune hypothyroidism and Down's syndrome. The female preponderance of thyroid autoimmunity is most likely due to sex steroid effects on the immune response, but an X chromosome–related genetic factor is also possible and may account for the high frequency of autoimmune

hypothyroidism in Turner's syndrome. Environmental susceptibility factors are poorly defined at present. A high iodine or low selenium intake and decreased exposure to microorganisms in childhood increase the risk of autoimmune hypothyroidism. Smoking cessation transiently increases incidence whereas alcohol intake seems protective. These factors may account for the increase in prevalence over the last two to three decades.

The thyroid lymphocytic infiltrate in autoimmune hypothyroidism is composed of activated T cells as well as B cells. Thyroid cell destruction is primarily mediated by the CD8+ cytotoxic T cells but local production of cytokines, such as tumor necrosis factor (TNF), interleukin-1 (IL-1), and interferon γ (IFN-γ), derived from the inflammatory infiltrate may render thyroid cells more susceptible to apoptosis mediated by death receptors, such as Fas, and by oxidative stress. These cytokines also impair thyroid cell function directly and induce the expression of other proinflammatory molecules by the thyroid cells themselves, such as cytokines, HLA class I and class II molecules, adhesion molecules, CD40, and nitric oxide. Administration of high concentrations of cytokines for therapeutic purposes (especially IFN-α) is associated with increased autoimmune thyroid disease, possibly through mechanisms similar to those in sporadic disease. Novel anticancer and immunomodulatory treatments, such as tyrosine kinase inhibitors and alemtuzumab, can also induce thyroid autoimmunity via their effects on T cell regulation.

Antibodies to TPO and thyroglobulin (Tg) are clinically useful markers of thyroid autoimmunity, but any pathogenic effect is restricted to a secondary role in amplifying an ongoing autoimmune response. TPO antibodies fix complement, and complement membrane-attack complexes are present in the thyroid in autoimmune hypothyroidism. However, transplacental passage of Tg or TPO antibodies has no effect on the fetal thyroid, which suggests that T cell–mediated injury is required to initiate autoimmune damage to the thyroid.

Up to 20% of patients with autoimmune hypothyroidism have antibodies against the TSH-R, which, in contrast to thyroid-stimulating immunoglobulin (TSI), do not stimulate the receptor but prevent the binding of TSH. These TSH-R-blocking antibodies, therefore, cause hypothyroidism and, especially in Asian patients, thyroid atrophy. Their transplacental passage may induce transient neonatal hypothyroidism. Rarely, patients have a mixture of TSI and TSH-R-blocking antibodies, and thyroid function can oscillate between hyperthyroidism and hypothyroidism as one or the other antibody becomes dominant. Predicting the course of disease in such individuals is difficult, and they require close monitoring of thyroid function. Bioassays can be used to document that TSH-R-blocking antibodies reduce the cyclic AMP–inducing effect of TSH on cultured TSH-R-expressing cells, but these assays are difficult to perform. Thyrotropin-binding inhibitory immunoglobulin (TBII) assays that measure the binding of antibodies to the receptor by competition with labeled TSH do not distinguish between TSI and TSH-R-blocking antibodies, but a positive result in a patient with spontaneous hypothyroidism is strong evidence for the presence of blocking antibodies. The use of these assays does not generally alter clinical management, although it may be useful to confirm the cause of transient neonatal hypothyroidism.

Clinical Manifestations The main clinical features of hypothyroidism are summarized in Table 376-3. The onset is usually insidious, and the patient may become aware of symptoms only when euthyroidism is restored. Patients with Hashimoto's thyroiditis may present because of goiter rather than symptoms of hypothyroidism. The goiter may not be large, but it is usually irregular and firm in consistency. Rarely uncomplicated Hashimoto's thyroiditis is associated with pain.

Patients with atrophic thyroiditis or the later stage of Hashimoto's thyroiditis present with symptoms and signs of hypothyroidism. The skin is dry, and there is decreased sweating, thinning of the epidermis, and hyperkeratosis of the stratum corneum. Increased dermal glycosaminoglycan content traps water, giving rise to skin thickening without pitting (*myxedema*). Typical features include a puffy face with edematous eyelids and nonpitting pretibial edema **(Fig. 376-1)**. There is pallor, often with a yellow tinge to the skin due to carotene

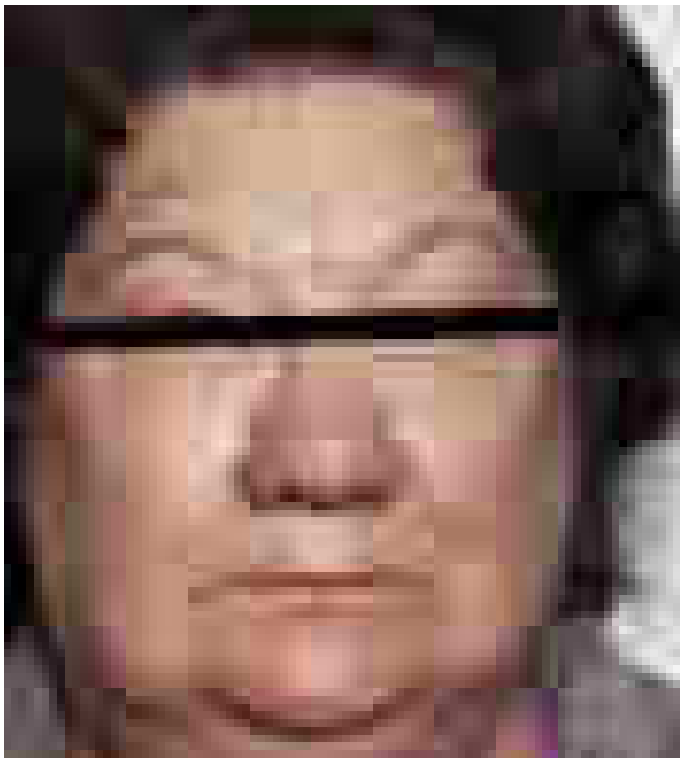


FIGURE 376-1 Facial appearance in hypothyroidism. Note puffy eyes and thickened skin.

accumulation. Nail growth is retarded, and hair is dry, brittle, difficult to manage, and falls out easily. In addition to diffuse alopecia, there is thinning of the outer third of the eyebrows, although this is not a specific sign of hypothyroidism.

Other common features include constipation and weight gain (despite a poor appetite). In contrast to popular perception, the weight gain is usually modest and due mainly to fluid retention in the myxedematous tissues. Libido is decreased in both sexes, and there may be oligomenorrhea or amenorrhea in long-standing disease, but menorrhagia may occur at an early stage. Fertility is reduced, and the incidence of miscarriage is increased. Prolactin levels are often modestly increased **(Chap. 373)** and may contribute to alterations in libido and fertility and cause galactorrhea.

Myocardial contractility and pulse rate are reduced, leading to a reduced stroke volume and bradycardia. Increased peripheral resistance may be accompanied by hypertension, particularly diastolic. Blood flow is diverted from the skin, producing cool extremities. Pericardial effusions occur in up to 30% of patients but rarely compromise cardiac function. Although alterations in myosin heavy chain isoform expression have been documented, cardiomyopathy is rare. Fluid may also accumulate in other serous cavities and in the middle ear, giving rise to conductive deafness. Pulmonary function is generally normal, but dyspnea may be caused by pleural effusion, impaired respiratory muscle function, diminished ventilatory drive, or sleep apnea.

Carpal tunnel and other entrapment syndromes are common, as is impairment of muscle function with stiffness, cramps, and pain. On examination, there may be slow relaxation of tendon reflexes and pseudomyotonia. Memory and concentration are impaired. Experimentally, positron emission tomography (PET) scans examining glucose metabolism in hypothyroid subjects show lower regional activity in the amygdala, hippocampus, and perigenual anterior cingulated cortex, among other regions, and this activity corrects after thyroxine replacement. Rare neurologic problems include reversible cerebellar ataxia, dementia, psychosis, and myxedema coma. *Hashimoto's encephalopathy* has been defined as a steroid-responsive syndrome associated with TPO antibodies, myoclonus, and slow-wave activity on electroencephalography, but the relationship with thyroid autoimmunity or hypothyroidism is not established. The hoarse voice and occasionally

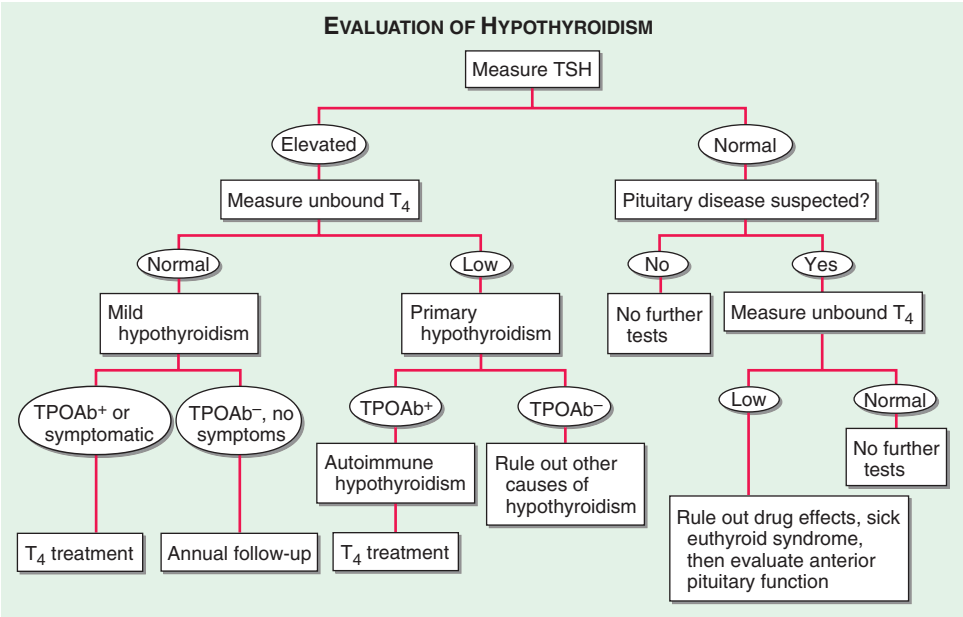


FIGURE 376-2 Evaluation of hypothyroidism. $TPOAb^+$, thyroid peroxidase antibodies present; $TPOAb^-$, thyroid peroxidase antibodies not present; TSH, thyroid-stimulating hormone.

clumsy speech of hypothyroidism reflect fluid accumulation in the vocal cords and tongue.

The features described above are the consequence of thyroid hormone deficiency. However, autoimmune hypothyroidism may be associated with signs or symptoms of other autoimmune diseases, particularly vitiligo, pernicious anemia, Addison's disease, alopecia areata, and type 1 diabetes mellitus. Less common associations include celiac disease, dermatitis herpetiformis, chronic active hepatitis, rheumatoid arthritis, systemic lupus erythematosus (SLE), myasthenia gravis, and Sjögren's syndrome. Thyroid-associated ophthalmopathy usually occurs in Graves' disease (see below), but in about 5% of patients it is associated with autoimmune hypothyroidism.

Autoimmune hypothyroidism is uncommon in children and usually presents with slow growth and delayed facial and dental maturation. The pituitary may be enlarged due to thyrotroph hyperplasia. Myopathy, with muscle swelling, is more common in children than in adults. In most cases, puberty is delayed, but precocious puberty sometimes occurs. There may be intellectual impairment if the onset is before 3 years and the hormone deficiency is severe.

Laboratory Evaluation A summary of the investigations used to determine the existence and cause of hypothyroidism is provided in Fig. 376-2. A normal TSH level excludes primary (but not secondary) hypothyroidism. If the TSH is elevated, an unbound T_4 level is needed to confirm the presence of clinical hypothyroidism, but T_4 is inferior to TSH when used as a screening test, because it will not detect subclinical hypothyroidism. Circulating unbound T_3 levels are normal in about 25% of patients, reflecting adaptive deiodinase responses to hypothyroidism. T_3 measurements are, therefore, not indicated.

Once clinical or subclinical hypothyroidism is confirmed, the etiology is usually easily established by demonstrating the presence of TPO and Tg antibodies, which are present in >95% of patients with autoimmune hypothyroidism. TBII can be found in 10–20% of patients, but measurement is not needed routinely. Other abnormal laboratory findings in hypothyroidism may include increased creatine phosphokinase, elevated cholesterol and triglycerides, and anemia (usually normocytic or macrocytic). Except when accompanied by iron deficiency, the anemia and other abnormalities gradually resolve with thyroxine replacement.

Differential Diagnosis An asymmetric goiter in Hashimoto's thyroiditis may be confused with a multinodular goiter (MNG) or thyroid carcinoma, in which thyroid antibodies may also be present. Ultrasound can be used to show the presence of a solitary lesion or an MNG rather than the heterogeneous thyroid enlargement typical of Hashimoto's thyroiditis. FNA biopsy is useful in the investigation of focal nodules. Other causes of hypothyroidism are discussed below and in Table 376-1 but rarely cause diagnostic confusion.

OTHER CAUSES OF HYPOTHYROIDISM

Iatrogenic hypothyroidism is a common cause of hypothyroidism and can often be detected by screening before symptoms develop. In the first 3–4 months after radioiodine treatment for Graves' disease, transient hypothyroidism may occur due to reversible radiation damage. Low-dose thyroxine treatment can be withdrawn if recovery occurs. Because TSH levels are suppressed by hyperthyroidism, unbound T_4 levels are a better measure of thyroid function than TSH in the months following radioiodine treatment. Mild hypothyroidism after subtotal thyroidectomy may also resolve after several months, as the gland remnant is stimulated by increased TSH levels.

Iodine deficiency is responsible for endemic goiter and cretinism but is an uncommon cause of adult hypothyroidism unless the iodine intake is very low or there are complicating factors, such as the consumption of thiocyanates in cassava or selenium deficiency. Although hypothyroidism due to iodine deficiency can be treated with thyroxine, public health measures to improve iodine intake should be advocated to eliminate this problem. Iodized salt or bread or a single bolus of oral or intramuscular iodized oil have all been used successfully.

Paradoxically, chronic iodine excess can also induce goiter and hypothyroidism. The intracellular events that account for this effect are unclear, but individuals with autoimmune thyroiditis are especially susceptible. Iodine excess is responsible for the hypothyroidism that occurs in up to 13% of patients treated with amiodarone (see below). Other drugs, particularly lithium, may also cause hypothyroidism. Transient hypothyroidism caused by thyroiditis is discussed below.

Secondary hypothyroidism is usually diagnosed in the context of other anterior pituitary hormone deficiencies; isolated TSH deficiency is very rare (Chap. 372). TSH levels may be low, normal, or even slightly increased in secondary hypothyroidism; the latter is due to secretion of immunoactive but bioinactive forms of TSH. The diagnosis is confirmed by detecting a low unbound T_4 level. The goal of treatment is to maintain T_4 levels in the upper half of the reference interval, because TSH levels cannot be used to monitor therapy.

TREATMENT

Hypothyroidism

CLINICAL HYPOTHYROIDISM

If there is no residual thyroid function, the daily replacement dose of levothyroxine is usually 1.6 μg/kg body weight (typically 100–150 μg), ideally taken at least 30 min before breakfast. In many patients, however, lower doses suffice until residual thyroid tissue is destroyed. In patients who develop hypothyroidism after the treatment of Graves' disease, there is often underlying autonomous function, necessitating lower replacement doses (typically 75–125 μg/d).

Adult patients under 60 years old without evidence of heart disease may be started on 50–100 μg levothyroxine (T_4) daily. The dose is adjusted on the basis of TSH levels, with the goal of treatment being a normal TSH, ideally in the lower half of the reference range. TSH responses are gradual and should be measured about

2 months after instituting treatment or after any subsequent change in levothyroxine dosage. The clinical effects of levothyroxine replacement are slow to appear. Patients may not experience full relief from symptoms until 3–6 months after normal TSH levels are restored. Adjustment of levothyroxine dosage is made in 12.5- or 25-μg increments if the TSH is high; decrements of the same magnitude should be made if the TSH is suppressed. Patients with a suppressed TSH of any cause, including T_4 overtreatment, have an increased risk of atrial fibrillation and reduced bone density.

Although desiccated animal thyroid preparations (thyroid extract USP) are available, they are not recommended because the ratio of T_3 to T_4 is nonphysiologic. The use of levothyroxine combined with liothyronine (triiodothyronine, T_3) has been investigated, but benefit has not been confirmed in prospective studies. There is no place for liothyronine alone as long-term replacement, because the short half-life necessitates three or four daily doses and is associated with fluctuating T_3 levels.

Once full replacement is achieved and TSH levels are stable, follow-up measurement of TSH is recommended at annual intervals. It is important to ensure ongoing adherence as patients do not feel any symptomatic difference after missing a few doses of levothyroxine, and this sometimes leads to self-discontinuation.

In patients of normal body weight who are taking ≥200 μg of levothyroxine per day, an elevated TSH level is often a sign of poor adherence to treatment. This is also the likely explanation for fluctuating TSH levels, despite a constant levothyroxine dosage. Such patients often have normal or high unbound T_4 levels, despite an elevated TSH, because they remember to take medication for a few days before testing; this is sufficient to normalize T_4, but not TSH levels. It is important to consider variable adherence, because this pattern of thyroid function tests is otherwise suggestive of disorders associated with inappropriate TSH secretion **(Chap. 375)**. Because T_4 has a long half-life (7 days), patients who miss a dose can be advised to take two doses of the skipped tablets at once. Other causes of increased levothyroxine requirements must be excluded, particularly malabsorption (e.g., celiac disease, small-bowel surgery, atrophic or *Helicobacter pylori*–related gastritis,), oral estrogen containing medications or selective estrogen receptor modulator therapy, ingestion with a meal, and drugs that interfere with T_4 absorption or metabolism such as bile acid sequestrants, ferrous sulfate, calcium supplements, selevamer, sucralfate, proton pump inhibitors, lovastatin, aluminum hydroxide, rifampicin, amiodarone, carbamazepine, phenytoin, and tyrosine kinase inhibitors.

SUBCLINICAL HYPOTHYROIDISM

By definition, subclinical hypothyroidism refers to biochemical evidence of thyroid hormone deficiency in patients who have few or no apparent clinical features of hypothyroidism. There are no universally accepted recommendations for the management of subclinical hypothyroidism, but levothyroxine is recommended if the patient is a woman who wishes to conceive or is pregnant, or when TSH levels are above 10 mIU/L. Otherwise, when TSH levels are below 10 mIU/L, a trial of treatment may be considered when patients have suggestive symptoms of hypothyroidism, positive TPO antibodies, or any evidence of heart disease. It is important to confirm that any elevation of TSH is sustained over a 3-month period before treatment is given. Treatment is administered by starting with a low dose of levothyroxine (25–50 μg/d) with the goal of normalizing TSH. If levothyroxine is not given, thyroid function should be evaluated annually.

SPECIAL TREATMENT CONSIDERATIONS

Rarely, levothyroxine replacement is associated with pseudotumor cerebri in children. Presentation appears to be idiosyncratic and occurs months after treatment has begun.

Because maternal hypothyroidism may both adversely affect fetal neural development and be associated with adverse gestational outcomes (miscarriage, preterm delivery), thyroid function should be monitored to preserve euthyroidism in women with a history or high risk of hypothyroidism. The presence of thyroid autoantibodies alone, in a euthyroid patient, is also associated with miscarriage and preterm delivery; large-scale trials are underway to establish whether levothyroxine therapy improves outcomes in this group. Prior to conception, levothyroxine therapy should be targeted to maintain a serum TSH in the normal range but <2.5 mIU/L for hypothyroid women. Subsequently, thyroid function should be evaluated immediately after pregnancy is confirmed and every 4 weeks during the first half of the pregnancy, with less frequent testing after 20 weeks' gestation (every 6–8 weeks depending on whether levothyroxine dose adjustment is ongoing). The levothyroxine dose may need to be increased by up to 45% during pregnancy. Women should increase levothyroxine from once daily dosing to nine doses per week as soon as pregnancy is confirmed, to anticipate this change. Thereafter dosage should be closely monitored with a goal TSH in the lower half of the trimester-specific normative range, if available, or <2.5 mIU/L. After delivery, levothyroxine doses typically return to prepregnancy levels. Pregnant women should be counseled to separate ingestion of prenatal vitamins and iron supplements from levothyroxine.

Elderly patients may require 20% less thyroxine than younger patients. In the elderly, especially patients with known coronary artery disease, the starting dose of levothyroxine is 12.5–25 μg/d with similar increments every 2–3 months until TSH is normalized. In some patients, it may be impossible to achieve full replacement despite optimal antianginal treatment. *Emergency surgery* is generally safe in patients with untreated hypothyroidism, although routine surgery in a hypothyroid patient should be deferred until euthyroidism is achieved.

Myxedema coma still has a 20–40% mortality rate, despite intensive treatment, and outcomes are independent of the T_4 and TSH levels. Clinical manifestations include reduced level of consciousness, sometimes associated with seizures, as well as the other features of hypothyroidism (Table 376-3). Hypothermia can reach 23°C (74°F). There may be a history of treated hypothyroidism with poor compliance, or the patient may be previously undiagnosed. Myxedema coma almost always occurs in the elderly and is usually precipitated by factors that impair respiration, such as drugs (especially sedatives, anesthetics, and antidepressants), pneumonia, congestive heart failure, myocardial infarction, gastrointestinal bleeding, or cerebrovascular accidents. Sepsis should also be suspected. Exposure to cold may also be a risk factor. Hypoventilation, leading to hypoxia and hypercapnia, plays a major role in pathogenesis; hypoglycemia and dilutional hyponatremia also contribute to the development of myxedema coma.

Levothyroxine can initially be administered as a single IV bolus of 200–400 μg, which serves as a loading dose, followed by a daily oral dose of 1.6 μg/kg/d, reduced by 25% if administered IV. If suitable IV preparation is not available, the same initial dose of levothyroxine can be given by nasogastric tube (although absorption may be impaired in myxedema). Because $T_4 \rightarrow T_3$ conversion is impaired in myxedema coma, there is a rationale for adding liothyronine (T_3) intravenously or via nasogastric tube to levothyroxine treatment, although excess liothyronine has the potential to provoke arrhythmias. An initial loading dose of 5–20 μg liothyronine should be followed by 2.5–10 μg 8 hourly, with lower doses for those at cardiovascular risk.

Supportive therapy should be provided to correct any associated metabolic disturbances. External warming is indicated only if the temperature is <30°C, as it can result in cardiovascular collapse **(Chap. 454)**. Space blankets should be used to prevent further heat loss. Parenteral hydrocortisone (50 mg every 6 h) should be administered, because there is impaired adrenal reserve in profound hypothyroidism. Any precipitating factors should be treated, including the early use of broad-spectrum antibiotics, pending the exclusion of infection. Ventilatory support with regular blood gas analysis is usually needed during the first 48 h. Hypertonic saline or IV glucose

may be needed if there is severe hyponatremia or hypoglycemia; hypotonic IV fluids should be avoided because they may exacerbate water retention secondary to reduced renal perfusion and inappropriate vasopressin secretion. The metabolism of most medications is impaired, and sedatives should be avoided if possible or used in reduced doses. Medication blood levels should be monitored, when available, to guide dosage.

FURTHER READING

Hanley P et al: Thyroid disorders in children and adolescents: A review. JAMA Pediatr 170:1008, 2016.

Jonklaas J et al: Guidelines for the treatment of hypothyroidism: Prepared by the American Thyroid Association Task Force on thyroid hormone replacement. Thyroid 24:1670, 2014.

Rugge JB, Bougatsos C, Chou R: Screening and treatment of thyroid dysfunction: An evidence review for the U. S. Preventive Services Task Force. Ann Intern Med 162:35, 2015.

377 Hyperthyroidism

J. Larry Jameson, Susan J. Mandel, Anthony P. Weetman

THYROTOXICOSIS

Thyrotoxicosis is defined as the state of thyroid hormone excess and is not synonymous with *hyperthyroidism*, which is the result of excessive thyroid function. However, the major etiologies of thyrotoxicosis are hyperthyroidism caused by Graves' disease, toxic multinodular goiter (MNG), and toxic adenomas. Other causes are listed in Table 377-1.

GRAVES' DISEASE

Epidemiology Graves' disease accounts for 60–80% of thyrotoxicosis. The prevalence varies among populations, reflecting genetic factors and iodine intake (high iodine intake is associated with an increased prevalence of Graves' disease). Graves' disease occurs in up to 2% of women but is one-tenth as frequent in men. The disorder rarely begins before adolescence and typically occurs between 20 and 50 years of age; it also occurs in the elderly.

TABLE 377-1 Causes of Thyrotoxicosis

Primary Hyperthyroidism
Graves' disease
Toxic multinodular goiter
Toxic adenoma
Functioning thyroid carcinoma metastases
Activating mutation of the TSH receptor
Activating mutation of $G_{s\alpha}$ (McCune-Albright syndrome)
Struma ovarii
Drugs: iodine excess (Jod-Basedow phenomenon)
Thyrotoxicosis without Hyperthyroidism
Subacute thyroiditis
Silent thyroiditis
Other causes of thyroid destruction: amiodarone, radiation, infarction of adenoma
Ingestion of excess thyroid hormone (thyrotoxicosis factitia) or thyroid tissue
Secondary Hyperthyroidism
TSH-secreting pituitary adenoma
Thyroid hormone resistance syndrome: occasional patients may have features of thyrotoxicosis
Chorionic gonadotropin-secreting tumors[a]
Gestational thyrotoxicosis[a]

[a]Circulating TSH levels are low in these forms of secondary hyperthyroidism.

Abbreviation: TSH, thyroid-stimulating hormone.

Pathogenesis As in autoimmune hypothyroidism, a combination of environmental and genetic factors, including polymorphisms in HLA-DR, the immunoregulatory genes *CTLA-4*, *CD25*, *PTPN22*, *FCRL3*, and *CD226*, as well as the gene encoding the thyroid-stimulating hormone receptor (TSH-R), contributes to Graves' disease susceptibility. The concordance for Graves' disease in monozygotic twins is 20–30%, compared to <5% in dizygotic twins. Indirect evidence suggests that stress is an important environmental factor, presumably operating through neuroendocrine effects on the immune system. Smoking is a minor risk factor for Graves' disease and a major risk factor for the development of ophthalmopathy. Sudden increases in iodine intake may precipitate Graves' disease, and there is a threefold increase in the occurrence of Graves' disease in the postpartum period. Graves' disease may occur during the immune reconstitution phase after highly active antiretroviral therapy (HAART) or alemtuzumab treatment.

The hyperthyroidism of Graves' disease is caused by thyroid-stimulating immunoglobulin (TSI) that are synthesized in the thyroid gland as well as in bone marrow and lymph nodes. Such antibodies can be detected by bioassays or by using the more widely available thyrotropin-binding inhibitory immunoglobulin (TBII) assays. The presence of TBII in a patient with thyrotoxicosis implies the existence of TSI, and these assays are useful in monitoring pregnant Graves' patients in whom high levels of TSI can cross the placenta and cause neonatal thyrotoxicosis. Other thyroid autoimmune responses, similar to those in autoimmune hypothyroidism (see above), occur concurrently in patients with Graves' disease. In particular, thyroid peroxidase (TPO) and thyroglobulin (Tg) antibodies occur in up to 80% of cases. Because the coexisting thyroiditis can also affect thyroid function, there is no direct correlation between the level of TSI and thyroid hormone levels in Graves' disease.

Cytokines appear to play a major role in thyroid-associated ophthalmopathy. There is infiltration of the extraocular muscles by activated T cells; the release of cytokines such as interferon γ (IFN-γ), tumor necrosis factor (TNF), and interleukin-1 (IL-1) results in fibroblast activation and increased synthesis of glycosaminoglycans that trap water, thereby leading to characteristic muscle swelling. Late in the disease, there is irreversible fibrosis of the muscles. Though the pathogenesis of thyroid-associated ophthalmopathy remains unclear, there is mounting evidence that the TSH-R is a shared autoantigen that is expressed in the orbit; this would explain the close association with autoimmune thyroid disease. Increased fat is an additional cause of retrobulbar tissue expansion. The increase in intraorbital pressure can lead to proptosis, diplopia, and optic neuropathy.

Clinical Manifestations Signs and symptoms include features that are common to any cause of thyrotoxicosis (Table 377-2) as well as those specific for Graves' disease. The clinical presentation depends on the severity of thyrotoxicosis, the duration of disease, individual susceptibility to excess thyroid hormone, and the patient's age. In the elderly, features of thyrotoxicosis may be subtle or masked, and patients may present mainly with fatigue and weight loss, a condition known as *apathetic thyrotoxicosis*.

TABLE 377-2 Signs and Symptoms of Thyrotoxicosis (Descending Order of Frequency)

SYMPTOMS	SIGNS[a]
Hyperactivity, irritability, dysphoria	Tachycardia; atrial fibrillation in the elderly
Heat intolerance and sweating	Tremor
Palpitations	Goiter
Fatigue and weakness	Warm, moist skin
Weight loss with increased appetite	Muscle weakness, proximal myopathy
Diarrhea	Lid retraction or lag
Polyuria	Gynecomastia
Oligomenorrhea, loss of libido	

[a]Excludes the signs of ophthalmopathy and dermopathy specific for Graves' disease.

Thyrotoxicosis may cause unexplained weight loss, despite an enhanced appetite, due to the increased metabolic rate. Weight gain occurs in 5% of patients, however, because of increased food intake. Other prominent features include hyperactivity, nervousness, and irritability, ultimately leading to a sense of easy fatigability in some patients. Insomnia and impaired concentration are common; apathetic thyrotoxicosis may be mistaken for depression in the elderly. Fine tremor is a frequent finding, best elicited by having patients stretch out their fingers while feeling the fingertips with the palm. Common neurologic manifestations include hyperreflexia, muscle wasting, and proximal myopathy without fasciculation. Chorea is rare. Thyrotoxicosis is sometimes associated with a form of hypokalemic periodic paralysis; this disorder is particularly common in Asian males with thyrotoxicosis, but it occurs in other ethnic groups as well.

The most common cardiovascular manifestation is sinus tachycardia, often associated with palpitations, occasionally caused by supraventricular tachycardia. The high cardiac output produces a bounding pulse, widened pulse pressure, and an aortic systolic murmur and can lead to worsening of angina or heart failure in the elderly or those with preexisting heart disease. Atrial fibrillation is more common in patients >50 years of age. Treatment of the thyrotoxic state alone converts atrial fibrillation to normal sinus rhythm in about half of patients, suggesting the existence of an underlying cardiac problem in the remainder.

The skin is usually warm and moist, and the patient may complain of sweating and heat intolerance, particularly during warm weather. Palmar erythema, onycholysis, and, less commonly, pruritus, urticaria, and diffuse hyperpigmentation may be evident. Hair texture may become fine, and a diffuse alopecia occurs in up to 40% of patients, persisting for months after restoration of euthyroidism. Gastrointestinal transit time is decreased, leading to increased stool frequency, often with diarrhea and occasionally mild steatorrhea. Women frequently experience oligomenorrhea or amenorrhea; in men, there may be impaired sexual function and, rarely, gynecomastia. The direct effect of thyroid hormones on bone resorption leads to osteopenia in long-standing thyrotoxicosis; mild hypercalcemia occurs in up to 20% of patients, but hypercalciuria is more common. There is a small increase in fracture rate in patients with a previous history of thyrotoxicosis.

In Graves' disease, the thyroid is usually diffusely enlarged to two to three times its normal size. The consistency is firm, but not nodular. There may be a thrill or bruit, best detected at the inferolateral margins of the thyroid lobes, due to the increased vascularity of the gland and the hyperdynamic circulation.

Lid retraction, causing a staring appearance, can occur in any form of thyrotoxicosis and is the result of sympathetic overactivity. However, Graves' disease is associated with specific eye signs that comprise *Graves' ophthalmopathy* **(Fig. 377-1*A*)**. This condition is also called *thyroid-associated ophthalmopathy*, because it occurs in the absence of hyperthyroidism in 10% of patients. Most of these individuals have autoimmune hypothyroidism or thyroid antibodies. The onset of Graves' ophthalmopathy occurs within the year before or after the diagnosis of thyrotoxicosis in 75% of patients but can sometimes precede or follow thyrotoxicosis by several years, accounting for some cases of euthyroid ophthalmopathy.

Some patients with Graves' disease have little clinical evidence of ophthalmopathy. However, the enlarged extraocular muscles typical of the disease, and other subtle features, can be detected in most patients when investigated by ultrasound or computed tomography (CT) imaging of the orbits. Unilateral signs are found in up to 10% of patients. The earliest manifestations of ophthalmopathy are usually a sensation of grittiness, eye discomfort, and excess tearing. About one-third of patients have proptosis, best detected by visualization of the sclera between the lower border of the iris and the lower eyelid, with the eyes in the primary position. Proptosis can be measured using an exophthalmometer. In severe cases, proptosis may cause corneal exposure and damage, especially if the lids fail to close during sleep. Periorbital edema, scleral injection, and chemosis are also frequent. In 5–10% of patients, the muscle swelling is so severe that diplopia results, typically, but not exclusively, when the patient looks up and laterally. The most serious manifestation is compression of the optic nerve at the apex of the orbit, leading to papilledema; peripheral field defects; and, if left untreated, permanent loss of vision.

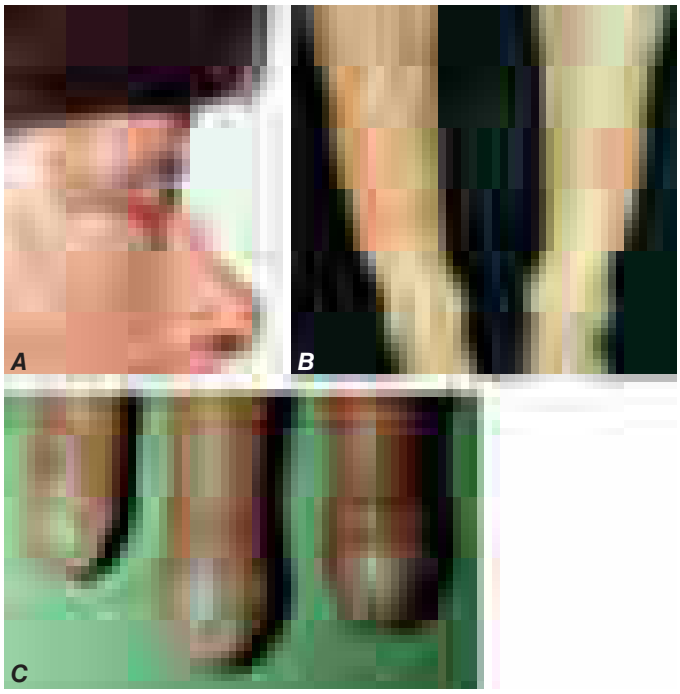


FIGURE 377-1 Features of Graves' disease. ***A.*** Ophthalmopathy in Graves' disease; lid retraction, periorbital edema, conjunctival injection, and proptosis are marked. ***B.*** Thyroid dermopathy over the lateral aspects of the shins. ***C.*** Thyroid acropachy.

The "NO SPECS" scoring system to evaluate ophthalmopathy is an acronym derived from the following changes:

0 = **N**o signs or symptoms
1 = **O**nly signs (lid retraction or lag), no symptoms
2 = **S**oft tissue involvement (periorbital edema)
3 = **P**roptosis (>22 mm)
4 = **E**xtraocular muscle involvement (diplopia)
5 = **C**orneal involvement
6 = **S**ight loss

Although useful as a mnemonic, the NO SPECS scheme is inadequate to describe the eye disease fully, and patients do not necessarily progress from one class to another; alternative scoring systems (e.g., the EUGOGO system developed by the European Group On Graves' Orbitopathy) that assess disease activity are preferable for monitoring and treatment purposes. When Graves' eye disease is active and severe, referral to an ophthalmologist is indicated and objective measurements are needed, such as lid-fissure width; corneal staining with fluorescein; and evaluation of extraocular muscle function (e.g., Hess chart), intraocular pressure and visual fields, acuity, and color vision.

Thyroid dermopathy occurs in <5% of patients with Graves' disease **(Fig. 377-1*B*)**, almost always in the presence of moderate or severe ophthalmopathy. Although most frequent over the anterior and lateral aspects of the lower leg (hence the term *pretibial myxedema*), skin changes can occur at other sites, particularly after trauma. The typical lesion is a noninflamed, indurated plaque with a deep pink or purple color and an "orange skin" appearance. Nodular involvement can occur, and the condition can rarely extend over the whole lower leg and foot, mimicking elephantiasis. *Thyroid acropachy* refers to a form of clubbing found in <1% of patients with Graves' disease **(Fig. 377-1*C*)**. It is so strongly associated with thyroid dermopathy that an alternative cause of clubbing should be sought in a Graves' patient without coincident skin and orbital involvement.

Laboratory Evaluation Investigations used to determine the existence and cause of thyrotoxicosis are summarized in **Fig. 377-2**. In

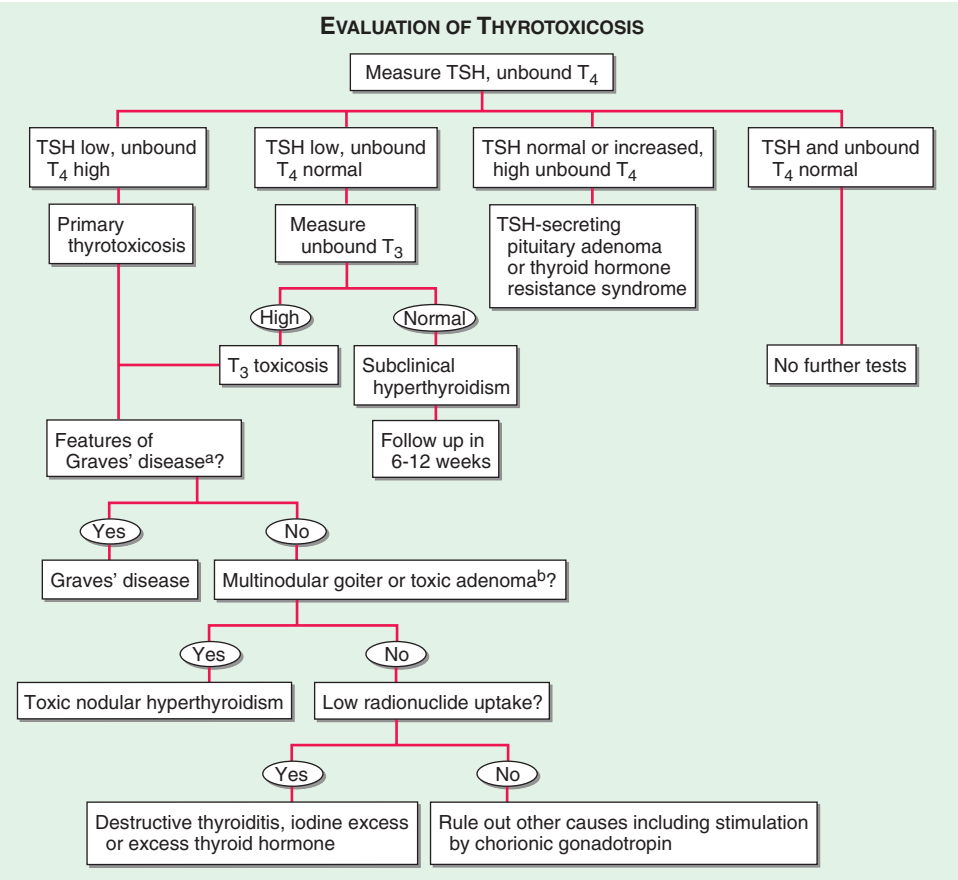


FIGURE 377-2 Evaluation of thyrotoxicosis. [a]Diffuse goiter, positive TPO antibodies or TRAb, ophthalmopathy, dermopathy. [b]Can be confirmed by radionuclide scan. TSH, thyroid-stimulating hormone.

Graves' disease, the TSH level is suppressed, and total and unbound thyroid hormone levels are increased. In 2–5% of patients (and more in areas of borderline iodine intake), only T_3 is increased (T_3 toxicosis). The converse state of T_4 toxicosis, with elevated total and unbound T_4 and normal T_3 levels, is occasionally seen when hyperthyroidism is induced by excess iodine, providing surplus substrate for thyroid hormone synthesis. Measurement of TPO antibodies or TBII may be useful if the diagnosis is unclear clinically but is not needed routinely. Associated abnormalities that may cause diagnostic confusion in thyrotoxicosis include elevation of bilirubin, liver enzymes, and ferritin. Microcytic anemia and thrombocytopenia may occur.

Differential Diagnosis Diagnosis of Graves' disease is straightforward in a patient with biochemically confirmed thyrotoxicosis, diffuse goiter on palpation, ophthalmopathy, and often a personal or family history of autoimmune disorders. For patients with thyrotoxicosis who lack these features, the diagnosis is generally established by a radionuclide (^{99m}Tc, ^{123}I, or ^{131}I) scan and uptake of the thyroid, which will distinguish the diffuse, high uptake of Graves' disease from destructive thyroiditis, ectopic thyroid tissue, and factitious thyrotoxicosis, as well as diagnosing a toxic adenoma or toxic MNG. Alternatively, TRAb measurement can be used to diagnose Graves' disease and color-flow Doppler ultrasonography may distinguish between hyperthyroidism (with increased blood flow) and destructive thyroiditis. In secondary hyperthyroidism due to a TSH-secreting pituitary tumor, there is also a diffuse goiter. The presence of a nonsuppressed TSH level and the finding of a pituitary tumor on CT or magnetic resonance imaging (MRI) scan suggest this diagnosis.

Clinical features of thyrotoxicosis can mimic certain aspects of other disorders, including panic attacks, mania, pheochromocytoma, and weight loss associated with malignancy. The diagnosis of thyrotoxicosis can be easily excluded if the TSH and unbound T_4 and T_3 levels are normal. A normal TSH also excludes Graves' disease as a cause of diffuse goiter.

Clinical Course Clinical features generally worsen without treatment; mortality was 10–30% before the introduction of satisfactory therapy. Some patients with mild Graves' disease experience spontaneous relapses and remissions. Rarely, there may be fluctuation between hypo- and hyperthyroidism due to changes in the functional activity of TSH-R antibodies. About 15% of patients who enter remission after treatment develop hypothyroidism 10–15 years later as a result of the destructive autoimmune process.

The clinical course of ophthalmopathy does not follow that of the thyroid disease, although thyroid dysfunction can worsen eye signs. Ophthalmopathy typically worsens over the initial 3–6 months, followed by a plateau phase over the next 12–18 months, and then some spontaneous improvement, particularly in the soft tissue changes. However, the course is more fulminant in up to 5% of patients, requiring intervention in the acute phase if there is optic nerve compression or corneal ulceration. Diplopia may appear late in the disease due to fibrosis of the extraocular muscles. Radioiodine treatment for hyperthyroidism worsens the eye disease in a small proportion of patients (especially smokers). Antithyroid drugs or surgery have no adverse effects on the clinical course of ophthalmopathy. Thyroid dermopathy, when it occurs, usually appears 1–2 years after the development of Graves' hyperthyroidism; it may improve spontaneously.

TREATMENT

Graves' Disease

The *hyperthyroidism* of Graves' disease is treated by reducing thyroid hormone synthesis, using an antithyroid drug, or reducing the amount of thyroid tissue with radioiodine (^{131}I) treatment or by thyroidectomy. Antithyroid drugs are the predominant therapy in many centers in Europe, Latin America, and Japan, whereas radioiodine is more often the first line of treatment in North America. These

differences reflect the fact that no single approach is optimal and that patients may require multiple treatments to achieve remission.

The main *antithyroid drugs* are thionamides; propylthiouracil, carbimazole (not available in the United States), and the active metabolite of the latter, methimazole. All inhibit the function of TPO, reducing oxidation and organification of iodide. These drugs also reduce thyroid antibody levels by mechanisms that remain unclear, and they appear to enhance spontaneous rates of remission. Propylthiouracil inhibits deiodination of $T_4 \rightarrow T_3$. However, this effect is of minor benefit, except in the most severe thyrotoxicosis, and is offset by the much shorter half-life of this drug (90 min) compared to methimazole (6 h). Due to the hepatotoxicity of propylthiouracil, the U.S. Food and Drug Administration (FDA) has limited indications for its use to the first trimester of pregnancy, the treatment of thyroid storm, and patients with minor adverse reactions to methimazole. If propylthiouracil is used, monitoring of liver function tests is recommended.

There are many variations of antithyroid drug regimens. The initial dose of carbimazole or methimazole is usually 10–20 mg every 8 or 12 h, but once-daily dosing is possible after euthyroidism is restored. Propylthiouracil is given at a dose of 100–200 mg every 6–8 h, and divided doses are usually given throughout the course. Lower doses of each drug may suffice in areas of low iodine intake. The starting dose of an antithyroid drug can be gradually reduced (titration regimen) as thyrotoxicosis improves. Less commonly, high doses may be given combined with levothyroxine supplementation (block-replace regimen) to avoid drug-induced hypothyroidism. The titration regimen is preferred to minimize the dose of antithyroid drug and provide an index of treatment response.

Thyroid function tests and clinical manifestations are reviewed 4–6 weeks after starting treatment, and the dose is titrated based on unbound T_4 levels. Most patients do not achieve euthyroidism until 6–8 weeks after treatment is initiated. TSH levels often remain suppressed for several months and therefore do not provide a sensitive index of treatment response. The usual daily maintenance doses of antithyroid drugs in the titration regimen are 2.5–10 mg of carbimazole or methimazole and 50–100 mg of propylthiouracil. In the block-replace regimen, the initial dose of antithyroid drug is held constant, and the dose of levothyroxine is adjusted to maintain normal unbound T_4 levels. When TSH suppression is alleviated, TSH levels can also be used to monitor therapy.

Maximum remission rates (up to 30–60% in some populations) are achieved by 12–18 months for the titration regimen and are higher in patients where TRAb levels are no longer detected, than in those with TRAb persistence. For unclear reasons, remission rates appear to vary in different geographic regions. Younger patients, males, smokers, and patients with a history of allergy, severe hyperthyroidism or large goiters are most likely to relapse when treatment stops, but outcomes are difficult to predict. All patients should be followed closely for relapse during the first year after treatment and at least annually thereafter.

The common minor side effects of antithyroid drugs are rash, urticaria, fever, and arthralgia (1–5% of patients). These may resolve spontaneously or after substituting an alternative antithyroid drug; rashes may respond to an antihistamine. Rare but major side effects include hepatitis (especially with propylthiouracil; avoid use in children) and cholestasis (methimazole and carbimazole); vasculitis; and, most important, agranulocytosis (<1%). It is essential that antithyroid drugs are stopped and not restarted if a patient develops major side effects. Written instructions should be provided regarding the symptoms of possible agranulocytosis (e.g., sore throat, fever, mouth ulcers) and the need to stop treatment pending an urgent complete blood count to confirm that agranulocytosis is not present. Management of agranulocytosis is described in **Chap. 98**. It is not useful to monitor blood counts prospectively, because the onset of agranulocytosis is idiosyncratic and abrupt.

Propranolol (20–40 mg every 6 h) or longer-acting selective β_1 receptor blockers such as atenolol may be helpful to control adrenergic symptoms, especially in the early stages before antithyroid drugs take effect. Beta blockers are also useful in patients with thyrotoxic periodic paralysis, pending correction of thyrotoxicosis. In consultation with a cardiologist, anticoagulation with warfarin should be considered in all patients with atrial fibrillation; there is often spontaneous reversion to sinus rhythm with control of hyperthyroidism, and long-term anticoagulation is not usually needed. Decreased warfarin doses are required when patients are thyrotoxic. If digoxin is used, increased doses are often needed in the thyrotoxic state.

Radioiodine causes progressive destruction of thyroid cells and can be used as initial treatment or for relapses after a trial of antithyroid drugs. There is a small risk of thyrotoxic crisis (see below) after radioiodine, which can be minimized by pretreatment with antithyroid drugs for at least a month before treatment. Antecedent treatment with an antithyroid drug and a beta blocker should be considered for all elderly patients or for those with cardiac problems. Carbimazole or methimazole must be stopped 2–3 days before radioiodine administration to achieve optimum iodine uptake, and can be restarted 3–7 days after radioiodine in those at risk of complications from worsening thyrotoxicosis. Propylthiouracil appears to have a prolonged radioprotective effect and should be stopped for a longer period before radioiodine is given, or a larger dose of radioiodine will be necessary.

Efforts to calculate an optimal dose of radioiodine that achieves euthyroidism without a high incidence of relapse or progression to hypothyroidism have not been successful. Some patients inevitably relapse after a single dose because the biologic effects of radiation vary between individuals, and hypothyroidism cannot be uniformly avoided even using accurate dosimetry. A practical strategy is to give a fixed dose based on clinical features, such as the severity of thyrotoxicosis, the size of the goiter (increases the dose needed), and the level of radioiodine uptake (decreases the dose needed). ^{131}I dosage generally ranges between 370 MBq (10 mCi) and 555 MBq (15 mCi). Most authorities favor an approach aimed at thyroid ablation (as opposed to euthyroidism), given that levothyroxine replacement is straightforward and most patients ultimately progress to hypothyroidism over 5–10 years, frequently with some delay in the diagnosis of hypothyroidism.

Certain radiation safety precautions are necessary in the first few days after radioiodine treatment, but the exact guidelines vary depending on local protocols. In general, patients need to avoid close, prolonged contact with children and pregnant women for 5–7 days because of possible transmission of residual isotope and exposure to radiation emanating from the gland. Rarely, there may be mild pain due to radiation thyroiditis 1–2 weeks after treatment. Hyperthyroidism can persist for 2–3 months before radioiodine takes full effect. For this reason, β-adrenergic blockers or antithyroid drugs can be used to control symptoms during this interval. Persistent hyperthyroidism can be treated with a second dose of radioiodine, usually 6 months after the first dose. The risk of hypothyroidism after radioiodine depends on the dosage but is at least 10–20% in the first year and 5% per year thereafter. Patients should be informed of this possibility before treatment and require close follow-up during the first year followed by annual thyroid function testing.

Pregnancy and breast-feeding are absolute contraindications to radioiodine treatment, but patients can conceive safely 6 months after treatment. The presence of ophthalmopathy, especially in smokers, requires caution. Prednisone, 30 mg/d, at the time of radioiodine treatment, tapered over 6–8 weeks may prevent exacerbation of ophthalmopathy, but radioiodine should generally be avoided in those with active moderate to severe eye disease. The overall risk of cancer after radioiodine treatment in adults is not increased. Although many physicians avoid radioiodine in children and adolescents because of the theoretical risks of malignancy, emerging evidence suggests that radioiodine can be used safely in older children.

Total or near-total thyroidectomy is an option for patients who relapse after antithyroid drugs and prefer this treatment to radioiodine. Some experts recommend surgery in young individuals, particularly when the goiter is very large. Careful control of thyrotoxicosis with antithyroid drugs, followed by potassium iodide (1–2 drops SSKI orally tid for 10 days), is needed prior to surgery to avoid thyrotoxic crisis and to reduce the vascularity of the gland. The major complications of surgery—bleeding, laryngeal edema, hypoparathyroidism, and damage to the recurrent laryngeal nerves—are unusual when the procedure is performed by highly experienced surgeons. Recurrence rates in the best series are <2%, but the rate of hypothyroidism is similar to that following radioiodine treatment, especially with the current trend away from subtotal thyroidectomy.

Antithyroid drugs should be used to manage Graves' disease in *pregnancy.* Because transplacental passage of these drugs may produce fetal hypothyroidism and goiter if the maternal dose is excessive, maternal antithyroid dose titration should target serum free or total T_4 levels at or just above the pregnancy reference range. If available, propylthiouracil should be used until 14–16 weeks' gestation because of the association of rare cases of methimazole/carbimazole embryopathy, including *aplasia cutis* and other defects, such as choanal atresia and tracheoesophageal fistulae. Because of the potential for teratogenic effects, recent recommendations suggest discontinuation of antithyroid medication in a newly pregnant woman with Graves' disease, who is euthyroid on a low dose of methimazole (<5–10 mg/day) or PTU (<100–200 mg/day), after evaluating recent thyroid function tests, disease history, goiter size, duration of therapy, and TRAb measurement. Following cessation, careful monitoring of maternal thyroid function tests is essential. On the other hand, for women at high risk of developing thyrotoxicosis if antithyroid drugs are discontinued (large goiter, requirement for higher antithyroid drug dosage), continued therapy is necessary, with PTU (if available) administration in the first trimester. But, because of its rare association with hepatotoxicity, propylthiouracil should be limited to the first trimester and then maternal therapy should be converted to methimazole (or carbimazole) at a ratio of 15–20 mg of propylthiouracil to 1 mg of methimazole. It is often possible to stop treatment in the last trimester because TSIs tend to decline in pregnancy. Nonetheless, the transplacental transfer of these antibodies if present at levels 3 times higher than the normative range rarely causes *fetal* or *neonatal thyrotoxicosis*. Poor intrauterine growth, a fetal heart rate of >160 beats/min, advanced bone age, fetal goiter, and high levels of maternal TSI after 26 weeks gestation may herald this complication. Antithyroid drugs given to the mother can be used to treat the fetus and may be needed for 1–3 months after delivery, until the maternal antibodies disappear from the baby's circulation. The postpartum period is a time of major risk for relapse of Graves' disease. Breast-feeding is safe with low doses of antithyroid drugs. Graves' disease in *children* is usually managed initially with methimazole or carbimazole (avoid propylthiouracil), often given as a prolonged course of the titration regimen. Surgery or radioiodine may be indicated for severe or relapsing disease.

Thyrotoxic crisis, or *thyroid storm*, is rare and presents as a life-threatening exacerbation of hyperthyroidism, accompanied by fever, delirium, seizures, coma, vomiting, diarrhea, and jaundice. The mortality rate due to cardiac failure, arrhythmia, or hyperthermia is as high as 30%, even with treatment. Thyrotoxic crisis is usually precipitated by acute illness (e.g., stroke, infection, trauma, diabetic ketoacidosis), surgery (especially on the thyroid), or radioiodine treatment of a patient with partially treated or untreated hyperthyroidism. Management requires intensive monitoring and supportive care, identification and treatment of the precipitating cause, and measures that reduce thyroid hormone synthesis. Large doses of propylthiouracil (500–1000 mg loading dose and 250 mg every 4 h) should be given orally or by nasogastric tube or per rectum; the drug's inhibitory action on $T_4 \rightarrow T_3$ conversion makes it the antithyroid drug of choice. If not available, methimazole can be used in doses of 20 mg every 6 h. One hour after the first dose of propylthiouracil, stable iodide (5 drops SSKI every 6 h) is given to block thyroid hormone synthesis via the Wolff-Chaikoff effect (the delay allows the antithyroid drug to prevent the excess iodine from being incorporated into new hormone). Propranolol should also be given to reduce tachycardia and other adrenergic manifestations (60–80 mg PO every 4 h; or 2 mg IV every 4 h). Although other β-adrenergic blockers can be used, high doses of propranolol decrease $T_4 \rightarrow T_3$ conversion, and the doses can be easily adjusted. Caution is needed to avoid acute negative inotropic effects, but controlling the heart rate is important, as some patients develop a form of high-output heart failure. Short-acting IV esmolol can be used to decrease heart rate while monitoring for signs of heart failure. Additional therapeutic measures include glucocorticoids (e.g., hydrocortisone 300 mg IV bolus, then 100 mg every 8 h), antibiotics if infection is present, cholestyramine to sequester thyroid hormones, cooling, oxygen, and IV fluids.

Ophthalmopathy requires no active treatment when it is mild or moderate, because there is usually spontaneous improvement. General measures include meticulous control of thyroid hormone levels, cessation of smoking, and an explanation of the natural history of ophthalmopathy. Discomfort can be relieved with artificial tears (e.g., hypromellose 0.3% or carbomer 0.2% ophthalmic gel) paraffin-based eye ointment, and the use of dark glasses with side frames. Periorbital edema may respond to a more upright sleeping position or a diuretic. Corneal exposure during sleep can be avoided by using patches or taping the eyelids shut. Minor degrees of diplopia improve with prisms fitted to spectacles. Some authorities also advocate selenium 100 μg bd. Severe ophthalmopathy, with optic nerve involvement or chemosis resulting in corneal damage, is an emergency requiring joint management with an ophthalmologist. Pulse therapy with IV methylprednisolone (e.g., 500 mg of methylprednisolone once weekly for 6 weeks, then 250 mg once weekly for 6 weeks) is preferable to oral glucocorticoids, which are used for moderately active disease. When glucocorticoids are ineffective, orbital decompression can be achieved by removing bone from any wall of the orbit, thereby allowing displacement of fat and swollen extraocular muscles. The transantral route is used most often because it requires no external incision. Proptosis recedes an average of 5 mm, but there may be residual or even worsened diplopia. Once the eye disease has stabilized, surgery may be indicated for relief of diplopia and correction of the appearance. External beam radiotherapy of the orbits has been used for many years, but the efficacy of this therapy remains unclear, and it is best reserved for those with moderately active disease who have failed or are not candidates for glucocorticoid therapy. Other immunosuppressive agents such as rituximab have shown some benefit, but their role is yet to be established.

Thyroid dermopathy does not usually require treatment, but it can cause cosmetic problems or interfere with the fit of shoes. Surgical removal is not indicated. If necessary, treatment consists of topical, high-potency glucocorticoid ointment under an occlusive dressing. Octreotide may be beneficial in some cases.

OTHER CAUSES OF THYROTOXICOSIS

Destructive thyroiditis (subacute or silent thyroiditis) typically presents with a short thyrotoxic phase due to the release of preformed thyroid hormones and catabolism of Tg (see "Subacute Thyroiditis," below). True hyperthyroidism is absent, as demonstrated by a low radionuclide uptake. Circulating Tg levels are typically increased. Other causes of thyrotoxicosis with low or absent thyroid radionuclide uptake include *thyrotoxicosis factitia*, iodine excess, and, rarely, ectopic thyroid tissue, particularly teratomas of the ovary (*struma ovarii*) and functional metastatic follicular carcinoma. Whole-body radionuclide studies can demonstrate ectopic thyroid tissue, and thyrotoxicosis factitia can be distinguished from destructive thyroiditis by the clinical features and low levels of Tg. Amiodarone treatment is associated with thyrotoxicosis in up to 10% of patients, particularly in areas of low iodine intake (see below).

TSH-secreting pituitary adenoma is a rare cause of thyrotoxicosis. It is characterized by the presence of an inappropriately normal or increased TSH level in a patient with hyperthyroidism, diffuse goiter, and elevated T_4 and T_3 levels **(Chap. 373)**. Elevated levels of the α-subunit of TSH, released by the TSH-secreting adenoma, support this diagnosis, which can be confirmed by demonstrating the pituitary tumor on MRI or CT scan. A combination of transsphenoidal surgery, sella irradiation, and octreotide may be required to normalize TSH, because many of these tumors are large and locally invasive at the time of diagnosis. Radioiodine or antithyroid drugs can be used to control thyrotoxicosis.

Thyrotoxicosis caused by *toxic MNG* and *hyperfunctioning solitary nodules* is discussed below.

THYROIDITIS

A clinically useful classification of thyroiditis is based on the onset and duration of disease **(Table 377-3)**.

ACUTE THYROIDITIS

Acute thyroiditis is rare and due to suppurative infection of the thyroid. In children and young adults, the most common cause is the presence of a piriform sinus, a remnant of the fourth branchial pouch that connects the oropharynx with the thyroid. Such sinuses are predominantly left-sided. A long-standing goiter and degeneration in a thyroid malignancy are risk factors in the elderly. The patient presents with thyroid pain, often referred to the throat or ears, and a small, tender goiter that may be asymmetric. Fever, dysphagia, and erythema over the thyroid are common, as are systemic symptoms of a febrile illness and lymphadenopathy.

The differential diagnosis of *thyroid pain* includes subacute or, rarely, chronic thyroiditis; hemorrhage into a cyst; malignancy including lymphoma; and, rarely, amiodarone-induced thyroiditis or amyloidosis. However, the abrupt presentation and clinical features of acute thyroiditis rarely cause confusion. The erythrocyte sedimentation rate (ESR) and white cell count are usually increased, but thyroid function is normal. Fine-needle aspiration (FNA) biopsy shows infiltration by polymorphonuclear leukocytes; culture of the sample can identify the organism. Caution is needed in immunocompromised patients as fungal, mycobacterial, or *Pneumocystis* thyroiditis can occur in this setting. Antibiotic treatment is guided initially by Gram stain and, subsequently, by cultures of the FNA biopsy. Surgery may be needed to drain an abscess, which can be localized by CT scan or ultrasound. Tracheal obstruction, septicemia, retropharyngeal abscess, mediastinitis, and jugular venous thrombosis may complicate acute thyroiditis but are uncommon with prompt use of antibiotics.

SUBACUTE THYROIDITIS

This is also termed *de Quervain's thyroiditis*, *granulomatous thyroiditis*, or *viral thyroiditis*. Many viruses have been implicated, including mumps, coxsackie, influenza, adenoviruses, and echoviruses, but attempts to identify the virus in an individual patient are often unsuccessful and do not influence management. The diagnosis of subacute thyroiditis is often overlooked because the symptoms can mimic pharyngitis. The peak incidence occurs at 30–50 years, and women are affected three times more frequently than men.

TABLE 377-3 Causes of Thyroiditis

Acute
Bacterial infection: especially *Staphylococcus*, *Streptococcus*, and *Enterobacter*
Fungal infection: *Aspergillus*, *Candida*, *Coccidioides*, *Histoplasma*, and *Pneumocystis*
Radiation thyroiditis after ^{131}I treatment
Amiodarone (may also be subacute or chronic)
Subacute
Viral (or granulomatous) thyroiditis
Silent thyroiditis (including postpartum thyroiditis)
Mycobacterial infection
Drug induced (interferon, amiodarone)
Chronic
Autoimmunity: focal thyroiditis, Hashimoto's thyroiditis, atrophic thyroiditis
Riedel's thyroiditis
Parasitic thyroiditis: echinococcosis, strongyloidiasis, cysticercosis
Traumatic: after palpation

Pathophysiology The thyroid shows a characteristic patchy inflammatory infiltrate with disruption of the thyroid follicles and multinucleated giant cells within some follicles. The follicular changes progress to granulomas accompanied by fibrosis. Finally, the thyroid returns to normal, usually several months after onset. During the initial phase of follicular destruction, there is release of Tg and thyroid hormones, leading to increased circulating T_4 and T_3 and suppression of TSH **(Fig. 377-3)**. During this destructive phase, radioactive iodine uptake is low or undetectable. After several weeks, the thyroid is depleted of stored thyroid hormone and a phase of hypothyroidism typically occurs, with low unbound T_4 (and sometimes T_3) and moderately increased TSH levels. Radioactive iodine uptake returns to normal or is even increased as a result of the rise in TSH. Finally, thyroid hormone and TSH levels return to normal as the disease subsides.

Clinical Manifestations The patient usually presents with a painful and enlarged thyroid, sometimes accompanied by fever. There may be features of thyrotoxicosis or hypothyroidism, depending on the phase of the illness. Malaise and symptoms of an upper respiratory tract infection may precede the thyroid-related features by several weeks. In other patients, the onset is acute, severe, and without obvious antecedent. The patient typically complains of a sore throat, and examination reveals a small goiter that is exquisitely tender. Pain is often referred to the jaw or ear. Complete resolution is the usual outcome, but late-onset permanent hypothyroidism occurs in 15% of cases, particularly in those with coincidental thyroid autoimmunity. A prolonged course over many months, with one or more relapses, occurs in a small percentage of patients.

Laboratory Evaluation As depicted in Fig. 377-3, thyroid function tests characteristically evolve through three distinct phases over about 6 months: (1) thyrotoxic phase, (2) hypothyroid phase, and (3) recovery phase. In the thyrotoxic phase, T_4 and T_3 levels are increased, reflecting their discharge from the damaged thyroid cells, and TSH is suppressed. The T_4/T_3 ratio is greater than in Graves' disease or thyroid autonomy, in which T_3 is often disproportionately increased. The diagnosis is confirmed by a high ESR and low uptake of radioiodine (<5%) or ^{99m}Tc pertechnetate (as compared to salivary gland pertechnetate concentration). The white blood cell count may be increased,

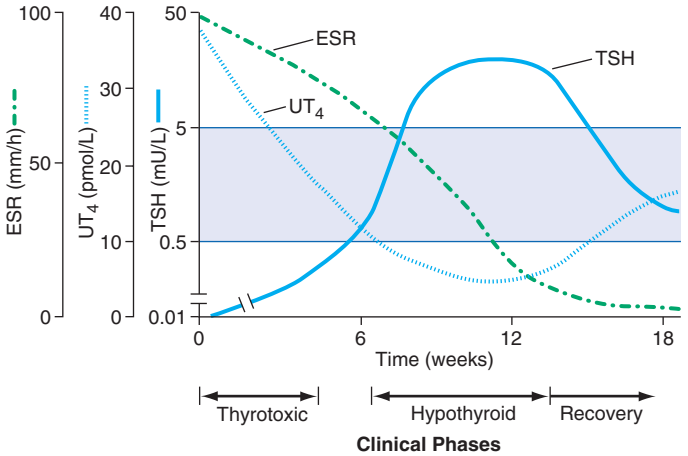


FIGURE 377-3 Clinical course of subacute thyroiditis. The release of thyroid hormones is initially associated with a thyrotoxic phase and suppressed thyroid-stimulating hormone (TSH). A hypothyroid phase then ensues, with low T_4 and TSH levels that are initially low but gradually increase. During the recovery phase, increased TSH levels combined with resolution of thyroid follicular injury lead to normalization of thyroid function, often several months after the beginning of the illness. ESR, erythrocyte sedimentation rate; UT_4, free or unbound T_4.

and thyroid antibodies are negative. If the diagnosis is in doubt, FNA biopsy may be useful, particularly to distinguish unilateral involvement from bleeding into a cyst or neoplasm.

TREATMENT

Subacute Thyroiditis

Relatively large doses of aspirin (e.g., 600 mg every 4–6 h) or nonsteroidal anti-inflammatory drugs (NSAIDs) are sufficient to control symptoms in many cases. If this treatment is inadequate, or if the patient has marked local or systemic symptoms, glucocorticoids should be given. The usual starting dose is 15–40 mg of prednisone, depending on severity. The dose is gradually tapered over 6–8 weeks, in response to improvement in symptoms and the ESR. If a relapse occurs during glucocorticoid withdrawal, the dosage should be increased and then withdrawn more gradually. Thyroid function should be monitored every 2–4 weeks using TSH and unbound T_4 levels. Symptoms of thyrotoxicosis improve spontaneously but may be ameliorated by β-adrenergic blockers; antithyroid drugs play no role in treatment of the thyrotoxic phase. Levothyroxine replacement may be needed if the hypothyroid phase is prolonged, but doses should be low enough (50–100 μg daily) to allow TSH-mediated recovery.

SILENT THYROIDITIS

Painless thyroiditis, or "*silent*" *thyroiditis*, occurs in patients with underlying autoimmune thyroid disease and has a clinical course similar to that of subacute thyroiditis. The condition occurs in up to 5% of women 3–6 months after pregnancy and is then termed *postpartum thyroiditis*. Typically, patients have a brief phase of thyrotoxicosis lasting 2–4 weeks, followed by hypothyroidism for 4–12 weeks, and then resolution; often, however, only one phase is apparent. The condition is associated with the presence of TPO antibodies antepartum, and it is three times more common in women with type 1 diabetes mellitus. As in subacute thyroiditis, the uptake of ^{99m}Tc pertechnetate or radioactive iodine is initially suppressed. In addition to the painless goiter, silent thyroiditis can be distinguished from subacute thyroiditis by a normal ESR and the presence of TPO antibodies. Glucocorticoid treatment is not indicated for silent thyroiditis. Severe thyrotoxic symptoms can be managed with a brief course of propranolol, 20–40 mg three or four times daily. Thyroxine replacement may be needed for the hypothyroid phase but should be withdrawn after 6–9 months, as recovery is the rule. Annual follow-up thereafter is recommended, because a proportion of these individuals develop permanent hypothyroidism. The condition may recur in subsequent pregnancies.

DRUG-INDUCED THYROIDITIS

Patients receiving cytokines, such as IFN-α or IL-2, or tyrosine kinase inhibitors may develop painless thyroiditis. IFN-α, which is used to treat chronic hepatitis B or C and hematologic and skin malignancies, causes thyroid dysfunction in up to 5% of treated patients. It has been associated with painless thyroiditis, hypothyroidism, and Graves' disease, and is most common in women with TPO antibodies prior to treatment. For discussion of amiodarone, see "Amiodarone Effects on Thyroid Function," below.

CHRONIC THYROIDITIS

Focal thyroiditis is present in 20–40% of euthyroid autopsy cases and is associated with serologic evidence of autoimmunity, particularly the presence of TPO antibodies. The most common clinically apparent cause of chronic thyroiditis is *Hashimoto's thyroiditis*, an autoimmune disorder that often presents as a firm or hard goiter of variable size (see above). *Riedel's thyroiditis* is a rare disorder that typically occurs in middle-aged women. It presents with an insidious, painless goiter with local symptoms due to compression of the esophagus, trachea, neck veins, or recurrent laryngeal nerves. Dense fibrosis disrupts normal gland architecture and can extend outside the thyroid capsule. Despite these extensive histologic changes, thyroid dysfunction is uncommon. The goiter is hard, nontender, often asymmetric, and fixed, leading to suspicion of a malignancy. Diagnosis requires open biopsy as FNA biopsy is usually inadequate. Treatment is directed to surgical relief of compressive symptoms. Tamoxifen may also be beneficial. There is an association between Riedel's thyroiditis and IgG4-related disease causing idiopathic fibrosis at other sites (retroperitoneum, mediastinum, biliary tree, lung, and orbit).

SICK EUTHYROID SYNDROME (NONTHYROIDAL ILLNESS)

Any acute, severe illness can cause abnormalities of circulating TSH or thyroid hormone levels in the absence of underlying thyroid disease, making these measurements potentially misleading. The major cause of these hormonal changes is the release of cytokines such as IL-6. Unless a thyroid disorder is strongly suspected, the routine testing of thyroid function should be avoided in acutely ill patients.

The most common hormone pattern in sick euthyroid syndrome (SES), also called nonthyroidal illness (NTI), is a decrease in total and unbound T_3 levels (low T_3 syndrome) with normal levels of T_4 and TSH. The magnitude of the fall in T_3 correlates with the severity of the illness. T_4 conversion to T_3 via peripheral 5′ (outer ring) deiodination is impaired, leading to increased reverse T_3 (rT_3). Since rT_3 is metabolized by 5′ deiodination, its clearance is also reduced. Thus, decreased clearance rather than increased production is the major basis for increased rT_3. Also, T_4 is alternately metabolized to the hormonally inactive T_3 sulfate. It is generally assumed that this low T_3 state is adaptive, because it can be induced in normal individuals by fasting. Teleologically, the fall in T_3 may limit catabolism in starved or ill patients.

Very sick patients may exhibit a dramatic fall in total T_4 and T_3 levels (low T_4 syndrome). With decreased tissue perfusion, muscle and liver expression of the type 3 deiodinase leads to accelerated T_4 and T_3 metabolism. This state has a poor prognosis. Another key factor in the fall in T_4 levels is altered binding to thyroxine-binding globulin (TBG). The commonly used free T_4 assays are subject to artifact when serum binding proteins are low and underestimate the true free T_4 level. Fluctuation in TSH levels also creates challenges in the interpretation of thyroid function in sick patients. TSH levels may range from <0.1 mIU/L in very ill patients, especially with dopamine or glucocorticoid therapy, to >20 mIU/L during the recovery phase of SES. The exact mechanisms underlying the subnormal TSH seen in 10% of sick patients and the increased TSH seen in 5% remain unclear but may be mediated by cytokines including IL-12 and IL-18.

Any severe illness can induce changes in thyroid hormone levels, but certain disorders exhibit a distinctive pattern of abnormalities. Acute liver disease is associated with an initial rise in total (but not unbound) T_3 and T_4 levels due to TBG release; these levels become subnormal with progression to liver failure. A transient increase in total and unbound T_4 levels, usually with a normal T_3 level, is seen in 5–30% of acutely ill psychiatric patients. TSH values may be transiently low, normal, or high in these patients. In the early stage of HIV infection, T_3 and T_4 levels rise, even if there is weight loss. T_3 levels fall with progression to AIDS, but TSH usually remains normal. Renal disease is often accompanied by low T_3 concentrations, but with normal rather than increased rT_3 levels, due to an unknown factor that increases uptake of rT_3 into the liver.

The diagnosis of SES is challenging. Historic information may be limited, and patients often have multiple metabolic derangements. Useful features to consider include previous history of thyroid disease and thyroid function tests, evaluation of the severity and time course of the patient's acute illness, documentation of medications that may affect thyroid function or thyroid hormone levels, and measurements of rT_3 together with unbound thyroid hormones and TSH. The diagnosis of SES is frequently presumptive, given the clinical context and pattern of laboratory values; only resolution of the test results with clinical recovery can clearly establish this disorder. Treatment of SES with thyroid hormone (T_4 and/or T_3) is controversial, but most authorities recommend monitoring the patient's thyroid function tests during recovery, without administering thyroid hormone, unless there is

historic or clinical evidence suggestive of hypothyroidism. Sufficiently large randomized controlled trials using thyroid hormone are unlikely to resolve this therapeutic controversy in the near future, because clinical presentations and outcomes are highly variable.

AMIODARONE EFFECTS ON THYROID FUNCTION

Amiodarone is a commonly used type III antiarrhythmic agent (Chap. 247). It is structurally related to thyroid hormone and contains 39% iodine by weight. Thus, typical doses of amiodarone (200 mg/d) are associated with very high iodine intake, leading to greater than fortyfold increases in plasma and urinary iodine levels. Moreover, because amiodarone is stored in adipose tissue, high iodine levels persist for >6 months after discontinuation of the drug. Amiodarone inhibits deiodinase activity, and its metabolites function as weak antagonists of thyroid hormone action. Amiodarone has the following effects on thyroid function: (1) acute, transient suppression of thyroid function; (2) hypothyroidism in patients susceptible to the inhibitory effects of a high iodine load; and (3) thyrotoxicosis that may be caused by either a Jod-Basedow effect from the iodine load, in the setting of MNG or incipient Graves' disease, or a thyroiditis-like condition.

The initiation of amiodarone treatment is associated with a transient decrease of T_4 levels, reflecting the inhibitory effect of iodine on T_4 release. Soon thereafter, most individuals escape from iodide-dependent suppression of the thyroid (Wolff-Chaikoff effect), and the inhibitory effects on deiodinase activity and thyroid hormone receptor action become predominant. These events lead to the following pattern of thyroid function tests: increased T_4, decreased T_3, increased rT_3, and a transient TSH increase (up to 20 mIU/L). TSH levels normalize or are slightly suppressed within 1–3 months.

The incidence of hypothyroidism from amiodarone varies geographically, apparently correlating with iodine intake. Hypothyroidism occurs in up to 13% of amiodarone-treated patients in iodine-replete countries, such as the United States, but is less common (<6% incidence) in areas of lower iodine intake, such as Italy or Spain. The pathogenesis appears to involve an inability of the thyroid gland to escape from the Wolff-Chaikoff effect in autoimmune thyroiditis. Consequently, amiodarone-associated hypothyroidism is more common in women and individuals with positive TPO antibodies. It is usually unnecessary to discontinue amiodarone for this side effect, because levothyroxine can be used to normalize thyroid function. TSH levels should be monitored, because T_4 levels are often increased for the reasons described above.

The management of amiodarone-induced thyrotoxicosis (AIT) is complicated by the fact that there are different causes of thyrotoxicosis and because the increased thyroid hormone levels exacerbate underlying arrhythmias and coronary artery disease. Amiodarone treatment causes thyrotoxicosis in 10% of patients living in areas of low iodine intake and in 2% of patients in regions of high iodine intake. There are two major forms of AIT, although some patients have features of both. Type 1 AIT is associated with an underlying thyroid abnormality (preclinical Graves' disease or nodular goiter). Thyroid hormone synthesis becomes excessive as a result of increased iodine exposure (Jod-Basedow phenomenon). Type 2 AIT occurs in individuals with no intrinsic thyroid abnormalities and is the result of drug-induced lysosomal activation leading to destructive thyroiditis with histiocyte accumulation in the thyroid; the incidence rises as cumulative amiodarone dosage increases. Mild forms of type 2 AIT can resolve spontaneously or can occasionally lead to hypothyroidism. Color-flow Doppler ultrasonography shows increased vascularity in type 1 AIT but decreased vascularity in type 2 AIT. Thyroid scintiscans are difficult to interpret in this setting because the high endogenous iodine levels diminish tracer uptake. However, the presence of normal or rarely increased uptake favors type 1 AIT.

In AIT, the drug should be stopped, if possible, although this is often impractical because of the underlying cardiac disorder. Discontinuation of amiodarone will not have an acute effect because of its storage and prolonged half-life. High doses of antithyroid drugs can be used in type 1 AIT but are often ineffective. Potassium perchlorate, 200 mg every 6 h, has been used to reduce thyroidal iodide content. Perchlorate treatment has been associated with agranulocytosis, although the risk appears relatively low with short-term use. Glucocorticoids, as administered for subacute thyroiditis, have modest benefit in type 2 AIT. Lithium blocks thyroid hormone release and can also provide some benefit. Near-total thyroidectomy rapidly decreases thyroid hormone levels and may be the most effective long-term solution if the patient can undergo the procedure safely.

FURTHER READING

Biondi B et al: The 2015 European Thyroid Association Guidelines on diagnosis and treatment of endogenous subclinical hyperthyroidism. Eur Thyroid J 4:149, 2015.

De Leo S et al: Hyperthyroidism. Lancet 388:906, 2016.

Hanley P et al: Thyroid disorders in children and adolescents: A review. JAMA Pediatr 170:1008, 2016.

Ross DS et al: 2016 American Thyroid Association guidelines for diagnosis and management of hyperthyroidism and other causes of thyrotoxicosis. Thyroid 26:1343, 2016.

Wiersinga WM: Advances in treatment of active, moderate-to-severe Graves' ophthalmopathy. Lancet Diabetes Endocrinol 5:134, 2017.

378 Thyroid Nodular Disease and Thyroid Cancer

J. Larry Jameson, Susan J. Mandel, Anthony P. Weetman

GOITER AND NODULAR THYROID DISEASE

Goiter refers to an enlarged thyroid gland. Biosynthetic defects, iodine deficiency, autoimmune disease, and nodular diseases can each lead to goiter, although by different mechanisms. Biosynthetic defects and iodine deficiency are associated with reduced efficiency of thyroid hormone synthesis, leading to increased thyroid-stimulating hormone (TSH), which stimulates thyroid growth as a compensatory mechanism to overcome the block in hormone synthesis. Graves' disease and Hashimoto's thyroiditis are also associated with goiter. In Graves' disease, the goiter results mainly from the TSH-R–mediated effects of thyroid-stimulating immunoglobulins. The goitrous form of Hashimoto's thyroiditis occurs because of acquired defects in hormone synthesis, leading to elevated levels of TSH and its consequent growth effects. Lymphocytic infiltration and immune system–induced growth factors also contribute to thyroid enlargement in Hashimoto's thyroiditis.

Thyroid nodular disease is characterized by the disordered growth of thyroid cells, which can be either hyperplastic or neoplastic. A patient may have a multinodular goiter (MNG) in which thyroid nodules (generally hyperplastic) replace the majority of the normal thyroid parenchyma; this presentation is more common in areas of borderline iodine deficiency. Or, the thyroid gland may be normal in size and contain discrete thyroid nodules. Because the management of goiter depends on the etiology, the detection of thyroid enlargement on physical examination should prompt further evaluation to identify its cause.

Nodular thyroid disease is common, occurring in about 3–7% of adults when assessed by physical examination. Using ultrasound, nodules are present in up to 50% of adults, with the majority being <1 cm in diameter. Thyroid nodules may be solitary or multiple, and they may be functional or nonfunctional.

DIFFUSE NONTOXIC (SIMPLE) GOITER

Etiology and Pathogenesis When diffuse enlargement of the thyroid occurs in the absence of nodules and hyperthyroidism, it is referred to as a *diffuse nontoxic goiter*. This is sometimes called *simple goiter*, because of the absence of nodules, or *colloid goiter*, because of the presence of uniform follicles that are filled with colloid. Worldwide, diffuse goiter is most commonly caused by iodine deficiency and is

termed *endemic goiter* when it affects >5% of the population. In nonendemic regions, *sporadic goiter* occurs, and the cause is usually unknown. Thyroid enlargement in teenagers is sometimes referred to as *juvenile goiter*. In general, goiter is more common in women than men, probably because of the greater prevalence of underlying autoimmune disease and the increased iodine demands associated with pregnancy.

In *iodine-deficient areas*, thyroid enlargement reflects a compensatory effort to trap iodide and produce sufficient hormone under conditions in which hormone synthesis is relatively inefficient. Somewhat surprisingly, TSH levels are usually normal or only slightly increased, suggesting increased sensitivity to TSH or activation of other pathways that lead to thyroid growth. Iodide appears to have direct actions on thyroid vasculature and may indirectly affect growth through vasoactive substances such as endothelins and nitric oxide. Endemic goiter may also be caused by exposure to environmental *goitrogens* such as cassava root, which contains a thiocyanate; vegetables of the Cruciferae family (known as cruciferous vegetables) (e.g., Brussels sprouts, cabbage, and cauliflower); and milk from regions where goitrogens are present in grass. Although relatively rare, inherited defects in thyroid hormone synthesis lead to a diffuse nontoxic goiter. Abnormalities at each step of hormone synthesis, including iodide transport (NIS), transgenic (Tg) synthesis, organification and coupling (thrombopoietin [TPO]), and the regeneration of iodide (dehalogenase), have been described.

CLINICAL MANIFESTATIONS AND DIAGNOSIS

If thyroid function is preserved, most goiters are asymptomatic. Examination of a diffuse goiter reveals a symmetrically enlarged, nontender, generally soft gland without palpable nodules. Goiter is defined, somewhat arbitrarily, as a lateral lobe with a volume greater than the thumb of the individual being examined. On ultrasound, total thyroid volume exceeding 30 mL is considered abnormal. If the thyroid is markedly enlarged, it can cause tracheal or esophageal compression. These features are unusual, however, in the absence of nodular disease and fibrosis. *Substernal goiter* may obstruct the thoracic inlet. *Pemberton's sign* refers to facial and neck congestion due to jugular venous obstruction when the arms are raised above the head, a maneuver that draws the thyroid into the thoracic inlet. Respiratory flow measurements and CT or MRI should be used to evaluate substernal goiter in patients with obstructive signs or symptoms.

Thyroid function tests should be performed in all patients with goiter to exclude thyrotoxicosis or hypothyroidism. It is not unusual, particularly in iodine deficiency, to find a low total T_4, with normal T_3 and TSH, reflecting enhanced $T_4 \rightarrow T_3$ conversion. A low TSH with a normal free T_3 and free T_4, particularly in older patients, suggests the possibility of thyroid autonomy or undiagnosed Graves' disease, and is termed *subclinical thyrotoxicosis*. The benefit of treatment (typically with radioiodine) in subclinical thyrotoxicosis, versus follow-up and implementing treatment if free T_3 or free T_4 levels become abnormal, is unclear, but treatment is increasingly recommended in the elderly to reduce the risk of atrial fibrillation and bone loss. TPO antibodies may be useful to identify patients at increased risk of autoimmune thyroid disease. Low urinary iodine levels (<50 μg/L) support a diagnosis of iodine deficiency. Thyroid scanning is not generally necessary but will reveal increased uptake in iodine deficiency and most cases of dyshormonogenesis.

TREATMENT

Diffuse Nontoxic (Simple) Goiter

Iodine replacement induces variable regression of goiter in iodine deficiency, depending on duration and the degree of hyperplasia, with accompanying fibrosis, and autonomous function that may have developed. Surgery is rarely indicated for diffuse goiter. Exceptions include documented evidence of tracheal compression or obstruction of the thoracic inlet, which are more likely to be associated with substernal MNGs (see below). Subtotal or near-total thyroidectomy for these or cosmetic reasons should be performed by an experienced surgeon to minimize complication rates. Surgery should be followed by replacement with levothyroxine, with the aim of keeping the TSH level at the lower end of the reference interval to prevent regrowth of the goiter.

NONTOXIC MULTINODULAR GOITER

Etiology and Pathogenesis Depending on the population studied, MNG or the presence of nodules in a thyroid of normal size occurs in up to 12% of adults. MNG should be distinguished from the presence of nodules in normal size thyroid gland (see "Approach to the Patient with Thyroid Nodules"). MNG is more common in women than men and increases in prevalence with age. It is more common in iodine-deficient regions but also occurs in regions of iodine sufficiency, reflecting multiple genetic, autoimmune, and environmental influences on the pathogenesis.

There is typically wide variation in nodule size. Histology reveals a spectrum of morphologies ranging from hypercellular, hyperplastic regions to cystic areas filled with colloid. Fibrosis is often extensive, and areas of hemorrhage or lymphocytic infiltration may be seen. Using molecular techniques, most nodules within an MNG are polyclonal in origin, suggesting a hyperplastic response to locally produced growth factors and cytokines. TSH, which is usually not elevated, may play a permissive or contributory role. Monoclonal neoplastic lesions may also occur, reflecting mutations in genes that confer a selective growth advantage to the progenitor cell.

Clinical Manifestations Most patients with nontoxic MNG are asymptomatic and euthyroid. MNG typically develops over many years and is detected on routine physical examination, when an individual notices an enlargement in the neck, or as an incidental finding on imaging. If the goiter is large enough, it can ultimately lead to compressive symptoms including difficulty swallowing, respiratory distress (tracheal compression), or plethora (venous congestion), but these symptoms are uncommon. Symptomatic MNGs are usually extraordinarily large and/or develop fibrotic areas that cause compression. Sudden pain in an MNG is usually caused by hemorrhage into a nodule but should raise the possibility of invasive malignancy. Hoarseness, reflecting laryngeal nerve involvement, also suggests malignancy.

Diagnosis On examination, thyroid architecture is distorted, and multiple nodules of varying size can be appreciated. Because many nodules are deeply embedded in thyroid tissue or reside in posterior or substernal locations, it is not possible to palpate all nodules. Pemberton's sign, characterized by facial suffusion when the patient's arms are elevated above the head, suggests that the goiter has increased pressure in the thoracic inlet. A TSH level should be measured to exclude subclinical hyper- or hypothyroidism, but thyroid function is usually normal. Tracheal deviation is common, but compression must usually exceed 70% of the tracheal diameter before there is significant airway compromise. Pulmonary function testing can be used to assess the functional effects of compression, which characteristically causes inspiratory stridor. CT or MRI can be used to evaluate the anatomy of the goiter and the extent of substernal extension or tracheal narrowing. A barium swallow may reveal the extent of esophageal compression. The risk of malignancy in MNG is similar to that in solitary nodules. Ultrasonography can be used to identify which nodules should be biopsied based on a combination of size and sonographic features **(Table 378-1)** **(Chap. 375)**. For nodules with more suspicious sonographic patterns (e.g., hypoechoic solid nodules with infiltrative borders), biopsy is recommended at a lower size cutoff than those with less suspicious imaging features **(Fig. 378-1)**.

TREATMENT

Nontoxic Multinodular Goiter

Most nontoxic MNGs can be managed conservatively. T_4 suppression is rarely effective for reducing goiter size and introduces the risk of subclinical or overt thyrotoxicosis, particularly if there is

TABLE 378-1 Grayscale Sonographic Features Associated with Thyroid Cancer

	MEDIAN SENSITIVITY [RANGE]	MEDIAN SPECIFICITY [RANGE]
Hypoechoic compared with surrounding thyroid	81% [48–90%]	53% [36–92%]
Marked hypoechogenicity	41% [27–59%]	94% [92–94%]
Microcalcifications	44% [26–73%]	89% [69–98%]
Irregular, microlobulated margins	55% [17–84%]	79% [62–85%]
Solid consistency	86% [78–91%]	48% [30–58%]
Taller than wide shape on transverse view	48% [33–84%]	92% [82–93%]

A

B

FIGURE 378-1 Sonographic patterns of thyroid nodules. ***A.*** High suspicion ultrasound pattern for thyroid malignancy (hypoechoic solid nodule with irregular borders and microcalcifications). ***B.*** Very low suspicion ultrasound pattern for thyroid malignancy (spongiform nodule with microcystic areas comprising over >50% of nodule volume).

underlying autonomy or if it develops during treatment. Contrast agents and other iodine-containing substances should be avoided because of the risk of inducing the *Jod-Basedow effect*, characterized by enhanced thyroid hormone production by autonomous nodules. Radioiodine has been used when surgery is contraindicated in areas where large nodular goiters are more prevalent (e.g., some areas of Europe and Brazil) because it can decrease MNG volume and may selectively ablate regions of autonomy. Dosage of ^{131}I depends on the size of the goiter and radioiodine uptake but is usually about 3.7 MBq (0.1 mCi) per gram of tissue, corrected for uptake (typical dose 370–1070 MBq [10–29 mCi]). Repeat treatment may be needed and effectiveness may be increased by concurrent administration of low-dose recombinant TSH (0.1 mg IM). It is possible to achieve a 40–50% reduction in goiter size in most patients. Earlier concerns about radiation-induced thyroid swelling and tracheal compression have diminished, as studies have shown this complication to be rare. When acute compression occurs, glucocorticoid treatment or surgery may be needed. Radiation-induced hypothyroidism is less common than after treatment for Graves' disease. However, posttreatment autoimmune thyrotoxicosis may occur in up to 5% of patients treated for nontoxic MNG. Surgery remains highly effective but is not without risk, particularly in older patients with underlying cardiopulmonary disease.

TOXIC MULTINODULAR GOITER

The pathogenesis of toxic MNG appears to be similar to that of nontoxic MNG; the major difference is the presence of functional autonomy in toxic MNG. The molecular basis for autonomy in toxic MNG remains unknown. As in nontoxic goiters, many nodules are polyclonal, whereas others are monoclonal and vary in their clonal origins. Genetic abnormalities known to confer functional autonomy, such as activating TSH-R or $G_{s\alpha}$ mutations (see below), are not usually found in the autonomous regions of toxic MNG goiter.

In addition to features of goiter, the clinical presentation of toxic MNG includes subclinical or mild overt hyperthyroidism. The patient is usually elderly and may present with atrial fibrillation or palpitations, tachycardia, nervousness, tremor, or weight loss. Recent exposure to iodine, from contrast dyes or other sources, may precipitate or exacerbate thyrotoxicosis. The TSH level is low. The uncombined T_4 level may be normal or minimally increased; T_3 is often elevated to a greater degree than T_4. Thyroid scan shows heterogeneous uptake with multiple regions of increased and decreased uptake; 24-h uptake of radioiodine may not be increased but is usually in the upper normal range.

Prior to definitive treatment of the hyperthyroidism, ultrasound imaging should be performed to assess the presence of discrete nodules corresponding to areas of decreased uptake ("cold" nodules). If present, fine-needle aspiration (FNA) may be indicated based on sonographic patterns and size cutoffs. The cytology results, if indeterminate or suspicious, may direct the therapy to surgery.

TREATMENT

Toxic Multinodular Goiter

Antithyroid drugs normalize thyroid function and are particularly useful in the elderly or ill patients with limited lifespan. In contrast to Graves' disease, spontaneous remission does not occur and so treatment is long-term. Radioiodine is generally the treatment of choice; it treats areas of autonomy as well as decreasing the mass of the goiter by ablating the functioning nodules. Sometimes, however, a degree of autonomy may persist, presumably because multiple autonomous regions may emerge after others are treated, and further radioiodine treatment may be necessary. Surgery provides definitive treatment of underlying thyrotoxicosis as well as goiter. Patients should be rendered euthyroid using an antithyroid drug before operation.

HYPERFUNCTIONING SOLITARY NODULE

A solitary, autonomously functioning thyroid nodule is referred to as *toxic adenoma*. The pathogenesis of this disorder has been unraveled by demonstrating the functional effects of mutations that stimulate the TSH-R signaling pathway. Most patients with solitary hyperfunctioning nodules have acquired somatic, activating mutations in the TSH-R **(Fig. 378-2)**. These mutations, located primarily in the receptor transmembrane domain, induce constitutive receptor coupling to $G_{s\alpha}$, increasing cyclic adenosine monophosphate (AMP) levels and leading to enhanced thyroid follicular cell proliferation and function. Less commonly, somatic mutations are identified in $G_{s\alpha}$. These mutations, which are similar to those seen in McCune-Albright syndrome **(Chap. 405)** or in a subset of somatotrope adenomas **(Chap. 373)**, impair guanosine triphosphate (GTP) hydrolysis, causing constitutive activation of the cyclic AMP signaling pathway. In most series, activating mutations in either the TSH-R or the $G_{s\alpha}$ subunit genes are identified in >90% of patients with solitary hyperfunctioning nodules.

Thyrotoxicosis is usually mild and is generally only detected when a nodule is >3 cm. The disorder is suggested by a subnormal TSH level; the presence of the thyroid nodule, often large enough to be palpable; and the absence of clinical features suggestive of Graves' disease or other causes of thyrotoxicosis. A thyroid scan provides a definitive diagnostic test, demonstrating focal uptake in the hyperfunctioning nodule and diminished uptake in the remainder of the gland, as activity of the normal thyroid is suppressed.

TREATMENT

Hyperfunctioning Solitary Nodule

Radioiodine ablation is usually the treatment of choice. Because normal thyroid function is suppressed, ^{131}I is concentrated in the hyperfunctioning nodule with minimal uptake and damage to normal thyroid tissue. Relatively large radioiodine doses (e.g., 370–1110 MBq [10–29.9 mCi] ^{131}I) have been shown to correct thyrotoxicosis in about 75% of patients within 3 months. Hypothyroidism occurs in <10% of those patients over the next 5 years. Surgical resection is also effective and is usually limited to lobectomy, thereby preserving thyroid function and minimizing risk of hypoparathyroidism or damage to the recurrent laryngeal nerves. Medical therapy using antithyroid drugs and beta blockers can normalize thyroid function but is not an optimal long-term treatment. Using ultrasound guidance, repeated percutaneous radiofrequency thermal ablation has been used successfully in some centers to ablate hyperfunctioning nodules, and this technique has also been used to reduce the size of nonfunctioning thyroid nodules.

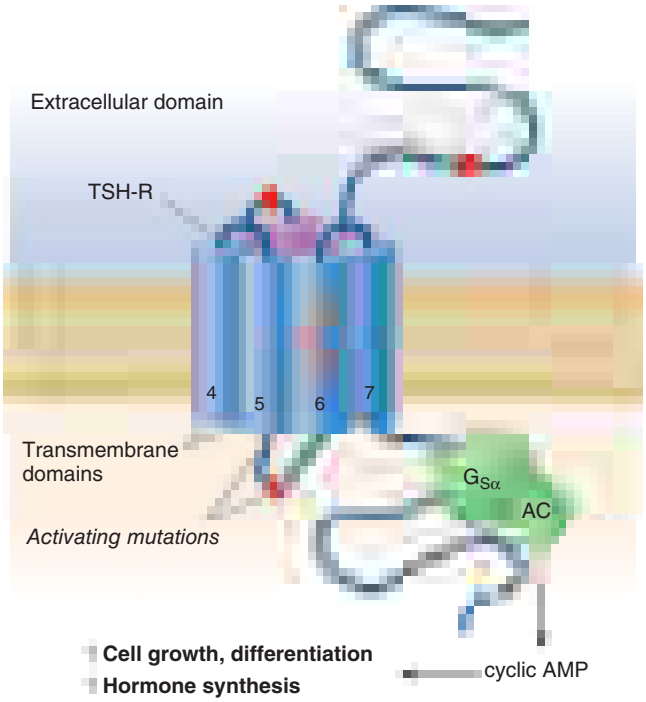


FIGURE 378-2 Activating mutations of the thyroid-stimulating hormone receptor (TSH-R). Mutations (*) that activate TSH-R reside mainly in transmembrane 5 and intracellular loop 3, although mutations have occurred in a variety of different locations. The effect of these mutations is to induce conformational changes that mimic TSH binding, thereby leading to coupling to stimulatory G protein ($G_{s\alpha}$) and activation of adenylate cyclase (AC), an enzyme that generates cyclic AMP.

BENIGN LESIONS

The various types of benign thyroid nodules are listed in **Table 378-2**. Benign nodules may be hyperplastic and reflect a combination of both macro- and microfollicular architecture or they may be neoplastic, encapsulated adenomas that generally have a more monotonous microfollicular pattern. If the adenoma is composed of oncocytic follicular cells arranged in a follicular pattern, this is termed a Hürthle cell adenoma. Hyperplastic nodules generally appear as mixed cystic/solid or spongiform lesions on ultrasound. The definition of spongiform requires the presence of microcystic areas comprising >50% of the nodule volume, with the concept that this microcystic sonographic pattern recapitulates the histology of macrofollicles containing colloid **(Fig. 378-1*B*)**. However, the majority of solid nodules (whether hypo-, iso-, or hyperechoic) are also benign. FNA, usually performed with ultrasound guidance, is the diagnostic procedure of choice to evaluate thyroid nodules (see the "Approach to the Patient with Thyroid Nodules" section). Pure thyroid cysts, <1% of all thyroid growths, consist of colloid and are benign as well. Cysts frequently recur, even

TABLE 378-2 Classification of Thyroid Growths

Benign	
Hyperplasia	
Colloid nodule	
Follicular epithelial cell adenomas	
Conventional	
Oncocytic (Hürthle cell)	
Malignant	**Approximate Prevalence, %**
Follicular epithelial cell	
Papillary carcinomas	80–85
Classic variant	
Follicular variant	
Diffuse sclerosing variant	
Tall cell, columnar cell variants	
Follicular carcinomas	2-5–7
Conventional	
Oncocytic (Hürthle cell)	
Poorly differentiated carcinomas	3–5
Anaplastic (undifferentiated) carcinomas	1
C cell origin(calcitonin-producing)	
Medullary thyroid cancer	<10
Sporadic	
Familial	
MEN 2	
Other malignancies	
Lymphomas	1
Metastases	
Breast, melanoma, lung, kidney	
Others	

Abbreviation: MEN, multiple endocrine neoplasia.

after repeated aspiration, and may require surgical excision if they are large. Ethanol ablation to sclerose the cyst has been used successfully for patients who are symptomatic.

TSH suppression with levothyroxine therapy does not decrease thyroid nodule size in iodine-sufficient populations. However, if there is relative iodine deficiency, both iodine and levothyroxine therapy have been demonstrated to decrease nodule volume. If levothyroxine is administered in this situation, the TSH should be maintained at or just below the lower limit of normal, but not frankly suppressed. If the nodule has not decreased in size after 6–12 months of therapy, treatment should be discontinued because little benefit is likely to accrue from long-term treatment; the risk of iatrogenic subclinical thyrotoxicosis should also be considered.

THYROID CANCER

Thyroid carcinoma is the most common malignancy of the endocrine system. Malignant tumors derived from the follicular epithelium are classified according to histologic features. Differentiated tumors, such as papillary thyroid cancer (PTC) or follicular thyroid cancer (FTC), are often curable, and the prognosis is good for patients identified with early-stage disease. In contrast, anaplastic thyroid cancer (ATC) is aggressive, responds poorly to treatment, and is associated with a bleak prognosis.

Over the last 30 years, the incidence of thyroid cancer has increased from 4.9 to 14.3 cases per 100,000 individuals in the United States, with over 65,000 cases diagnosed in 2015. However, disease-specific mortality has not changed. The increased incidence is predominantly attributable to small T1 papillary cancer tumors (<2 cm), and has led experts to consider that thyroid cancer is being overdiagnosed, suggesting that cancers are being detected that would otherwise be unlikely to harm a patient. The concept of cancer overdiagnosis is predicated upon the presence of a disease reservoir (the autopsy prevalence of PTC is ~25%), activities leading to disease detection (increased diagnostic imaging with incidental detection of nodules), and a mismatch in the directional rate between diagnosis and mortality (thyroid cancer disease-specific mortality not changed in 40 years). Similar trends have been observed worldwide, especially in those countries with higher proportion of privately financed healthcare, leading to increased resource utilization including imaging. The 20-year disease-specific mortality for low risk thyroid cancer is 1%.

Current trends in thyroid cancer care focus on: (1) avoiding overdiagnosis by limiting FNA by sonographic risk stratification with size cut offs; (2) limiting surgery, radioiodine, and subsequent surveillance for low risk tumors; (3) identifying patients at higher recurrence risk for more aggressive treatment and monitoring. Prognosis is worse in older persons (>65 years). Thyroid cancer is twice as common in women as men, but male gender is associated with a worse prognosis. Additional important risk factors include a history of childhood (before age 18) head or neck irradiation, evidence for local tumor fixation or gross metastatic involvement of lymph nodes, and the presence of distant metastases **(Table 378-3)**.

Several unique features of thyroid cancer facilitate its management: (1) thyroid nodules are amenable to biopsy by FNA; (2) iodine radioisotopes can be used to diagnose (^{123}I and ^{131}I) and potentially treat (^{131}I) differentiated thyroid cancer, reflecting the unique uptake of this anion by the thyroid gland; and (3) serum markers allow the detection of residual or recurrent disease, including the use of Tg levels for PTC and FTC, and calcitonin for medullary thyroid cancer (MTC).

TABLE 378-3 Risk Factors for Thyroid Carcinoma in Patients with Thyroid Nodule from History and Physical Examination

History of head and neck irradiation before the age of 18, including, mantle radiation for Hodgkin's disease, and brain radiation for childhood leukemia or other cranial malignancies	Family history of papillary thyroid cancer in 2 or more first degree relatives, MEN 2, or other genetic syndromes associated with thyroid malignancy (e.g., Cowden's syndrome, familial polyposis, Carney complex, PTEN [phosphatase and tensin homolog] hamartoma tumor)
Exposure to ionizing radiation from fallout in childhood or adolescence	Vocal cord paralysis, hoarse voice
Age <20 or >65 years	Nodule fixed to adjacent structures
Rapidly enlarging neck mass	Lateral cervical lymphadenopathy
Male gender	

Abbreviation: MEN, multiple endocrine neoplasia.

TABLE 378-4 Thyroid Cancer Classification[a]

Papillary or Follicular Thyroid Cancers		
	<45 years	>45 years
Stage I	Any T, any N, M0	T1, N0, M0
Stage II	Any T, any N, M1	T2, N0, M0
Stage III	—	T3, N0, M0 T1–T3, N1a, M0
Stage IVA	—	T4a, any N, M0 T1–T3, N1b, M0
Stage IVB		T4b, any N, M0
Stage IVC		Any T, any N, M1
Anaplastic Thyroid Cancer		
Stage IV	All cases are stage IV	
Medullary Thyroid Cancer		
Stage I	T1, N0, M0	
Stage II	T2 or T3, N0, M0	
Stage III	T1–T3, N1a, M0	
Stage IVA	T4a, any N, M0 T1–T3, N1b, M0	
Stage IVB	T4b, any N, M0	
Stage IVC	Any T, any N, M1	

[a]Criteria include: T, the size and extent of the primary tumor (T1a ≤1 cm; T1b >1 cm but ≤2 cm; T2 >2 cm but ≤4 cm; T3 >4 cm or any tumor with extension into perithyroidal soft tissue or sternothyroid muscle; T4a invasion into subcutaneous soft tissues, larynx, trachea, esophagus, or recurrent laryngeal nerve; T4b invasion into prevertebral fascia or encasement of carotid artery or mediastinal vessels); N, the absence (N0) or presence (N1a level IV central compartment; N1b levels II–V lateral compartment, upper mediastinal or retro/parapharyngeal) of regional node involvement; M, the absence (M0) or presence (M1) of distant metastases.

Note that updated TNM classification was released in early 2018.

Source: American Joint Committee on Cancer staging system for thyroid cancers using the TNM classification, 7th edition.

CLASSIFICATION

Thyroid neoplasms can arise in each of the cell types that populate the gland, including thyroid follicular cells, calcitonin-producing C cells, lymphocytes, and stromal and vascular elements, as well as metastases from other sites (Table 378-2). The American Joint Committee on Cancer (AJCC) has designated a staging system using the tumor, node, metastasis (TNM) classification, which is most commonly used **(Table 378-4)**. Revised guidelines released in 2018 changed the age cutoffs from <45 to <55 years, as well as some of the staging criteria (see Further Reading).

PATHOGENESIS AND GENETIC BASIS

Radiation Early studies of the pathogenesis of thyroid cancer focused on the role of external radiation, which predisposes to chromosomal breaks, leading to genetic rearrangements and loss of tumor-suppressor genes. External radiation of the mediastinum, face, head, and neck region was administered in the past to treat an array of conditions, including acne and enlargement of the thymus, tonsils, and adenoids. Radiation exposure increases the risk of benign and malignant thyroid nodules, is associated with multicentric cancers, and shifts the incidence of thyroid cancer to an earlier age group. Radiation from nuclear fallout also increases the risk of thyroid cancer. Children seem more predisposed to the effects of radiation than adults.

TSH and Growth Factors Many differentiated thyroid cancers express TSH receptors and, therefore, remain responsive to TSH. Higher serum TSH levels, even within normal range, are associated

with increased thyroid cancer risk in patients with thyroid nodules. These observations provide the rationale for T_4 suppression of TSH in patients with thyroid cancer. Residual expression of TSH receptors also allows TSH-stimulated uptake of ^{131}I therapy (see below).

Oncogenes and Tumor-Suppressor Genes Thyroid cancers are monoclonal in origin, consistent with the idea that they originate as a consequence of mutations that confer a growth advantage to a single cell. In addition to increased rates of proliferation, some thyroid cancers exhibit impaired apoptosis and features that enhance invasion, angiogenesis, and metastasis. Thyroid neoplasms have been analyzed for a variety of genetic alterations, but without clear evidence of an ordered acquisition of somatic mutations as they progress from the benign to the malignant state. On the other hand, certain mutations, such as *RET/PTC* and *PAX8-PPARγ1* rearrangements, are relatively specific for thyroid neoplasia.

As described above, activating mutations of the TSH-R and the $G_{s\alpha}$ subunit are associated with autonomously functioning nodules. Although these mutations induce thyroid cell growth, this type of nodule is almost always benign.

Activation of the RET-RAS-BRAF signaling pathway is seen in up to 70% of PTCs, although the types of mutations are heterogeneous. A variety of rearrangements involving the *RET* gene on chromosome 10 bring this receptor tyrosine kinase under the control of other promoters, leading to receptor overexpression. *RET* rearrangements occur in 20–40% of PTCs in different series and were observed with increased frequency in tumors developing after the Chernobyl radiation accident. Rearrangements in PTC have also been observed for another tyrosine kinase gene, *TRK1*, which is located on chromosome 1. To date, the identification of PTC with *RET* or *TRK1* rearrangements has not proven useful for predicting prognosis or treatment responses. *BRAF V600E* mutations appear to be the most common genetic alteration in PTC. These mutations activate the kinase, which stimulates the mitogen-activated protein kinase (MAPK) cascade. *RAS* mutations, which also stimulate the MAPK cascade, are found in about 20–30% of thyroid neoplasms (*NRAS* > *HRAS* > *KRAS*), including both PTC follicular variant and FTC. Of note, simultaneous *RET*, *BRAF*, and *RAS* mutations rarely occur in the same tumor, suggesting that activation of the MAPK cascade is critical for tumor development, independent of the step that initiates the cascade.

RAS mutations also occur in FTCs. In addition, a rearrangement of the thyroid developmental transcription factor PAX8 with the nuclear receptor PPARγ is identified in a significant fraction of FTCs. Overall, about 70% of follicular cancers have mutations or genetic rearrangements. Loss of heterozygosity of 3p or 11q, consistent with deletions of tumor-suppressor genes, is also common in FTCs.

Most of the mutations seen in differentiated thyroid cancers have also been detected in ATCs. *TERT* promoter mutations occur in <10% of differentiated PTC but are more common in ATC. *BRAF* mutations are seen in up to 50% of ATCs. Mutations in *CTNNB1*, which encodes β-catenin, occur in about two-thirds of ATCs, but not in PTC or FTC. Mutations of the tumor-suppressor *P53* also play an important role in the development of ATC. Because *P53* plays a role in cell cycle surveillance, DNA repair, and apoptosis, its loss may contribute to the rapid acquisition of genetic instability as well as poor treatment responses **(Chap. 68)**.

The role of molecular diagnostics in the clinical management of thyroid cancer is under investigation. In principle, analyses of specific mutations might aid in classification, prognosis, or choice of treatment. Although *BRAF V600E* mutations are associated with loss of iodine uptake by tumor cells, there is no clear evidence to date that this information alters clinical decision-making. Higher recurrence rates have been variably reported in patients with *BRAF*-positive PTC, but the impact on survival rates is unclear.

MTC, when associated with multiple endocrine neoplasia (MEN) type 2, harbors an inherited mutation of the *RET* gene. Unlike the rearrangements of *RET* seen in PTC, the mutations in MEN 2 are point mutations that induce constitutive activity of the tyrosine kinase **(Chap. 381)**. MTC is preceded by hyperplasia of the C cells, raising the likelihood that as-yet-unidentified "second hits" lead to cellular transformation. A subset of sporadic MTC contains somatic mutations that activate *RET*.

WELL-DIFFERENTIATED THYROID CANCER

Papillary PTC is the most common type of thyroid cancer, accounting for 80–85% of well-differentiated thyroid malignancies. Microscopic PTC is present in up to 25% of thyroid glands at autopsy, but most of these lesions are very small (several millimeters) and are not clinically significant. Characteristic cytologic features of PTC help make the diagnosis by FNA or after surgical resection; these include, large, clear nuclei with powdery chromatin (described as an "orphan Annie eye" appearance) with nuclear grooves and prominent nucleoli. The histologic finding of these cells arranged in either papillary structures versus follicles distinguishes the classic and follicular variants of PTC, respectively.

PTC may be multifocal and invade locally within the thyroid gland as well as through the thyroid capsule and into adjacent structures in the neck. It has a propensity to spread via the lymphatic system but can metastasize hematogenously as well, particularly to bone and lung. Because of the relatively slow growth of the tumor, a significant burden of pulmonary metastases may accumulate, sometimes with remarkably few symptoms. The prognostic implication of lymph node spread depends upon the volume of metastatic disease. Micrometastases, defined as <2 mm of cancer in a lymph node, do not affect prognosis. However, gross metastatic involvement of multiple 2–3 cm lymph nodes indicates a 25–30% chance of recurrence, and may increase mortality in older patients. The staging of PTC by the TNM system is outlined in Table 378-4. Most papillary cancers are identified in the early stages (>80% stages I or II) and have an excellent prognosis, with survival curves similar to expected survival **(Fig. 378-3)**. Mortality is markedly increased in stage IV disease, especially in the presence of distant metastases (stage IV-C), but this group comprises only about 1% of patients. The treatment of PTC is described below.

Follicular The incidence of FTC varies widely in different parts of the world; it is more common in iodine-deficient regions. Currently, FTC accounts for only about 5% of all thyroid cancers diagnosed in the United States. FTC is difficult to diagnose by FNA because the distinction between benign and malignant follicular neoplasms requires histology because the nuclear features of follicular adenomas and carcinomas do not differ. Rather, follicular carcinoma is diagnosed by the presence of capsular and/or vascular invasion. FTC tends to spread by hematogenous routes leading to bone, lung, and central nervous system metastases. Mortality rates associated with angioinvasive FTC are less favorable than for PTC, in part because a larger proportion of patients present with stage IV disease. Poor prognostic features include

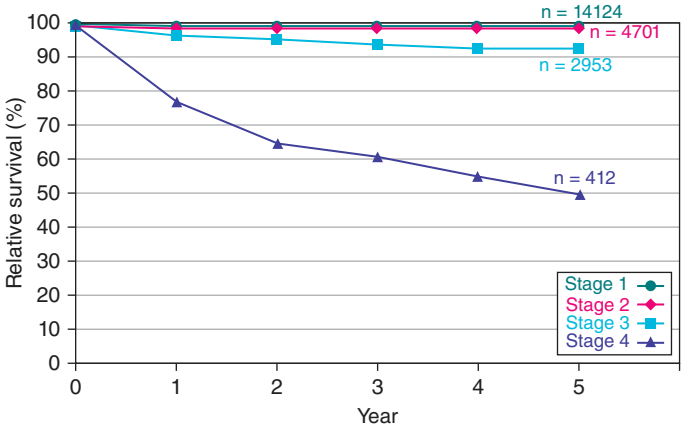


FIGURE 378-3 Survival rates of patients with different stages of papillary cancer. *(Adapted with permission from SB Edge, DR Byrd: Thyroid, in CC Compton, AB Fritz, FL Greene, A Trotti [eds]: AJCC Cancer Staging Manual, 7th ed. New York, Springer, 2010, pp. 87–92.)*

distant metastases, age >50 years, primary tumor size >4 cm, Hürthle cell histology, and the presence of marked vascular invasion.

TREATMENT

Well-Differentiated Thyroid Cancer Surgery

All well-differentiated thyroid cancers >1cm (T1b or larger) should be surgically excised although active surveillance may be an option for small intrathyroidal micropapillary thyroid cancers (T1a) without metastases. In addition to removing the primary lesion, surgery allows accurate histologic diagnosis and staging. Because there is no compelling evidence that bilateral thyroid surgery improves survival, the initial surgical procedure may be either a unilateral (lobectomy) or bilateral (near total thyroidectomy) procedure for patients with intrathyroidal cancers >1 cm and <4 cm (T1b and T2 tumors) in the absence of metastatic disease. For patients at high risk for recurrence, bilateral surgery allows administration of radioiodine for remnant ablation and potential treatment of iodine-avid metastases, if indicated, as well as for monitoring of serum Tg levels. Therefore, near-total thyroidectomy is appropriate for tumors >4 cm or in the presence of metastases or clinical evidence of extrathyroidal invasion. In addition, for patients found to have a high risk tumor after lobectomy based upon aggressive pathology features (e.g., vascular invasion or a less differentiated subtype), completion surgery should be performed. Surgical complication rates are acceptably low if the surgeon is highly experienced in the procedure. Preoperative sonography should be performed in all patients to assess the central and lateral cervical lymph node compartments for suspicious adenopathy, which if present, should undergo FNA and be removed, as indicated, at surgery.

TSH SUPPRESSION THERAPY

Because most tumors are still TSH-responsive, levothyroxine suppression of TSH is a mainstay of thyroid cancer treatment. Although TSH suppression clearly provides therapeutic benefit, there are no prospective studies that define the optimal level of TSH suppression. The degree of TSH suppression should be individualized based on a patient's risk of recurrence. It should be adjusted over time as surveillance blood tests and imaging confirm absence of disease or, alternatively, indicate possible residual/recurrent cancer. For patients at low risk of recurrence, TSH should be maintained in the lower normal limit (0.5–2.0 mIU/L). For patients either at intermediate or high risk of recurrence, TSH levels should be kept to 0.1 to 0.5 mIU/L and <0.1 mIU/L, respectively, if there are no strong contraindications to mild thyrotoxicosis. TSH should be <0.1 mIU/L for those with known metastatic disease.

RADIOIODINE TREATMENT

After near-total thyroidectomy, <1 gm of thyroid tissue remains in the thyroid bed. Postsurgical radioablation of the remnant thyroid eliminates residual normal thyroid, facilitating the use of Tg determinations. In addition, well-differentiated thyroid cancer often incorporates radioiodine, although less efficiently than normal thyroid follicular cells. Radioiodine uptake is determined primarily by expression of the NIS and is stimulated by TSH, requiring expression of the TSH-R. The retention time for radioactivity is influenced by the extent to which the tumor retains differentiated functions such as iodide trapping and organification. Consequently, for patients at higher risk of recurrence and for those with known distant metastatic disease, ^{131}I therapy may provide an adjuvant role and potentially treat residual tumor cells.

Indications Not all patients benefit from radioiodine therapy. Neither recurrence nor survival rates are improved in stage I patients with T1 tumors (≤2 cm) confined to the thyroid. No benefit has been demonstrated for larger (>2 cm but <4 cm) low-risk tumors, such as minimally invasive follicular cancer or encapsulated PTC follicular variant. However, in higher risk patients (larger tumors, more aggressive variants of papillary cancer, tumor vascular invasion, extrathyroidal invasion, presence of large-volume lymph node metastases), radioiodine reduces recurrence and may increase survival for older patients.

^{131}I Thyroid Ablation and Treatment As noted above, the decision to use ^{131}I for thyroid ablation should be coordinated with the surgical approach, because radioablation is much more effective when there is minimal remaining normal thyroid tissue. Radioiodine is administered after iodine depletion (patient follows a low-iodine diet for 1–2 weeks) and in the presence of elevated serum TSH levels to stimulate uptake of the isotope into both the remnant and potentially any residual tumor. To achieve high serum TSH levels, there are two approaches. A patient may be withdrawn from thyroid hormone so that endogenous TSH is secreted and, ideally, the serum TSH level is >25 mIU/L at the time of ^{131}I therapy. A typical strategy is to treat the patient for several weeks postoperatively with liothyronine (25 µg qd or bid), followed by thyroid hormone withdrawal for 2 weeks. Alternatively, recombinant human TSH (rhTSH) is administered as two daily consecutive injections (0.9 mg) with administration of ^{131}I 24 h after the second injection. The patient can continue to take levothyroxine and remains euthyroid. Both approaches have equal success in achieving remnant ablation.

A pretreatment scanning dose of ^{131}I (usually 111 MBq [3 mCi]) or ^{123}I (74 MBq [2 mCi]) can reveal the amount of residual tissue and provides guidance about the dose needed to accomplish ablation. However, because of concerns about radioactive "stunning" that impairs subsequent treatment, there is a trend to avoid pretreatment scanning with ^{131}I and use either ^{123}I or proceed directly to ablation, unless there is suspicion that the amount of residual tissue will alter therapy or that there is distant metastatic disease. In the United States, outpatient doses of up to 6475 MBq (175 mCi) can be given at most centers. The administered dose depends on the indication for therapy with lower doses of 1100 MBq (30 mCi) given for remnant ablation but higher doses of up to 5500 MBq (150 mCi) used as adjuvant therapy when residual disease is suspected or present. A whole-body scanning (WBS) following radioiodine treatment is used to confirm the ^{131}I uptake in the remnant and to identify possible metastatic disease.

Surveillance Testing Serum thyroglobulin is a sensitive marker of residual/recurrent thyroid cancer after ablation of the residual postsurgical thyroid tissue. Current Tg assays have functional sensitivities as low as 0.1 ng/mL, as opposed to older assays with functional sensitivities of 1–2 ng/mL, reducing the number of patients with truly undetectable serum Tg levels. Because the vast majority of PTC recurrences are in cervical lymph nodes, a neck ultrasound should be performed about 6 months after thyroid ablation; ultrasound has been shown to be more sensitive than WBS in this scenario.

In low-risk patients who have no clinical evidence of residual disease after ablation, negative cervical sonography, and a basal Tg <0.2 ng/mL on levothyroxine, the risk of structural recurrence is <3% at 5 years, and the frequency of follow-up testing can be decreased to annual TSH and Tg testing, with only periodic ultrasound examination.

The use of WBS is reserved for patients with known iodine-avid metastases or those with elevated serum thyroglobulin levels and negative imaging with ultrasound, chest CT, neck cross-sectional imaging and positron emission tomography (PET) CT who may require additional ^{131}I therapy.

In addition to radioiodine, external beam radiotherapy is also used to treat gross residual neck disease or specific metastatic lesions, particularly when they cause bone pain or threaten neurologic injury (e.g., vertebral metastases).

New Potential Therapies Kinase inhibitors are being explored as a means to target pathways known to be active in thyroid cancer, including the RAS, BRAF, RET, EGFR, VEGFR, and angiogenesis pathways. A multicenter randomized controlled trial of the

multikinase inhibitor sorafenib in 417 patients with progressive metastatic thyroid cancer reported a doubling of progression-free survival to 10.8 months in the treatment group compared with the placebo group. Ongoing trials are exploring whether differentiation protocols with kinase inhibitors or other approaches might enhance radioiodine uptake and efficacy.

ANAPLASTIC AND OTHER FORMS OF THYROID CANCER

Anaplastic Thyroid Cancer As noted above, ATC is a poorly differentiated and aggressive cancer. The prognosis is poor, and most patients die within 6 months of diagnosis. Because of the undifferentiated state of these tumors, the uptake of radioiodine is usually negligible, but it can be used therapeutically if there is residual uptake. Chemotherapy has been attempted with multiple agents, including anthracyclines and paclitaxel, but it is usually ineffective. External beam radiation therapy can be attempted and continued if tumors are responsive. Recent data demonstrate survival benefit with immune checkpoint inhibition therapy.

Thyroid Lymphoma Lymphoma in the thyroid gland often arises in the background of Hashimoto's thyroiditis. A rapidly expanding thyroid mass suggests the possibility of this diagnosis. Diffuse large-cell lymphoma is the most common type in the thyroid. Biopsies reveal sheets of lymphoid cells that can be difficult to distinguish from small-cell lung cancer or ATC. These tumors are often highly sensitive to external radiation. Surgical resection should be avoided as initial therapy because it may spread disease that is otherwise localized to the thyroid. If staging indicates disease outside of the thyroid, treatment should follow guidelines used for other forms of lymphoma (Chap. 104).

MEDULLARY THYROID CARCINOMA

MTC can be sporadic or familial and accounts for about 5% of thyroid cancers. There are three familial forms of MTC: MEN 2A, MEN 2B, and familial MTC without other features of MEN (Chap. 381). In general, MTC is more aggressive in MEN 2B than in MEN 2A, and familial MTC is more aggressive than sporadic MTC. Elevated serum calcitonin provides a marker of residual or recurrent disease. All patients with MTC should be tested for *RET* mutations, because genetic counseling and testing of family members can be offered to those individuals who test positive for mutations.

The management of MTC is primarily surgical. Prior to surgery, pheochromocytoma should be excluded in all patients with a RET mutation. Unlike tumors derived from thyroid follicular cells, these tumors do not take up radioiodine. External radiation treatment and targeted kinase inhibitors may provide palliation in patients with advanced disease (Chap. 381).

APPROACH TO THE PATIENT

Thyroid Nodules

Palpable thyroid nodules are found in about 5% of adults, but the prevalence varies considerably worldwide. Given this high prevalence rate, practitioners may identify thyroid nodules on physical examination. However, the increased usage of diagnostic medical imaging (e.g., carotid ultrasound, cervical spine MRI) has led to an increased frequency of incidental nodule detection, accounting for the majority of patients currently presenting for nodule evaluation. The main goal of this evaluation is to identify, in a cost-effective manner, the small subgroup of individuals with malignant lesions that have the potential to be clinically significant.

Nodules are more common in iodine-deficient areas, in women, and with aging. Most palpable nodules are >1 cm in diameter, but the ability to feel a nodule is influenced by its location within the gland (superficial versus deeply embedded), the anatomy of the patient's neck, and the experience of the examiner. More sensitive methods of detection, such as CT, thyroid ultrasound, and pathologic studies, reveal thyroid nodules in up to 50% of glands in individuals aged >50 years. The presence of these thyroid incidentalomas has led to much debate about how to detect nodules and which nodules to investigate further.

An approach to the evaluation of a solitary nodule is outlined in Fig. 378-4. Most patients with thyroid nodules have normal thyroid function tests. Nonetheless, thyroid function should be assessed by measuring a TSH level, which may be suppressed by one or more autonomously functioning nodules. If the TSH is suppressed, a radionuclide scan is indicated to determine if the identified nodule is "hot," as lesions with increased uptake are almost never malignant and FNA is unnecessary. Otherwise, the next step in evaluation is performance of a thyroid ultrasound for three reasons: (1) Ultrasound will confirm if the palpable nodule is indeed a nodule. About 15% of "palpable" nodules are not confirmed on imaging, and therefore, no further evaluation is required. (2) Ultrasound will assess if there are additional nonpalpable nodules for which FNA may be recommended based on imaging features and size. (3) Ultrasound will characterize the imaging pattern of the nodule, which, combined with the nodule's size, facilitate decision-making about FNA. Numerous studies have demonstrated consistent risk estimates for thyroid cancer based upon certain sonographic patterns. For example, a spongiform nodule has a <3% chance of cancer and observation rather than FNA is reasonable, whereas 10–20% of solid hypoechoic nodules with smooth borders are malignant and FNA is recommended at a size cutoff of 1 cm (Table 378-1, Fig. 378-1). Evidence-based guidelines from both the American Thyroid Association (ATA) and the American Association of Clinical Endocrinologists provide recommendations for nodule FNA based on sonographic patterns and size cut offs, with lower size cut offs for nodules with more suspicious ultrasound patterns. Given what is known about the prevalence and generally indolent behavior of small thyroid cancers <1 cm, the 2015 ATA guidelines do not recommend FNA for any nodule <1 cm unless metastatic cervical lymph nodes are present.

FNA biopsy, ideally performed with ultrasound guidance, is the best diagnostic test when performed by physicians familiar with the procedure and when the results are interpreted by experienced cytopathologists. The technique is particularly useful for detecting PTC. However, the distinction between benign and malignant follicular lesions is often not possible using cytology alone because of the absence of characteristic nuclear features in follicular carcinoma. In several large studies, FNA biopsies yielded the following cytology diagnoses: 65% benign, 5% malignant or suspicious for malignancy, 10% nondiagnostic or yielding insufficient material for diagnosis, and 20% indeterminate. The Bethesda System is now widely used to provide more uniform terminology for reporting thyroid nodule FNA cytology results. This six-tiered classification system with the respective estimated malignancy rates is shown in Table 378-5. Specifically, the Bethesda System subcategorized cytology specimens previously labeled as indeterminate into three categories: atypia or follicular lesion of undetermined significance (AUS/FLUS), follicular neoplasm, and suspicious for malignancy.

Cytology results indicative of malignancy generally mandate surgery, after performing preoperative sonography to evaluate the cervical lymph nodes. Nondiagnostic cytology specimens most often result from cystic lesions but may also occur in fibrous long-standing nodules. Ultrasound-guided FNA is indicated when a repeat FNA is necessary. Repeat FNA will yield a diagnostic cytology in about 50% of cases. Benign nodules may be monitored by ultrasound for growth, and repeat FNA may be considered if the nodule enlarges. The use of levothyroxine to suppress serum TSH is not effective in shrinking nodules in iodine-replete populations, and therefore, levothyroxine should not be used. The three indeterminate cytology classifications introduced by the Bethesda System are associated with different risks of malignancy (Table 378-5). For nodules with suspicious for malignancy cytology, surgery is

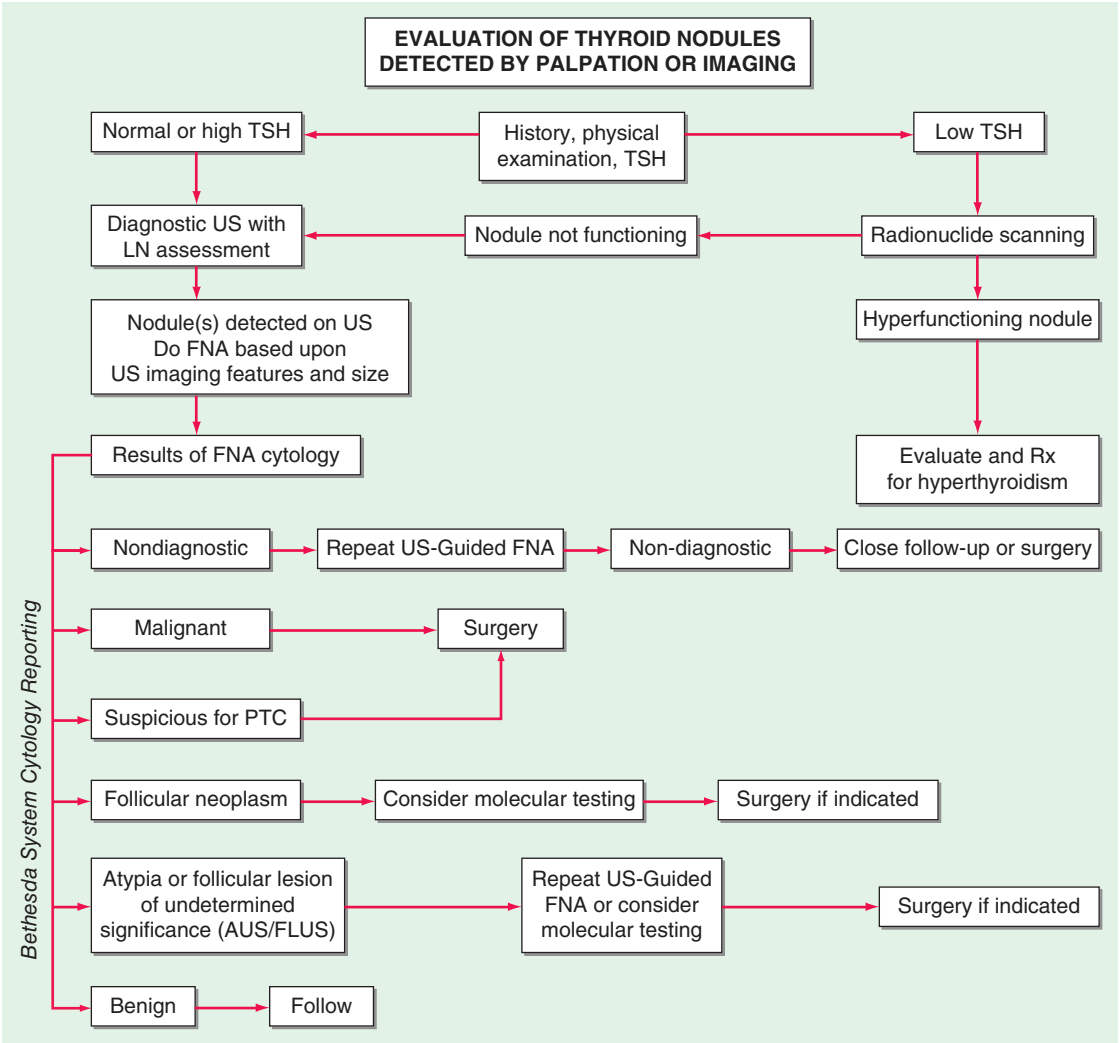


FIGURE 378-4 Approach to the patient with a thyroid nodule. See text and references for details. FNA, fine-needle aspiration; LN, lymph node; PTC, papillary thyroid cancer; TSH, thyroid-stimulating hormone; US, ultrasound.

recommended after ultrasound assessment of cervical lymph nodes. Options to be discussed with the patient include lobectomy versus total thyroidectomy.

On the other hand, the majority of nodules with AUS/FLUS and follicular neoplasm cytology results are benign; only 10–30% are malignant. The traditional approach for these patients is diagnostic lobectomy for histopathologic diagnosis. Therefore, up to 85% of patients undergo surgery for benign nodules. A high-sensitivity (~90%) novel molecular test using gene expression profiling technology may reduce the need for unnecessary surgery in these two groups. In a multicenter trial of over 265 such nodules, a negative gene expression classifier test reduced the risk of malignancy to about 6%, leading to clinical recommendations for follow-up rather than surgery. In addition, based upon results from next generation sequencing, molecular diagnostic panels, which include point mutations, small insertions/deletions, and gene fusions, are currently under investigation with the two goals: (1) identification and risk stratification of thyroid cancers based upon a positive result; (2) reduction in cancer risk to an acceptable level for nonsurgical surveillance based upon a negative result.

The evaluation of a thyroid nodule is stressful for most patients. They are concerned about the possibility of thyroid cancer, whether verbalized or not. It is constructive, therefore, to review the diagnostic approach and to reassure patients when no malignancy is found. When a suspicious lesion or thyroid cancer is identified, the generally favorable prognosis and available treatment options can be reassuring.

TABLE 378-5 Bethesda Classification for Thyroid Cytology

DIAGNOSTIC CATEGORY	RISK OF MALIGNANCY
I. Nondiagnostic or unsatisfactory	1–5%
II. Benign	2–4%
III. Atypia or follicular lesion of unknown significance (AUS/FLUS)	5–15%
IV. Follicular neoplasm	15–30%
V. Suspicious for malignancy	60–75%
VI. Malignant	97–100%

FURTHER READING

Davies L, Welch HG: Current thyroid cancer trends in the United States. JAMA Otolaryngol Head Neck Surg 140:317, 2014.

Fagin JA, Wells SA: Biologic and clinical perspectives on thyroid cancer. N Engl J Med 375:1054, 2016.

Haugen BR et al: 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer. Thyroid 26:1, 2016.

Tuttle RM, Haugen B, Perrier ND: Updated American Joint Committee and Cancer/Tumor-Node-Metastasis Staging System for differentiated and anaplastic thyroid cancer (eighth edition): What changed and why? Thyroid 27:751, 2017.

379 Disorders of the Adrenal Cortex

Wiebke Arlt

The adrenal cortex produces three classes of corticosteroid hormones: glucocorticoids (e.g., cortisol), mineralocorticoids (e.g., aldosterone), and adrenal androgen precursors (e.g., dehydroepiandrosterone [DHEA]) **(Fig. 379-1)**. Glucocorticoids and mineralocorticoids act through specific nuclear receptors, regulating aspects of the physiologic stress response as well as blood pressure and electrolyte homeostasis. Adrenal androgen precursors are converted in the gonads and peripheral target cells to sex steroids that act via nuclear androgen and estrogen receptors.

Disorders of the adrenal cortex are characterized by deficiency or excess of one or several of the three major corticosteroid classes. Hormone deficiency can be caused by inherited glandular or enzymatic disorders or by destruction of the pituitary or adrenal gland by autoimmune disorders, infection, infarction, or iatrogenic events such as surgery or hormonal suppression. Hormone excess is usually the result of neoplasia, leading to increased production of adrenocorticotropic hormone (ACTH) by the pituitary or neuroendocrine cells (ectopic ACTH) or increased production of glucocorticoids, mineralocorticoids, or adrenal androgen precursors by adrenal nodules. Adrenal nodules are increasingly identified incidentally during abdominal imaging performed for other reasons.

ADRENAL ANATOMY AND DEVELOPMENT

The normal adrenal glands weigh 6–11 g each. They are located above the kidneys and have their own blood supply. Arterial blood flows initially to the subcapsular region and then meanders from the outer cortical zona glomerulosa through the intermediate zona fasciculata to the inner zona reticularis and eventually to the adrenal medulla. The right suprarenal vein drains directly into the vena cava, while the left suprarenal vein drains into the left renal vein.

During early embryonic development, the adrenals originate from the urogenital ridge and then separate from gonads and kidneys at about the sixth week of gestation. Concordant with the time of sexual differentiation (seventh to ninth week of gestation, **Chap. 383**), the

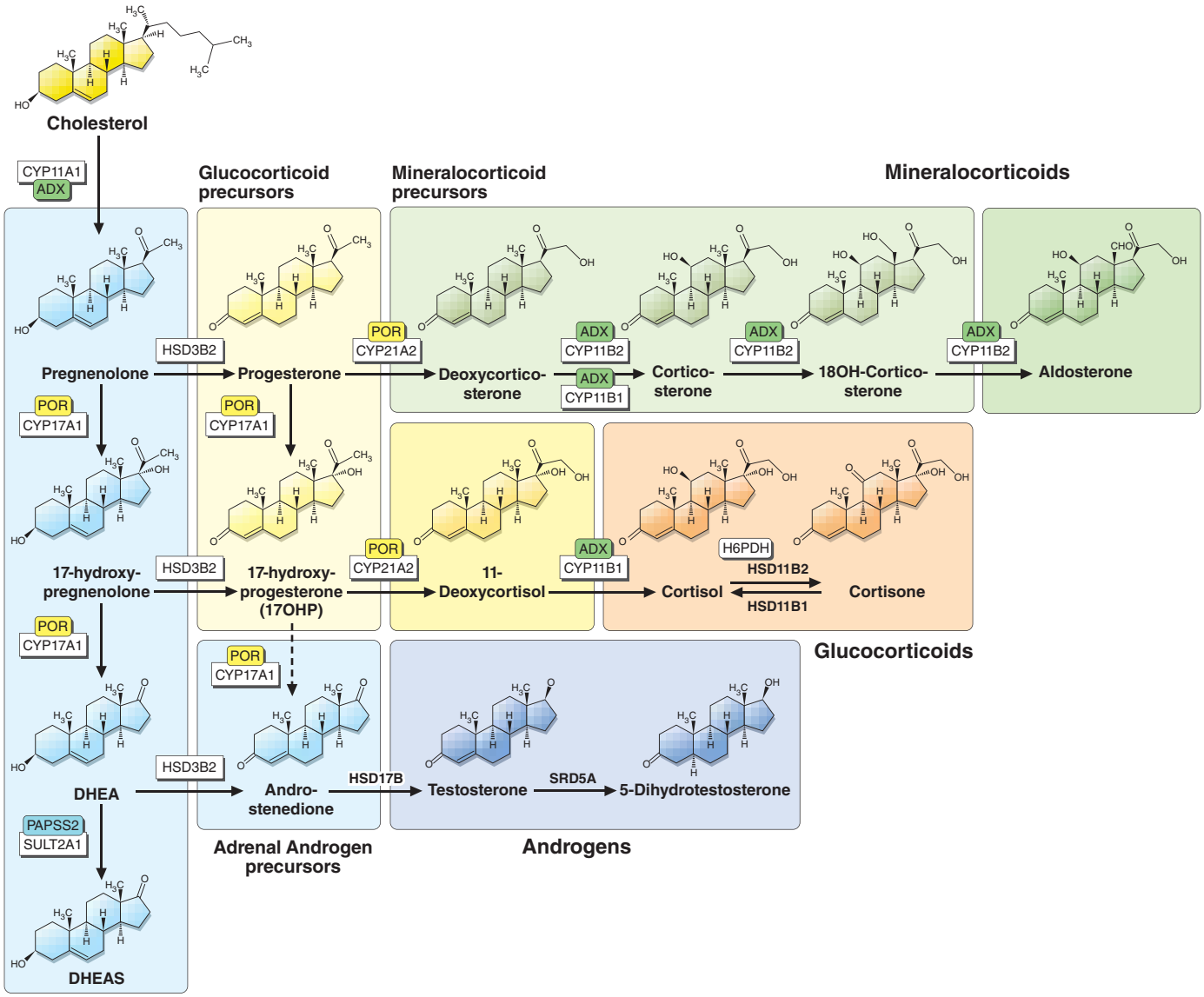


FIGURE 379-1 Adrenal steroidogenesis. ADX, adrenodoxin; CYP11A1, side chain cleavage enzyme; CYP11B1, 11β-hydroxylase; CYP11B2, aldosterone synthase; CYP17A1, 17α-hydroxylase/17,20 lyase; CYP21A2, 21-hydroxylase; DHEA, dehydroepiandrosterone; DHEAS, dehydroepiandrosterone sulfate; H6PDH, hexose-6-phosphate dehydrogenase; HSD11B1, 11β-hydroxysteroid dehydrogenase type 1; HSD11B2, 11β-hydroxysteroid dehydrogenase type 2; HSD17B, 17β-hydroxysteroid dehydrogenase; HSD3B2, 3β-hydroxysteroid dehydrogenase type 2; PAPSS2, PAPS synthase type 2; POR, P450 oxidoreductase; SRD5A, 5α-reductase; SULT2A1, DHEA sulfotransferase.

adrenal cortex starts to produce cortisol and the adrenal sex steroid precursor DHEA. The orphan nuclear receptors SF1 (steroidogenic factor 1; encoded by the gene *NR5A1*) and DAX1 (dosage-sensitive sex reversal gene 1; encoded by the gene *NR0B1*), among others, play a crucial role during this period of development, as they regulate a multitude of adrenal genes involved in steroidogenesis.

REGULATORY CONTROL OF STEROIDOGENESIS

Production of glucocorticoids and adrenal androgens is under the control of the hypothalamic-pituitary-adrenal (HPA) axis, whereas mineralocorticoids are regulated by the renin-angiotensin-aldosterone (RAA) system.

Glucocorticoid synthesis is under inhibitory feedback control by the hypothalamus and the pituitary **(Fig. 379-2)**. Hypothalamic release of corticotropin-releasing hormone (CRH) occurs in response to endogenous or exogenous stress. CRH stimulates the cleavage of the 241–amino acid polypeptide proopiomelanocortin (POMC) by pituitary-specific prohormone convertase 1 (PC1), yielding the 39–amino acid peptide ACTH. ACTH is released by the corticotrope cells of the anterior pituitary and acts as the pivotal regulator of adrenal cortisol synthesis, with additional short-term effects on mineralocorticoid and adrenal androgen synthesis. The release of CRH, and subsequently ACTH, occurs in a pulsatile fashion that follows a circadian rhythm under the control of the hypothalamus, specifically its suprachiasmatic nucleus (SCN), with additional regulation by a complex network of cell-specific clock genes. Reflecting the pattern of ACTH secretion, adrenal cortisol secretion exhibits a distinct circadian rhythm, starting to rise in the early morning hours prior to awakening, with peak levels in the morning and low levels in the evening **(Fig. 379-3)**.

Diagnostic tests assessing the HPA axis make use of the fact that it is regulated by negative feedback. Glucocorticoid excess is diagnosed by employing a dexamethasone suppression test. Dexamethasone, a potent synthetic glucocorticoid, suppresses CRH/ACTH by binding hypothalamic-pituitary glucocorticoid receptors (GRs) and, therefore, results in downregulation of endogenous cortisol synthesis. Various versions of the dexamethasone suppression test are described in detail in **Chap. 373**. If cortisol production is autonomous (e.g., adrenal nodule), ACTH is already suppressed and dexamethasone has little additional effect. If cortisol production is driven by an ACTH-producing pituitary adenoma, dexamethasone suppression is ineffective at low doses but usually induces suppression at high doses. If cortisol production is driven by an ectopic source of ACTH, the tumors are usually resistant to dexamethasone suppression. Thus, the dexamethasone suppression test is useful to establish the diagnosis of Cushing's syndrome and to assist with the differential diagnosis of cortisol excess.

Conversely, to assess glucocorticoid deficiency, ACTH stimulation of cortisol production is used. The ACTH peptide contains 39 amino acids but the first 24 are sufficient to elicit a physiologic response. The standard ACTH stimulation test involves administration of cosyntropin (ACTH 1-24), 0.25 mg IM or IV, and collection of blood samples at 0, 30, and 60 min for cortisol. A normal response is defined as a cortisol level >20 μg/dL (>550 nmol/L) 30–60 min after cosyntropin stimulation. A low-dose (1 μg cosyntropin IV) version of this test has been advocated; however, it has no superior diagnostic value and is more cumbersome to carry out. Alternatively, an insulin tolerance test (ITT) can be used to assess adrenal function. It involves injection of insulin to induce hypoglycemia, which represents a strong stress signal that triggers hypothalamic CRH release and activation of the entire HPA axis. The ITT involves administration of regular insulin 0.1 U/kg IV (dose should be lower if hypopituitarism is likely) and collection of blood samples at 0, 30, 60, and 120 min for glucose, cortisol, and growth hormone (GH), if also assessing the GH axis. Oral or IV glucose is administered after the patient has achieved symptomatic hypoglycemia (usually glucose <40 mg/dL). A normal response is defined as a cortisol >20 μg/dL and GH >5.1 μg/L. The ITT requires careful clinical monitoring and sequential measurements of glucose. It is contraindicated in patients with coronary disease, cerebrovascular disease, or seizure disorders, which has made the short cosyntropin test the commonly accepted first-line test.

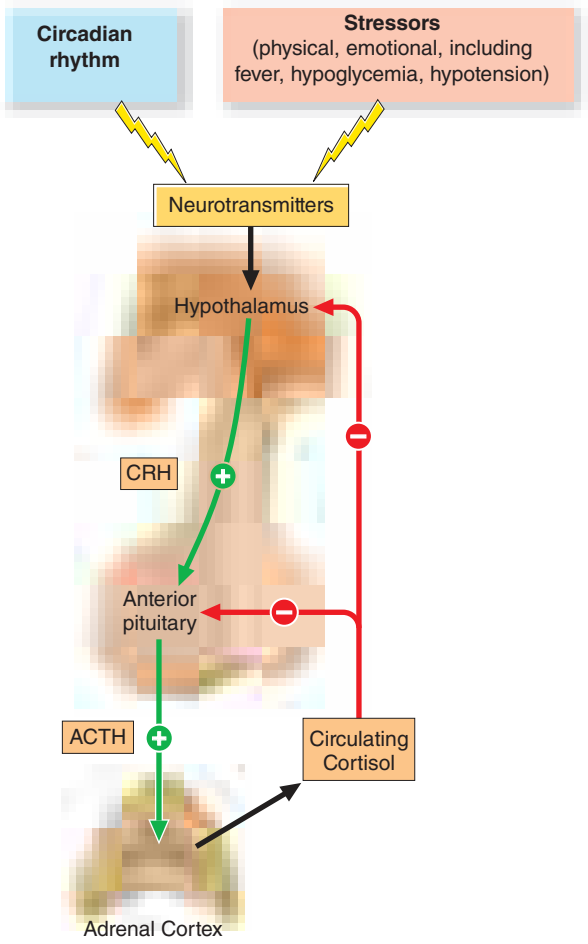


FIGURE 379-2 Regulation of the hypothalamic-pituitary-adrenal (HPA) axis. ACTH, adrenocorticotropic hormone; CRH, corticotropin-releasing hormone.

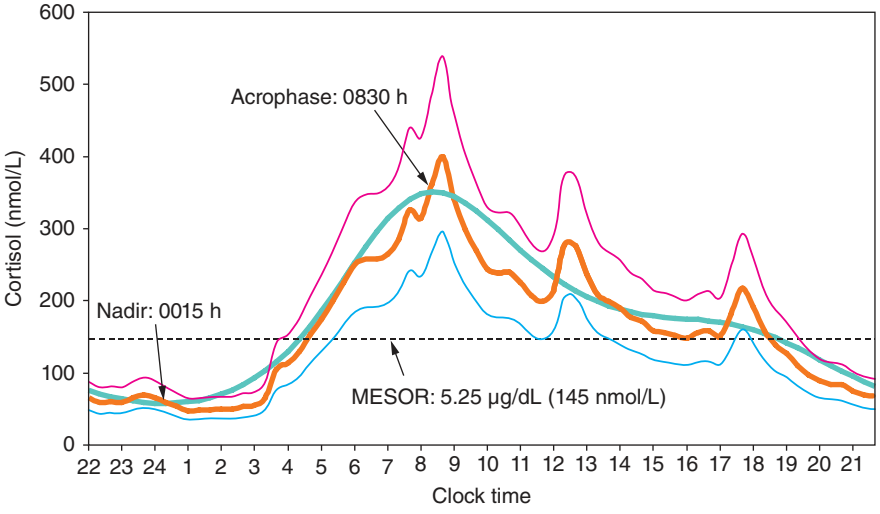


FIGURE 379-3 Physiologic cortisol circadian rhythm. Circulating cortisol concentrations (geometrical mean ± standard deviation values and fitted cosinor) drop under the rhythm-adjusted mean (MESOR) in the early evening hours, with nadir levels around midnight and a rise in the early morning hours; peak levels are observed ~8:30 AM (acrophase). *(Modified after M Debono et al: Modified-release hydrocortisone to provide circadian cortisol profiles. J Clin Endocrinol Metab 94:1548, 2009.)*

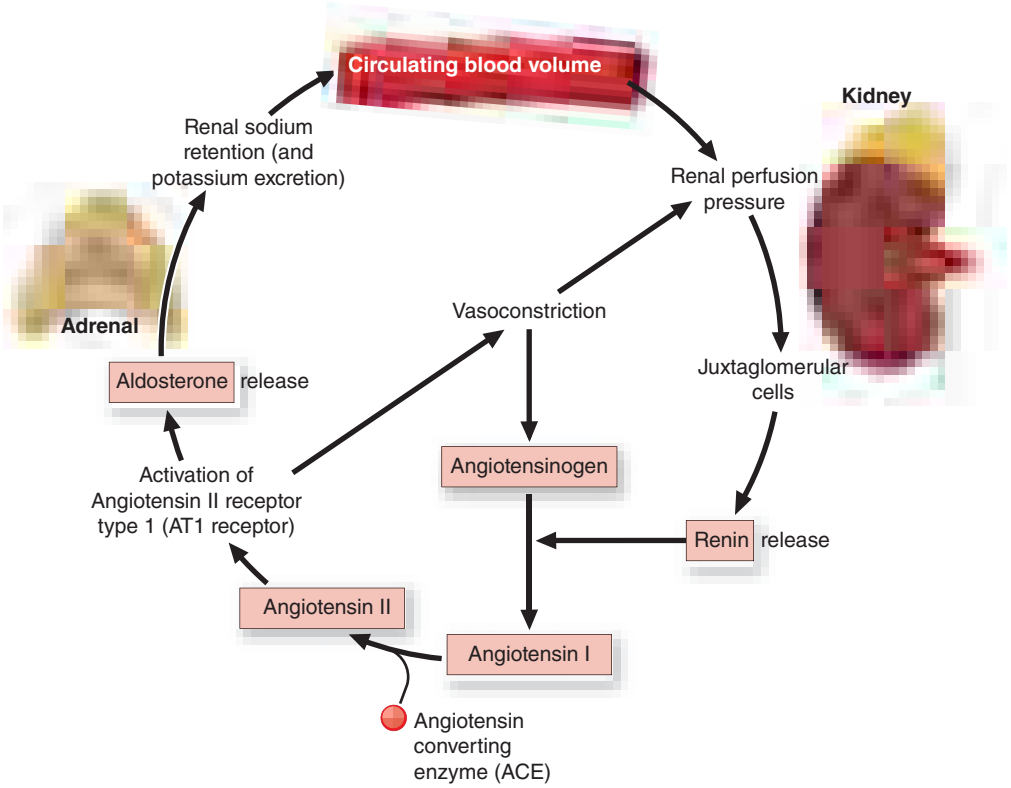


FIGURE 379-4 Regulation of the renin-angiotensin-aldosterone (RAA) system.

Mineralocorticoid production is controlled by the RAA regulatory cycle, which is initiated by the release of renin from the juxtaglomerular cells in the kidney, resulting in cleavage of angiotensinogen to angiotensin I in the liver **(Fig. 379-4)**. Angiotensin-converting enzyme (ACE) cleaves angiotensin I to angiotensin II, which binds and activates the angiotensin II receptor type 1 (AT1 receptor [AT1R]), resulting in increased adrenal aldosterone production and vasoconstriction. Aldosterone enhances sodium retention and potassium excretion, and increases the arterial perfusion pressure, which in turn regulates renin release. Because mineralocorticoid synthesis is primarily under the control of the RAA system, hypothalamic-pituitary damage does not significantly impact the capacity of the adrenal to synthesize aldosterone.

Similar to the HPA axis, the assessment of the RAA system can be used for diagnostic purposes. If mineralocorticoid excess is present, there is a counter-regulatory downregulation of plasma renin (see below for testing). Conversely, in mineralocorticoid deficiency, plasma renin is markedly increased. Physiologically, oral or IV sodium loading results in suppression of aldosterone, a response that is attenuated or absent in patients with autonomous mineralocorticoid excess.

STEROID HORMONE SYNTHESIS, METABOLISM, AND ACTION

ACTH stimulation is required for the initiation of steroidogenesis. The ACTH receptor MC2R (melanocortin 2 receptor) interacts with the MC2R-accessory protein MRAP, and the complex is transported to the adrenocortical cell membrane, where it binds to ACTH **(Fig. 379-5)**. ACTH stimulation generates cyclic AMP (cAMP), which upregulates the protein kinase A (PKA) signaling pathway. Inactive PKA is a tetramer of two regulatory and two catalytic subunits that is dissociated by cAMP into a dimer of two regulatory subunits bound to cAMP and two free and active catalytic subunits. PKA activation impacts steroidogenesis in three distinct ways: (1) increases the import of cholesterol esters; (2) increases the activity of hormone-sensitive lipase, which cleaves cholesterol esters to cholesterol for import into the mitochondrion; and (3) increases the availability and phosphorylation of CREB (cAMP response element binding), a transcription factor that enhances transcription of CYP11A1 and other enzymes required for glucocorticoid synthesis.

Adrenal steroidogenesis occurs in a zone-specific fashion, with mineralocorticoid synthesis occurring in the outer zona glomerulosa, glucocorticoid synthesis in the zona fasciculata, and adrenal androgen synthesis in the inner zona reticularis (Fig. 379-1). All steroidogenic pathways require cholesterol import into the mitochondrion, a process initiated by the action of the steroidogenic acute regulatory (StAR) protein, which shuttles cholesterol from the outer to the inner mitochondrial membrane. The majority of steroidogenic enzymes are cytochrome P450 (CYP) enzymes, which are either located in the mitochondrion (side chain cleavage enzyme, CYP11A1; 11β-hydroxylase, CYP11B1; aldosterone synthase, CYP11B2) or in the endoplasmic reticulum membrane (17α-hydroxylase, CYP17A1; 21-hydroxylase, CYP21A2; aromatase, CYP19A1). These enzymes require electron donation via specific redox cofactor enzymes, P450 oxidoreductase (POR), and adrenodoxin/adrenodoxin reductase (ADX/ADR) for the microsomal and mitochondrial CYP enzymes, respectively. In addition, the short-chain dehydrogenase 3β-hydroxysteroid dehydrogenase type 2 (3β-HSD2), also termed Δ4, Δ5 isomerase, plays a major role in adrenal steroidogenesis.

The cholesterol side chain cleavage enzyme CYP11A1 generates pregnenolone. Glucocorticoid synthesis requires conversion of pregnenolone to progesterone by 3β-HSD2, followed by conversion to 17-hydroxyprogesterone (17OHP) by CYP17A1, further hydroxylation at carbon 21 by CYP21A2, and eventually, 11β-hydroxylation by CYP11B1 to generate active cortisol (Fig. 379-1). Mineralocorticoid synthesis also requires progesterone, which is first converted to deoxycorticosterone (DOC) by CYP21A2 and then converted via corticosterone and 18-hydroxycorticosterone to aldosterone in three steps catalyzed by CYP11B2. For adrenal androgen synthesis, pregnenolone undergoes conversion by CYP17A1, which uniquely catalyzes two enzymatic reactions. Via its 17α-hydroxylase activity, CYP17A1 converts pregnenolone to 17-hydroxypregnenolone, followed by generation of the universal sex steroid precursor DHEA via CYP17A1 17,20 lyase activity. The majority of DHEA is secreted by the adrenal in the form of its sulfate ester, DHEAS, generated by DHEA sulfotransferase (SULT2A1).

Following its release from the adrenal, cortisol circulates in the bloodstream mainly bound to cortisol-binding globulin (CBG) and to a lesser extent to albumin, with only a minor fraction circulating as

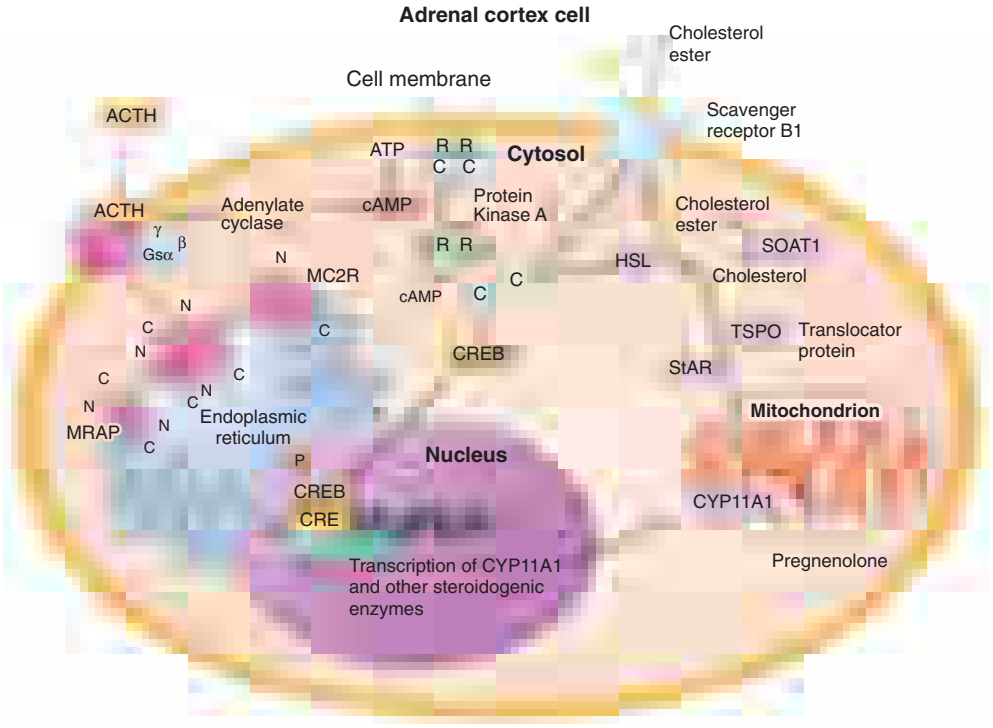


FIGURE 379-5 ACTH effects on adrenal steroidogenesis. ACTH, adrenocorticotropic hormone; binding protein; HSL, hormone-sensitive lipase; MRAP, MC2R-accessory protein; protein kinase A catalytic subunit (C; *PRKACA*), PKA regulatory subunit (R; *PRKAR1A*); SOAT1, sterol O-acyltransferase 1; StAR, steroidogenic acute regulatory (protein); TSPO, translocator protein.

free, unbound hormone. Free cortisol is thought to enter cells directly, not requiring active transport. In addition, in a multitude of peripheral target tissues of glucocorticoid action, including adipose, liver, muscle, and brain, cortisol is generated from inactive cortisone within the cell by the enzyme 11β-hydroxysteroid dehydrogenase type 1 (11β-HSD1) (Fig. 379-6). Thereby, 11β-HSD1 functions as a tissue-specific prereceptor regulator of glucocorticoid action. For the conversion of inactive cortisone to active cortisol, 11β-HSD1 requires nicotinamide adenine dinucleotide phosphate (NADPH [reduced form]), which is provided by the enzyme hexose-6-phosphate dehydrogenase (H6PDH). Like the catalytic domain of 11β-HSD1, H6PDH is located in the lumen of the endoplasmic reticulum, and converts glucose-6-phosphate (G6P) to 6-phosphogluconate (6PGL), thereby regenerating NADP+ to NADPH, which drives the activation of cortisol from cortisone by 11β-HSD1.

In the cytosol of target cells, cortisol binds and activates the GR, which results in dissociation of heat shock proteins (HSPs) from the receptor and subsequent dimerization (Fig. 379-6). Cortisol-bound GR dimers translocate to the nucleus and activate glucocorticoid response elements (GREs) in the DNA sequence, thereby enhancing transcription of glucocorticoid-regulated genes (GR transactivation). However, cortisol-bound GR can also form heterodimers with transcription factors such as AP-1 or NF-κB, resulting in transrepression

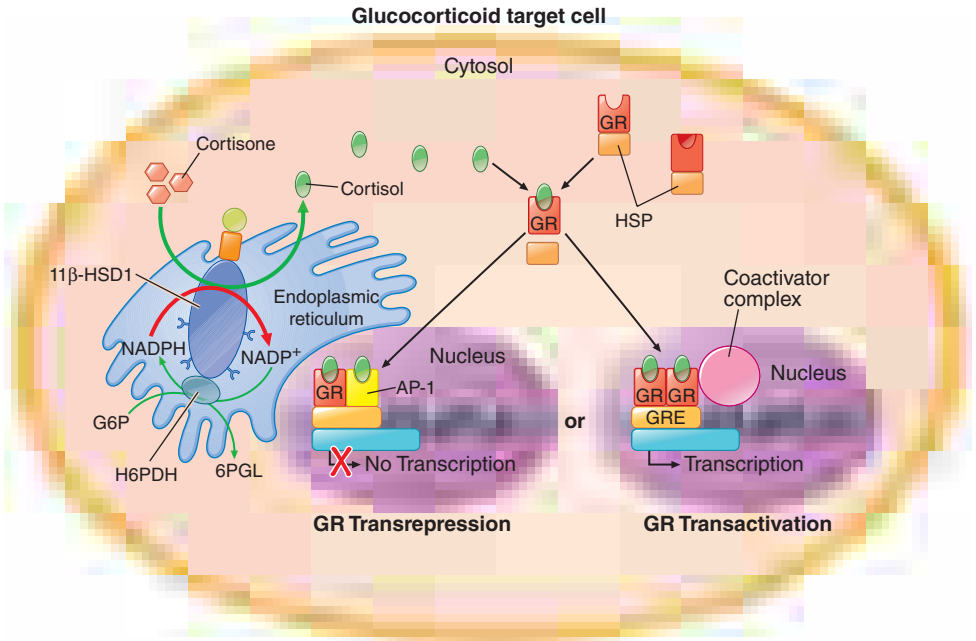


FIGURE 379-6 Prereceptor activation of cortisol and glucocorticoid receptor (GR) action. AP-1, activator protein-1; G6P, glucose-6-phosphate; GREs, glucocorticoid response elements; HSPs, heat shock proteins; NADPH, nicotinamide adenine dinucleotide phosphate (reduced form); 6PGL, 6-phosphogluconate.

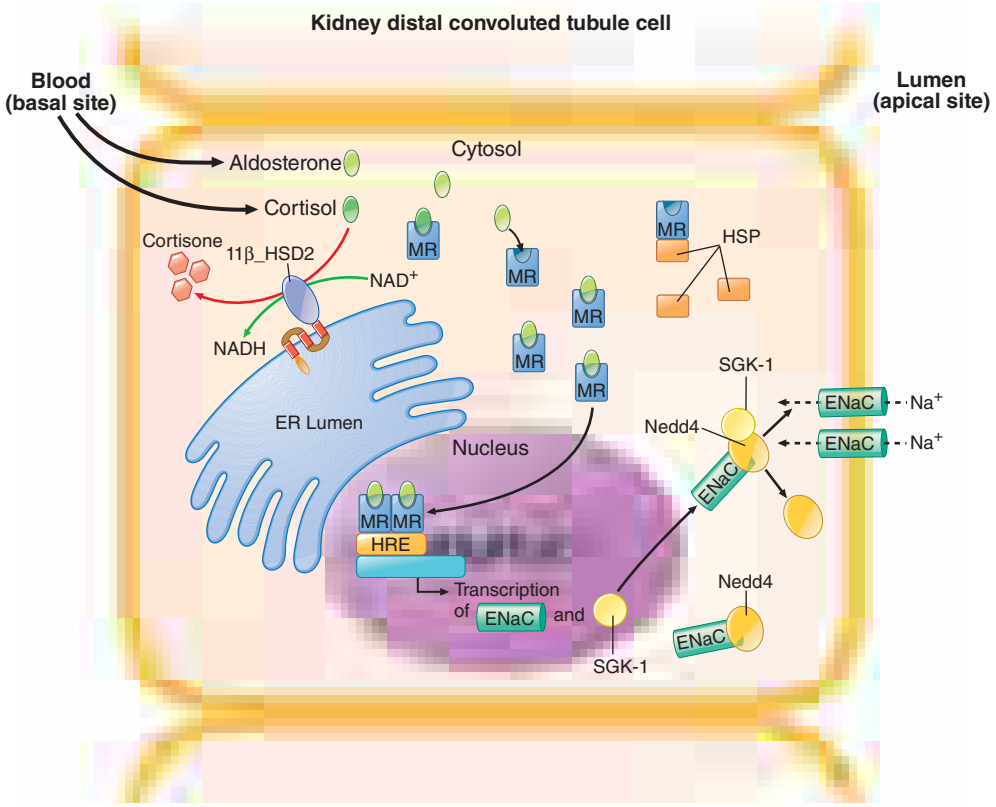


FIGURE 379-7 Prereceptor inactivation of cortisol and mineralocorticoid receptor action. ENaC, epithelial sodium channel; HRE, hormone response element; NADH, nicotinamide adenine dinucleotide; SGK-1, serum glucocorticoid-inducible kinase-1.

of proinflammatory genes, a mechanism of major importance for the anti-inflammatory action of glucocorticoids. It is important to note that corticosterone also exerts glucocorticoid activity, albeit much weaker than cortisol itself. However, in rodents, corticosterone is the major glucocorticoid, and in patients with 17-hydroxylase deficiency, lack of cortisol can be compensated for by higher concentrations of corticosterone that accumulates as a consequence of the enzymatic block.

Cortisol is inactivated to cortisone by the microsomal enzyme 11β-hydroxysteroid dehydrogenase type 2 (11β-HSD2) **(Fig. 379-7)**, mainly in the kidney, but also in the colon, salivary glands, and other target tissues. Cortisol and aldosterone bind the mineralocorticoid receptor (MR) with equal affinity; however, cortisol circulates in the bloodstream at about a 1000-fold higher concentration. Thus, only rapid inactivation of cortisol to cortisone by 11β-HSD2 prevents MR activation by excess cortisol, thereby acting as a tissue-specific modulator of the MR pathway. In addition to cortisol and aldosterone, deoxycorticosterone (DOC) (Fig. 379-1) also exerts mineralocorticoid activity. DOC accumulation due to 11β-hydroxylase deficiency or due to tumor-related excess production can result in mineralocorticoid excess.

Aldosterone synthesis in the adrenal zona glomerulosa cells is driven by the enzyme aldosterone synthase (CYP11B2). The binding of angiotensin II to the AT1 receptor causes glomerulosa cell membrane depolarization by increasing intracellular sodium through inhibition of sodium potassium (Na^+/K^+) ATPase enzymes as well as potassium channels. This drives an increase in intracellular calcium by opening of voltage-dependent calcium channels or inhibition of calcium (Ca^{2+}) ATPase enzymes. Consequently, the calcium signaling pathway is triggered, resulting in upregulation of CYP11B2 transcription **(Fig. 379-8)**.

Analogous to cortisol action via the GR, aldosterone (or cortisol) binding to the MR in the kidney tubule cell dissociates the heat shock protein (HSP)–receptor complex, allowing homodimerization of the MR, and translocation of the hormone-bound MR dimer to the nucleus (Fig. 379-7). The activated MR enhances transcription of the epithelial sodium channel (ENaC) and serum glucocorticoid-inducible kinase 1 (SGK-1). In the cytosol, interaction of ENaC with Nedd4 prevents cell surface expression of ENaC. However, SGK-1 phosphorylates serine residues within the Nedd4 protein, reduces the interaction between Nedd4 and ENaC, and consequently, enhances the trafficking of ENaC to the cell surface, where it mediates sodium retention.

CUSHING'S SYNDROME

(See also Chap. 373) Cushing's syndrome reflects a constellation of clinical features that result from chronic exposure to excess glucocorticoids of any etiology. The disorder can be ACTH-dependent (e.g., pituitary corticotrope adenoma, ectopic secretion of ACTH by nonpituitary tumor) or ACTH-independent (e.g., adrenocortical adenoma, adrenocortical carcinoma [ACC], nodular adrenal hyperplasia), as well as iatrogenic (e.g., administration of exogenous glucocorticoids to treat various inflammatory conditions). The term *Cushing's disease* refers specifically to Cushing's syndrome caused by a pituitary corticotrope adenoma.

Epidemiology Cushing's syndrome is generally considered a rare disease. It occurs with an incidence of 1–2 per 100,000 population per year. However, it is debated whether mild cortisol excess may be more prevalent among patients with features of Cushing's such as centripetal obesity, type 2 diabetes, and osteoporotic vertebral fractures, recognizing that these are relatively nonspecific and common in the population.

In the overwhelming majority of patients, Cushing's syndrome is caused by an ACTH-producing corticotrope adenoma of the pituitary **(Table 379-1)**, as initially described by Harvey Cushing in 1912. Cushing's disease more frequently affects women, with the exception of prepubertal cases, where it is more common in boys. By contrast, ectopic ACTH syndrome is more frequently identified in men. Only 10% of patients with Cushing's syndrome have a primary, adrenal cause of their disease (e.g., autonomous cortisol excess independent of ACTH), and most of these patients are women. However, overall, the medical use of glucocorticoids for immunosuppression, or for the treatment of inflammatory disorders, is the most common cause of Cushing's syndrome.

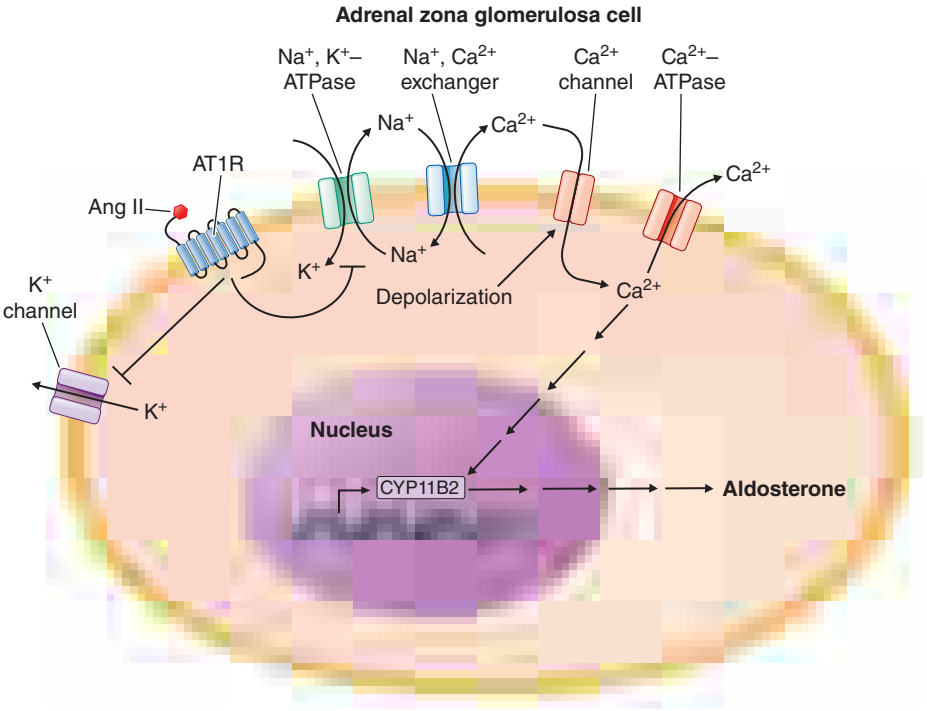


FIGURE 379-8 Regulation of adrenal aldosterone synthesis. AngII, angiotensin II; AT1R, angiotensin II receptor type 1; CYP11B2, aldosterone synthase. *(Modified after F Beuschlein: Regulation of aldosterone secretion: from physiology to disease. Eur J Endocrinol 168:R85, 2013.)*

Etiology In at least 90% of patients with Cushing's disease, ACTH excess is caused by a corticotrope pituitary microadenoma, often only a few millimeters in diameter. Pituitary macroadenomas (i.e., tumors >1 cm in size) are found in only 5–10% of patients. Pituitary corticotrope adenomas usually occur sporadically but very rarely can be found in the context of multiple endocrine neoplasia type 1 (MEN 1) **(Chap. 381)**. Pituitary adenomas causative of Cushing's disease frequently harbor mutations in the deubiquitinase USP8, which leads to constitutive activation of EGF signaling and consequent upregulated expression of the ACTH precursor POMC. USP8 mutations are found more frequently in adults (41 vs 17% in children) and in women (43 vs 17% in men) with Cushing's disease.

Ectopic ACTH production is predominantly caused by occult carcinoid tumors, most frequently in the lung, but also in thymus or pancreas. Because of their small size, these tumors are often difficult to locate. Advanced small-cell lung cancer can cause ectopic ACTH production. In rare cases, ectopic CRH and/or ACTH production has been found to originate from medullary thyroid carcinoma or pheochromocytoma, the latter co-secreting catecholamines and ACTH.

TABLE 379-1 Causes of Cushing's Syndrome

CAUSES OF CUSHING'S SYNDROME	FEMALE:MALE RATIO	%
ACTH-Dependent Cushing's		**90**
Cushing's disease (= ACTH-producing pituitary adenoma)	4:1	75
Ectopic ACTH syndrome (due to ACTH secretion by bronchial or pancreatic carcinoid tumors, small-cell lung cancer, medullary thyroid carcinoma, pheochromocytoma, and others)	1:1	15
ACTH-Independent Cushing's	**4:1**	**10**
Adrenocortical adenoma		5–10
Adrenocortical carcinoma		1
Rare causes: macronodular adrenal hyperplasia; primary pigmented nodular adrenal disease (micro- and/or macronodular); McCune-Albright syndrome		<1

Abbreviation: ACTH, adrenocorticotropic hormone.

The majority of patients with ACTH-independent cortisol excess harbor a cortisol-producing adrenal adenoma; intratumor, i.e., somatic mutations in the PKA catalytic subunit PRKACA have been identified as cause of disease in 40% of these tumors. ACCs may also cause ACTH-independent disease and are often large, with excess production of several corticosteroid classes.

A rare but notable cause of adrenal cortisol excess is macronodular adrenal hyperplasia with low circulating ACTH, but with evidence for autocrine stimulation of cortisol production via intraadrenal ACTH production. These hyperplastic nodules are often also characterized by ectopic expression of G protein–coupled receptors not usually found in the adrenal, including receptors for luteinizing hormone, vasopressin, serotonin, interleukin 1, catecholamines, or gastric inhibitory peptide (GIP), the cause of food-dependent Cushing's. Activation of these receptors results in upregulation of PKA signaling, as physiologically occurs with ACTH, with a subsequent increase in cortisol production. A combination of germline and somatic mutations in the tumor-suppressor gene *ARMC5* have been identified as a prevalent cause of Cushing's due to bilateral macronodular adrenal hyperplasia; these patients often present with biochemical evidence of Cushing's but lack specific clinical signs, which develop slowly over decades and accelerate cardiovascular risk. Constitutively activating mutations in the PKA catalytic subunit PRKACA are found as somatic mutations in one-third of cortisol-producing adrenocortical adenomas and, as germline mutations, can also represent a rare cause of macronodular adrenal hyperplasia associated with cortisol excess.

Germline mutations in one of the regulatory subunits of PKA, PRKAR1A, are found in patients with primary pigmented nodular adrenal disease (PPNAD) as part of *Carney's complex*, an autosomal dominant multiple neoplasia condition associated with cardiac myxomas, hyperlentiginosis, Sertoli cell tumors, and PPNAD. PPNAD can present as micronodular or macronodular hyperplasia, or both. Phosphodiesterases can influence intracellular cAMP and can thereby impact PKA activation. Mutations in PDE11A and PDE8B have been identified in patients with bilateral adrenal hyperplasia and Cushing's, with and without evidence of PPNAD.

Another rare cause of ACTH-independent Cushing's is *McCune-Albright syndrome*, also associated with polyostotic fibrous dysplasia, unilateral café-au-lait spots, and precocious puberty. McCune-Albright syndrome is caused by activating mutations in the stimulatory G protein alpha subunit 1, GNAS-1 (guanine nucleotide-binding protein alpha stimulating activity polypeptide 1), and such mutations have also been found in bilateral macronodular hyperplasia without other McCune-Albright features and, in rare instances, also in isolated cortisol-producing adrenal adenomas (Table 379-1; **Chap. 405**).

Clinical Manifestations Glucocorticoids affect almost all cells of the body, and thus signs of cortisol excess impact multiple physiologic systems **(Table 379-2)**, with upregulation of gluconeogenesis, lipolysis, and protein catabolism causing the most prominent features. In addition, excess glucocorticoid secretion overcomes the ability of 11β-HSD2 to rapidly inactivate cortisol to cortisone in the kidney, thereby exerting mineralocorticoid actions, manifest as diastolic hypertension, hypokalemia, and edema. Excess glucocorticoids also interfere with central regulatory systems, leading to suppression of gonadotropins with subsequent hypogonadism and amenorrhea, and suppression of the hypothalamic-pituitary-thyroid axis, resulting in decreased thyroid-stimulating hormone (TSH) secretion.

TABLE 379-2 Signs and Symptoms of Cushing's Syndrome

BODY COMPARTMENT/ SYSTEM	SIGNS AND SYMPTOMS
Body fat	Weight gain, central obesity, rounded face, fat pad on back of neck ("buffalo hump")
Skin	Facial plethora, thin and brittle skin, easy bruising, broad and purple stretch marks, acne, hirsutism
Bone	Osteopenia, osteoporosis (vertebral fractures), decreased linear growth in children
Muscle	Weakness, proximal myopathy (prominent atrophy of gluteal and upper leg muscles with difficulty climbing stairs or getting up from a chair)
Cardiovascular system	Hypertension, hypokalemia, edema, atherosclerosis
Metabolism	Glucose intolerance/diabetes, dyslipidemia
Reproductive system	Decreased libido, in women amenorrhea (due to cortisol-mediated inhibition of gonadotropin release)
Central nervous system	Irritability, emotional lability, depression, sometimes cognitive defects; in severe cases, paranoid psychosis
Blood and immune system	Increased susceptibility to infections, increased white blood cell count, eosinopenia, hypercoagulation with increased risk of deep vein thrombosis and pulmonary embolism

The majority of clinical signs and symptoms observed in Cushing's syndrome are relatively nonspecific and include features such as obesity, diabetes, diastolic hypertension, hirsutism, and depression that are commonly found in patients who do not have Cushing's. Therefore, careful clinical assessment is an important aspect of evaluating suspected cases. A diagnosis of Cushing's should be considered when several clinical features are found in the same patient, in particular when more specific features are found. These include fragility of the skin, with easy bruising and broad (>1 cm), purplish striae (Fig. 379-9), and signs of proximal myopathy, which becomes most obvious when trying to stand up from a chair without the use of hands or when climbing stairs. Clinical manifestations of Cushing's do not differ substantially among the different causes of Cushing's. In ectopic ACTH syndrome, hyperpigmentation of the knuckles, scars, or skin areas exposed to increased friction can be observed (Fig. 379-9) and is caused by stimulatory effects of excess ACTH and other POMC cleavage products on melanocyte pigment production. Furthermore, patients with ectopic ACTH syndrome, and some with ACC as the cause of Cushing's, may have a more brisk onset and rapid progression of clinical signs and symptoms.

Patients with Cushing's syndrome can be acutely endangered by deep vein thrombosis, with subsequent pulmonary embolism due to a hypercoagulable state associated with Cushing's. The majority of patients also experience psychiatric symptoms, mostly in the form of anxiety or depression, but acute paranoid or depressive psychosis may also occur. Even after cure, long-term health may be affected by persistently impaired health-related quality of life and increased risk of cardiovascular disease and osteoporosis with vertebral fractures, depending on the duration and degree of exposure to significant cortisol excess.

Diagnosis The most important first step in the management of patients with suspected Cushing's syndrome is to establish the correct diagnosis. Most mistakes in clinical management, leading to unnecessary imaging or surgery, are made because the diagnostic protocol is not followed (Fig. 379-10). This protocol requires establishing the diagnosis of Cushing's beyond doubt prior to employing any tests used for the differential diagnosis of the condition. In principle, after excluding exogenous glucocorticoid use as the cause of clinical signs and symptoms, suspected cases should be tested if there are multiple and progressive features of Cushing's, particularly features with a potentially higher discriminatory value. Exclusion of Cushing's is also indicated in patients with incidentally discovered adrenal masses.

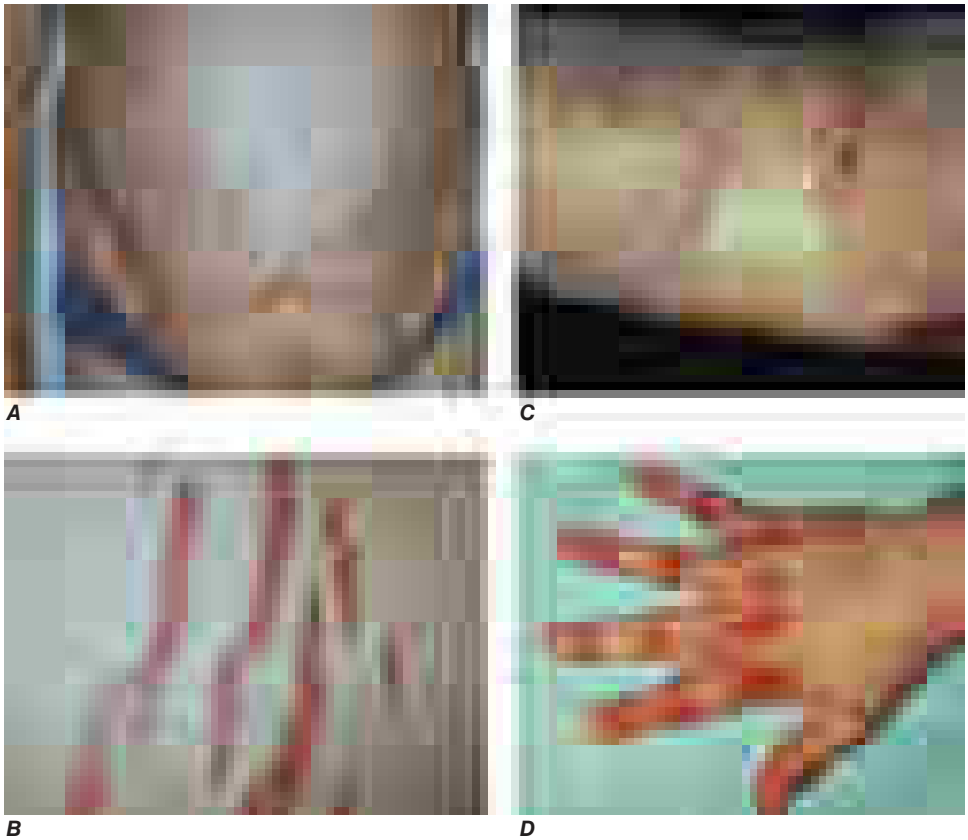


FIGURE 379-9 Clinical features of Cushing's syndrome. *A.* Note central obesity and broad, purple stretch marks (***B.*** close-up). ***C.*** Note thin and brittle skin in an elderly patient with Cushing's syndrome. ***D.*** Hyperpigmentation of the knuckles in a patient with ectopic adrenocorticotropic hormone (ACTH) excess.

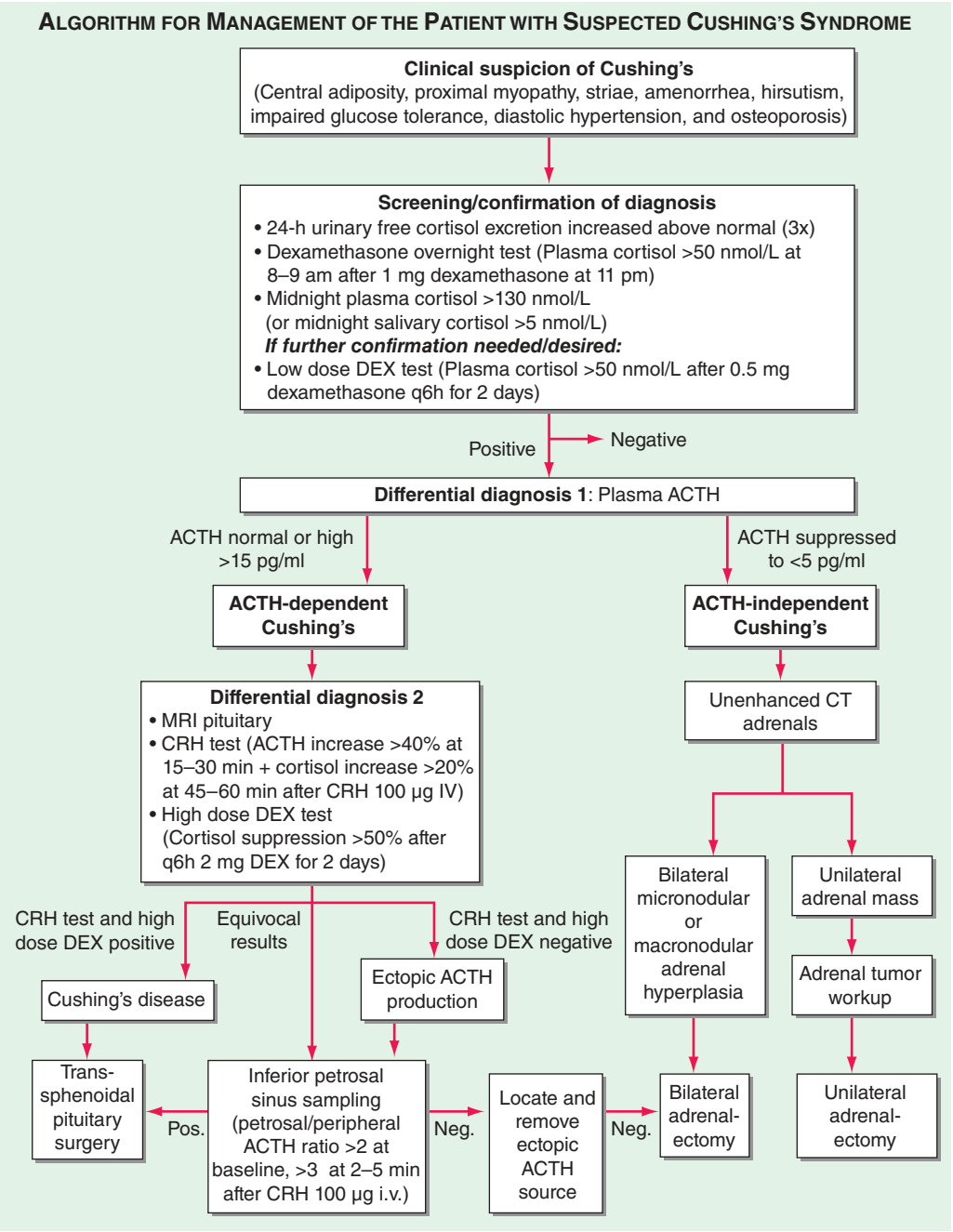


FIGURE 379-10 Management of the patient with suspected Cushing's syndrome. ACTH, adrenocorticotropic hormone; CRH, corticotropin-releasing hormone; CT, computed tomography; DEX, dexamethasone; MRI, magnetic resonance imaging.

A diagnosis of Cushing's can be considered as established if the results of several tests are consistently suggestive of Cushing's. These tests may include increased 24-h urinary free cortisol excretion in three separate collections, failure to appropriately suppress morning cortisol after overnight exposure to dexamethasone, and evidence of loss of diurnal cortisol secretion with high levels at midnight, the time of the physiologically lowest secretion (Fig. 379-10). Factors potentially affecting the outcome of these diagnostic tests have to be excluded such as incomplete 24-h urine collection or rapid inactivation of dexamethasone due to concurrent intake of CYP3A4-inducing drugs (e.g., antiepileptics, rifampicin). Concurrent intake of oral contraceptives that raise CBG and thus total cortisol can cause failure to suppress after dexamethasone. If in doubt, testing should be repeated after 4–6 weeks off estrogens. Patients with pseudo-Cushing states, i.e., alcohol-related, and those with cyclic Cushing's may require further testing to safely confirm or exclude the diagnosis of Cushing's. In addition, the biochemical assays employed can affect the test results, with specificity representing a common problem with antibody-based assays for the measurement of urinary free cortisol. These assays have been greatly improved by the introduction of highly specific tandem mass spectrometry.

Differential Diagnosis The evaluation of patients with confirmed Cushing's should be carried out by an endocrinologist and begins with the differential diagnosis of ACTH-dependent and ACTH-independent cortisol excess (Fig. 379-10). Generally, plasma ACTH levels are suppressed in cases of autonomous adrenal cortisol excess, as a consequence of enhanced negative feedback to the hypothalamus and pituitary. By contrast, patients with ACTH-dependent Cushing's have normal or increased plasma ACTH, with very high levels being found in some patients with ectopic ACTH syndrome. Importantly, imaging should only be used after it is established whether the cortisol excess is ACTH-dependent or ACTH-independent, because nodules in the pituitary or the adrenal are a common finding in

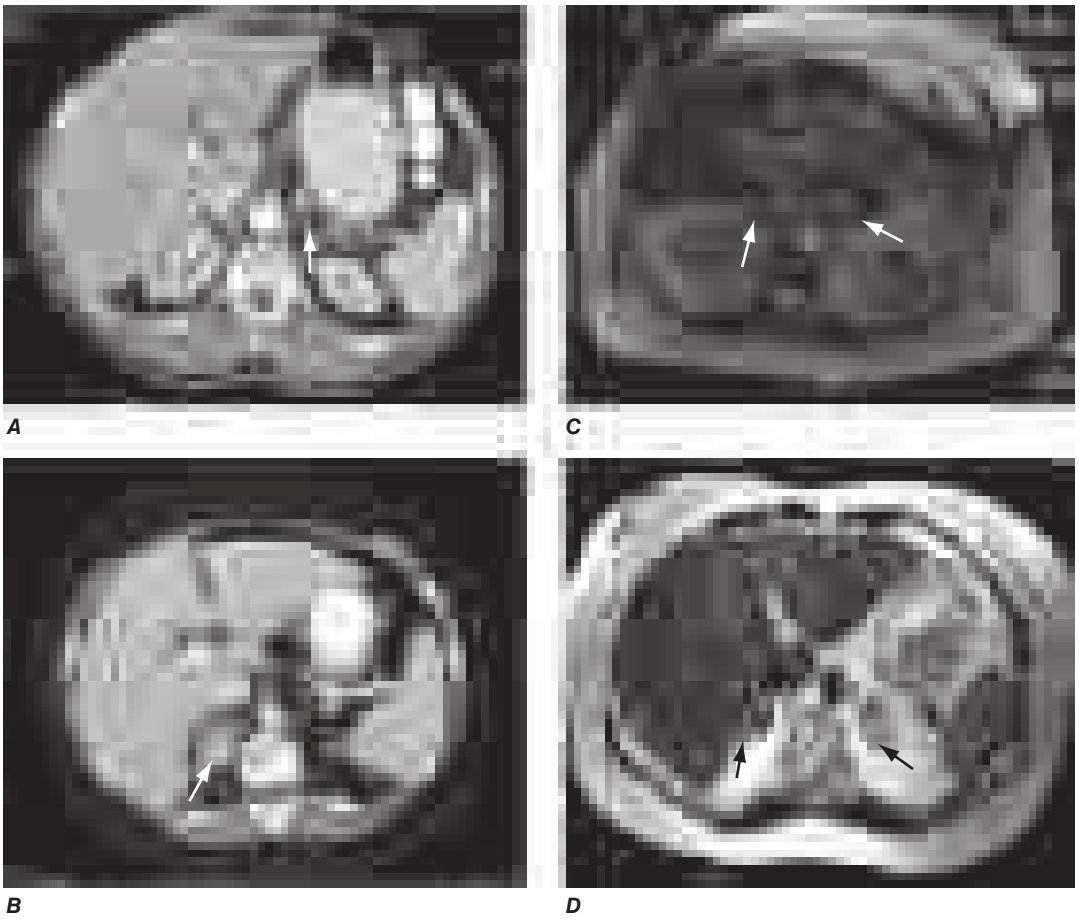


FIGURE 379-11 Adrenal imaging in Cushing's syndrome. ***A.*** Adrenal computed tomography (CT) showing normal bilateral adrenal morphology (*arrows*). ***B.*** CT scan depicting a right adrenocortical adenoma (*arrow*) causing Cushing's syndrome. ***C.*** Magnetic resonance imaging (MRI) showing bilateral adrenal hyperplasia due to excess adrenocorticotropic hormone stimulation in Cushing's disease. ***D.*** MRI showing bilateral macronodular hyperplasia causing Cushing's syndrome.

the general population. In patients with confirmed ACTH-independent excess, adrenal imaging is indicated **(Fig. 379-11)**, preferably using an unenhanced computed tomography (CT) scan. This allows assessment of adrenal morphology and determination of precontrast tumor density in Hounsfield units (HUs), which helps to distinguish between benign and malignant adrenal lesions.

For ACTH-dependent cortisol excess **(Chap. 373)**, a magnetic resonance image (MRI) of the pituitary is the investigation of choice, but it may not show an abnormality in up to 40% of cases because of small tumors below the sensitivity of detection. Characteristically, pituitary corticotrope adenomas fail to enhance following gadolinium administration on T1-weighted MRI images. In all cases of confirmed ACTH-dependent Cushing's, further tests are required for the differential diagnosis of pituitary Cushing's disease and ectopic ACTH syndrome. These tests exploit the fact that most pituitary corticotrope adenomas still display regulatory features, including residual ACTH suppression by high-dose glucocorticoids and CRH responsiveness. In contrast, ectopic sources of ACTH are typically resistant to dexamethasone suppression and unresponsive to CRH (Fig. 379-10). However, it should be noted that a small minority of ectopic ACTH-producing tumors exhibit dynamic responses similar to pituitary corticotrope tumors. If the two tests show discordant results, or if there is any other reason for doubt, the differential diagnosis can be further clarified by performing bilateral inferior petrosal sinus sampling (IPSS) with concurrent blood sampling for ACTH in the right and left inferior petrosal sinus and a peripheral vein. An increased central/peripheral plasma ACTH ratio >2 at baseline and >3 at 2–5 min after CRH injection is indicative of Cushing's disease (Fig. 379-10), with very high sensitivity and specificity. Of note, the results of the IPSS cannot be reliably used for lateralization (i.e., prediction of the location of the tumor within the pituitary), because there is broad interindividual variability in the venous drainage of the pituitary region. Importantly, no cortisol-lowering agents should be used prior to IPSS.

If the differential diagnostic testing indicates ectopic ACTH syndrome, then further imaging should include high-resolution, fine-cut CT scanning of the chest and abdomen for scrutiny of the lung, thymus, and pancreas. If no lesions are identified, an MRI of the chest can be considered because carcinoid tumors usually show high signal intensity on T2-weighted images. Furthermore, octreotide scintigraphy can be helpful in some cases because ectopic ACTH-producing tumors often express somatostatin receptors. Depending on the suspected cause, patients with ectopic ACTH syndrome should also undergo blood sampling for fasting gut hormones, chromogranin A, calcitonin, and biochemical exclusion of pheochromocytoma.

TREATMENT

Cushing's Syndrome

Overt Cushing's is associated with a poor prognosis if left untreated. In ACTH-independent disease, treatment consists of surgical removal of the adrenal tumor. For smaller tumors, a minimally invasive approach can be used, whereas for larger tumors and those suspected of malignancy, an open approach is preferred.

In Cushing's disease, the treatment of choice is selective removal of the pituitary corticotrope tumor, usually via an endoscopic transsphenoidal approach. This results in an initial cure rate of 70–80% when performed by a highly experienced surgeon. However, even after initial remission following surgery, long-term follow-up is important because late relapse occurs in a significant number of patients. If pituitary disease recurs, there are several options, including second surgery, radiotherapy, stereotactic radiosurgery, and

bilateral adrenalectomy. These options need to be applied in a highly individualized fashion.

In some patients with very severe, overt Cushing's (e.g., difficult to control hypokalemic hypertension or acute psychosis), it may be necessary to introduce medical therapy to rapidly control the cortisol excess during the period leading up to surgery. Similarly, patients with metastasized, glucocorticoid-producing carcinomas may require long-term antiglucocorticoid drug treatment. In case of ectopic ACTH syndrome, in which the tumor cannot be located, one must carefully weigh whether drug treatment or bilateral adrenalectomy is the most appropriate choice, with the latter facilitating immediate cure but requiring life-long corticosteroid replacement. In this instance, it is paramount to ensure regular imaging follow-up for identification of the ectopic ACTH source.

Oral agents with established efficacy in Cushing's syndrome are metyrapone and ketoconazole. Metyrapone inhibits cortisol synthesis at the level of 11β-hydroxylase (Fig. 379-1), whereas the antimycotic drug ketoconazole inhibits the early steps of steroidogenesis. Typical starting doses are 500 mg tid for metyrapone (maximum dose, 6 g) and 200 mg tid for ketoconazole (maximum dose, 1200 mg). Mitotane, a derivative of the insecticide o,p'DDD, is an adrenolytic agent that is also effective for reducing cortisol. Because of its side effect profile, it is most commonly used in the context of ACC, but low-dose treatment (500–1000 mg/d) has also been used in benign Cushing's. In severe cases of cortisol excess, etomidate, an agent that potently blocks 11β-hydroxylase and aldosterone synthase, can be used to lower cortisol. It is administered by continuous IV infusion in low, nonanesthetic doses.

After the successful removal of an ACTH- or cortisol-producing tumor, the HPA axis will remain suppressed. Thus, hydrocortisone replacement needs to be initiated at the time of surgery and slowly tapered following recovery, to allow physiologic adaptation to normal cortisol levels. Depending on degree and duration of cortisol excess, the HPA axis may require many months or even years to resume normal function and sometimes does not recover. Generally, ectopic ACTH syndrome shows the best recovery rate (80%) and adrenal Cushing's has the lowest (40%), with Cushing's disease intermediate (60%).

MINERALOCORTICOID EXCESS

Epidemiology Following the first description of a patient with an aldosterone-producing adrenal adenoma (*Conn's syndrome*), mineralocorticoid excess was thought to represent a rare cause of hypertension. However, in studies systematically screening all patients with hypertension, a much higher prevalence is now recognized, ranging from 5 to 12%. The prevalence is higher when patients are preselected for hypokalemic hypertension.

Etiology The most common cause of mineralocorticoid excess is primary aldosteronism, reflecting excess production of aldosterone by the adrenal zona glomerulosa. Bilateral micronodular hyperplasia is somewhat more common than unilateral adrenal adenomas **(Table 379-3)**. Somatic mutations in channels and enzymes responsible for increasing sodium and calcium influx in adrenal zona glomerulosa cells have been identified as prevalent causes of aldosterone-producing adrenal adenomas (Table 379-3) and, in the case of germline mutations, also of primary aldosteronism due to bilateral macronodular adrenal hyperplasia. However, bilateral adrenal hyperplasia as a cause of mineralocorticoid excess is usually micronodular, but can also contain larger nodules that might be mistaken for a unilateral adenoma. In rare instances, primary aldosteronism is caused by an ACC. Carcinomas should be considered in younger patients and in those with larger tumors, because benign aldosterone-producing adenomas usually measure <2 cm in diameter.

A rare cause of aldosterone excess is glucocorticoid-remediable aldosteronism (GRA), which is caused by a chimeric gene resulting from crossover of promoter sequences between the *CYP11B1* and *CYP11B2* genes that are involved in glucocorticoid and mineralocorticoid synthesis, respectively (Fig. 379-1). This rearrangement brings *CYP11B2* transcription under the control of ACTH receptor signaling; consequently, aldosterone production is regulated by ACTH rather than by renin. The family history can be helpful because there may be evidence for dominant transmission of hypertension. Recognition of the disorder is important because it can be associated with early-onset hypertension and strokes. In addition, glucocorticoid suppression can reduce aldosterone production.

Other rare causes of mineralocorticoid excess are listed in Table 379-3. An important cause is excess binding and activation of the

TABLE 379-3 Causes of Mineralocorticoid Excess

CAUSES OF MINERALOCORTICOID EXCESS	MECHANISM	%
Primary Aldosteronism		
Adrenal (Conn's) adenoma	Autonomous aldosterone excess can be caused by somatic (intratumor) mutations in the potassium channel GIRK4 (encoded by *KCNJ5*; identified as cause of disease in 40% of aldosterone-producing adenomas; rare germline mutations can cause bilateral macronodular adrenal hyperplasia). Further causes include somatic mutations affecting the α-subunit of the Na^+/K^+-ATPase (encoded by *ATP1A1*), the plasma membrane calcium-transporting ATPase 3 (encoded by *ATP2B3*), and somatic mutations in *CACNA1D* or *CACNA1H* encoding the voltage-gated calcium channel CaV 1.3 and CaV3.2, respectively. All mutations result in upregulation of CYP11B2 and hence aldosterone synthesis.	60
Bilateral (micronodular) adrenal hyperplasia	Autonomous aldosterone excess, mostly micronodular and rarely macronodular, with germline *KCNJ5* mutations being a rare cause.	60
Glucocorticoid-remediable hyperaldosteronism (dexamethasone-suppressible hyperaldosteronism)	Crossover between the *CYP11B1* and *CYP11B2* genes results in ACTH-driven aldosterone production	<1
Other Causes (Rare)		**<1**
Syndrome of apparent mineralocorticoid excess (SAME)	Mutations in *HSD11B2* result in lack of renal inactivation of cortisol to cortisone, leading to excess activation of the MR by cortisol	
Cushing's syndrome	Cortisol excess overcomes the capacity of HSD11B2 to inactivate cortisol to cortisone, consequently flooding the MR	
Glucocorticoid resistance	Upregulation of cortisol production due to GR mutations results in flooding of the MR by cortisol	
Adrenocortical carcinoma	Autonomous aldosterone and/or DOC excess	
Congenital adrenal hyperplasia	Accumulation of DOC due to mutations in *CYP11B1* or *CYP17A1*	
Progesterone-induced hypertension	Progesterone acts as an abnormal ligand due to mutations in the MR gene	
Liddle's syndrome	Mutant ENaC β or γ subunits resulting in reduced degradation of ENaC keeping the membrane channel in open conformation for longer, enhancing mineralocorticoid action	

Abbreviations: ACTH, adrenocorticotropic hormone; DOC, deoxycorticosterone; ENaC, epithelial sodium channel; GR, glucocorticoid receptor; HSD11B2, 11β-hydroxysteroid dehydrogenase type 2; MR, mineralocorticoid receptor.

MR by a steroid other than aldosterone. Cortisol acts as a potent mineralocorticoid if it escapes efficient inactivation to cortisone by 11β-HSD2 in the kidney (Fig. 379-7). This can be caused by inactivating mutations in the *HSD11B2* gene resulting in the syndrome of apparent mineralocorticoid excess (SAME) that characteristically manifests with severe hypokalemic hypertension in childhood. However, milder mutations may cause normokalemic hypertension manifesting in adulthood (type II SAME). Inhibition of 11β-HSD2 by excess licorice ingestion also results in hypokalemic hypertension, as does overwhelming of 11β-HSD2 conversion capacity by cortisol excess in Cushing's syndrome. DOC also binds and activates the MR and can cause hypertension if its circulating concentrations are increased. This can arise through autonomous DOC secretion by an ACC, but also when DOC accumulates as a consequence of an adrenal enzymatic block, as seen in congenital adrenal hyperplasia (CAH) due to CYP11B1 (11β-hydroxylase) or CYP17A1 (17α-hydroxylase) deficiency (Fig. 379-1). Progesterone can cause hypokalemic hypertension in rare individuals who harbor a MR mutation that enhances binding and activation by progesterone; physiologically, progesterone normally exerts antimineralocorticoid activity. Finally, excess mineralocorticoid activity can be caused by mutations in the β or γ subunits of the ENaC, disrupting its interaction with Nedd4 (Fig. 379-7), and thereby decreasing receptor internalization and degradation. The constitutively active ENAC drives hypokalemic hypertension, resulting in an autosomal dominant disorder termed *Liddle's syndrome*.

Clinical Manifestations Excess activation of the MR leads to potassium depletion and increased sodium retention, with the latter causing an expansion of extracellular and plasma volume. Increased ENaC activity also results in hydrogen depletion that can cause metabolic alkalosis. Aldosterone also has direct effects on the vascular system, where it increases cardiac remodeling and decreases compliance. Aldosterone excess may cause direct damage to the myocardium and the kidney glomeruli, in addition to secondary damage due to systemic hypertension.

The clinical hallmark of mineralocorticoid excess is hypokalemic hypertension; serum sodium tends to be normal due to the concurrent fluid retention, which in some cases can lead to peripheral edema. Hypokalemia can be exacerbated by thiazide drug treatment, which leads to increased delivery of sodium to the distal renal tubule, thereby driving potassium excretion. Severe hypokalemia can be associated with muscle weakness, overt proximal myopathy, or even hypokalemic paralysis. Severe alkalosis contributes to muscle cramps and, in severe cases, can cause tetany.

Of note, patients with primary aldosteronism show increased rates of osteoporosis, type 2 diabetes, and cognitive dysfunction. A significant proportion of patients suffer from concurrent mild autonomous cortisol excess (MACE), termed *Connshing syndrome*.

Diagnosis Diagnostic screening for mineralocorticoid excess is not currently recommended for all patients with hypertension, but should be restricted to those who exhibit hypertension associated with drug resistance, hypokalemia, an adrenal mass, or onset of disease before the age of 40 years **(Fig. 379-12)**. The accepted screening test is concurrent measurement of plasma renin and aldosterone with subsequent calculation of the aldosterone-renin ratio (ARR) (Fig. 379-12); serum potassium needs to be normalized prior to testing. Stopping antihypertensive medication can be cumbersome, particularly in patients with severe hypertension. Thus, for practical purposes, in the first instance the patient can remain on the usual antihypertensive medications, with the exception that MR antagonists need to be ceased at least 4 weeks prior to ARR measurement. The remaining antihypertensive drugs usually do not affect the outcome of ARR testing, except that beta blocker treatment can cause false-positive results and ACE/AT1R inhibitors can cause false-negative results in milder cases **(Table 379-4)**.

ARR screening is positive if the ratio is >750 pmol/L per ng/mL per hour, with a concurrently high normal or increased aldosterone (Fig. 379-12). If one relies on the ARR only, the likelihood of a false-positive ARR becomes greater when renin levels are very low. The characteristics of the biochemical assays are also important. Some labs measure plasma renin activity, whereas others measure plasma renin concentrations. Antibody-based assays for the measurement of serum aldosterone lack the reliability of tandem mass spectrometry assays, but these are not yet ubiquitously available.

Diagnostic confirmation of mineralocorticoid excess in a patient with positive ARR screening result should be undertaken by an endocrinologist as the tests lack optimized validation. The most straightforward is the saline infusion test, which involves the IV administration of 2 L of physiologic saline over a 4-h period. Failure of aldosterone to suppress <140 pmol/L (5 ng/dL) is indicative of autonomous mineralocorticoid excess. Alternative tests are the oral sodium loading test (300 mmol NaCl/d for 3 days) or the fludrocortisone suppression test (0.1 mg q6h with 30 mmol NaCl q8h for 4 days); the latter can be difficult because of the risk of profound hypokalemia and increased hypertension. In patients with overt hypokalemic hypertension, strongly positive ARR, and concurrently increased aldosterone levels, confirmatory testing is usually not necessary.

Differential Diagnosis and Treatment After the diagnosis of hyperaldosteronism is established, the next step is to use adrenal imaging to further assess the cause. Fine-cut CT scanning of the adrenal region is the method of choice because it provides excellent visualization of adrenal morphology. CT will readily identify larger tumors suspicious of malignancy but may miss lesions <5 mm. The differentiation between bilateral micronodular hyperplasia and a unilateral adenoma is only required if a surgical approach is feasible and desired. Consequently, selective adrenal vein sampling (AVS) should only be carried out in surgical candidates with either no obvious lesion on CT or evidence of a unilateral lesion in patients >40 years, because the latter patients have a high likelihood of harboring a coincidental, endocrine-inactive adrenal adenoma (Fig. 379-12). AVS is used to compare aldosterone levels in the inferior vena cava and between the right and left adrenal veins. AVS requires concurrent measurement of cortisol to document correct placement of the catheter in the adrenal veins and should demonstrate a cortisol gradient >3 between the vena cava and each adrenal vein. Lateralization is confirmed by an aldosterone/cortisol ratio that is at least twofold higher on one side than the other. AVS is a complex procedure that requires a highly skilled interventional radiologist. Even then, the right adrenal vein can be difficult to cannulate correctly, which, if not achieved, invalidates the procedure. There is also no agreement as to whether the two adrenal veins should be cannulated simultaneously or successively and whether ACTH stimulation enhances the diagnostic value of AVS.

Patients >40 years with confirmed mineralocorticoid excess and a unilateral lesion on CT can go straight to surgery, which is also indicated in patients with confirmed lateralization documented by a valid AVS procedure. Laparoscopic adrenalectomy is the preferred approach. Patients who are not surgical candidates, or with evidence of bilateral hyperplasia based on CT or AVS, should be treated medically (Fig. 379-12). Medical treatment, which can also be considered prior to surgery to avoid postsurgical hypoaldosteronism, consists primarily of the MR antagonist spironolactone. It can be started at 12.5–50 mg bid and titrated up to a maximum of 400 mg/d to control blood pressure and normalize potassium. Side effects include menstrual irregularity, decreased libido, and gynecomastia. The more selective MR antagonist eplerenone can also be used. Doses start at 25 mg bid, and it can be titrated up to 200 mg/d. Another useful drug is the sodium channel blocker amiloride (5–10 mg bid).

In patients with normal adrenal morphology and family history of early-onset, severe hypertension, a diagnosis of GRA should be considered and can be evaluated using genetic testing. Treatment of GRA consists of administering dexamethasone, using the lowest dose possible to control blood pressure. Some patients also require additional MR antagonist treatment.

The diagnosis of nonaldosterone-related mineralocorticoid excess is based on documentation of suppressed renin and suppressed aldosterone in the presence of hypokalemic hypertension. This testing is best carried out by employing urinary steroid metabolite profiling by gas

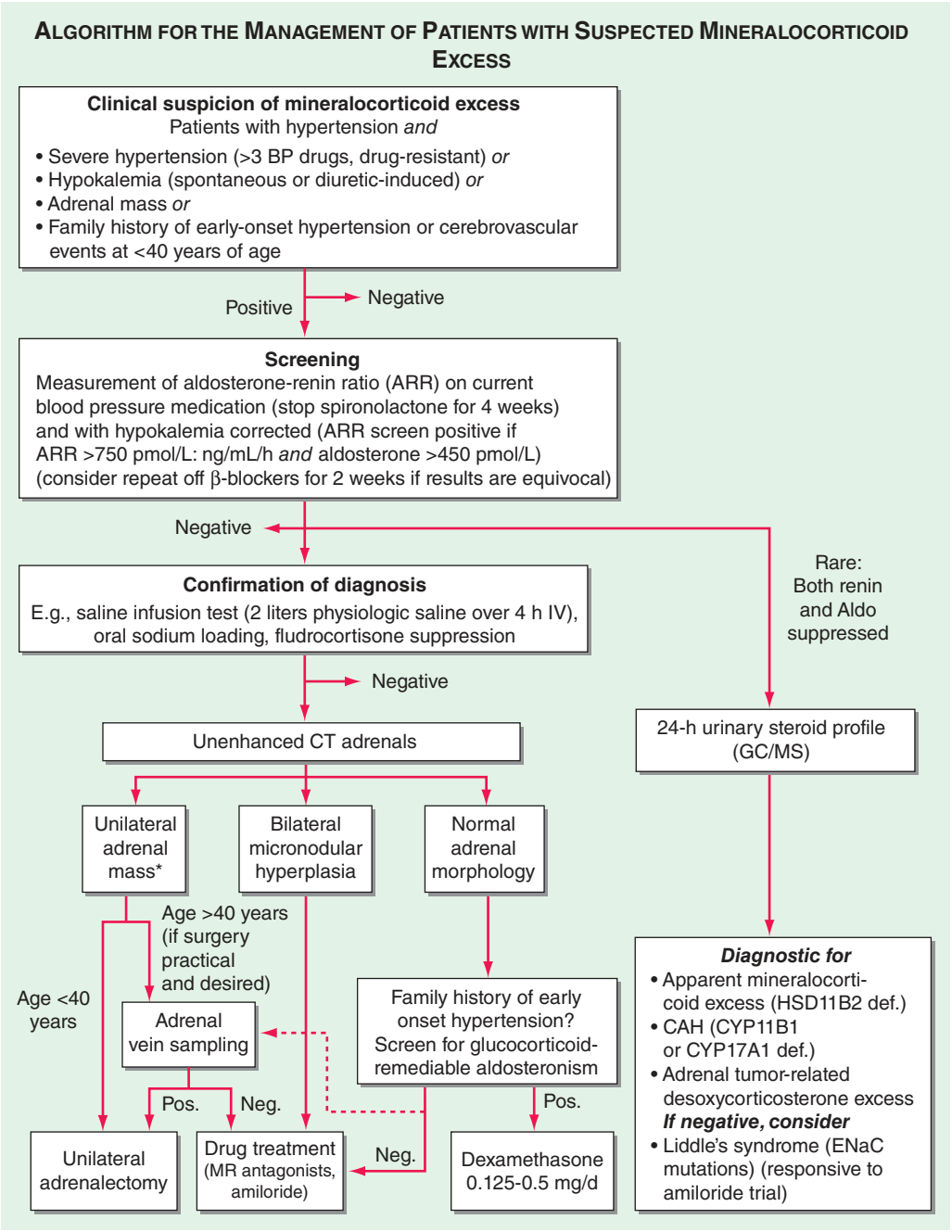


FIGURE 379-12 Management of patients with suspected mineralocorticoid excess. *Perform adrenal tumor workup (see Fig. 379-13). BP, blood pressure; CAH, congenital adrenal hyperplasia; CT, computed tomography; GC/MS, gas chromatography/mass spectrometry; PRA, plasma renin activity.

TABLE 379-4 Effects of Antihypertensive Drugs on the Aldosterone-Renin Ratio (ARR)

DRUG	EFFECT ON RENIN	EFFECT ON ALDOSTERONE	NET EFFECT ON ARR
β Blockers	↓	↑	↑
α_1 Blockers	→	→	→
α_2 Sympathomimetics	→	→	→
ACE inhibitors	↑	↓	↓
AT1R blockers	↑	↓	↓
Calcium antagonists	→	→	→
Diuretics	(↑)	(↑)	→/(↓)

Abbreviations: ACE, angiotensin-converting enzyme; AT1R, angiotensin II receptor type 1.

chromatography/mass spectrometry (GC/MS). An increased free cortisol over free cortisone ratio is suggestive of SAME and can be treated with dexamethasone. Steroid profiling by GC/MS also detects the steroids associated with CYP11B1 and CYP17A1 deficiency or the irregular steroid secretion pattern in a DOC-producing ACC (Fig. 379-12). If the GC/MS profile is normal, then Liddle's syndrome should be considered. It is very sensitive to amiloride treatment but will not respond to MR antagonist treatment, because the defect is due to a constitutively active ENaC.

APPROACH TO THE PATIENT: INCIDENTALLY DISCOVERED ADRENAL MASS

Epidemiology Incidentally discovered adrenal masses, commonly termed adrenal "incidentalomas," are common, with a prevalence of 2–5% in the general population as documented in CT and autopsy

series. The prevalence increases with age, with 1% of 40-year-olds and 7% of 70-year-olds harboring an adrenal mass. The widespread use of cross-sectional imaging has also increased the recognized prevalence.

Etiology Most solitary adrenal tumors are monoclonal neoplasms. Several genetic syndromes, including MEN 1 (*MEN1*), MEN 2 (*RET*), Carney's complex (*PRKAR1A*), and McCune-Albright (*GNAS1*), can have adrenal tumors as one of their features. Somatic mutations in *MEN1, GNAS1*, and *PRKAR1A* have been identified in a small proportion of sporadic adrenocortical adenomas. Aberrant expression of membrane receptors (GIP, α- and β-adrenergic, luteinizing hormone, vasopressin V1, and interleukin 1 receptors) has been identified in some sporadic cases of macronodular adrenocortical hyperplasia.

The majority of adrenal nodules are endocrine-inactive adrenocortical adenomas. However, larger series suggest that up to 25% of adrenal nodules are hormonally active, due to a cortisol- or aldosterone-producing adrenocortical adenoma or a pheochromocytoma associated with catecholamine excess **(Table 379-5)**. ACC is rare but is the cause of an adrenal mass in 5% of patients. However, metastases originating from another solid tissue tumor are an additional cause of adrenal incidentaloma, and have a higher incidence in patients undergoing imaging for tumor staging or follow-up monitoring (Table 379-5).

Differential Diagnosis and Treatment Patients with an adrenal mass >1 cm require a diagnostic evaluation. Two key questions need to be addressed: (1) Does the tumor autonomously secrete hormones that could have a detrimental effect on health? (2) Is the adrenal mass benign or malignant?

Hormone secretion by an adrenal mass occurs along a continuum, with a gradual increase in clinical manifestations in parallel with hormone levels. Exclusion of catecholamine excess from a pheochromocytoma arising from the adrenal medulla is a mandatory part of the diagnostic workup **(Fig. 379-13)**. Furthermore, autonomous cortisol resulting in Cushing's syndrome requires exclusion and, in patients with hypertension or low serum potassium, also primary aldosteronism. Adrenal incidentalomas can be associated with MACE, and patients usually lack overt clinical features of Cushing's syndrome.

TABLE 379-5 Classification of Unilateral Adrenal Masses

MASS	APPROXIMATE PREVALENCE (%)
Benign	
Adrenocortical adenoma	
Endocrine-inactive	60–85
Cortisol-producing	5–10
Aldosterone-producing	2–5
Pheochromocytoma	5–10
Adrenal myelolipoma	<1
Adrenal ganglioneuroma	<0.1
Adrenal hemangioma	<0.1
Adrenal cyst	<1
Adrenal hematoma/hemorrhagic infarction	<1
Indeterminate	
Adrenocortical oncocytoma	<1
Malignant	
Adrenocortical carcinoma	2–5
Malignant pheochromocytoma	<1
Adrenal neuroblastoma	<0.1
Lymphomas (including primary adrenal lymphoma)	<1
Metastases (most frequent: breast, lung)	1–2

Note: Bilateral adrenal enlargement/masses may be caused by congenital adrenal hyperplasia, bilateral macronodular hyperplasia, bilateral hemorrhage (due to antiphospholipid syndrome or sepsis-associated Waterhouse-Friderichsen syndrome), granuloma, amyloidosis, or infiltrative disease including tuberculosis.

Nonetheless, they may exhibit one or more components of the metabolic syndrome (e.g., obesity, type 2 diabetes, or hypertension). There is ongoing debate about the optimal treatment for these patients. Overproduction of adrenal androgen precursors, DHEA and its sulfate, is rare and most frequently seen in the context of ACC, as are increased levels of steroid precursors such as 17OHP.

For the differentiation of benign from malignant adrenal masses, imaging is relatively sensitive, although specificity is suboptimal. Unenhanced CT is the procedure of choice for imaging the adrenal glands (Fig. 379-11). A diagnosis of ACC, pheochromocytoma, and benign adrenal myelolipoma becomes more likely with increasing diameter of the adrenal mass. However, size alone is of poor predictive value, with only 80% sensitivity and 60% specificity for the differentiation of benign from malignant masses when using a 4-cm cut-off. Metastases are rare but are found with similar frequency in adrenal masses of all sizes. Tumor density on unenhanced CT is of additional diagnostic value, as many adrenocortical adenomas are lipid rich and thus present with low attenuation values (i.e., densities of <10 Hounsfield Units [HUs]). However, similar numbers of adrenocortical adenomas are lipid poor and present with higher HU, making it difficult to differentiate them from ACCs, as well as also pheochromocytomas, both of which invariably have high attenuation values (i.e., densities >20 HU on precontrast scans). Generally, benign lesions are rounded and homogenous, whereas most malignant lesions appear lobulated and inhomogeneous. Pheochromocytoma and adrenomyelolipoma may also exhibit lobulated and inhomogeneous features. MRI also allows for the visualization of the adrenal glands with somewhat lower resolution than CT. However, because it does not involve exposure to ionizing radiation, it is preferred in children, young adults, and during pregnancy. MRI has a valuable role in the characterization of indeterminate adrenal lesions using chemical shift analysis, with malignant tumors rarely showing loss of signal on opposed-phase MRI; however, this may also be observed in a proportion of benign adrenocortical adenomas.

Fine-needle aspiration (FNA) or CT-guided biopsy of an adrenal mass is very rarely indicated. FNA of a pheochromocytoma can cause a life-threatening hypertensive crisis. FNA of an ACC violates the tumor capsule and can cause needle track metastasis. FNA should only be considered in a patient with a history of nonadrenal malignancy and a newly detected adrenal mass, after careful exclusion of pheochromocytoma, and if the outcome will influence therapeutic management. It is important to recognize that in 25% of patients with a previous history of nonadrenal malignancy, a newly detected mass on CT is not a metastasis. While FNA can diagnose extra-adrenal malignancies, it has very limited ability to differentiate between benign and malignant adrenocortical lesions, and hence should not be used for diagnosis of ACC.

Adrenal masses associated with confirmed hormone excess or suspected malignancy are usually treated surgically (Fig. 379-13) or, if adrenalectomy is not feasible or desired, with medication. Preoperative exclusion of glucocorticoid excess is particularly important for the prediction of postoperative suppression of the contralateral adrenal gland, which requires glucocorticoid replacement peri- and postoperatively. If the initial decision is for observation, imaging and biochemical testing should be repeated 6–12 months after the first assessment. However, this may be performed earlier in patients with borderline imaging or hormonal findings. Adrenal masses associated with normal biochemistry at diagnosis and a tumor radiodensity of <10 HU on unenhanced CT can be safely assumed to represent a benign adenoma and do not require further follow-up.

ADRENOCORTICAL CARCINOMA

ACC is a rare malignancy with an annual incidence of 1–2 per million population. ACC is generally considered a highly malignant tumor; however, it presents with broad interindividual variability with regard to biologic characteristics and clinical behavior. Somatic mutations in the tumor-suppressor gene *TP53* are found in 25% of apparently sporadic ACC. Germline *TP53* mutations are the cause of the Li-Fraumeni syndrome associated with multiple solid organ cancers including ACC

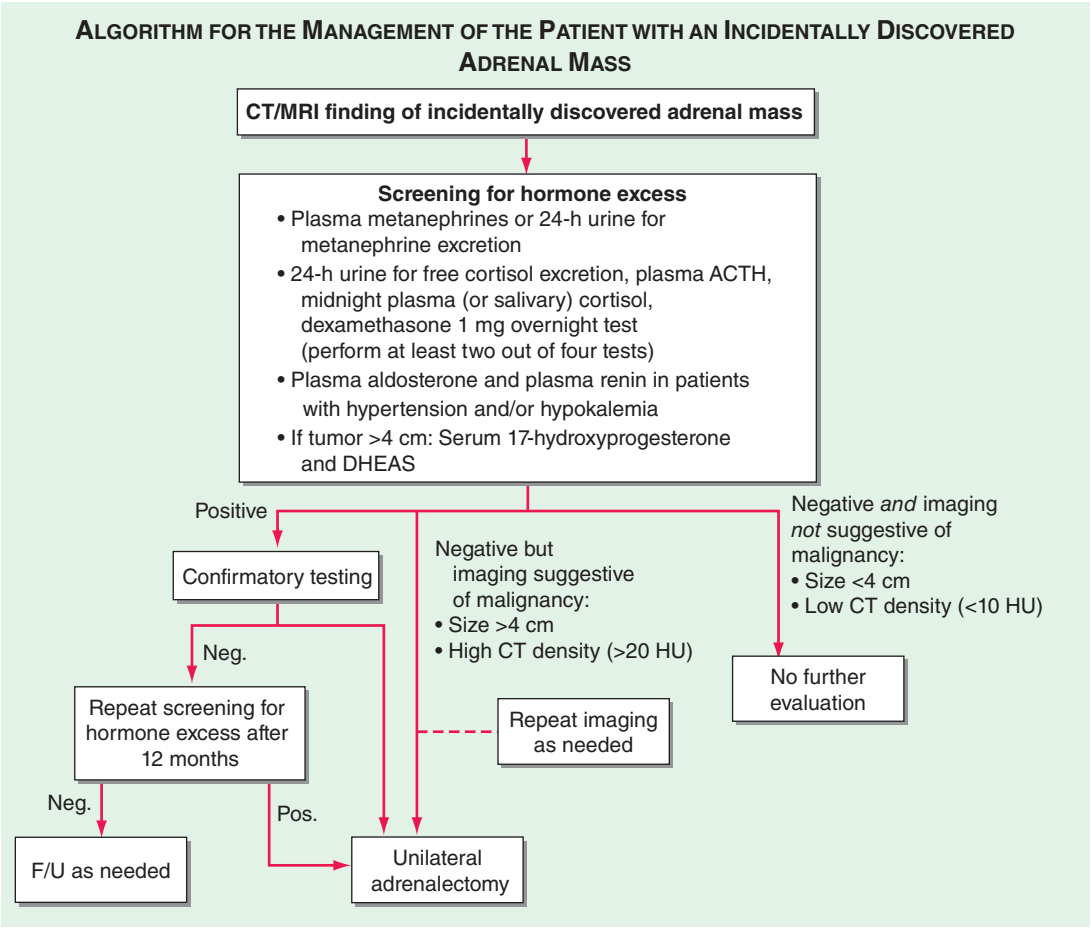


FIGURE 379-13 Management of the patient with an incidentally discovered adrenal mass. CT, computed tomography; F/U, follow-up; MRI, magnetic resonance imaging.

and are found in 25% of pediatric ACC cases; the *TP53* mutation R337H is found in almost all pediatric ACC in Brazil. Other genetic changes identified in ACC include alterations in the Wnt/β-catenin pathway and in the insulin-like growth factor 2 (IGF2) cluster; IGF2 overexpression is found in 90% of ACC.

Patients with large adrenal tumors suspicious of malignancy should be managed by a multidisciplinary specialist team, including an endocrinologist, an oncologist, a surgeon, a radiologist, and a histopathologist. FNA is not indicated in suspected ACC: first, cytology and also histopathology of a core biopsy cannot differentiate between benign and malignant primary adrenal masses; second, FNA violates the tumor capsule and may even cause needle canal metastasis. Even when the entire tumor specimen is available, the histopathologic differentiation between benign and malignant lesions is a diagnostic challenge. The most common histopathologic classification is the Weiss score, taking into account high nuclear grade; mitotic rate (>5/HPF); atypical mitosis; <25% clear cells; diffuse architecture; and presence of necrosis, venous invasion, and invasion of sinusoidal structures and tumor capsule. The presence of three or more elements suggests ACC.

Although 60–70% of ACCs show biochemical evidence of steroid overproduction, in many patients, this is not clinically apparent due to the relatively inefficient steroid production by the adrenocortical cancer cells. Excess production of glucocorticoids and adrenal androgen precursors are most common. Mixed excess production of several corticosteroid classes by an adrenal tumor is generally indicative of malignancy.

Tumor staging at diagnosis **(Table 379-6)** has important prognostic implications and requires scanning of the chest and abdomen for local organ invasion, lymphadenopathy, and metastases. Intravenous contrast medium is necessary for maximum sensitivity for hepatic metastases. An adrenal origin may be difficult to determine on standard axial CT imaging if the tumors are large and invasive, but CT reconstructions and MRI are more informative **(Fig. 379-14)** using multiple planes and different sequences. Vascular and adjacent organ invasion is diagnostic of malignancy. 18-Fluoro-2-deoxy-D-glucose positron emission tomography (18-FDG PET) is highly sensitive for the detection of malignancy and can be used to detect small metastases or local recurrence that may not be obvious on CT (Fig. 379-14). However, FDG PET has limited specificity and therefore cannot be used for differentiating benign from malignant adrenal lesions. Metastasis in ACC most frequently occurs to liver and lung.

There is no established grading system for ACC, and the Weiss score carries no prognostic value; the most important prognostic histopathologic parameter is the Ki67 proliferation index, with Ki67 <10% indicative of slow to moderate growth velocity, whereas a Ki67 ≥10% is

TABLE 379-6 Classification System for Staging of Adrenocortical Carcinoma

ENSAT STAGE	TNM STAGE	TNM DEFINITIONS
I	**T1,N0,M0**	T1, tumor ≤5 cm N0, no positive lymph node M0, no distant metastases
II	**T2,N0,M0**	T2, tumor >5 cm N0, no positive lymph node M0, no distant metastases
III	**T1–T2,N1,M0** **T3–T4,N0–N1,M0**	N1, positive lymph node(s) M0, no distant metastases T3, tumor infiltration into surrounding tissue T4, tumor invasion into adjacent organs *or* venous tumor thrombus in vena cava or renal vein
IV	**T1–T4,N0–N1,M1**	M1, presence of distant metastases

Abbreviations: ENSAT, European Network for the Study of Adrenal Tumors; TNM, tumor, node, metastasis.

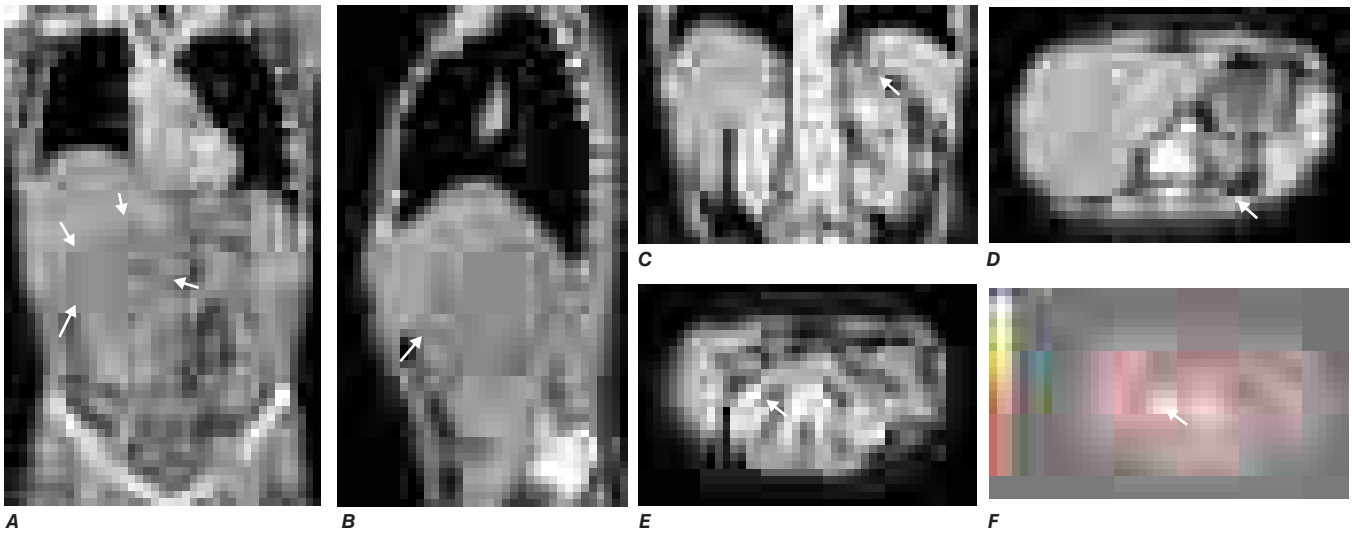


FIGURE 379-14 Imaging in adrenocortical carcinoma (ACC). Magnetic resonance imaging scan with (***A***) frontal and (***B***) lateral views of a right ACC that was detected incidentally. Computed tomography (CT) scan with (***C***) coronal and (***D***) transverse views depicting a right-sided ACC. Note the irregular border and inhomogeneous structure. CT scan (***E***) and positron emission tomography/CT (***F***) visualizing a peritoneal metastasis of an ACC in close proximity to the right kidney (*arrow*).

associated with poor prognosis including high risk of recurrence and rapid progression.

Cure of ACC can only be achieved by early detection and complete surgical removal. Capsule violation during primary surgery, metastasis at diagnosis, and primary treatment in a nonspecialist center and by a nonspecialist surgeon are major determinants of poor survival. If the primary tumor invades adjacent organs, en bloc removal of kidney and spleen should be considered to reduce the risk of recurrence and regional lymph node dissection may further reduce this risk. Surgery can also be considered in a patient with metastases if there is severe tumor-related hormone excess. This indication needs to be carefully weighed against surgical risk, including thromboembolic complications, and the resulting delay in the introduction of other therapeutic options. Patients with confirmed ACC and successful removal of the primary tumor should receive adjuvant treatment with mitotane (o,p'DDD), particularly in patients with a high risk of recurrence as determined by tumor size >8 cm, histopathologic signs of vascular invasion, capsule invasion or violation, and a Ki67 proliferation index ≥10%. Adjuvant mitotane should be continued for at least 2 years, if the patient can tolerate side effects. Regular monitoring of plasma mitotane levels is mandatory (therapeutic range 14–20 mg/L; neurotoxic complications more frequent at >20 mg/L). Mitotane is usually started at 500 mg tid, with stepwise increases to a maximum dose of 2000 mg tid in days (high-dose saturation) or weeks (low-dose saturation) as tolerated. Once therapeutic range plasma mitotane levels are achieved, the dose can be tapered to maintenance doses mostly ranging from 1000 to 1500 mg tid. Mitotane treatment results in disruption of cortisol synthesis and thus requires glucocorticoid replacement; glucocorticoid replacement dose should be at least double of that usually used in adrenal insufficiency (i.e., 20 mg tid) because mitotane induces hepatic CYP3A4 activity resulting in rapid inactivation of glucocorticoids. Mitotane also increases circulating CBG, thereby decreasing the available free cortisol fraction. Single metastases can be addressed surgically or with radiofrequency ablation as appropriate. If the tumor recurs or progresses during mitotane treatment, cytotoxic chemotherapy should be considered; the established first-line chemotherapy regimen is the combination of cisplatin, etoposide, and doxorubicin plus continuing mitotane. Painful bone metastasis responds to irradiation. Overall survival in ACC is still poor, with 5-year survival rates of 30–40% and a median survival of 15 months in metastatic ACC.

ADRENAL INSUFFICIENCY

Epidemiology The prevalence of well-documented, permanent adrenal insufficiency is 5 in 10,000 in the general population. Hypothalamic-pituitary origin of disease is most frequent, with a prevalence of 3 in 10,000, whereas primary adrenal insufficiency has a prevalence of 2 in 10,000. Approximately one-half of the latter cases are acquired, mostly caused by autoimmune destruction of the adrenal glands; the other one-half are genetic, most commonly caused by distinct enzymatic blocks in adrenal steroidogenesis affecting glucocorticoid synthesis (i.e., CAH.)

Adrenal insufficiency arising from suppression of the HPA axis as a consequence of exogenous glucocorticoid treatment is much more common, occurring in 0.5–2% of the population in developed countries.

Etiology *Primary adrenal insufficiency* is most commonly caused by autoimmune adrenalitis. Isolated autoimmune adrenalitis accounts for 30–40%, whereas 60–70% develop adrenal insufficiency as part of autoimmune polyglandular syndromes (APSs) **(Chap. 381) (Table 379-7)**. APS1, also termed APECED (autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy), is the underlying cause in 10% of patients affected by APS. APS1 is transmitted in an autosomal recessive manner and is caused by mutations in the autoimmune regulator gene *AIRE*. Associated autoimmune conditions overlap with those seen in APS2, but may also include total alopecia, primary hypoparathyroidism, and, in rare cases, lymphoma. APS1 patients invariably develop chronic mucocutaneous candidiasis, usually manifest in childhood, and preceding adrenal insufficiency by years or decades. The much more prevalent APS2 is of polygenic inheritance, with confirmed associations with the *HLA-DR3* gene region in the major histocompatibility complex and distinct gene regions involved in immune regulation (*CTLA-4, PTPN22, CLEC16A*). Coincident autoimmune disease most frequently includes thyroid autoimmune disease, vitiligo, and premature ovarian failure. Less commonly, additional features may include type 1 diabetes and pernicious anemia caused by vitamin B_{12} deficiency.

X-linked adrenoleukodystrophy has an incidence of 1:20,000 males and is caused by mutations in the *X-ALD* gene encoding the peroxisomal membrane transporter protein ABCD1; its disruption results in accumulation of very long chain (>24 carbon atoms) fatty acids. Approximately 50% of cases manifest in early childhood with rapidly progressive white matter disease (cerebral ALD); 35% present during adolescence or in early adulthood with neurologic features indicative of myelin and peripheral nervous system involvement (adrenomyeloneuropathy [AMN]). In the remaining 15%, adrenal insufficiency is the sole manifestation of disease. Of note, distinct mutations manifest with variable penetrance and phenotypes within affected families.

Rarer causes of adrenal insufficiency involve destruction of the adrenal glands as a consequence of infection, hemorrhage, or infiltration (Table 379-7); tuberculous adrenalitis is still a frequent cause of disease in developing countries. Adrenal metastases rarely cause adrenal insufficiency, and this occurs only with bilateral, bulky metastases.

TABLE 379-7 Causes of Primary Adrenal Insufficiency

DIAGNOSIS	GENE	ASSOCIATED FEATURES
Autoimmune polyglandular syndrome 1 (APS1)	*AIRE*	Hypoparathyroidism, chronic mucocutaneous candidiasis, other autoimmune disorders, rarely lymphomas
Autoimmune polyglandular syndrome 2 (APS2)	Associations with HLA-DR3, CTLA-4	Hypothyroidism, hyperthyroidism, premature ovarian failure, vitiligo, type 1 diabetes mellitus, pernicious anemia
Isolated autoimmune adrenalitis	Associations with HLA-DR3, CTLA-4	
Congenital adrenal hyperplasia (CAH)	*CYP21A2, CYP11B1, CYP17A1, HSD3B2, POR*	See Table 379-10 (see also **Chap. 383**)
Congenital lipoid adrenal hyperplasia (CLAH)	*STAR, CYP11A1*	46,XY DSD, gonadal failure (see also **Chap. 383**)
Adrenal hypoplasia congenita (AHC)	*NR0B1 (DAX-1), NR5A1 (SF-1)*	46,XY DSD, gonadal failure (see also **Chap. 383**)
Adrenoleukodystrophy (ALD), adrenomyeloneuropathy (AMN)	*ABCD1*	Demyelination of central nervous system (ALD) or spinal cord and peripheral nerves (AMN)
Familial glucocorticoid deficiency	*MC2R*	Tall stature
	MRAP	None
	STAR	None
	NNT	None
	TXNRD2	None
	MCM4	Growth retardation, natural killer cell deficiency
Triple A syndrome	*AAAS*	Alacrima, achalasia, neurologic impairment
Smith-Lemli-Opitz syndrome	*SLOS*	Cholesterol synthesis disorder associated with mental retardation, craniofacial malformations, growth failure
Kearns-Sayre syndrome	Mitochondrial DNA deletions	Progressive external ophthalmoplegia, pigmentary retinal degeneration, cardiac conduction defects, gonadal failure, hypoparathyroidism, type 1 diabetes,
IMAGe syndrome	*CDKN1C*	Intrauterine growth retardation, metaphyseal dysplasia, genital anomalies
MIRAGE syndrome	*SAMD9*	Myelodysplasia, infection, restriction of growth, genital phenotypes, and enteropathy
Sphingosine-1-phosphate lyase deficiency	*SGPL1*	Steroid-resistant nephrotic syndrome, immunodeficiency, neurological defects, ichthyosis, primary hypothyroidism, cryptorchidism
Adrenal infections		Tuberculosis, HIV, CMV, cryptococcosis, histoplasmosis, coccidioidomycosis
Adrenal infiltration		Metastases, lymphomas, sarcoidosis, amyloidosis, hemochromatosis
Adrenal hemorrhage		Meningococcal sepsis (Waterhouse-Friderichsen syndrome), primary antiphospholipid syndrome
Drug-induced		Mitotane, aminoglutethimide, abiraterone, trilostane, etomidate, ketoconazole, suramin, RU486
Bilateral adrenalectomy		E.g., in the management of Cushing's or after bilateral nephrectomy

Abbreviations: AIRE, autoimmune regulator; CMV, cytomegalovirus; DSD, disordered sex development; MC2R, ACTH receptor; MCM4, mini chromosome maintenance-deficient 4 homologue; MRAP, MC2R-accessory protein; NNT, nicotinamide nucleotide transhydrogenase.

Inborn causes of primary adrenal insufficiency other than CAH are rare, causing <1% of cases. However, their elucidation provides important insights into adrenal gland development and physiology. Mutations causing primary adrenal insufficiency (Table 379-7) include factors regulating adrenal development and steroidogenesis (DAX-1, SF-1), cholesterol synthesis, import and cleavage (DHCR7, StAR, CYP11A1), elements of the adrenal ACTH response pathway (MC2R, MRAP) (Fig. 379-5), and factors involved in redox regulation (NNT, TXNRD2) and DNA repair (MCM4, CDKN1C).

Secondary adrenal insufficiency is the consequence of dysfunction of the hypothalamic-pituitary component of the HPA axis **(Table 379-8)**. Excluding iatrogenic suppression, the overwhelming majority of cases are caused by pituitary or hypothalamic tumors or their treatment by surgery or irradiation **(Chap. 373)**. Rarer causes include pituitary apoplexy, either as a consequence of an infarcted pituitary adenoma or transient reduction in the blood supply of the pituitary during surgery or after rapid blood loss associated with parturition, also termed Sheehan's syndrome. Isolated ACTH deficiency is rarely caused by autoimmune disease or pituitary infiltration (Table 379-8). Mutations in the ACTH precursor POMC or in factors regulating pituitary development are genetic causes of ACTH deficiency (Table 379-8).

Clinical Manifestations In principle, the clinical features of primary adrenal insufficiency (Addison's disease) are characterized by the loss of both glucocorticoid and mineralocorticoid secretion **(Table 379-9)**. In secondary adrenal insufficiency, only glucocorticoid deficiency is present, as the adrenal itself is intact and thus still amenable to regulation by the RAA system. Adrenal androgen secretion is disrupted in both primary and secondary adrenal insufficiency (Table 379-9). Hypothalamic-pituitary disease can lead to additional clinical manifestations due to involvement of other endocrine axes (thyroid, gonads, growth hormone, prolactin) or visual impairment with bitemporal hemianopia caused by chiasmal compression. It is important to recognize that iatrogenic adrenal insufficiency caused by exogenous glucocorticoid suppression of the HPA axis may result in all symptoms associated with glucocorticoid deficiency (Table 379-9), if exogenous glucocorticoids are stopped abruptly. However, patients will appear clinically cushingoid as a result of the preceding overexposure to glucocorticoids.

Chronic adrenal insufficiency manifests with relatively nonspecific signs and symptoms, such as fatigue and loss of energy, often resulting in delayed or missed diagnoses (e.g., as depression or anorexia). A distinguishing feature of primary adrenal insufficiency is hyperpigmentation, which is caused by excess ACTH stimulation of melanocytes. Hyperpigmentation is most pronounced in skin areas exposed to increased friction or shear stress and is increased by sunlight **(Fig. 379-15)**. Conversely, in secondary adrenal insufficiency, the skin has an alabaster-like paleness due to lack of ACTH secretion.

Hyponatremia is a characteristic biochemical feature in primary adrenal insufficiency and is found in 80% of patients at presentation. Hyperkalemia is present in 40% of patients at initial diagnosis. Hyponatremia is primarily caused by mineralocorticoid deficiency but

TABLE 379-8 Causes of Secondary Adrenal Insufficiency

DIAGNOSIS	GENE	ASSOCIATED FEATURES
Pituitary tumors (endocrine active and inactive adenomas, very rare: carcinoma)		Depending on tumor size and location: visual field impairment (bilateral hemianopia), hyperprolactinemia, secondary hypothyroidism, hypogonadism, growth hormone deficiency
Other mass lesions affecting the hypothalamic-pituitary region		Craniopharyngioma, meningioma, ependymoma, metastases
Pituitary irradiation		Radiotherapy administered for pituitary tumors, brain tumors, or craniospinal irradiation in leukemia
Autoimmune hypophysitis		Often associated with pregnancy; may present with panhypopituitarism or isolated ACTH deficiency; can be associated with autoimmune thyroid disease, more rarely with vitiligo, premature ovarian failure, type 1 diabetes, pernicious anemia
Pituitary apoplexy/hemorrhage		Hemorrhagic infarction of large pituitary adenomas or pituitary infarction consequent to traumatic major blood loss (e.g., surgery or pregnancy: Sheehan's syndrome)
Pituitary infiltration		Tuberculosis, actinomycosis, sarcoidosis, histiocytosis X, granulomatosis with polyangiitis (Wegener's), metastases
Drug-induced		Chronic glucocorticoid excess (endogenous or exogenous)
Congenital isolated ACTH deficiency	*TBX19* (Tpit)	
Combined pituitary hormone deficiency (CPHD)	*PROP-1*	Progressive development of CPHD in the order GH, PRL, TSH, LH/FSH, ACTH
	HESX1	CPHD and septo-optic dysplasia
	LHX3	CPHD and limited neck rotation, sensorineural deafness
	LHX4	CPHD and cerebellar abnormalities
	SOX3	CPHD and variable mental retardation
Proopiomelanocortin (POMC) deficiency	*POMC*	Early-onset obesity, red hair pigmentation

Abbreviations: ACTH, adrenocorticotropic hormone; GH, growth hormone; LH/FSH, luteinizing hormone/follicle-stimulating hormone; PRL, prolactin; TSH, thyroid-stimulating hormone.

TABLE 379-9 Signs and Symptoms of Adrenal Insufficiency

Signs and Symptoms Caused by Glucocorticoid Deficiency
Fatigue, lack of energy
Weight loss, anorexia
Myalgia, joint pain
Fever
Normochromic anemia, lymphocytosis, eosinophilia
Slightly increased TSH (due to loss of feedback inhibition of TSH release)
Hypoglycemia (more frequent in children)
Low blood pressure, postural hypotension
Hyponatremia (due to loss of feedback inhibition of AVP release)
Signs and Symptoms Caused by Mineralocorticoid Deficiency (Primary Adrenal Insufficiency Only)
Abdominal pain, nausea, vomiting
Dizziness, postural hypotension
Salt craving
Low blood pressure, postural hypotension
Increased serum creatinine (due to volume depletion)
Hyponatremia
Hyperkalemia
Signs and Symptoms Caused by Adrenal Androgen Deficiency
Lack of energy
Dry and itchy skin (in women)
Loss of libido (in women)
Loss of axillary and pubic hair (in women)
Other Signs and Symptoms
Hyperpigmentation (primary adrenal insufficiency only) (due to excess of proopiomelanocortin [POMC]-derived peptides)
Alabaster-colored pale skin (secondary adrenal insufficiency only) (due to deficiency of POMC-derived peptides)

Abbreviations: AVP, arginine vasopressin; TSH, thyroid-stimulating hormone.

can also occur in secondary adrenal insufficiency due to diminished inhibition of antidiuretic hormone (ADH) release by cortisol, resulting in mild syndrome of inappropriate secretion of antidiuretic hormone (SIADH). Glucocorticoid deficiency also results in slightly increased TSH concentrations that normalize within days to weeks after initiation of glucocorticoid replacement.

Acute adrenal insufficiency, also termed adrenal crisis, usually occurs after a prolonged period of nonspecific complaints and is more frequently observed in patients with primary adrenal insufficiency, due to the loss of both glucocorticoid and mineralocorticoid secretion. Postural hypotension may progress to hypovolemic shock. Adrenal insufficiency may mimic features of acute abdomen with abdominal tenderness, nausea, vomiting, and fever. In some cases, the primary presentation may resemble neurologic disease, with decreased responsiveness, progressing to stupor and coma. An adrenal crisis can be triggered by an intercurrent illness, surgical or other stress, or increased glucocorticoid inactivation (e.g., hyperthyroidism). Prospective data indicate 8.3 adrenal crises and 0.5 adrenal crisis-related deaths per 100 patient years.

Diagnosis The diagnosis of adrenal insufficiency is established by the short cosyntropin test, a safe and reliable tool with excellent predictive diagnostic value **(Fig. 379-16)**. The cut-off for failure is usually defined at cortisol levels of <450–500 nmol/L (16–18 μg/dL) sampled 30–60 min after ACTH stimulation; the exact cut-off is dependent on the locally available assay, with generally lower cut-offs for mass spectrometry-based assays. During the early phase of HPA disruption (e.g., within 4 weeks of pituitary insufficiency), patients may still respond to exogenous ACTH stimulation. In this circumstance, the ITT is an alternative choice but is more invasive and should be carried out only under a specialist's supervision (see above). Induction of hypoglycemia is contraindicated in individuals with diabetes mellitus, cardiovascular disease, or history of seizures. Random serum cortisol measurements are of limited diagnostic value, because baseline cortisol levels may be coincidentally low due to the physiologic diurnal rhythm of cortisol secretion (Fig. 379-3). Similarly, many patients with secondary adrenal insufficiency have relatively normal baseline cortisol levels but fail to mount an appropriate cortisol response to ACTH, which can only be revealed by stimulation testing. Importantly, tests to establish the diagnosis of adrenal insufficiency should never delay treatment. Thus, in a patient with suspected adrenal crisis, it is reasonable to draw baseline cortisol levels, provide replacement therapy, and defer formal stimulation testing until a later time.

Once adrenal insufficiency is confirmed, measurement of plasma ACTH is the next step, with increased or inappropriately low levels defining primary and secondary origin of disease, respectively (Fig. 379-16). In primary adrenal insufficiency, increased plasma renin will confirm the presence of mineralocorticoid deficiency. At initial presentation, patients with primary adrenal insufficiency should undergo screening for steroid autoantibodies as a marker of

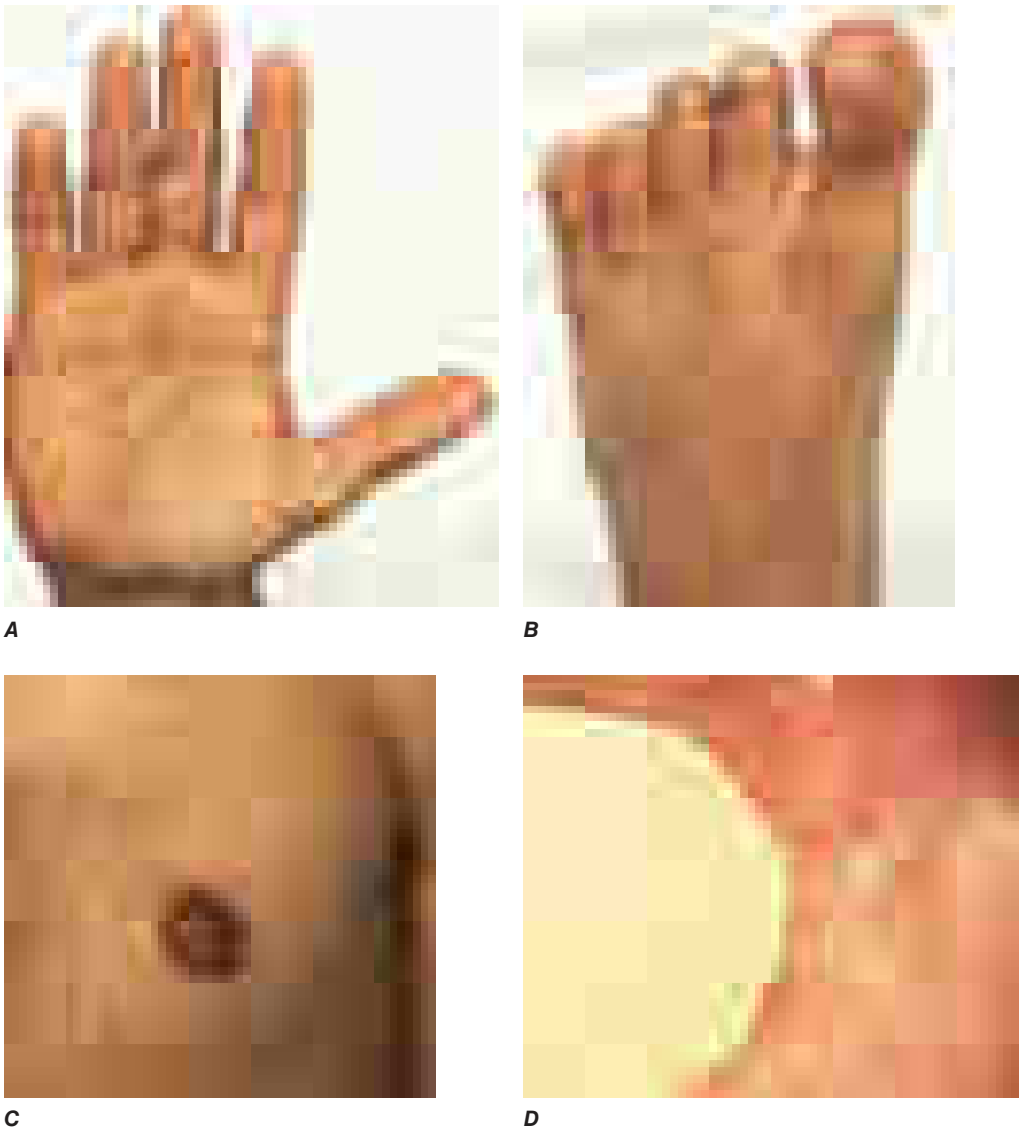


FIGURE 379-15 Clinical features of Addison's disease. Note the hyperpigmentation in areas of increased friction including (***A***) palmar creases, (***B***) dorsal foot, (***C***) nipples and axillary region, and (***D***) patchy hyperpigmentation of the oral mucosa.

autoimmune adrenalitis. If these tests are negative, adrenal imaging by CT is indicated to investigate possible hemorrhage, infiltration, or masses. In male patients with negative autoantibodies in the plasma, very long-chain fatty acids should be measured to exclude X-ALD. Patients with inappropriately low ACTH, in the presence of confirmed cortisol deficiency, should undergo hypothalamic-pituitary imaging by MRI. Features suggestive of preceding pituitary apoplexy, such as sudden-onset severe headache or history of previous head trauma, should be carefully explored, particularly in patients with no obvious MRI lesion.

TREATMENT

Acute Adrenal Insufficiency

Acute adrenal insufficiency requires immediate initiation of rehydration, usually carried out by saline infusion at initial rates of 1 L/h with continuous cardiac monitoring. Glucocorticoid replacement should be initiated by bolus injection of 100 mg hydrocortisone, followed by the administration of 200 mg hydrocortisone over 24 h, preferably by continuous infusion or alternatively by bolus IV or IM injections. Mineralocorticoid replacement can be initiated once the daily hydrocortisone dose has been reduced to <50 mg because at higher doses hydrocortisone provides sufficient stimulation of MRs.

Glucocorticoid replacement for the treatment of chronic adrenal insufficiency should be administered at a dose that replaces the physiologic daily cortisol production, which is usually achieved by the oral administration of 15–25 mg hydrocortisone in two to three divided doses. Pregnancy may require an increase in hydrocortisone dose by 50% during the last trimester. In all patients, at least one-half of the daily dose should be administered in the morning. Currently available glucocorticoid preparations fail to mimic the physiologic cortisol secretion rhythm (Fig. 379-3). Long-acting glucocorticoids such as prednisolone or dexamethasone are not preferred because they result in increased glucocorticoid exposure due to extended GR activation at times of physiologically low cortisol secretion. There are no well-established dose equivalencies, but as a guide, equipotency can be assumed for 1 mg hydrocortisone, 1.6 mg cortisone acetate, 0.2 mg prednisolone, 0.25 mg prednisone, and 0.025 mg dexamethasone.

Monitoring of glucocorticoid replacement is mainly based on the history and examination for signs and symptoms suggestive of glucocorticoid over- or underreplacement, including assessment of body weight and blood pressure. Plasma ACTH, 24-h urinary free cortisol, or serum cortisol day curves reflect whether hydrocortisone has been taken or not, but do not convey reliable information about replacement quality. In patients with isolated primary adrenal insufficiency, monitoring should include screening for autoimmune

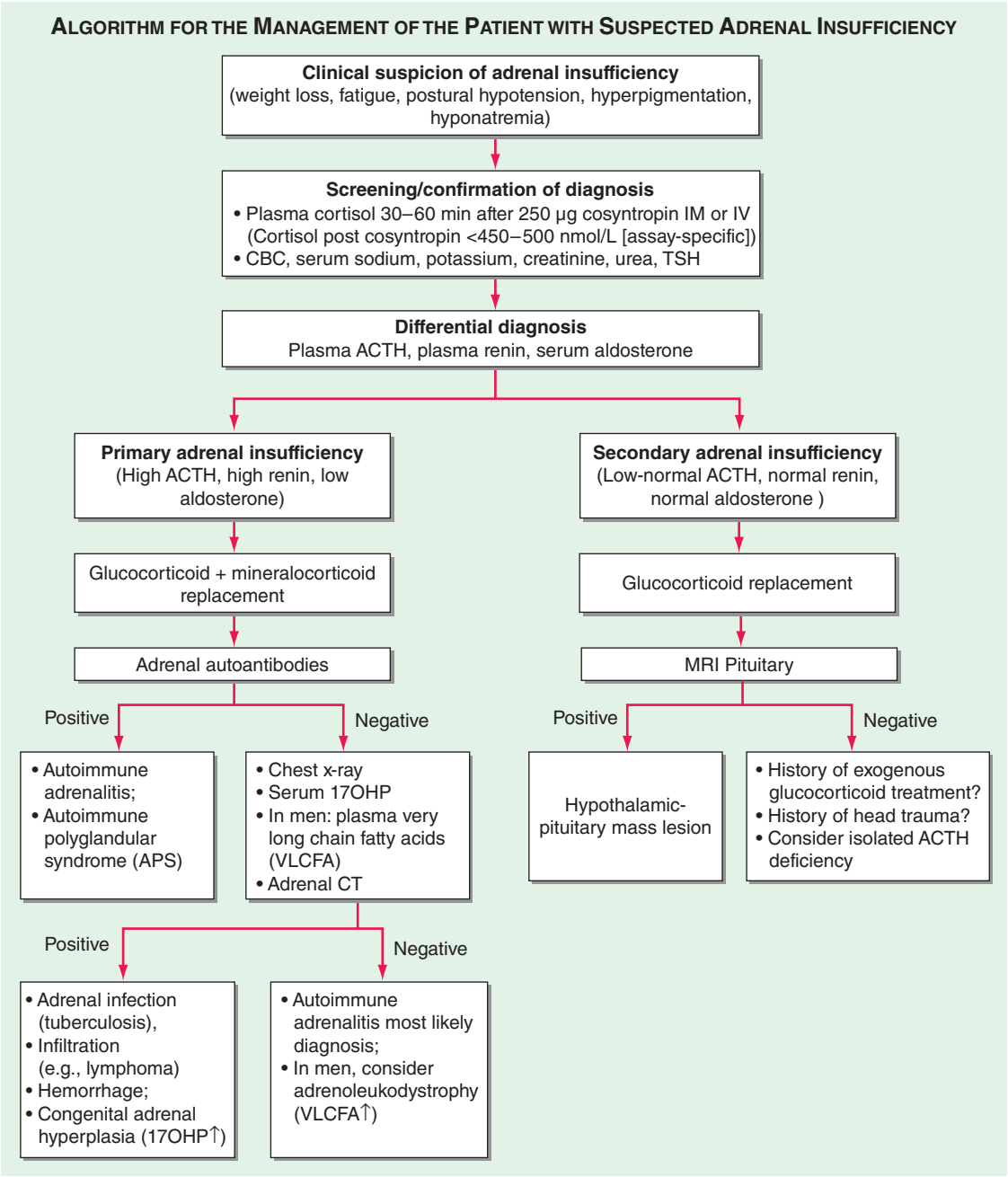


FIGURE 379-16 Management of the patient with suspected adrenal insufficiency. ACTH, adrenocorticotropic hormone; CBC, complete blood count; MRI, magnetic resonance imaging; PRA, plasma renin activity; TSH, thyroid-stimulating hormone.

thyroid disease, and female patients should be made aware of the possibility of premature ovarian failure. Supraphysiologic glucocorticoid treatment with doses equivalent to 30 mg hydrocortisone or more will affect bone metabolism, and these patients should undergo regular bone mineral density evaluation. All patients with adrenal insufficiency need to be instructed about the requirement for stress-related glucocorticoid dose adjustments. These generally consist of doubling the routine oral glucocorticoid dose in the case of intercurrent illness with fever and bed rest and the need for IV hydrocortisone injection at a daily dose of 100 mg in cases of prolonged vomiting, surgery, or trauma. All patients, but in particular those living or traveling in regions with delayed access to acute health care, should carry a hydrocortisone self-injection emergency kit, in addition to their usual steroid emergency cards and bracelets, and should receive training in its use.

Mineralocorticoid replacement in primary adrenal insufficiency should be initiated at a dose of 100–150 µg fludrocortisone. The adequacy of treatment can be evaluated by measuring blood pressure, sitting and standing, to detect a postural drop indicative of hypovolemia. In addition, serum sodium, potassium, and plasma renin should be measured regularly. Renin levels should be kept in the upper normal reference range. Changes in glucocorticoid dose may also impact on mineralocorticoid replacement as cortisol also binds the MR; 40 mg hydrocortisone is equivalent to 100 µg fludrocortisone. In patients living or traveling in areas with hot or tropical weather conditions, the fludrocortisone dose should be increased by 50–100 µg during the summer. Mineralocorticoid dose may also need to be adjusted during pregnancy, due to the antimineralocorticoid activity of progesterone, but this is less often required than hydrocortisone dose adjustment. Plasma renin cannot serve as a monitoring tool during pregnancy, because renin rises physiologically during gestation.

Adrenal androgen replacement is an option in patients with lack of energy, despite optimized glucocorticoid and mineralocorticoid replacement. It may also be indicated in women with features of androgen deficiency, including loss of libido. Adrenal

TABLE 379-10 Variants of Congenital Adrenal Hyperplasia

VARIANT	GENE	IMPACT ON STEROID SYNTHESIS	DIAGNOSTIC MARKER STEROIDS IN SERUM (AND URINE)
21-Hydroxylase deficiency (21OHD)	*CYP21A2*	Glucocorticoid deficiency, mineralocorticoid deficiency, adrenal androgen excess	17-Hydroxyprogesterone, 21-deoxycortisol (pregnanetriol, 17-hydroxypregnanolone, pregnanetriolone)
11β-Hydroxylase deficiency (11OHD)	*CYP11B1*	Glucocorticoid deficiency, mineralocorticoid excess, adrenal androgen excess	11-Deoxycortisol, 11-deoxycorticosterone (tetrahydro-11-deoxycortisol, tetrahydro-11-deoxycorticosterone)
17α-Hydroxylase deficiency (17OHD)	*CYP17A1*	(Glucocorticoid deficiency), mineralocorticoid excess, androgen deficiency	11-Deoxycorticosterone, corticosterone, pregnenolone, progesterone (tetrahydro-11-deoxycorticosterone, tetrahydrocorticosterone, pregnenediol, pregnanediol)
3β-Hydroxysteroid dehydrogenase deficiency (3bHSDD)	*HSD3B2*	Glucocorticoid deficiency, (mineralocorticoid deficiency), adrenal androgen excess (females and males), gonadal androgen deficiency (males)	17-Hydroxypregnanolone (pregnanetriol)
P450 oxidoreductase deficiency (PORD)	*POR*	Glucocorticoid deficiency, (mineralocorticoid excess), prenatal androgen excess and postnatal androgen deficiency, skeletal malformations	Pregnenolone, progesterone, 17-hydroxyprogesterone (pregnanediol, pregnanetriol)

androgen replacement can be achieved by once-daily administration of 25–50 mg DHEA. Treatment is monitored by measurement of DHEAS, androstenedione, testosterone, and sex hormone–binding globulin (SHBG) 24 h after the last DHEA dose.

CONGENITAL ADRENAL HYPERPLASIA

(See also Chap. 383) CAH is caused by mutations in genes encoding steroidogenic enzymes involved in glucocorticoid synthesis (*CYP21A2, CYP17A1, HSD3B2, CYP11B1*) or in the cofactor enzyme P450 oxidoreductase that serves as an electron donor to CYP21A2 and CYP17A1 (Fig. 379-1). Invariably, patients affected by CAH exhibit glucocorticoid deficiency. Depending on the exact step of enzymatic block, they may also have excess production of mineralocorticoids or deficient production of sex steroids **(Table 379-10)**. The diagnosis of CAH is readily established by measurement of the steroids accumulating before the distinct enzymatic block, either in serum or in urine, preferably by the use of mass spectrometry–based assays (Table 379-10).

Mutations in *CYP21A2* are the most prevalent cause of CAH, responsible for 90–95% of cases. 21-Hydroxylase deficiency disrupts glucocorticoid and mineralocorticoid synthesis (Fig. 379-1), resulting in diminished negative feedback via the HPA axis. This leads to increased pituitary ACTH release, which drives increased synthesis of adrenal androgen precursors and subsequent androgen excess. The degree of impairment of glucocorticoid and mineralocorticoid secretion depends on the severity of mutations. Major loss-of-function mutations result in combined glucocorticoid and mineralocorticoid deficiency (classic CAH, neonatal presentation), whereas less severe mutations affect glucocorticoid synthesis only (simple virilizing CAH, neonatal or early childhood presentation). The mildest mutations result in the least severe clinical phenotype, nonclassic CAH, usually presenting during adolescence and early adulthood and with preserved glucocorticoid production.

Androgen excess is present in all patients and manifests with broad phenotypic variability, ranging from severe virilization of the external genitalia in neonatal girls (e.g., 46,XX disordered sex development [DSD]) to hirsutism and oligomenorrhea resembling a polycystic ovary syndrome phenotype in young women with nonclassic CAH. In countries without neonatal screening for CAH, boys with classic CAH usually present with life-threatening adrenal crisis in the first few weeks of life (salt-wasting crisis); a simple-virilizing genotype manifests with precocious pseudopuberty and advanced bone age in early childhood, whereas men with nonclassic CAH are usually detected only through family screening.

Glucocorticoid treatment is more complex than for other causes of primary adrenal insufficiency as it not only needed to replace missing glucocorticoids but also to control the increased ACTH drive and subsequent androgen excess. Current treatment is hampered by the lack of glucocorticoid preparations that mimic the diurnal cortisol secretion profile, resulting in a prolonged period of ACTH stimulation and subsequent androgen production during the early morning hours. In childhood, optimization of growth and pubertal development are important goals of glucocorticoid treatment, in addition to prevention of adrenal crisis and treatment of 46,XX DSD. In adults, the focus shifts to preserving fertility and preventing side effects of glucocorticoid overtreatment, namely, the metabolic syndrome and osteoporosis. Fertility can be compromised in women due to oligomenorrhea/amenorrhea with chronic anovulation as a consequence of androgen excess. Men may develop so-called testicular adrenal rest tumors **(Fig. 379-17)**. These consist of hyperplastic cells with shared adrenal and gonadal characteristics located in the rete testis and should not be confused with testicular tumors. Testicular adrenal rest tissue can compromise sperm production and induce testicular fibrosis that may be irreversible.

TREATMENT

Congenital Adrenal Hyperplasia

Hydrocortisone is a good treatment option for the prevention of adrenal crisis, but longer acting prednisolone may be needed to control androgen excess. In children, hydrocortisone is given in divided doses at 1–1.5 times the normal cortisol production rate (about 10–13 mg/m^2 per day). In adults, if hydrocortisone does not suffice, intermediate-acting glucocorticoids (e.g., prednisone) may be given, using the lowest dose necessary to suppress excess androgen production. For achieving fertility, dexamethasone treatment may be required, but should be only given for the shortest possible time period to limit adverse metabolic side effects. Biochemical monitoring should include androstenedione and testosterone, aiming for the normal sex-specific reference range. 17OHP is a useful marker of overtreatment, indicated by 17OHP levels within the normal range of healthy controls. Glucocorticoid overtreatment may suppress the hypothalamic-pituitary-gonadal axis. Thus, treatment needs to be carefully titrated against clinical features of disease control. Stress dose glucocorticoids should be given at double or triple the daily dose for surgery, acute illness, or severe trauma. Poorly controlled CAH can result in adrenocortical hyperplasia, which gave the disease its name, and may present as macronodular hyperplasia subsequent to long-standing ACTH excess (Fig. 379-17). The nodular areas can develop autonomous adrenal androgen production and may be unresponsive to glucocorticoid treatment.

Mineralocorticoid requirements change during life and are higher in children, explained by relative mineralocorticoid resistance that diminishes with ongoing maturation of the kidney. Children with CAH usually receive mineralocorticoid and salt replacement. However, young adults with CAH should undergo reassessment of their mineralocorticoid reserve. Plasma renin should be regularly monitored and kept within the upper half of the normal reference range.

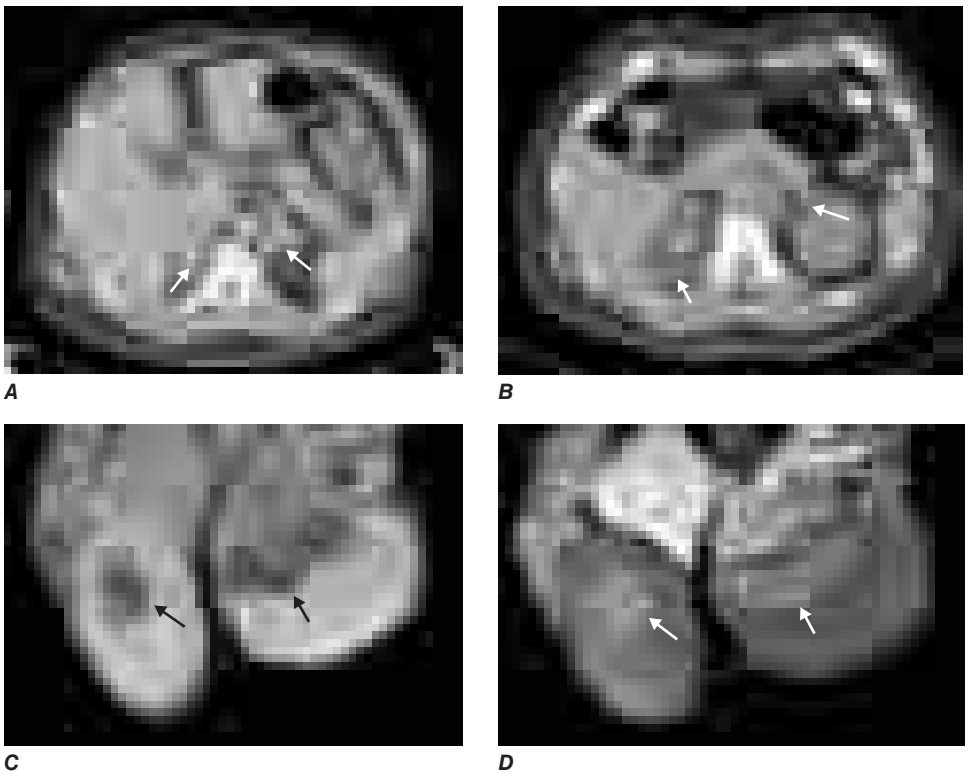


FIGURE 379-17 Imaging in congenital adrenal hyperplasia (CAH). Adrenal computed tomography scans showing homogenous bilateral hyperplasia in a young patient with classic CAH (***A***) and macronodular bilateral hyperplasia (***B***) in a middle-aged patient with classic CAH with longstanding poor disease control. Magnetic resonance imaging scan with T1-weighted (***C***) and T2-weighted (***D***) images showing bilateral testicular adrenal rest tumors (*arrows*) in a young patient with salt-wasting CAH. *(Courtesy of N. Reisch.)*

FURTHER READING

Arlt W et al: Steroid metabolome analysis reveals prevalent glucocorticoid excess in primary aldosteronism. JCI Insight 2017 April 20;2(8). pii: 93136. doi: 10.1172/jci.insight.93136. [Epub ahead of print]

Bancos I et al: Diagnosis and management of adrenal insufficiency. Lancet Diabetes Endocrinol 3:216, 2015.

Berr CM et al: Time to recovery of adrenal function after curative surgery for Cushing's syndrome depends on etiology. J Clin Endocrinol Metab 100:1300, 2015.

Bornstein SR et al: Diagnosis and treatment of primary adrenal insufficiency: An Endocrine Society Clinical Practice Guideline. J Clin Endocrinol Metab 101:364, 2016.

Dinnes J et al: Management of endocrine disease: Imaging for the diagnosis of malignancy in incidentally discovered adrenal masses: A systematic review and meta-analysis. Eur J Endocrinol 175:R51, 2016.

El-Maouche D et al: Congenital adrenal hyperplasia. Lancet 2017 May 30. pii: S0140-6736(17)31431-9. doi: 10.1016/S0140-6736(17)31431-9. [Epub ahead of print]

Fassnacht M et al: Management of adrenal incidentalomas: European Society of Endocrinology Clinical Practice Guideline in collaboration with the European Network for the Study of Adrenal Tumors. Eur J Endocrinol 175:G1, 2016.

Lodish M, Stratakis CA: A genetic and molecular update on adrenocortical causes of Cushing syndrome. Nat Rev Endocrinol 12:255, 2016.

Loriaux DL: Diagnosis and differential diagnosis of Cushing's syndrome. N Engl J Med 376:1451, 2017.

Mallappa A et al: A phase 2 study of Chronocort, a modified-release formulation of hydrocortisone, in the treatment of adults with classic congenital adrenal hyperplasia. J Clin Endocrinol Metab 100:1137, 2015.

Nieman LK et al: Treatment of Cushing's syndrome: An Endocrine Society Clinical Practice Guideline. J Clin Endocrinol Metab 100: 2807, 2015.

Turcu AF et al: Single-dose study of a corticotropin-releasing factor receptor-1 antagonist in women with 21-hydroxylase deficiency. J Clin Endocrinol Metab 101:1174, 2016.

Williams TA et al: Outcomes after adrenalectomy for unilateral primary aldosteronism: An international consensus on outcome measures and analysis of remission rates in an international cohort. Lancet Diabetes Endocrinol 5:689, 2017.

Young WF Jr: The management of primary aldosteronism: Case detection, diagnosis, and treatment: An endocrine society clinical practice guideline. J Clin Endocrinol Metab 101:1889, 2016.

380 Pheochromocytoma

Hartmut P. H. Neumann

Pheochromocytomas and paragangliomas are catecholamine-producing tumors derived from the sympathetic or parasympathetic nervous system. These tumors may arise sporadically or be inherited as features of multiple endocrine neoplasia type 2 (MEN 2), von Hippel–Lindau (VHL) disease, or several other pheochromocytoma-associated syndromes. The diagnosis of pheochromocytomas identifies a potentially correctable cause of hypertension, and their removal can prevent hypertensive crises that can be lethal. The clinical presentation is variable, ranging from an adrenal incidentaloma to a hypertensive crisis with associated cerebrovascular or cardiac complications.

EPIDEMIOLOGY

Pheochromocytoma was first described in 1800 by Charles Sugrue from Cork, Ireland, and the histological findings were first reported by Felix Fraenkel and Max Schottelius from Freiburg, Germany, in 1886. Pheochromocytoma is estimated to occur in 2–8 of 1 million persons

per year, and ~0.1% of hypertensive patients harbor a pheochromocytoma. The mean age at diagnosis is ~40 years, although the tumors can occur from early childhood until late in life. The classic "rule of tens" for pheochromocytomas states that ~10% are bilateral, 10% are extra-adrenal, and 10% are malignant.

ETIOLOGY AND PATHOGENESIS

Pheochromocytomas and paragangliomas are well-vascularized tumors that arise from cells derived from the sympathetic (e.g., adrenal medulla or sympathetic trunk) or parasympathetic (e.g., carotid body, glomus tympanicum, glomus jugulare, glomus vagale) paraganglia **(Fig. 380-1)**. The name *pheochromocytoma* reflects the formerly used black-colored staining caused by chromaffin oxidation of catecholamines; although a variety of terms have been used to describe these tumors, most clinicians use this designation to describe symptomatic catecholamine-producing tumors, including those in extra-adrenal retroperitoneal, pelvic, and thoracic sites. The term *paraganglioma* is used to describe catecholamine-producing tumors in the skull base and neck; these tumors may secrete little or no catecholamine. In contrast to common clinical parlance, the World Health Organization (WHO) restricts the term *pheochromocytoma* to adrenal tumors and applies the term *paraganglioma* to tumors at all other sites.

The etiology of sporadic pheochromocytomas and paragangliomas is unknown. However, 25–33% of patients have an inherited condition, including germline mutations in the classically recognized *RET* (*re*arranged during *t*ransfection), *VHL, NF1* (neurofibromatosis type 1), *SDHB, SDHC,* and *SDHD* (subunits of SDH) genes or in the more recently recognized *SDHA, SDHAF2, TMEM127* (transmembrane protein 127), *MAX* (myc-associated factor X), *FH* (fumarate hydratase), *PDH1, PDH2* (pyruvate dehydrogenase), *HIF1alpha* (hypoxia-inducible factor), *MDH2* (malate dehydrogenase), and *KIF1Bß* (kinesin family member) genes. Biallelic gene inactivation, a characteristic of tumor suppressor genes has been demonstrated for the *VHL, NF1, SDHx, TMEM127, MAX, FH, PDH1, PDH2, MDH2,* and *KIF1Bß* genes. In contrast, *RET* is a protooncogene, and mutations activate receptor tyrosine kinase activity. Succinate dehydrogenase (SDH) is an enzyme of the Krebs cycle and the mitochondrial respiratory chain. The VHL protein is a component of an ubiquitin E3 ligase. *VHL* mutations reduce protein degradation, resulting in upregulation of components involved in cell cycle progression, glucose metabolism, and oxygen sensing. Currently, pheochromocytoma and paraganglioma susceptibility genes are attributed to two clusters. Cluster 1 mutations are associated with pseudohypoxia and aberrant VEGF signaling (*VHL, PHD, FH,* or *SDHx* [subunits of SDH]) whereas cluster 2 mutations are associated with abnormal activation of kinase signaling pathways (*RET, NF1, TMEM127, MAX,* or *KIF1Bβ*).

TABLE 380-1 Clinical Features Associated with Pheochromocytoma, Listed by Frequency of Occurrence

1. Headaches	10. Weight loss
2. Profuse sweating	11. Paradoxical response to antihypertensive drugs
3. Palpitations and tachycardia	12. Polyuria and polydipsia
4. Hypertension, sustained or paroxysmal	13. Constipation
5. Anxiety and panic attacks	14. Orthostatic hypotension
6. Pallor	15. Dilated cardiomyopathy
7. Nausea	16. Erythrocytosis
8. Abdominal pain	17. Elevated blood sugar
9. Weakness	18. Hypercalcemia

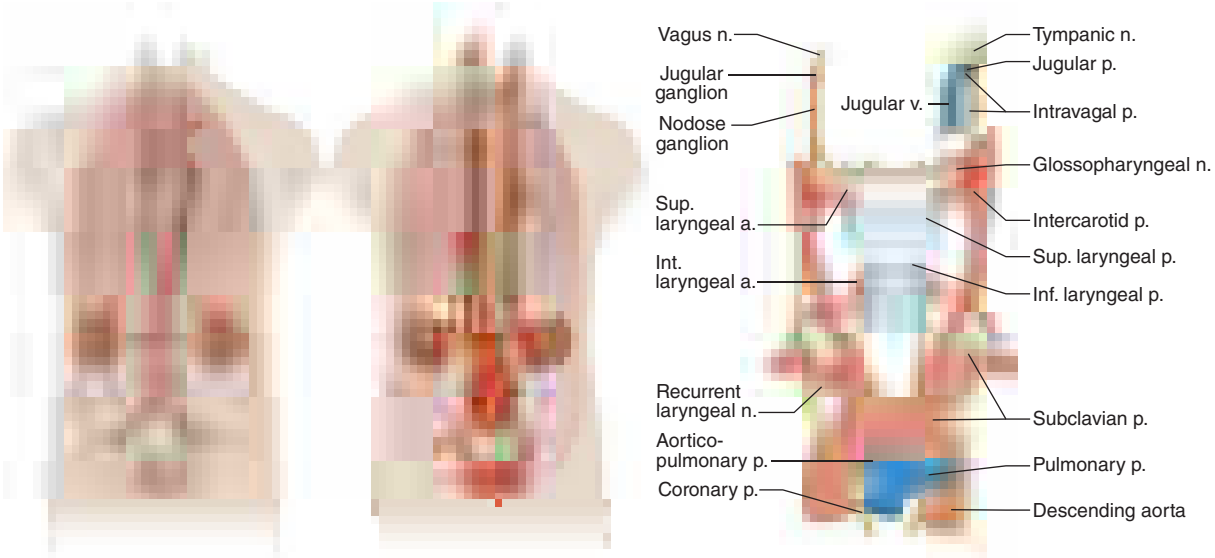
CLINICAL FEATURES

Its clinical presentation is so variable that pheochromocytoma has been termed "the great masquerader" **(Table 380-1)**. Among the presenting manifestations, episodes of palpitation, headache, and profuse sweating are typical, and these manifestations constitute a classic triad. The presence of all three manifestations in association with hypertension makes pheochromocytoma a likely diagnosis. However, a pheochromocytoma can be asymptomatic for years, and some tumors grow to a considerable size, before patients note symptoms.

The dominant sign is hypertension. Classically, patients have episodic hypertension, but sustained hypertension is also common. Catecholamine crises can lead to heart failure, pulmonary edema, arrhythmias, and intracranial hemorrhage. During episodes of hormone release, which can occur at widely divergent intervals, patients are anxious and pale, and they experience tachycardia and palpitations. These paroxysms generally last <1 h and may be precipitated by surgery, positional changes, exercise, pregnancy, urination (particularly with bladder pheochromocytomas), and various medications (e.g., tricyclic antidepressants, opiates, metoclopramide).

DIAGNOSIS

The diagnosis is based on documentation of catecholamine excess by biochemical testing and localization of the tumor by imaging. These two criteria are of equal importance, although measurement



A Adrenal pheochromocytoma

B Extra-adrenal pheochromocytoma

C Head and neck paraganglioma

FIGURE 380-1 The paraganglial system and topographic sites (in red) of pheochromocytomas and paragangliomas. *(Parts A and B from WM Manger, RW Gifford: Clinical and experimental pheochromocytoma. Cambridge, Blackwell Science, 1996; Part C from GG Glenner, PM Grimley: Tumors of the Extra-adrenal Paraganglion System [Including Chemoreceptors], Atlas of Tumor Pathology, 2nd Series, Fascicle 9. Washington, DC, AFIP, 1974.)*

TABLE 380-2 Biochemical and Imaging Methods Used for Diagnosis of Pheochromocytoma and Paraganglioma

DIAGNOSTIC METHOD	SENSITIVITY	SPECIFICITY
24-h urinary tests		
Catecholamines	+++	+++
Fractionated metanephrines	++++	++
Total metanephrines	+++	++++
Plasma tests		
Catecholamines	+++	++
Free metanephrines	++++	+++
Imaging		
CT	++++	+++
MRI	++++	+++
MIBG scintigraphy	+++	++++
Somatostatin receptor scintigraphy[a]	++	++
^{18}Fluoro-DOPA PET/CT	+++	++++

[a]Values are particularly high in head and neck paragangliomas.

Abbreviations: MIBG, metaiodobenzylguanidine; PET/CT, positron emission tomography plus CT. For the biochemical tests, the ratings correspond globally to sensitivity and specificity rates as follows: ++, <85%; +++, 85–95%; and ++++, >95%.

of catecholamines or metanephrines (their methylated metabolites) is traditionally the first step in diagnosis.

Biochemical Testing Pheochromocytomas and paragangliomas synthesize and store catecholamines, which include norepinephrine (noradrenaline), epinephrine (adrenaline), and dopamine. Elevated plasma and urinary levels of catecholamines and metanephrines form the cornerstone of diagnosis. The characteristic fluctuations in the hormonal activity of tumors results in considerable variation in serial catecholamine measurements. However, most tumors continuously leak O-methylated metabolites, which are detected by measurement of metanephrines.

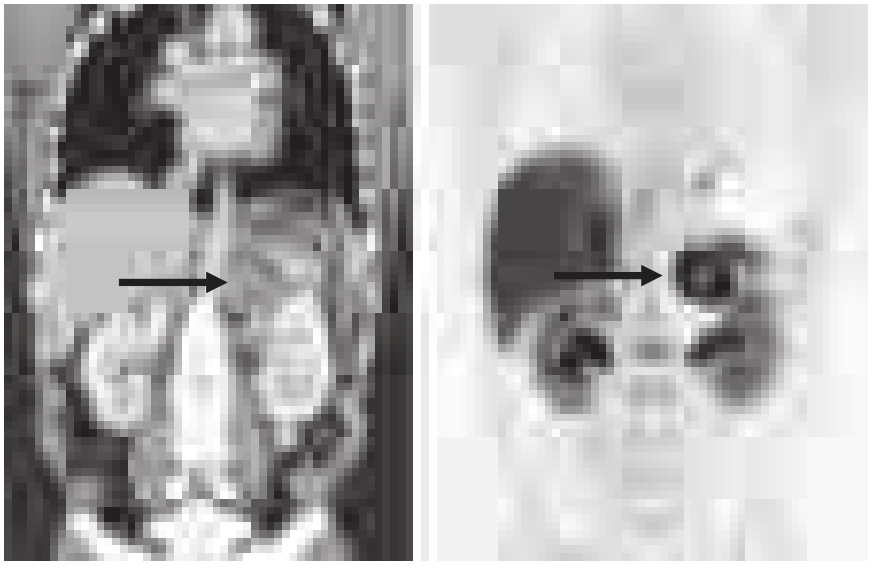
Catecholamines and metanephrines can be measured by different methods, including high-performance liquid chromatography, enzyme-linked immunosorbent assay, and liquid chromatography/mass spectrometry. When pheochromocytoma is suspected on clinical grounds (i.e., when values are three times the upper limit of normal), this diagnosis is highly likely regardless of the assay used. However, as summarized in Table 380-2, the sensitivity and specificity of available biochemical tests vary greatly, and these differences are important in assessing patients with borderline elevations of different compounds. Urinary tests for metanephrines (total or fractionated) and catecholamines are widely available and are used commonly for initial evaluation. Among these tests, those for the fractionated metanephrines and catecholamines are the most sensitive. Plasma tests are more convenient and include measurements of catecholamines and metanephrines. Measurements of plasma metanephrine are the most sensitive and are less susceptible to false-positive elevations from stress, including venipuncture. Although the incidence of false-positive test results has been reduced by the introduction of newer assays, physiologic stress responses and medications that increase catecholamine levels still can confound testing. Because the tumors are relatively rare, borderline elevations are likely to represent false-positive results. In this circumstance, it is important to exclude dietary or drug-related factors (withdrawal of levodopa or use of sympathomimetics, diuretics, tricyclic antidepressants, alpha and beta blockers) that might cause false-positive results and then to repeat testing or perform a clonidine suppression test (i.e., the measurement of plasma normetanephrine 3 h after oral administration of 300 μg of clonidine). Other pharmacologic tests, such as the phentolamine test and the glucagon provocation test, are of relatively low sensitivity and are not recommended.

Diagnostic Imaging A variety of methods have been used to localize pheochromocytomas and paragangliomas (Table 380-2, Figs. 380-2, 380-3, and 380-4). CT and MRI are similar in sensitivity and should be performed with contrast. T2-weighted MRI with gadolinium contrast is optimal for detecting pheochromocytomas and is somewhat better than CT for imaging extraadrenal pheochromocytomas and paragangliomas. About 5% of adrenal incidentalomas, which usually are detected by CT or MRI, prove to be pheochromocytomas upon endocrinologic evaluation.

Tumors also can be localized by procedures using radioactive tracers, including ^{131}I- or ^{123}I-metaiodobenzylguanidine (MIBG) scintigraphy, ^{111}In-somatostatin analogue scintigraphy, ^{18}F-DOPA positron emission tomography (PET), ^{68}Ga-DOTATATE PET, or ^{18}F-fluorodeoxyglucose (FDG) PET (Fig. 380-2*B* and 380-4*A* and *B*). These agents are particularly useful in the documentation of hereditary syndromes but also in malignant pheochromocytoma, because uptake is exhibited also in paragangliomas and metastases.

Pathology Pheochromocytomas and paragangliomas are found at the classical sites of the adrenal medulla (Fig. 380-2) and paraganglia (Fig. 380-3). Histologically the tumors often show a characteristic "Zellballen" pattern, consisting of nests of neuroendocrine chief cells with peripheral glial-like sustentacular cells. However, a broad spectrum of architectural and cytological features can be seen. Immunohistochemistry is positive for chromogranin and synaptophysin in the chief cells and S-100 in the sustentacular cells (Fig. 380-5*A-D*). Increasingly, staining with antibodies against the proteins encoded by susceptibility genes for hereditary pheochromocytomas, such as SDHB, is used to histologically demonstrate defects of these proteins, thereby making germline mutations more likely (Fig. 380-5*E* and *F*).

Differential Diagnosis When the possibility of a pheochromocytoma is being entertained, other disorders to consider include essential hypertension, anxiety attacks, use of cocaine or amphetamines, mastocytosis or carcinoid syndrome (usually without hypertension), intracranial lesions, clonidine withdrawal, autonomic epilepsy, and factitious crises (usually from use of sympathomimetic amines). When an asymptomatic adrenal mass is identified, likely diagnoses other than pheochromocytoma include a nonfunctioning adrenal adenoma, an aldosteronoma, and a cortisol-producing adenoma (Cushing's syndrome).



A ***B***

FIGURE 380-2 Typical pheochromocytoma (adrenal unilateral). ***A.*** MRI ***B.*** ^{18}F-DOPA positron emission tomography (PET), tumor marked by arrows. *(Part A was provided courtesy of Dr. Tobias Krauss, Freiburg. Part B was provided courtesy of Dr. Juri Ruf, Freiburg.)*

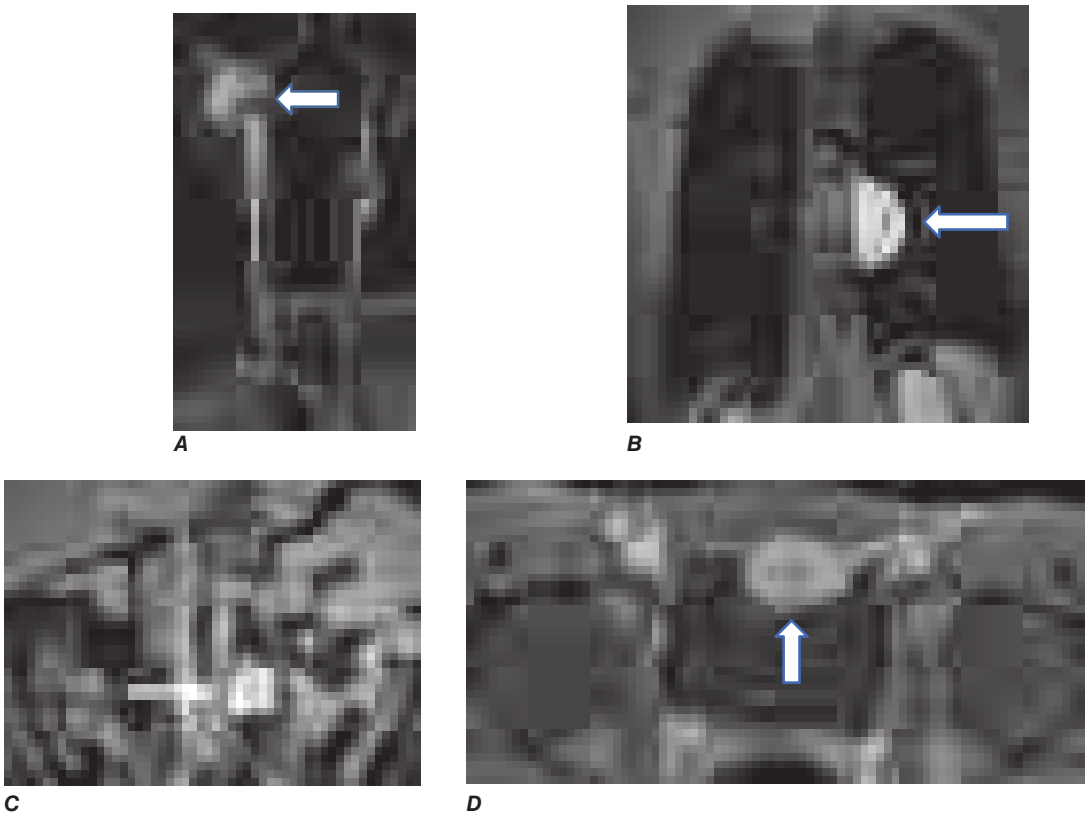


FIGURE 380-3 Paragangliomas (Extraadrenal pheochromocytomas). *A*. Carotid body tumor. ***B.*** Thoracic tumor. ***C***. Interaorto-caval tumor. ***D.*** Pelvic tumor at the anterior wall of the urinary bladder. Tumors marked by arrows. *(Part A was provided courtesy of Dr. Carsten Boedeker, Stralsund. Parts B and D were provided courtesy of Dr Tobias Krauss, Freiburg. Part C was provided courtesy of Dr Martin Walz, Essen.)*

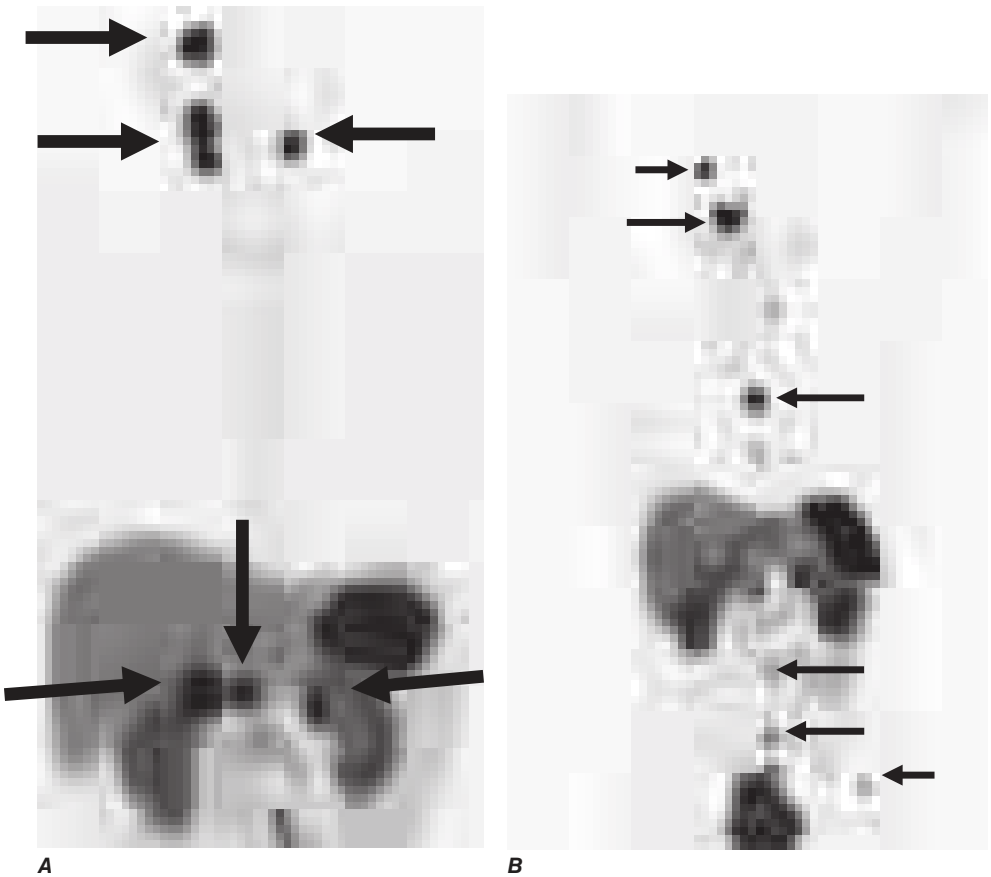


FIGURE 380-4 Multiple and metastatic pheochromocytoma. *A*. Paraganglioma syndrome. A patient with the SDHD W5X mutation and PGL1 ^{68}Ga-DOTATATE positron emission tomography (PET) demonstrating tumor uptake in the right jugular glomus, the right and left carotid body, both adrenal glands and an interaorto-caval paraganglion (arrows). Note the physiologic accumulation of the radiopharmaceutical agent in the kidneys and the liver. ***B.*** ^{18}F-DOPA PET of a patient with metastatic pheochromocytoma. Several metastases marked by arrows. *(Parts A and B were provided courtesy of Dr. Juri Ruf, Freiburg.)*

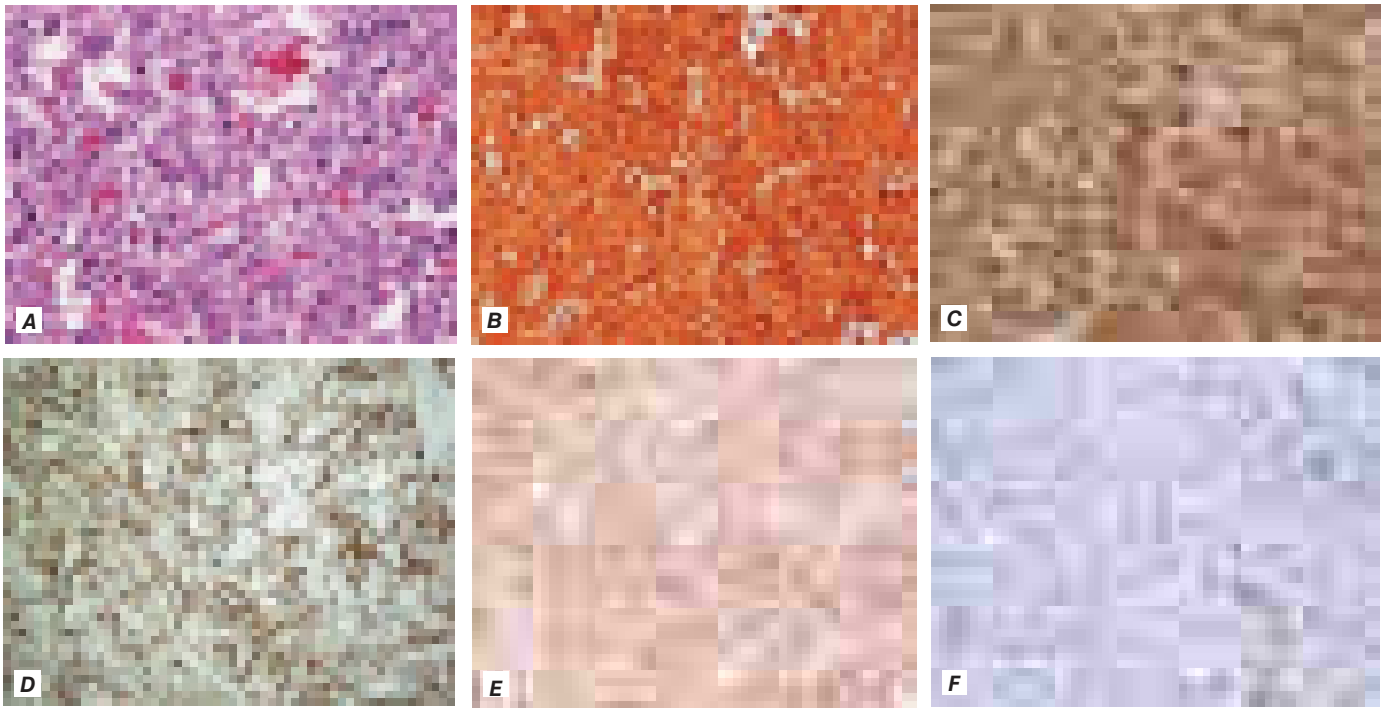


FIGURE 380-5 Histology and immunohistochemistry of pheochromocytoma. ***A.*** hematoxylin & eosin, ***B.*** chromogranin, ***C.*** synaptophysin, ***C*** and ***B*** stain chief cells ***D***: S-100 stains sustentacular cells. ***E***, ***F.*** Immunohistochemistry with SDHB antibody: positive staining (granular cytoplasmic staining) indicates intact SDHB (***E***), whereas negative staining (endothelial cells positive as internal control) (***F***) indicates structurally changed or absent SDHB due to a germline mutation in the *SDHB* gene, which was confirmed by molecular genetic analysis of a blood sample. *(Parts A–D and F were provided courtesy of Dr Helena Leijon, Helsinki. Part E was provided courtesy of Dr. Kurt Werner Schmid, Essen.)*

TREATMENT

Pheochromocytoma

Complete tumor removal, the ultimate therapeutic goal, can be achieved by partial or total adrenalectomy. It is important to preserve the normal adrenal cortex in order to prevent Addison's disease, particularly in hereditary disorders in which bilateral pheochromocytomas are most likely. Preoperative preparation of the patient has to be considered, and blood pressure should be consistently <160/90 mmHg. Classically, blood pressure has been controlled by α-adrenergic blockers (oral phenoxybenzamine, 0.5–4 mg/kg of body weight). Because patients are volume-constricted, liberal salt intake and hydration are necessary to avoid severe orthostasis. Oral prazosin or intravenous phentolamine can be used to manage paroxysms while adequate alpha blockade is awaited. Beta blockers (e.g., 10 mg of propranolol three or four times per day) can then be added. Other antihypertensives, such as calcium channel blockers or angiotensin-converting enzyme inhibitors, have also been used effectively.

Surgery should be performed by teams of surgeons and anesthesiologists with experience in the management of pheochromocytomas. Blood pressure can be labile during surgery, particularly at the outset of intubation or when the tumor is manipulated. Nitroprusside infusion is useful for intraoperative hypertensive crises, and hypotension usually responds to volume infusion. The latter side effect can, however, be avoided in normotensive pheochromocytoma patients by having on stand-by intraoperative nitroprusside, which has been shown to be safe and avoids postoperative hypotension often caused by alpha blockers; the long-lasting guideline for obligatory preoperative treatment with alpha blockers is being reconsidered.

Minimally invasive techniques (laparoscopy or retroperitoneoscopy) have become the standard approaches in pheochromocytoma surgery. They are associated with fewer complications, a faster recovery, and optimal cosmetic results. Extraadrenal abdominal and most thoracic pheochromocytomas can also be removed endoscopically. Postoperatively, catecholamine normalization should be documented. An adrenocorticotropic hormone (ACTH) test should be used to exclude cortisol deficiency when bilateral adrenal cortex–sparing surgery has been performed.

Head and neck paragangliomas are a challenge for surgeons, since damage of adjacent tissue, mainly vessels or cranial nerves is a frequent permanent side effect. Careful consideration of best management is important, and radiotherapy may be an alternative, especially for large head and neck paragangliomas.

MALIGNANT PHEOCHROMOCYTOMA

About 5–10% of pheochromocytomas and paragangliomas are malignant. The diagnosis of malignant pheochromocytoma is problematic. The typical histologic criteria of cellular atypia, presence of mitoses, and invasion of vessels or adjacent tissues are insufficient for the diagnosis of malignancy in pheochromocytoma. Thus, the term *malignant pheochromocytoma* is restricted to tumors with lymph node or distant metastases, the latter most commonly found by nuclear medicine imaging in lungs, bone, or liver locations suggesting a vascular pathway of spread **(Fig. 380-4*B*)**. Because hereditary syndromes are associated with multifocal tumor sites, these features should be anticipated in patients with germline mutations, especially of *RET, VHL, SDHD*, or *SDHB*. However, distant metastases also occur in these syndromes, especially in carriers of *SDHB* mutations.

Treatment of malignant pheochromocytoma or paraganglioma is challenging. Options include tumor mass reduction, alpha blockers for symptoms, chemotherapy, nuclear medicine radiotherapy and stereotactic radiation. The first-line choice is nuclear medicine therapy for scintigraphically documented metastases, preferably with ^{131}I-MIBG in 100–300 mCi doses over 3–6 cycles. Other options for radionuclide treatment are somatostatin receptor ligands, e.g., DOTATOC labeled with ^{90}Yttrium or ^{177}Lutetium, both for palliative outcomes. Averbuch's chemotherapy protocol includes dacarbazine (600 mg/m^2 on days 1 and 2), cyclophosphamide (750 mg/m^2 on day 1), and vincristine (1.4 mg/m^2 on day 1), all repeated every 21 days for 3–6 cycles. Palliation (stable disease to shrinkage) is achieved in about one-half of patients. Due to increasing insights in the genetics of pheochromocytoma, and their molecular pathways, new targeted chemotherapeutic options such as sunitinib and temozolomide/thalidomide are under

development. The prognosis of metastatic pheochromocytoma or paraganglioma is variable, with 5-year survival rates of 30–60%.

PHEOCHROMOCYTOMA IN PREGNANCY

Pheochromocytomas occasionally are diagnosed in pregnancy. Endoscopic removal, preferably in the fourth to sixth month of gestation, is possible and can be followed by uneventful childbirth. Regular screening in families with inherited pheochromocytomas provides an opportunity to identify and remove asymptomatic tumors in women of reproductive age.

PHEOCHROMOCYTOMA-ASSOCIATED SYNDROMES

About 25–33% of patients with a pheochromocytoma or paraganglioma have an inherited syndrome. At diagnosis, patients with inherited syndromes are a mean of ~15 years younger than patients with sporadic tumors.

The best-known pheochromocytoma-associated syndrome is the autosomal dominant disorder MEN 2 (see Chap. 381). Both types of MEN 2 (2A and 2B) are caused by mutations in *RET*, which encodes a tyrosine kinase. The locations of *RET* mutations correlate with the severity of disease and the type of MEN 2 (see Chap. 381). MEN 2A is characterized by medullary thyroid carcinoma (MTC), pheochromocytoma, and hyperparathyroidism; MEN 2B also includes MTC and pheochromocytoma as well as multiple mucosal neuromas, marfanoid habitus, and other developmental disorders, though it typically lacks hyperparathyroidism. MTC is found in virtually all patients with MEN 2, but pheochromocytoma occurs in only ~50% of these patients. Nearly all pheochromocytomas in MEN 2 are benign and located in the adrenals, often bilaterally. Pheochromocytoma may be symptomatic before MTC. Prophylactic thyroidectomy is being performed in many carriers of *RET* mutations; pheochromocytomas should be excluded before any surgery in these patients.

The *paraganglioma syndromes* (PGLs) have been classified by genetic analyses of families with head and neck paragangliomas. The susceptibility genes encode subunits of the enzyme SDH, a component in the Krebs cycle and the mitochondrial electron transport chain. SDH is formed by four subunits (A–D). Mutations of *SDHA* (PGL5), *SDHB* (PGL4), *SDHC* (PGL3), *SDHD* (PGL1), and *SDHAF2* (PGL2) predispose to the PGLs. The transmission of the disease in carriers of *SDHA*, *SDHB*, and *SDHC* germline mutations is autosomal dominant. In contrast, in virtually all *SDHD* and *SDHAF2* families, only the progeny of affected fathers develop tumors if they inherit the mutation. PGL1 is most common, followed by PGL4; PGL2, PGL3, and PGL5 are rare. Adrenal, extraadrenal abdominal, and thoracic pheochromocytomas, which are components of PGL1, PGL4, and PGL5, are rare in PGL3 and absent in PGL2 (Fig. 380-4*A*). About one-third of patients with PGL4 develop metastases.

VHL is an autosomal dominant disorder that predisposes to retinal and cerebellar hemangioblastomas, which also occur in the brainstem and spinal cord (Fig. 380-6). Other important features of VHL are clear

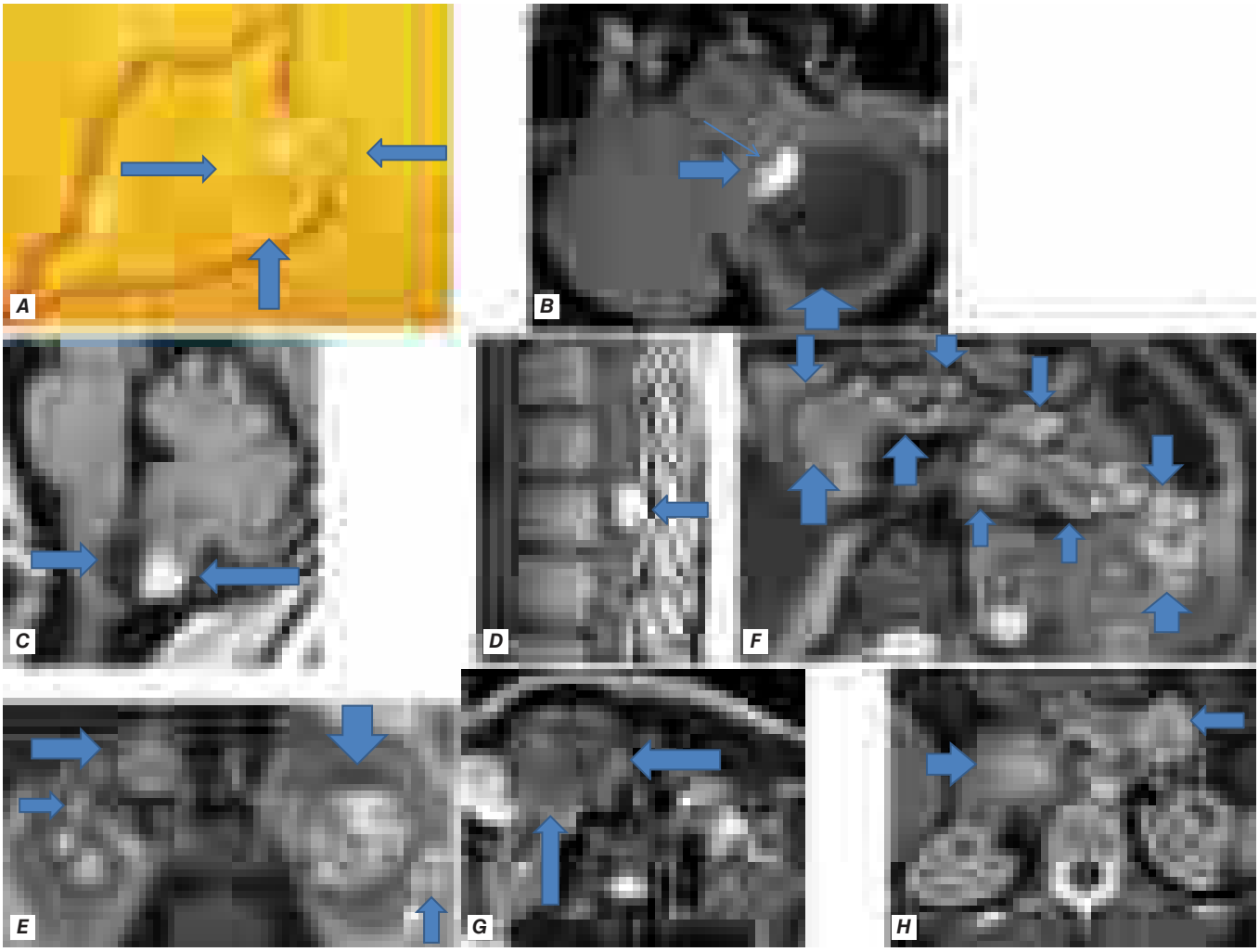


FIGURE 380-6 Von Hippel–Lindau disease. Tumors and cysts marked by arrows. ***A.*** Retinal angioma (*arrows* with a pair of feeding vessels). All subsequent panels show findings on MRI: ***B–D.*** Hemangioblastomas of the cerebellum (large cyst and a solid mural tumor) (***B***) in brainstem (in part cystic) (***C***) and spinal cord (thoracic) (***D***). ***E.*** Bilateral renal clear cell carcinomas with two tumors on each side ***F.*** Multiple pancreatic cysts. ***G.*** Microcystic serous pancreatic cystadenoma (with multiple tiny spaces). ***H.*** Two pancreatic islet cell tumors. *(Part A was provided courtesy of Dr. Dieter Schmidt. Part B was provided courtesy of Dr. Christian Taschner, Freiburg. Part C was provided courtesy of Dr. Sven Glaesker, Brussels. Part D was provided courtesy of Dr. Jan-Helge Klingler, Freiburg. Part E was provided courtesy of Dr. Cordula Jilg, Freiburg. Parts F–H were provided courtesy of Dr Tobias Krauss, Freiburg.)*

cell renal carcinomas, pancreatic neuroendocrine tumors, endolymphatic sac tumors of the inner ear, cystadenomas of the epididymis and broad ligament, and multiple pancreatic or renal cysts. Although the *VHL* gene can be inactivated by all types of mutations, patients with pheochromocytoma predominantly have missense mutations. About 20–30% of patients with VHL have pheochromocytomas, but in some families the incidence can reach 90%. The recognition of pheochromocytoma as a VHL-associated feature provides an opportunity to diagnose retinal, central nervous system, renal, and pancreatic tumors at a stage when effective treatment may still be possible.

NF1 was the first described pheochromocytoma-associated syndrome. The *NF1* gene functions as a tumor suppressor by regulating the Ras signaling cascade. Classic features of neurofibromatosis include multiple neurofibromas, café au lait spots, axillary freckling of the skin, and Lisch nodules of the iris. Pheochromocytomas occur in only ~1% of these patients and are located predominantly in the adrenals. Malignant pheochromocytoma is not uncommon.

Other familial pheochromocytoma has been attributed to hereditary, mainly adrenal tumors in patients with germline mutations in the genes *TMEM127* and *MAX*. Transmission is also autosomal dominant, and mutations of *MAX*, like those of *SDHD*, cause tumors only if inherited from the father.

GUIDELINES FOR GENETIC SCREENING OF PATIENTS WITH PHEOCHROMOCYTOMA OR PARAGANGLIOMA

Effective preventive medicine for pheochromocytoma and pheochromocytoma-associated diseases requires management according to identified germline mutations in susceptibility genes. In addition to family history, general features suggesting an inherited syndrome include young age, multifocal tumors, extraadrenal tumors, and malignant tumors (Table 380-3 and Fig. 380-7). Because of the relatively high prevalence of familial syndromes among patients who present with pheochromocytoma or paraganglioma, it is useful to identify germline

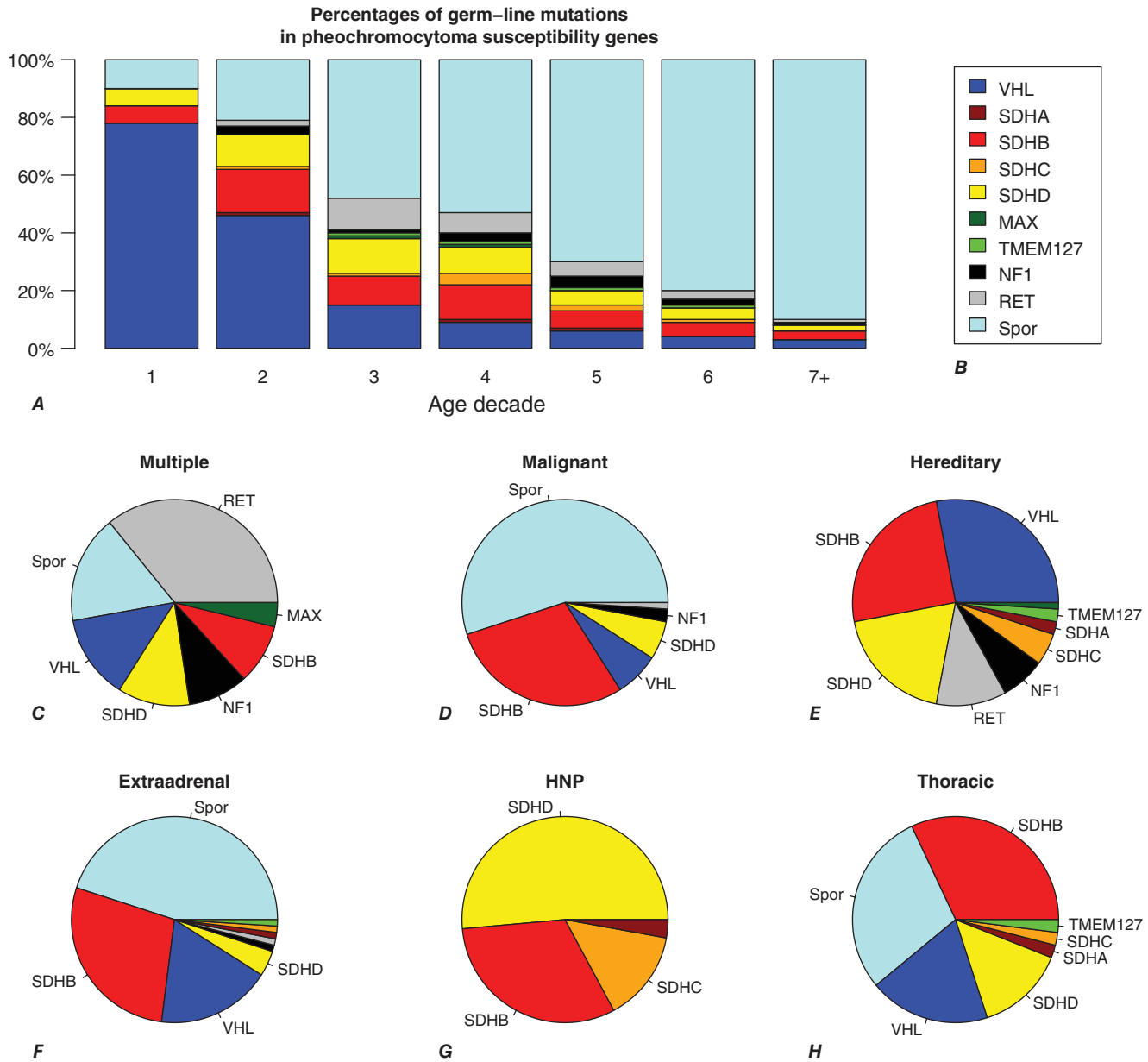


FIGURE 380-7 Mutation distribution in the *VHL*, *RET*, *SDHB*, *SDHC*, *SDHD*, and *NF1* genes in 2796 patients with pheochromocytomas and paragangliomas from the European-American Pheochromocytoma-Paraganglioma Registry based in Freiburg, Germany, and updated as of January 1, 2017. ***A.*** Correlation with age. The bars depict the frequency of sporadic (spor) or various inherited forms of pheochromocytoma in different age groups. The inherited disorders are much more common among younger individuals presenting with pheochromocytoma. ***B.*** Percentages of mutated genes in hereditary pheochromocytomas and paragangiomas. ***C–G.*** Germline mutations according to multiple (***C***), malignant (***D***), hereditary (***E***), extraadrenal retroperitoneal (***F***), head and neck paragangliomas (***G***), thoracic (***H***). *(Data from the Freiburg International Pheochromocytoma and Paraganglioma Registry, 2017. Figures courtesy of Dr. Charis Eng, Cleveland; Dr. Ulrich Wellner, Luebeck; Dr. Birke Bausch, Freiburg; Dr. Giuseppe Opcher, Padova; and Frederic Castinetti, Marseille.)*

TABLE 380-3 Patterns of Occurrence in Inherited Pheochromocytoma and Paraganglioma–Associated Syndromes

MUTATED GENE	ADRENAL TUMORS	HEAD AND NECK TUMORS	EXTRAADRENAL RETROPERITONEAL OR PELVIC TUMORS	THORACIC TUMORS	MULTIPLE TUMORS	BILATERAL ADRENAL TUMORS	METASTATIC TUMORS	FAMILY HISTORY IN PROBANDS FOR COMPONENTS OF THE GIVEN SYNDROME
MAX	100	<1	9	<1	82	73	9	25
NF1	98	<1	4	<1	26	26	2	15
RET	100	<1	<1	<1	61	61	<1	37
SDHA	31	44	27	3	9	4	12	3
SDHB	26	44	38	6	14	2	27	13
SDHC	3	94	<1	<1	11	<1	<1	11
SDHD	20	87	17	10	71	7	8	46
VHL	94	<1	20	3	50	57	4	37
TMEM127	95	8	1	<1	41	37	10	10

Note: Frequencies in percent of clinical characteristics of pheochromocytomas/paragangliomas of patients with germline mutations of the genes MAX, NF1, RET, SDHA, SDHB, SDHC, SDHD, VHL, and TMEM127.

mutations even in patients without a known family history. A first step is to search for clinical features of inherited syndromes and to obtain an in-depth, multigenerational family history. Each of these syndromes exhibits autosomal dominant transmission with variable penetrance, but a proband with a mother affected by paraganglial tumors is not predisposed to PLG1 and PGL2 (*SDHD* and *SDHAF2* mutation carrier). Cutaneous neurofibromas, café au lait spots, and axillary freckling suggest neurofibromatosis. Germline mutations in *NF1* have not been reported in patients with sporadic pheochromocytomas. Thus, *NF1* testing does not need to be performed in the absence of other clinical features of neurofibromatosis. A personal or family history of MTC or an elevation of serum calcitonin strongly suggests MEN 2 and should prompt testing for *RET* mutations. A history of visual impairment or tumors of the cerebellum, kidney, brainstem, or spinal cord suggests the possibility of VHL. A personal and/or family history of head and neck paraganglioma suggests PGL1 or PGL4.

A single adrenal pheochromocytoma in a patient with an otherwise unremarkable history may still be associated with mutations of *VHL, RET, SDHB*, or *SDHD* (in decreasing order of frequency). Two-thirds of extraadrenal tumors are associated with one of these syndromes, and multifocal tumors occur with decreasing frequency in carriers of *RET, SDHD, VHL, SDHB*, and *MAX* mutations. About 30% of head and neck paragangliomas are associated with germline mutations of one of the SDH subunit genes (most often *SDHD*) and are rare in carriers of *VHL, RET, MAX*, and *TMEM127* mutations. Immunohistochemistry is helpful in the preselection of hereditary pheochromocytoma. Negative immunostaining with antibodies to SDHB (Fig. 380-5*F*), TMEM127, and MAX may predict mutations of the *SDHx* (PGL1-5), *TMEM127*, and *MAX* genes, respectively.

New technologies with whole genome sequence analysis are expected to replace targeted Sanger sequencing. It should soon be possible to search for germline mutations in a set of genes, such that all susceptibility genes for pheochromocytoma associated syndromes could be analyzed in one procedure. Of note, many sequencing protocols do not detect large deletions of one or more exons.

Once the underlying syndrome is diagnosed, the benefit of genetic testing can be extended to relatives. For this purpose, it is necessary to identify the germline mutation in the proband and, after genetic counseling, to perform DNA sequence analyses of the responsible gene in relatives to determine whether they are affected. Other family members may benefit when individuals who carry a germline mutation are biochemically screened for paraganglial tumors.

Asymptomatic paraganglial tumors, now often detected in patients with hereditary tumors and their relatives, are challenging to manage. Watchful waiting strategies have been introduced. Head and neck paragangliomas—mainly carotid body, jugular, and vagal tumors—are increasingly treated by radiation, since surgery is frequently associated with permanent palsy of cranial nerves II, VII, IX, X, XI, and XII. Nevertheless, tympanic paragangliomas are symptomatic early, and most of these tumors can easily be resected, with subsequent improvement of hearing and alleviation of tinnitus.

FURTHER READING

Bausch B et al: Characterization of endolymphatic sac tumors and von Hippel-Lindau disease in the International Endolymphatic Sac Tumor Registry. Head Neck 38: Suppl 1:E673, 2016.

Bausch B et al: Systematic clinical characterization of the pheochromocytoma and paraganglioma susceptibility genes *SDHA, TMEM127, MAX* and *SDHAF2*—towards gene-informed prevention. JAMA Oncol 2017 Apr 6. doi: 10.1001/jamaoncol.2017.0223. [Epub ahead of print].

Castinetti F et al: Outcomes of adrenal-sparing surgery or total adrenalectomy in phaeochromocytoma associated with multiple endocrine neoplasia type 2: An international retrospective population-based study. Lancet Oncol 15:648, 2014.

Groeben H et al: A closed case series of phaeochromotyoma and paraganglioma surgery—Time to rethink the concept of a perioperative α–receptor blockade? Br J Anaesth 118:182, 2017.

Groeben H et al: Perioperative α-receptor blockade in phaeochromocytoma surgery: An observational case series. Br J Anaesth 118:182, 2017.

Suarez C et al: Modern trends in the management of head and neck paragangliomas. Eur Arch Otorhinolaryngol 272:3595, 2015.

Von Dobschuetz E et al: A registry-based study of thyroid paraganglioma: Histological and genetic characteristics. Endocrine-Related Cancer 22:191, 2015.

381 Multiple Endocrine Neoplasia

R.V. Thakker

Multiple endocrine neoplasia (MEN) is characterized by a predilection for tumors involving two or more endocrine glands. Four major forms of MEN are recognized and referred to as MEN types 1–4 (MEN 1–4) **(Table 381-1)**. Each type of MEN is inherited as an autosomal dominant syndrome or may occur sporadically, that is, without a family history. However, this distinction between familial and sporadic forms is often difficult because family members with the disease may have died before symptoms developed. In addition to MEN 1–4, at least six other syndromes are associated with multiple endocrine and other organ neoplasias (MEONs) **(Table 381-2)**. These MEONs include the hyperparathyroidism-jaw tumor (HPT-JT) syndrome, Carney complex, von Hippel-Lindau disease **(Chap. 380)**, neurofibromatosis type 1 **(Chap. 86)**, Cowden's syndrome (CWS), and McCune-Albright syndrome (MAS) **(Chap. 405)**; all of these are inherited as autosomal dominant disorders, except for MAS, which is caused by mosaic expression of a postzygotic somatic cell mutation (Table 381-2).

TABLE 381-1 Multiple Endocrine Neoplasia (MEN) Syndromes

TYPE (CHROMOSOMAL LOCATION)	TUMORS (ESTIMATED PENETRANCE)	GENE AND MOST FREQUENTLY MUTATED CODONS
MEN 1 (11q13)	Parathyroid adenoma (90%) Enteropancreatic tumor (30–70%) • Gastrinoma (>50%) • Insulinoma (10–30%) • Nonfunctioning and PPoma (20–55%) • Glucagonoma (<3%) • VIPoma (<1%) Pituitary adenoma (15–50%) • Prolactinoma (60%) • Somatotrophinoma (25%) • Corticotropinoma (<5%) • Nonfunctioning (<5%) Associated tumors • Adrenal cortical tumor (20–70%) • Pheochromocytoma (<1%) • Bronchopulmonary NET (2%) • Thymic NET (2%) • Gastric NET (10%) • Lipomas (>33%) • Angiofibromas (85%) • Collagenomas (70%) • Meningiomas (8%)	*MEN1* 83/84, 4-bp del (≈4%) 119, 3-bp del (≈3%) 209-211, 4-bp del (≈8%) 418, 3-bp del (≈4%) 514-516, del or ins (≈7%) Intron 4 ss (≈10%)
MEN 2 (10 cen-10q11.2)		
MEN 2A	MTC (90%) Pheochromocytoma (>50%) Parathyroid adenoma (10–25%)	*RET* 634, e.g., Cys → Arg (~85%)
MTC only	MTC (100%)	*RET* 618, missense (>50%)
MEN 2B (also known as MEN 3)	MTC (>90%) Pheochromocytoma (>50%) Associated abnormalities (40–50%) • Mucosal neuromas • Marfanoid habitus • Medullated corneal nerve fibers • Megacolon	*RET* 918, Met → Thr (>95%)
MEN 4 (12p13)	Parathyroid adenoma[a] Pituitary adenoma[a] Reproductive organ tumors[a] (e.g., testicular cancer, neuroendocrine cervical carcinoma) ?Adrenal + renal tumors[a]	*CDKN1B*; no common mutations identified to date

[a]Insufficient numbers reported to provide prevalence information.

Note: Autosomal dominant inheritance of the MEN syndromes has been established.

Abbreviations: del, deletion; ins, insertion; MTC, medullary thyroid cancer; NET, neuroendocrine tumor; PPoma, pancreatic polypeptide–secreting tumor; VIPoma, vasoactive intestinal polypeptide–secreting tumor.

Source: Reproduced from RV Thakker et al: J Clin Endocrinol Metab 97:2990, 2012.

TABLE 381-2 Multiple Endocrine and Other Organ Neoplasia Syndromes (MEONs)

DISEASE[a]	GENE PRODUCT	CHROMOSOMAL LOCATION
Hyperparathyroidism-jaw tumor (HPT-JT)	Parafibromin	1q31.2
Carney complex		
CNC1	PPKAR1A	17q24.2
CNC2	?[b]	2p16
von Hippel-Lindau disease (VHL)	pVHL (elongin)	3p25
Neurofibromatosis type 1 (NF1)	Neurofibromin	17q11.2
Cowden's syndrome (CWS)		
CWS1	PTEN	10q23.31
CWS2	SDHB	1p36.13
CWS3	SDHD	11q23.1
CWS4	KLLN	10q23.31
CWS5	PIK3CA	3q26.32
CWS6	AKT1	14q32.33
CWS7	SEC23B	20p11.23
McCune-Albright syndrome (MAS)	Gsα	20q13.32

[a]The inheritance for these disorders is autosomal dominant, except MAS, which is due to mosaicism that results from the postzygotic somatic cell mutation of the *GNAS1* gene, encoding Gsα. [b]?, unknown.

A diagnosis of a MEN or MEON syndrome may be established in an individual by one of three criteria: (1) clinical features (two or more of the associated tumors [or lesions] in an individual); (2) familial pattern (one of the associated tumors [or lesions] in a first-degree relative of a patient with a clinical diagnosis of the syndrome); and (3) genetic analysis (a germline mutation in the associated gene in an individual, who may be clinically affected or asymptomatic). Mutational analysis in MEN and MEON syndromes is helpful in clinical practice to: (1) confirm the clinical diagnosis; (2) identify family members who harbor the mutation and require screening for relevant tumor detection and early/appropriate treatment; and (3) identify the ~50% of family members who do not harbor the germline mutation and can, therefore, be alleviated of the anxiety of developing associated tumors. This latter aspect also helps to reduce health care costs by reducing the need for unnecessary biochemical and radiologic investigations.

MULTIPLE ENDOCRINE NEOPLASIA TYPE 1

Clinical Manifestations MEN type 1 (MEN 1), which is also referred to as Wermer's syndrome, is characterized by the triad of tumors involving the parathyroids, pancreatic islets, and anterior pituitary. In addition, adrenal cortical tumors, carcinoid tumors usually of the foregut, meningiomas, facial angiofibromas, collagenomas, and lipomas may also occur in some patients with MEN 1. Combinations of the affected glands and their pathologic features (e.g., hyperplastic adenomas of the parathyroid glands) may differ in members of the same family and even between identical twins. In addition, a nonfamilial (e.g., sporadic) form occurs in 8–14% of patients with MEN 1, and molecular genetic studies have confirmed the occurrence of de novo mutations of the *MEN1* gene in ~10% of patients with MEN 1. The prevalence of MEN 1 is ~0.25% based on randomly chosen postmortem studies but is 1–18% among patients with primary hyperparathyroidism, 16–38% among patients with pancreatic islet tumors, and <3% among patients with pituitary tumors. The disorder affects all age groups, with a reported age range of 5–81 years, with clinical and biochemical manifestations developing in the vast majority by the fifth decade. The clinical manifestations of MEN 1 are related to the sites of tumors and their hormonal products. In the absence of treatment, endocrine tumors are associated with an earlier mortality in patients with MEN 1, with a 50% probability of death by the age of 50 years. The cause of death is usually a malignant tumor, often from a pancreatic neuroendocrine tumor (NET) or foregut carcinoid. In addition, the treatment outcomes of patients with MEN 1–associated tumors are not as successful as those in patients with non–MEN 1 tumors. This is because MEN 1–associated tumors, with the exception of pituitary NETs, are usually multiple, making it difficult to achieve a successful

TABLE 381-3 Biochemical and Radiological Screening in Multiple Endocrine Neoplasia Type 1

TUMOR	AGE TO BEGIN (YEARS)	BIOCHEMICAL TEST (PLASMA OR SERUM) ANNUALLY	IMAGING TEST (TIME INTERVAL)
Parathyroid	8	Calcium, PTH	None
Pancreatic NETs			
Gastrinoma	20	Gastrin (± gastric pH)	None
Insulinoma	5	Fasting glucose, insulin	None
Other pancreatic NET	<10	Chromogranin A; pancreatic polypeptide, glucagon, vasoactive intestinal peptide	MRI, CT, or EUS (annually)
Anterior pituitary	5	Prolactin, IGF-I	MRI (every 3 years)
Adrenal	<10	None unless symptoms or signs of functioning tumor and/or tumor >1 cm identified on imaging	MRI or CT (annually with pancreatic imaging)
Thymic and bronchial carcinoid	15	None	CT or MRI (every 1–2 years)

Abbreviations: CT, computed tomography; EUS, endoscopic ultrasound; IGF-I, insulin-like growth factor I; MRI, magnetic resonance imaging; PTH, parathyroid hormone.

Source: Reproduced from RV Thakker et al: J Clin Endocrinol Metab 97:2990, 2012.

surgical cure. Occult metastatic disease is also more prevalent in MEN 1, and the tumors may be larger, more aggressive, and resistant to treatment.

Parathyroid Tumors **(See also Chap. 403)** Primary hyperparathyroidism occurs in ~90% of patients and is the most common feature of MEN 1. Patients may have asymptomatic hypercalcemia or vague symptoms associated with hypercalcemia (e.g., polyuria, polydipsia, constipation, malaise, or dyspepsia). Nephrolithiasis and osteitis fibrosa cystica (less commonly) may also occur. Biochemical investigations reveal hypercalcemia, usually in association with elevated circulating parathyroid hormone (PTH) **(Table 381-3)**. The hypercalcemia is usually mild, and severe hypercalcemia or parathyroid cancer is a rare occurrence. Additional differences in the primary hyperparathyroidism of patients with MEN 1, as opposed to those without MEN 1, include an earlier age at onset (20–25 vs 55 years) and an equal male-to-female ratio (1:1 vs 1:3). Preoperative imaging (e.g., neck ultrasound with ^{99m}Tc-sestamibi parathyroid scintigraphy) is of limited benefit because all parathyroid glands may be affected, and neck exploration may be required irrespective of preoperative localization studies.

TREATMENT

Parathyroid Tumors

Surgical removal of the abnormally overactive parathyroids in patients with MEN 1 is the definitive treatment. However, it is controversial whether to perform subtotal (e.g., removal of 3.5 glands) or total parathyroidectomy with or without autotransplantation of parathyroid tissue in the forearm, and whether surgery should be performed at an early or late stage. Minimally invasive parathyroidectomy is not recommended because all four parathyroid glands are usually affected with multiple adenomas or hyperplasia. Surgical experience should be taken into account given the variability in pathology in MEN 1. Calcimimetics (e.g., cinacalcet), which act via the calcium-sensing receptor, have been used to treat primary hyperparathyroidism in some patients when surgery is unsuccessful or contraindicated.

Pancreatic Tumors **(See also Chap. 80)** The incidence of pancreatic islet cell tumors, which are NETs, in patients with MEN 1 ranges from 30 to 80% in different series. Most of these tumors (Table 381-1) produce excessive amounts of hormone (e.g., gastrin, insulin, glucagon, vasoactive intestinal polypeptide [VIP]) and are associated with distinct clinical syndromes, although some are nonfunctioning or nonsecretory. These pancreatic islet cell tumors have an earlier age at onset in patients with MEN 1 than in patients without MEN 1.

Gastrinoma Gastrin-secreting tumors (gastrinomas) are associated with marked gastric acid production and recurrent peptic ulcerations, a combination referred to as the Zollinger-Ellison syndrome. Gastrinomas occur more often in patients with MEN 1 who are aged >30 years. Recurrent severe multiple peptic ulcers, which may perforate, and cachexia are major contributors to the high mortality. Patients with Zollinger-Ellison syndrome may also suffer from diarrhea and steatorrhea. The diagnosis is established by demonstration of an elevated fasting serum gastrin concentration in association with increased basal gastric acid secretion (Table 381-3). However, the diagnosis of Zollinger-Ellison syndrome may be difficult in hypercalcemic MEN 1 patients, because hypercalcemia can also cause hypergastrinemia. Ultrasonography, endoscopic ultrasonography, computed tomography (CT), nuclear magnetic resonance imaging (MRI), selective abdominal angiography, venous sampling, and somatostatin receptor scintigraphy are helpful in localizing the tumor prior to surgery. Gastrinomas represent >50% of all pancreatic NETs in patients with MEN 1, and ~20% of patients with gastrinomas will be found to have MEN 1. Gastrinomas, which may also occur in the duodenal mucosa, are the major cause of morbidity and mortality in patients with MEN 1. Most MEN 1 gastrinomas are malignant and metastasize before a diagnosis is established.

TREATMENT

Gastrinoma

Medical treatment of patients with MEN 1 and Zollinger-Ellison syndrome is directed toward reducing basal acid output to <10 mmol/L. Parietal cell H^+-K^+-adenosine triphosphatase (ATPase) inhibitors (e.g., omeprazole or lansoprazole) reduce acid output and are the drugs of choice for gastrinomas. Some patients may also require additional treatment with the histamine H_2 receptor antagonists, cimetidine or ranitidine. The role of surgery in the treatment of gastrinomas in patients with MEN 1 is controversial. The goal of surgery is to reduce the risk of distant metastatic disease and improve survival. For a nonmetastatic gastrinoma situated in the pancreas, surgical excision is often effective. However, the risk of hepatic metastases increases with tumor size, such that 25–40% of patients with pancreatic NETs >4 cm develop hepatic metastases, and 50–70% of patients with tumors 2–3 cm in size have lymph node metastases. Survival in MEN 1 patients with gastrinomas <2.5 cm in size is 100% at 15 years, but 52% at 15 years, if metastatic disease is present. The presence of lymph node metastases does not appear to adversely affect survival. Surgery for gastrinomas that are >2–2.5 cm has been recommended, because the disease-related survival in these patients is improved following surgery. In addition, duodenal gastrinomas, which occur more frequently in patients with MEN 1, have been treated successfully with surgery. However, in most patients with MEN 1, gastrinomas are multiple or extrapancreatic, and with the exception of duodenal gastrinomas, surgery is rarely successful. For example, the results of one study revealed that only ~15% of patients with MEN 1 were free of disease immediately after surgery,

and at 5 years, this number had decreased to ~5%; the respective outcomes in patients without MEN 1 were better, at 45% and 40%. Given these findings, most specialists recommend a nonsurgical management for gastrinomas in MEN 1, except as noted earlier for smaller, isolated lesions. Treatment of disseminated gastrinomas is difficult. Chemotherapy with streptozotocin and 5-fluorouracil; hormonal therapy with octreotide or lanreotide, which are human somatostatin analogues; hepatic artery embolization; administration of human leukocyte interferon; and removal of all resectable tumor have been successful in some patients.

Insulinoma These β islet cell insulin-secreting tumors represent 10–30% of all pancreatic tumors in patients with MEN 1. Patients with an insulinoma present with hypoglycemic symptoms (e.g., weakness, headaches, sweating, faintness, seizures, altered behavior, weight gain) that typically develop after fasting or exertion and improve after glucose intake. The most reliable test is a supervised 72-h fast. Biochemical investigations reveal increased plasma insulin concentrations in association with hypoglycemia (Table 381-3). Circulating concentrations of C peptide and proinsulin, which are also increased, are useful in establishing the diagnosis. It also is important to demonstrate the absence of sulfonylureas in plasma and urine samples obtained during the investigation of hypoglycemia (Table 381-3). Surgical success is greatly enhanced by preoperative localization by endoscopic ultrasonography, CT scanning, or celiac axis angiography. Additional localization methods may include preoperative and perioperative percutaneous transhepatic portal venous sampling, selective intraarterial stimulation with hepatic venous sampling, and intraoperative direct pancreatic ultrasonography. Insulinomas occur in association with gastrinomas in 10% of patients with MEN 1, and the two tumors may arise at different times. Insulinomas occur more often in patients with MEN 1 who are aged <40 years, and some arise in individuals aged <20 years. In contrast, in patients without MEN 1, insulinomas generally occur in those aged >40 years. Insulinomas may be the first manifestation of MEN 1 in 10% of patients, and ~4% of patients with insulinomas will have MEN 1.

TREATMENT

Insulinoma

Medical treatment, which consists of frequent carbohydrate meals and diazoxide or octreotide, is not always successful, and surgery is the optimal treatment. Surgical treatment, which ranges from enucleation of a single tumor to a distal pancreatectomy or partial pancreatectomy, has been curative in many patients. Chemotherapy may include streptozotocin, 5-fluorouracil, and doxorubicin. Hepatic artery embolization has been used for metastatic disease.

Glucagonoma These glucagon-secreting pancreatic NETs occur in <3% of patients with MEN 1. The characteristic clinical manifestations of a skin rash (necrolytic migratory erythema), weight loss, anemia, and stomatitis may be absent. The tumor may have been detected in an asymptomatic patient with MEN 1 undergoing pancreatic imaging or by the finding of glucose intolerance and hyperglucagonemia.

TREATMENT

Glucagonoma

Surgical removal of the glucagonoma is the treatment of choice. However, treatment may be difficult because ~50–80% of patients have metastases at the time of diagnosis. Medical treatment with somatostatin analogues (e.g., octreotide or lanreotide) or chemotherapy with streptozotocin and 5-fluorouracil has been successful in some patients, and hepatic artery embolization has been used to treat metastatic disease.

Vasoactive Intestinal Peptide (VIP) Tumors (VIPomas) VIPomas have been reported in only a few patients with MEN 1. This clinical syndrome is characterized by watery diarrhea, hypokalemia, and achlorhydria (WDHA syndrome), which is also referred to as the Verner-Morrison syndrome, or the VIPoma syndrome. The diagnosis is established by excluding laxative and diuretic abuse, confirming a stool volume in excess of 0.5–1.0 L/d during a fast, and documenting a markedly increased plasma VIP concentration.

TREATMENT

Vipomas

Surgical management of VIPomas, which are mostly located in the tail of the pancreas, can be curative. However, in patients with unresectable tumor, somatostatin analogues, such as octreotide and lanreotide, may be effective. Streptozotocin with 5-fluorouracil may be beneficial, along with hepatic artery embolization for the treatment of metastases.

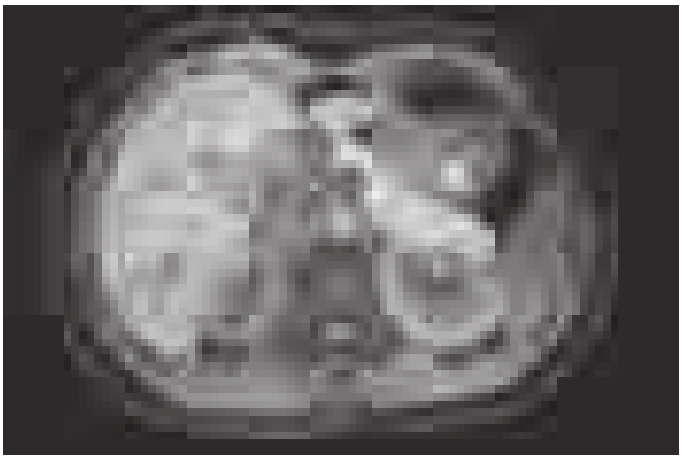
Pancreatic Polypeptide-Secreting Tumors (PPomas) and Nonfunctioning Pancreatic NETs PPomas are found in a large number of patients with MEN 1. No pathologic sequelae of excessive polypeptide (PP) secretion are apparent, and the clinical significance of PP is unknown. Many PPomas may have been unrecognized or classified as nonfunctioning pancreatic NETs, which likely represent the most common enteropancreatic NET associated with MEN 1 **(Fig. 381-1)**. The absence of both a clinical syndrome and specific biochemical abnormalities may result in a delayed diagnosis of nonfunctioning pancreatic NETs, which are associated with a worse prognosis than other functioning tumors, including insulinoma and gastrinoma. The optimum screening method and its timing interval for nonfunctioning pancreatic NETs remain to be established. At present, endoscopic ultrasound likely represents the most sensitive method of detecting small pancreatic tumors, but somatostatin receptor scintigraphy is the most reliable method for detecting metastatic disease (Table 381-3).

TREATMENT

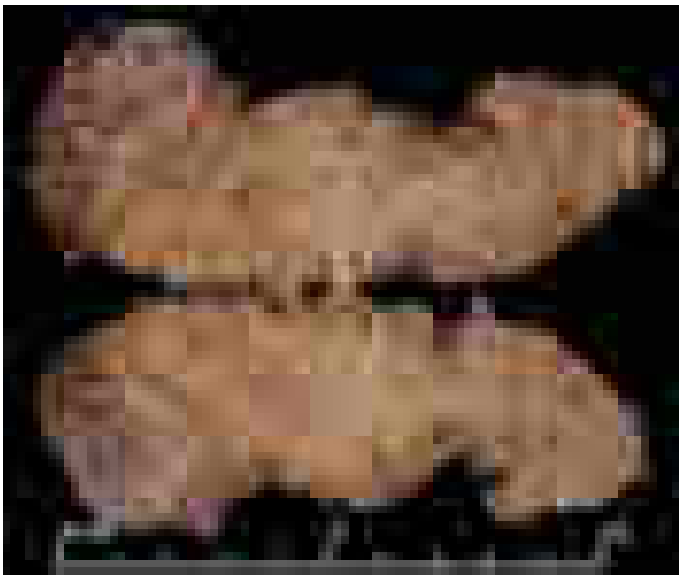
PPomas and Nonfunctioning Pancreatic NETs

The management of nonfunctioning pancreatic NETs in the asymptomatic patient is controversial. One recommendation is to undertake surgery irrespective of tumor size after biochemical assessment is complete. Alternatively, other experts recommend surgery based on tumor size, using either >1 cm or >2 cm at different centers. Pancreatoduodenal surgery is successful in removing the tumors in 80% of patients, but >40% of patients develop complications, including diabetes mellitus, frequent steatorrhea, early and late dumping syndromes, and other gastrointestinal symptoms. However, ~50–60% of patients treated surgically survive >5 years. When considering these recommendations, it is important to consider that occult metastatic disease (e.g., tumors not detected by imaging investigations) is likely to be present in a substantial proportion of these patients at the time of presentation. Inhibitors of tyrosine kinase receptors (TKRs) and of the mammalian target of rapamycin (mTOR) signaling pathway have been reported to be effective in treating pancreatic NETs and in doubling the progression-free survival time.

Other Pancreatic NETs NETs secreting growth hormone–releasing hormone (GHRH), GHRHomas, have been reported rarely in patients with MEN 1. It is estimated that ~33% of patients with GHRHomas have other MEN 1–related tumors. GHRHomas may be diagnosed by demonstrating elevated serum concentrations of growth hormone and GHRH. More than 50% of GHRHomas occur in the lung, 30% occur in the pancreas, and 10% are found in the small intestine. Somatostatinomas secrete somatostatin, a peptide that inhibits the secretion of a variety of hormones, resulting in hyperglycemia,



A



B

FIGURE 381-1 Pancreatic nonfunctioning neuroendocrine tumor (NET) in a 14-year-old patient with multiple endocrine neoplasia type 1 (MEN 1). ***A.*** An abdominal magnetic resonance imaging scan revealed a low-intensity >2.0 cm (anteroposterior maximal diameter) tumor within the neck of pancreas. There was no evidence of invasion of adjacent structures or metastases. The tumor is indicated by *white dashed circle*. ***B.*** The pancreatic NET was removed by surgery, and macroscopic examination confirmed the location of the tumor (*white dashed circles*) in the neck of the pancreas. Immunohistochemistry showed the tumor to immunostain for chromogranin A, but not gastrointestinal peptides or menin, thereby confirming that it was a nonsecreting NET due to loss of menin expression. *(Part A adapted with permission from PJ Newey et al: J Clin Endocrinol Metab 10:3640, 2009.)*

cholelithiasis, low acid output, steatorrhea, diarrhea, abdominal pain, anemia, and weight loss. Although 7% of pancreatic NETs secrete somatostatin, the clinical features of somatostatinoma syndrome are unusual in patients with MEN 1.

Pituitary Tumors (See also Chap. 373) Pituitary tumors occur in 15–50% of patients with MEN 1 (Table 381-1). These occur as early as 5 years of age or as late as the ninth decade. MEN 1 pituitary adenomas are more frequent in women than men and significantly are macroadenomas (i.e., diameter >1 cm). Moreover, about one-third of these pituitary tumors show invasive features such as infiltration of tumor cells into surrounding normal juxtatumoral pituitary tissue. However, no specific histologic parameters differentiate between MEN 1 and non–MEN 1 pituitary tumors. Approximately 60% of MEN 1–associated pituitary tumors secrete prolactin, <25% secrete growth hormone, 5% secrete adrenocorticotropic hormone (ACTH), and the remainder appear to be nonfunctioning, with some secreting glycoprotein subunits (Table 381-1). However, pituitary tumors derived from MEN 1 patients may exhibit immunoreactivity to several hormones. In particular, there is a greater frequency of somatolactotrope tumors. Prolactinomas are the first manifestation of MEN 1 in ~15% of patients, whereas somatotrope tumors occur more often in patients aged >40 years. Fewer than 3% of patients with anterior pituitary tumors will have MEN 1. Clinical manifestations are similar to those in patients with sporadic pituitary tumors without MEN 1 and depend on the hormone secreted and the size of the pituitary tumor. Thus, patients may have symptoms of hyperprolactinemia (e.g., amenorrhea, infertility, and galactorrhea in women, or impotence and infertility in men) or have features of acromegaly or Cushing's disease. In addition, enlarging pituitary tumors may compress adjacent structures such as the optic chiasm or normal pituitary tissue, causing visual disturbances and/or hypopituitarism. In asymptomatic patients with MEN 1, periodic biochemical monitoring of serum prolactin and insulin-like growth factor I (IGF-I) levels, as well as MRI of the pituitary, can lead to early identification of pituitary tumors (Table 381-3). In patients with abnormal results, hypothalamic-pituitary testing should characterize the nature of the pituitary lesion and its effects on the secretion of other pituitary hormones.

TREATMENT

Pituitary Tumors

Treatment of pituitary tumors in patients with MEN 1 consists of therapies similar to those used in patients without MEN 1 and includes appropriate medical therapy (e.g., bromocriptine or cabergoline for prolactinoma; or octreotide or lanreotide for somatotrope tumors) or selective transsphenoidal adenomectomy, if feasible, with radiotherapy reserved for residual unresectable tumor tissue. Pituitary tumors in MEN 1 patients may be more aggressive and less responsive to medical or surgical treatments.

Associated Tumors Patients with MEN 1 may also develop carcinoid tumors, adrenal cortical tumors, facial angiofibromas, collagenomas, thyroid tumors, and lipomatous tumors.

Carcinoid Tumors (See also Chap. 80) Carcinoid tumors occur in >3% of patients with MEN 1 (Table 381-1). The carcinoid tumor may be located in the bronchi, gastrointestinal tract, pancreas, or thymus. At the time of diagnosis, most patients are asymptomatic and do not have clinical features of the carcinoid syndrome. Importantly, no hormonal or biochemical abnormality (e.g., plasma chromogranin A) is consistently observed in individuals with thymic or bronchial carcinoid tumors. Thus, screening for these tumors is dependent on radiologic imaging. The optimum method for screening has not been established. CT and MRI are sensitive for detecting thymic and bronchial tumors (Table 381-3), although repeated CT scanning raises concern about exposure to repeated doses of ionizing radiation. Octreotide scintigraphy may also reveal some thymic and bronchial carcinoids, although there is insufficient evidence to recommend its routine use. Gastric carcinoids, of which the type II gastric enterochromaffin-like (ECL) cell carcinoids (ECLomas) are associated with MEN 1 and Zollinger-Ellison syndrome, may be detected incidentally at the time of gastric endoscopy for dyspeptic symptoms in MEN 1 patients. These tumors, which may be found in >10% of MEN 1 patients, are usually multiple and sized <1.5 cm. Bronchial carcinoids in patients with MEN 1 occur predominantly in women (male-to-female ratio, 1:4). In contrast, thymic carcinoids in European patients with MEN 1 occur predominantly in men (male-to-female ratio, 20:1), with cigarette smokers having a higher risk for these tumors; thymic carcinoids in Japanese patients with MEN 1 have a less marked sex difference (male-to-female ratio 2:1). The course of thymic carcinoids in MEN 1 appears to be particularly aggressive. The presence of thymic tumors in patients with MEN 1 is associated with a median survival after diagnosis of ~9.5 years, with 70% of patients dying as a direct result of the tumor.

TREATMENT

Carcinoid Tumors

If resectable, surgical removal of carcinoid tumors is the treatment of choice. For unresectable tumors and those with metastatic disease, treatment with radiotherapy or chemotherapeutic agents (e.g., cisplatin, etoposide) may be used. In addition, somatostatin analogues, such as octreotide or lanreotide, have resulted in symptom improvement and regression of some tumors. Little is known about the malignant potential of gastric type II ECLomas, but treatment with somatostatin analogues, such as octreotide or lanreotide, has resulted in regression of these ECLomas.

Adrenocortical Tumors **(See also Chap. 379)** Asymptomatic adrenocortical tumors occur in 20–70% of patients with MEN 1 depending on the radiologic screening methods used (Table 381-1). Most of these tumors, which include cortical adenomas, hyperplasia, multiple adenomas, nodular hyperplasia, cysts, and carcinomas, are nonfunctioning. Indeed, <10% of patients with enlarged adrenal glands have hormonal hypersecretion, with primary hyperaldosteronism and ACTH-independent Cushing's syndrome being encountered most commonly. Occasionally, hyperandrogenemia may occur in association with adrenocortical carcinoma. Pheochromocytoma in association with MEN 1 is rare. Biochemical investigation (e.g., plasma renin and aldosterone concentrations, low-dose dexamethasone suppression test, urinary catecholamines, and/or metanephrines) should be undertaken in those with symptoms or signs suggestive of functioning adrenal tumors or in those with tumors >1 cm. Adrenocortical carcinoma occurs in ~1% of MEN 1 patients but increases to >10% for adrenal tumors >1 cm.

TREATMENT

Adrenocortical Tumors

Consensus has not been reached about the management of MEN 1–associated nonfunctioning adrenal tumors, because the majority are benign. However, the risk of malignancy increases with size, particularly for tumors with a diameter >4 cm. Indications for surgery for adrenal tumors include size >4 cm in diameter, atypical or suspicious radiologic features (e.g., increased Hounsfield unit on unenhanced CT scan) and size of 1–4 cm in diameter, or significant measurable growth over a 6-month period. The treatment of functioning (e.g., hormone-secreting) adrenal tumors is similar to that for tumors occurring in non–MEN 1 patients.

Meningioma Central nervous system (CNS) tumors, including ependymomas, schwannomas, and meningiomas, have been reported in MEN 1 patients (Table 381-1). Meningiomas are found in <10% of patients with other clinical manifestations of MEN 1 (e.g., primary hyperparathyroidism) for >15 years. The majority of meningiomas are not associated with symptoms, and 60% do not enlarge. The treatment of MEN 1–associated meningiomas is similar to that in non–MEN 1 patients.

Lipomas Subcutaneous lipomas occur in >33% of patients with MEN 1 (Table 381-1) and are frequently multiple. In addition, visceral, pleural, or retroperitoneal lipomas may occur in patients with MEN 1. Management is conservative. However, when surgically removed for cosmetic reasons, they typically do not recur.

Facial Angiofibromas and Collagenomas The occurrence of multiple facial angiofibromas in patients with MEN 1 may range from >20 to >90%, and occurrence of collagenomas may range from 0 to >70% (Table 381-1). These cutaneous findings may allow presymptomatic diagnosis of MEN 1 in the relatives of a patient with MEN 1. Treatment for these cutaneous lesions is usually not required.

Thyroid Tumors Thyroid tumors, including adenomas, colloid goiters, and carcinomas, have been reported to occur in >25% of patients with MEN 1. However, the prevalence of thyroid disorders in the general population is high, and it has been suggested that the association of thyroid abnormalities in patients with MEN 1 may be incidental. The treatment of thyroid tumors in MEN 1 patients is similar to that for non–MEN 1 patients.

Genetics and Screening The *MEN1* gene is located on chromosome 11q13 and consists of 10 exons, which encode a 610–amino acid protein, menin, that regulates transcription, genome stability, cell division, and proliferation. The pathophysiology of MEN 1 follows the Knudson two-hit hypothesis with a tumor-suppressor role for menin. Inheritance of a germline *MEN1* mutation predisposes an individual to developing a tumor that arises following a somatic mutation, which may be a point mutation or more commonly a deletion, leading to loss of heterozygosity (LOH) in the tumor DNA. The germline mutations of the *MEN1* gene are scattered throughout the entire 1830-bp coding region and splice sites, and there is no apparent correlation between the location of *MEN1* mutations and clinical manifestations of the disorder, in contrast with the situation in patients with MEN 2 (Table 381-1). More than 10% of *MEN1* germline mutations arise de novo and may be transmitted to subsequent generations. Some families with MEN 1 mutations develop parathyroid tumors as the sole endocrinopathy, and this condition is referred to as familial isolated hyperparathyroidism (FIHP). However, between 5 and 25% of patients with MEN 1 do not harbor germline mutations or deletions of the *MEN1* gene. Such patients with MEN 1–associated tumors but without *MEN1* mutations may represent phenocopies or have mutations involving other genes. Other genes associated with MEN 1–like features include *CDC73*, which encodes parafibromin, whose mutations result in the HPT-JT syndrome; the calcium-sensing receptor gene (*CaSR*), whose mutations result in familial benign hypocalciuric hypercalcemia (FBHH); and the aryl hydrocarbon receptor interacting protein gene (*AIP*), a tumor suppressor located on chromosome 11q13 whose mutations are associated with familial isolated pituitary adenomas (FIPA). Genetic testing to determine the *MEN1* mutation status in symptomatic family members within a MEN 1 kindred, as well as to all index cases (e.g., patients) with two or more endocrine tumors, is advisable. If an *MEN1* mutation is not identified in the index case with two or more endocrine tumors, clinical and genetic tests for other disorders such as HPT-JT syndrome, FBHH, FIPA, MEN 2, or MEN 4 should be considered, because these patients may represent phenocopies for MEN 1.

The current guidelines recommend that *MEN1* mutational analysis should be undertaken in: (1) an index case with two or more MEN 1–associated endocrine tumors (e.g., parathyroid, pancreatic, or pituitary tumors); (2) asymptomatic first-degree relatives of a known *MEN1* mutation carrier; and (3) first-degree relatives of a *MEN1* mutation carrier with symptoms, signs, or biochemical or radiologic evidence for one or more MEN 1–associated tumors. In addition, *MEN1* mutational analysis should be considered in patients with suspicious or atypical MEN 1. This would include individuals with parathyroid adenomas before the age of 30 years or multigland parathyroid disease; individuals with gastrinoma or multiple pancreatic NETs at any age; or individuals who have two or more MEN 1–associated tumors that are not part of the classical triad of parathyroid, pancreatic islet, and anterior pituitary tumors (e.g., parathyroid tumor plus adrenal tumor). Family members, including asymptomatic individuals who have been identified to harbor a *MEN1* mutation, will require biochemical and radiologic screening (Table 381-3). In contrast, relatives who do not harbor the *MEN1* mutation have a risk of developing MEN 1–associated endocrine tumors that is similar to that of the general population; thus, relatives without the *MEN1* mutation do not require repeated screening.

Mutational analysis in asymptomatic individuals should be undertaken at the earliest opportunity and, if possible, in the first decade of life because tumors have developed in some children by the age of 5 years. Appropriate biochemical and radiologic investigations (Table 381-3) aimed at detecting the development of tumors should then be undertaken in affected individuals. Mutant gene carriers should

TABLE 381-4 Recommendations for Tests and Surgery in Men 2 and Men 3[a]

		RECOMMENDED AGE (YEARS) FOR TEST/INTERVENTION				
RET MUTATION, EXON (EX) LOCATION, AND CODON INVOLVED	RISK[b]	*RET* MUTATIONAL ANALYSIS	FIRST SERUM CALCITONIN AND NECK ULTRASOUND	PROPHYLACTIC THYROIDECTOMY	SCREENING FOR PHEOCHROMOCYTOMA	SCREENING FOR PHPT
Ex8 (533)[c]; Ex10 (609, 611, 618, 620)[c]; Ex11 (630, 631, 666)[c]; Ex13 (768, 790)[c]; Ex14 (804)[c]; Ex15 (891)[c]; EX16 (912)[c]	+	<3–5	5	<5[d]	16[e]	16
Ex11 (634)[c]; Ex15 (883)[c]	++	<3	<3	<5[f]	11[e]	11
Ex15 (883)[g]; Ex16 (918)[g]	+++	ASAP and by <1	ASAP and by <0.5–1	ASAP and by <1	11[e]	—[h]

[a]Adapted from American Thyroid Association Guidelines, RT Kloos et al: Thyroid 6:565, 2009; and revised American Thyroid Association Guidelines, S Wells et al: Thyroid 25:567, 2015. [b]Risk for early development of metastasis and aggressive growth of medullary thyroid cancer: +++, highest; ++, high; and + moderate. [c]Mutations associated with MEN 2A (or medullary thyroid carcinoma only). [d] Timing of surgery to be based on elevation of serum calcitonin and/or joint discussion with pediatrician, surgeon and parent/family. Later surgery may be appropriate if serum calcitonin and neck ultrasound are normal. [e]Prescence of pheochromocytoma must be excluded prior to any surgical intervention, and also in women with *RET* mutation, who are planning pregnancy or a pregnant. [f]Surgery earlier than 5 years based on elevation of serum calcitonin. Optimal timing of surgery should be decided by the surgeon and pediatrician, in consultation with the child's parent. [g]Mutations associated with MEN 2B (MEN 3). [h]Not required because PHPT is not a feature of MEN 2B (MEN 3).

Abbreviations: ASAP, as soon as possible; MEN, multiple endocrine neoplasia; PHPT, primary hyperparathyroidism.

undergo biochemical screening at least once per annum and also have baseline pituitary and abdominal imaging (e.g., MRI or CT), which should then be repeated at 1- to 3-year intervals (Table 381-3). Screening should commence after 5 years of age and should continue for life because the disease may develop as late as the eighth decade. The screening history and physical examination elicit the symptoms and signs of hypercalcemia; nephrolithiasis; peptic ulcer disease; neuroglycopenia; hypopituitarism; galactorrhea and amenorrhea in women; acromegaly; Cushing's disease; and visual field loss and the presence of subcutaneous lipomas, angiofibromas, and collagenomas. Biochemical screening should include measurements of serum calcium, PTH, gastrointestinal hormones (e.g., gastrin, insulin with a fasting glucose, glucagon, VIP, PP), chromogranin A, prolactin, and IGF-I in all individuals. More specific endocrine function tests should be undertaken in individuals who have symptoms or signs suggestive of a specific clinical syndrome. Biochemical screening for the development of MEN 1 tumors in asymptomatic members of families with MEN 1 is of great importance to reduce morbidity and mortality from the associated tumors.

MULTIPLE ENDOCRINE NEOPLASIA TYPE 2 AND TYPE 3

Clinical Manifestations MEN type 2 (MEN 2), which is also called Sipple's syndrome, is characterized by the association of medullary thyroid carcinoma (MTC), pheochromocytomas, and parathyroid tumors (Table 381-1). Three clinical variants of MEN 2 are recognized: MEN 2A, MEN 2B, and MTC only. MEN 2A, which is often referred to as MEN 2, is the most common variant. In MEN 2A, MTC is associated with pheochromocytomas in 50% of patients (may be bilateral) and with parathyroid tumors in 20% of patients. MEN 2A may rarely occur in association with Hirschsprung's disease, caused by the absence of autonomic ganglion cells in the terminal hindgut, resulting in colonic dilatation, severe constipation, and obstruction. MEN 2A may also be associated with cutaneous lichen amyloidosis, which is a pruritic lichenoid lesion that is usually located on the upper back. MEN 2B, which is also referred to as MEN 3, represents 5% of all cases of MEN 2 and is characterized by the occurrence of MTC and pheochromocytoma in association with a Marfanoid habitus; mucosal neuromas of the lips, tongue, and eyelids; medullated corneal fibers; and intestinal autonomic ganglion dysfunction leading to multiple diverticulae and megacolon. Parathyroid tumors do not usually occur in MEN 2B. MTC only (FMTC) is a variant in which MTC is the sole manifestation of the syndrome. However, the distinction between FMTC and MEN 2A is difficult and should only be considered if there are at least four family members aged >50 years who are affected by MTC but not pheochromocytomas or primary hyperparathyroidism. All of the MEN 2 variants are due to mutations of the rearranged during transfection (*RET*) protooncogene, which encodes a TKR. Moreover, there is a correlation between the locations of *RET* mutations and MEN 2 variants. Thus, ~95% of MEN 2A patients have mutations involving the cysteine-rich extracellular domain, with mutations of codon 634 accounting for ~85% of MEN 2A mutations; FMTC patients also have mutations of the cysteine-rich extracellular domain, with most mutations occurring in codon 618. In contrast, ~95% of MEN 2B/MEN 3 patients have mutations of codon 918 of the intracellular tyrosine kinase domain (Table 381-1 and Table 381-4).

Medullary Thyroid Carcinoma MTC is the most common feature of MEN 2A and MEN 2B and occurs in almost all affected individuals. MTC represents 5–10% of all thyroid gland carcinomas, and 20% of MTC patients have a family history of the disorder. The use of *RET* mutational analysis to identify family members at risk for hereditary forms of MTC has altered the presentation of MTC from that of symptomatic tumors to a preclinical disease for which prophylactic thyroidectomy (Table 381-4) is undertaken to improve the prognosis and ideally result in cure. However, in patients who do not have a known family history of MEN 2A, FMTC, or MEN 2B, and therefore have not had *RET* mutational analysis, MTC may present as a palpable mass in the neck, which may be asymptomatic or associated with symptoms of pressure or dysphagia in >15% of patients. Diarrhea occurs in 30% of patients and is associated either with elevated circulating concentrations of calcitonin or tumor-related secretion of serotonin and prostaglandins. Some patients may also experience flushing. In addition, ectopic ACTH production by MTC may cause Cushing's syndrome. The diagnosis of MTC relies on the demonstration of hypercalcitoninemia (>90 pg/mL in the basal state); stimulation tests using IV pentagastrin (0.5 mg/kg) and or calcium infusion (2 mg/kg) are rarely used now, reflecting improvements in the assay for calcitonin. Neck ultrasonography with fine-needle aspiration of the nodules can confirm the diagnosis. Radionucleotide thyroid scans may reveal MTC tumors as "cold" nodules. Radiography may reveal dense irregular calcification within the involved portions of the thyroid gland and in lymph nodes involved with metastases. Positron emission tomography (PET) may help to identify the MTC and metastases (Fig. 381-2). Metastases of MTC usually occur to the cervical lymph nodes in the early stages and to the mediastinal nodes, lung, liver, trachea, adrenal, esophagus, and bone in later stages. Elevations in serum calcitonin concentrations are often the first sign of recurrence or persistent disease, and the serum calcitonin doubling time is useful for determining prognosis. MTC can have an aggressive clinical course, with early metastases and death in ~10% of patients. A family history of aggressive MTC or MEN 2B may be elicited.

TREATMENT

Medullary Thyroid Carcinoma

Individuals with *RET* mutations who do not have clinical manifestations of MTC should be offered prophylactic surgery between the ages of <1 and 5 years. The timing of surgery will depend on

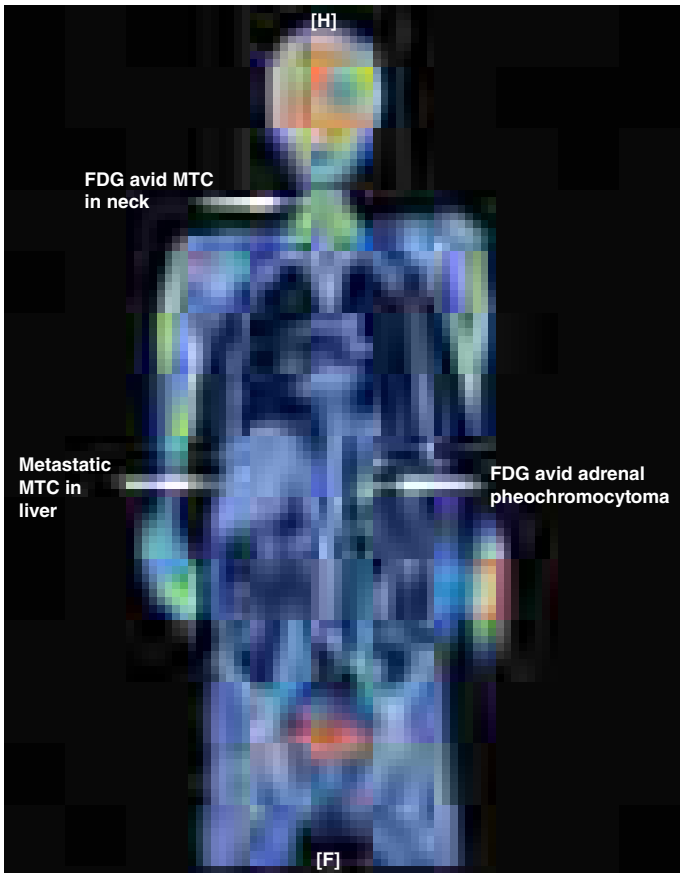


FIGURE 381-2 Fluorodeoxyglucose (FDG) positron emission tomography scan in a patient with multiple endocrine neoplasia type 2A, showing medullary thyroid cancer (MTC) with hepatic and skeletal (left arm) metastasis and a left adrenal pheochromocytoma. Note the presence of excreted FDG compound in the bladder. *(Reproduced with permission from A Naziat et al: Clin Endocrinol [Oxf] 78:966, 2013.)*

the type of *RET* mutation and its associated risk for early development, metastasis, and aggressive growth of MTC (Table 381-4). Such patients should have a total thyroidectomy with a systematic central neck dissection to remove occult nodal metastasis, although the value of undertaking a central neck dissection has been subject to debate. Prophylactic thyroidectomy, with lifelong thyroxine replacement, has dramatically improved outcomes in patients with MEN 2 and MEN 3, such that ~90% of young patients with *RET* mutations who had a prophylactic thyroidectomy have no evidence of persistent or recurrent MTC at 7 years after surgery. In patients with clinically evident MTC, a total thyroidectomy with bilateral central resection is recommended, and an ipsilateral lateral neck dissection should be undertaken if the primary tumor is >1 cm in size or there is evidence of nodal metastasis in the central neck. Surgery is the only curative therapy for MTC. The 10-year survival in patients with metastatic MTC is ~20%. For inoperable MTC or metastatic disease, the TKR inhibitors, such as vandetanib and cabozantinib, have improved the progression-free survival times. Other types of chemotherapy are of limited efficacy, but radiotherapy may help to palliate local disease.

Pheochromocytoma (See also Chap. 380) These noradrenaline- and adrenaline-secreting tumors occur in >50% of patients with MEN 2A and MEN 2B and are a major cause of morbidity and mortality. Patients may have symptoms and signs of catecholamine secretion (e.g., headaches, palpitations, sweating, poorly controlled hypertension), or they may be asymptomatic with detection through biochemical screening based on a history of familial MEN 2A, MEN 2B, or MTC. Pheochromocytomas in patients with MEN 2A and MEN 2B differ significantly in distribution when compared with patients without MEN 2A and MEN 2B. Extra-adrenal pheochromocytomas, which occur in 10% of patients without MEN 2A and MEN 2B, are observed rarely in patients with MEN 2A and MEN 2B. Malignant pheochromocytomas are much less common in patients with MEN 2A and MEN 2B. The biochemical and radiologic investigation of pheochromocytoma in patients with MEN 2A and MEN 2B is similar to that in non–MEN 2 patients and includes the measurement of plasma (obtained from supine patients) and urinary free fractionated metanephrines (e.g., normetanephrine and metanephrines measured separately), CT or MRI scanning, radionuclide scanning with meta-iodo-(^{123}I or ^{131}I)-benzyl guanidine (MIBG), and PET using (^{18}F)-fluorodopamine or (^{18}F)-fluoro-2-dexoxy-D-glucose (Fig. 381-2).

TREATMENT

Pheochromocytoma

Surgical removal of pheochromocytoma, using α and β adrenoreceptor blockade before and during the operation, is the recommended treatment. Other antihypertensive agents, including calcium channel blockers, are sometimes required for adequate blood pressure control. Endoscopic adrenal-sparing surgery, which decreases postoperative morbidity, hospital stay, and expense, as opposed to open surgery, has become the method of choice.

Parathyroid Tumors (See also Chap. 403) Parathyroid tumors occur in 10–25% of patients with MEN 2A. However, >50% of these patients do not have hypercalcemia. The presence of abnormally enlarged parathyroids, which are unusually hyperplastic, is often seen in the normocalcemic patient undergoing thyroidectomy for MTC. The biochemical investigation and treatment of hypercalcemic patients with MEN 2A is similar to that of patients with MEN 1.

Genetics and Screening To date, ~50 different *RET* mutations have been reported, and these are located in exons 5, 8, 10, 11, 13, 14, 15, and 16. *RET* germline mutations are detected in >95% of MEN 2A, FMTC, and MEN 2B families, with Cys634Arg being most common in MEN 2A, Cys618Arg being most common in FMTC, and Met918Thr being most common in MEN 2B (Tables 381-1 and 381-4). Between 5 and 10% of patients with MTC or MEN 2A–associated tumors have de novo *RET* germline mutations, and ~50% of patients with MEN 2B have de novo *RET* germline mutations. These de novo *RET* germline mutations always occur on the paternal allele. Approximately 5% of patients with sporadic pheochromocytoma have a germline *RET* mutation, but such germline *RET* mutations do not appear to be associated with sporadic primary hyperparathyroidism. Thus, *RET* mutational analysis should be performed in: (1) all patients with MTC who have a family history of tumors associated with MEN 2, FMTC, or MEN 3, such that the diagnosis can be confirmed and genetic testing offered to asymptomatic relatives; (2) all patients with MTC and pheochromocytoma without a known family history of MEN 2 or MEN 3; (3) all patients with MTC, but without a family history of MEN 2, FMTC, or MEN 3, because these patients may have a de novo germline *RET* mutations; (4) all patients with bilateral pheochromocytoma; and (5) patients with unilateral pheochromocytoma, particularly if this occurs with increased calcitonin levels.

Screening for MEN 2/MEN 3–associated tumors in patients with *RET* germline mutations should be undertaken annually and include serum calcitonin measurements, a neck ultrasound for MTC, plasma and 24-h urinary fractionated metanephrines for pheochromocytoma, and albumin-corrected serum calcium or ionized calcium with PTH for primary hyperparathyroidism. In patients with MEN 2–associated *RET* mutations, screening for MTC should begin by 1–5 years, for pheochromocytoma by 11–16 years, and for primary hyperparathyroidism by 11–16 years of age (Table 381-4).

MULTIPLE ENDOCRINE NEOPLASIA TYPE 4

Clinical Manifestations Patients with MEN 1–associated tumors, such as parathyroid adenomas, pituitary adenomas, and

pancreatic NETs, occurring in association with gonadal, adrenal, renal, and thyroid tumors have been reported to have mutations of the gene encoding the 196–amino acid cyclin-dependent kinase inhibitor (CK1) p27 kip1 (*CDNKIB*). Such families with MEN 1–associated tumors and *CDNKIB* mutations are designated to have MEN 4 (Table 381-1). The investigations and treatments for the MEN 4–associated tumors are similar to those for MEN 1 and non–MEN 1 tumors.

Genetics and Screening To date, 13 different MEN 4–associated mutations of *CDKNIB*, which is located on chromosome 12p13, have been reported, and all of these are associated with a loss of function. These MEN 4 patients may represent ~3% of the 5–10% of patients with MEN 1 who do not have mutations of the *MEN1* gene. Germline *CDKNIB* mutations may rarely be found in patients with sporadic (i.e., nonfamilial) forms of primary hyperparathyroidism.

HYPERPARATHYROIDISM-JAW TUMOR SYNDROME (SEE ALSO CHAP. 403)

Clinical Manifestations Hyperparathyroidism-jaw tumor (HPT-JT) syndrome is an autosomal dominant disorder characterized by the development of parathyroid tumors (15% are carcinomas) and fibro-osseous jaw tumors. In addition, some patients may also develop Wilms' tumors, renal cysts, renal hamartomas, renal cortical adenomas, papillary RCCs, pancreatic adenocarcinomas, uterine tumors, testicular mixed germ cell tumors with a major seminoma component, and Hürthle cell thyroid adenomas. The parathyroid tumors may occur in isolation and without any evidence of jaw tumors, and this may cause confusion with other hereditary hypercalcemic disorders, such as MEN 1. However, genetic testing to identify the causative mutation will help to establish the correct diagnosis. The investigation and treatment for HPT-JT–associated tumors are similar to those in non-HPT-JT patients, except that early parathyroidectomy is advisable because of the increased frequency of parathyroid carcinoma.

Genetics and Screening The gene that causes HPT-JT is located on chromosome 1q31.2 and encodes a 531–amino acid protein, parafibromin (Table 381-2). Parafibromin is also referred to as cell division cycle protein 73 (CDC73) and has a role in transcription. Genetic testing in families helps to identify mutation carriers who should be periodically screened for the development of tumors (Table 381-5).

VON HIPPEL-LINDAU DISEASE (SEE ALSO CHAP. 380)

Clinical Manifestations von Hippel-Lindau (VHL) disease is an autosomal dominant disorder characterized by hemangioblastomas of the retina and CNS; cysts involving the kidneys, pancreas, and epididymis; renal cell carcinoma (RCC); pheochromocytomas; and pancreatic islet cell tumors. The retinal and CNS hemangioblastomas are benign vascular tumors that may be multiple; those in the CNS may cause symptoms by compressing adjacent structures and/or increasing intracranial pressure. In the CNS, the cerebellum and spinal cord are the most frequently involved sites. The renal abnormalities consist of cysts and carcinomas, and the lifetime risk of RCC in VHL is 70%. The endocrine tumors in VHL consist of pheochromocytomas and pancreatic islet cell tumors. The clinical presentation of pheochromocytoma in VHL disease is similar to that in sporadic cases, except that there is a higher frequency of bilateral or multiple tumors, which may involve extra-adrenal sites in VHL disease. The most frequent pancreatic lesions in VHL are multiple cyst-adenomas, which rarely cause clinical disease. However, nonsecreting pancreatic islet cell tumors occur in <10% of VHL patients, who are usually asymptomatic. The pancreatic tumors in these patients are often detected by regular screening using abdominal imaging. Pheochromocytomas should be investigated and treated as described earlier for MEN 2. The pancreatic islet cell tumors frequently become malignant, and early surgery is recommended.

Genetics and Screening The *VHL* gene, which is located on chromosome 3p26-p25, is widely expressed in human tissues and encodes a 213–amino acid protein (pVHL) (Table 381-2). A wide variety of germline *VHL* mutations have been identified. *VHL* acts as a tumor-suppressor gene. A correlation between the type of mutation and the clinical phenotype has been reported; large deletions and protein-truncating mutations are associated with a low incidence of pheochromocytomas, whereas some missense mutations in VHL patients are associated with pheochromocytoma (referred to as VHL type 2C). Other missense mutations may be associated with hemangioblastomas and RCC but not pheochromocytoma (referred to as VHL type 1), whereas distinct missense mutations are associated with hemangioblastomas, RCC, and pheochromocytoma (VHL type 2B). VHL type 2A, which refers to the occurrence of hemangioblastomas and pheochromocytoma without RCC, is associated with rare missense mutations. The basis for these complex genotype-phenotype relationships remains to be elucidated. One major function of pVHL, which is also referred to as elongin, is to downregulate the expression of vascular endothelial growth factor (VEGF) and other hypoxia-inducible mRNAs. Thus, pVHL, in complex with other proteins, regulates the expression of hypoxia-inducible factors (HIF-1 and HIF-2) such that loss of functional pVHL leads to a stabilization of the HIF protein complexes, resulting in VEGF overexpression and tumor angiogenesis. Screening for the development of pheochromocytomas and pancreatic islet cell tumors is as described earlier for MEN 2 and MEN 1, respectively (Tables 381-3 and 381-4).

NEUROFIBROMATOSIS

Clinical Manifestations Neurofibromatosis type 1 (NF1), which is also referred to as von Recklinghausen's disease, is an autosomal dominant disorder characterized by the following manifestations: neurologic (e.g., peripheral and spinal neurofibromas); ophthalmologic (e.g., optic gliomas and iris hamartomas such as Lisch nodules); dermatologic (e.g., café au lait macules); skeletal (e.g., scoliosis, macrocephaly, short stature, and pseudoarthrosis); vascular (e.g., stenoses of renal and intracranial arteries); and endocrine (e.g., pheochromocytoma, carcinoid tumors, and precocious puberty). Neurofibromatosis type 2 (NF2) is also an autosomal dominant disorder but is characterized by the development of bilateral vestibular schwannomas (acoustic neuromas) that lead to deafness, tinnitus, or vertigo. Some patients with NF2 also develop meningiomas, spinal schwannomas, peripheral nerve neurofibromas, and café au lait macules. Endocrine abnormalities are not found in NF2 and are associated solely with NF1. Pheochromocytomas, carcinoid tumors, and precocious puberty occur in ~1% of patients with NF1, and growth hormone deficiency has also been reported. The features of pheochromocytomas in NF1 are similar to those in non-NF1 patients, with 90% of tumors being located within the adrenal medulla and the remaining 10% at an extra-adrenal location, which often involves the para-aortic region. Primary carcinoid tumors are often periampullary and may also occur in the ileum but rarely in the pancreas, thyroid, or lungs. Hepatic metastases are associated with symptoms of the carcinoid syndrome, which include flushing,

TABLE 381-5 HPT-JT Screening Guidelines

TUMOR[a]	TEST	FREQUENCY[b]
Parathyroid	Serum Ca, PTH	6–12 months
Ossifying jaw fibroma	Panoramic jaw x-ray with neck shielding[c]	5 years
Renal	Abdominal MRI[c,d]	5 years
Uterine	Ultrasound (transvaginal or transabdominal) and additional imaging ± D&C if indicated[e]	Annual

[a]Screening for most common HPT-JT–associated tumors is considered. Assessment for other reported tumor types may be indicated (e.g., pancreatic, thyroid, testicular tumors). [b]Frequency of repeating test after baseline tests performed. [c]X-rays and imaging involving ionizing radiation should ideally be avoided to minimize risk of generating subsequent mutations. [d]Ultrasound scan recommended if MRI unavailable. [e]Such selective pelvic imaging should be considered after obtaining a detailed menstrual history.

Abbreviations: Ca, calcium; D&C, dilation and curettage; HPT-JT, hyperparathyroidism-jaw tumor syndrome; MRI, magnetic resonance imaging; PTH, parathyroid hormone.

Source: Reproduced from PJ Newey et al: Hum Mutat 31:295, 2010.

diarrhea, bronchoconstriction, and tricuspid valve disease. Precocious puberty is usually associated with the extension of an optic glioma into the hypothalamus with resultant early activation of gonadotropin-releasing hormone secretion. Growth hormone deficiency has also been observed in some NF1 patients, who may or may not have optic chiasmal gliomas, but it is important to note that short stature is frequent in the absence of growth hormone deficiency in patients with NF1. The investigation and treatment for tumors are similar to those undertaken for each respective tumor type in non-NF1 patients.

Genetics and Screening The *NF1* gene, which is located on chromosome 17q11.2 and acts as a tumor suppressor, consists of 60 exons that span more than 350 kb of genomic DNA (Table 381-2). Mutations in *NF1* are of diverse types and are scattered throughout the exons. The NF1 gene product is the protein neurofibromin, which has homologies to the p120GAP (GTPase activating protein) and acts on p21ras by converting the active GTP bound form to its inactive GDP form. Mutations of *NF1* impair this downregulation of the p21ras signaling pathways, which in turn results in abnormal cell proliferation. Screening for the development of pheochromocytomas and carcinoid tumors is as described earlier for MEN 2 and MEN 1, respectively (Tables 381-3 and 381-4).

CARNEY COMPLEX

Clinical Manifestations Carney complex (CNC) is an autosomal dominant disorder characterized by spotty skin pigmentation (usually of the face, labia, and conjunctiva), myxomas (usually of the eyelids and heart, but also the tongue, palate, breast, and skin), psammomatous melanotic schwannomas (usually of the sympathetic nerve chain and upper gastrointestinal tract), and endocrine tumors that involve the adrenals, Sertoli cells, somatotropes, thyroid, and ovary. Cushing's syndrome, the result of primary pigmented nodular adrenal disease (PPNAD), is the most common endocrine manifestation of CNC and may occur in one-third of patients. Patients with CNC and Cushing's syndrome often have an atypical appearance by being thin (as opposed to having truncal obesity). In addition, they may have short stature, muscle and skin wasting, and osteoporosis. These patients often have levels of urinary free cortisol that are normal or increased only marginally. Cortisol production may fluctuate periodically with days or weeks of hypercortisolism; this pattern is referred to as "periodic Cushing's syndrome." Patients with Cushing's syndrome usually have loss of the circadian rhythm of cortisol production. Acromegaly, the result of a somatotrope tumor, affects ~10% of patients with CNC. Testicular tumors may also occur in one-third of patients with CNC. These may either be large-cell calcifying Sertoli cell tumors, adrenocortical rests, or Leydig cell tumors. The Sertoli cell tumors occasionally may be estrogen-secreting and lead to precocious puberty or gynecomastia. Some patients with CNC have been reported to develop thyroid follicular tumors, ovarian cysts, or breast duct adenomas.

Genetics and Screening CNC type 1 (CNC1) is due to mutations of the protein kinase A (PKA) regulatory subunit 1 α (R1α) (*PPKAR1A*), a tumor suppressor, whose gene is located on chromosome 17q.24.2 (Table 381-2). The gene causing CNC type 2 (CNC2) is located on chromosome 2p16 and has not yet been identified. It is interesting to note, however, that some tumors do not show LOH of 2p16 but instead show genomic instability, suggesting that this CNC gene may not be a tumor suppressor. Screening and treatment of these endocrine tumors are similar to those described earlier for patients with MEN 1 and MEN 2 (Tables 381-3 and 381-4).

COWDEN'S SYNDROME

Clinical Manifestations Multiple hamartomatous lesions, especially of the skin, mucous membranes (e.g., buccal, intestinal, and colonic), breast, and thyroid, are characteristic of Cowden's syndrome (CWS), which is an autosomal dominant disorder. Thyroid abnormalities occur in two-thirds of patients with CWS, and these usually consist of multinodular goiters or benign adenomas, although <10% of patients may have a follicular thyroid carcinoma. Breast abnormalities occur in >75% of patients and consist of either fibrocystic disease or adenocarcinomas. The investigation and treatment for CWS tumors are similar to those undertaken for non-CWS patients.

Genetics and Screening CWS is genetically heterogenous, and seven types (CWS1–7) are recognized (Table 381-2). CWS1 is due to mutations of the phosphate and tensin homologue deleted on chromosome 10 (*PTEN*) gene, located on chromosome 10q23.31. CWS2 is caused by mutations of the succinate dehydrogenase subunit B (*SDHB*) gene, located on chromosome 1p36.13; and CWS3 is caused by mutations of the *SDHD* gene, located on chromosome 11q13.1. *SDHB* and *SDHD* mutations are also associated with pheochromocytoma. CWS4 is caused by hypermethylation of the Killin (*KLLN*) gene, the promoter of which shares the same transcription site as *PTEN* on chromosome 10q23.31. CWS5 is caused by mutations of the phosphatidylinositol 3-kinase catalytic alpha (*PIK3CA*) gene on chromosome 3q26.32. CWS6 is caused by mutations of the V-Akt murine thymoma viral oncogene homolog 1 (*AKT1*) gene on chromosome 14q32.33, and CWS7 is caused by mutations of the saccharomyces cerevisiae homology of B (SEC23B) gene and chromosome 20p11.23. Screening for thyroid abnormalities entails neck ultrasonography and fine-needle aspiration with analysis of cell cytology.

MCCUNE-ALBRIGHT SYNDROME (SEE ALSO CHAP. 405)

Clinical Manifestations McCune-Albright syndrome (MAS) is characterized by the triad of polyostotic fibrous dysplasia, which may be associated with hypophosphatemic rickets; café au lait skin pigmentation; and peripheral precocious puberty. Other endocrine abnormalities include thyrotoxicosis, which may be associated with a multinodular goiter, somatotrope tumors, and Cushing's syndrome (due to adrenal tumors). Investigation and treatment for each endocrinopathy are similar to those used in patients without MAS.

Genetics and Screening MAS is a disorder of mosaicism that results from postzygotic somatic cell mutations of the G protein α-stimulating subunit (Gsα), encoded by the *GNAS1* gene, located on chromosome 20q13.32 (Table 381-2). The Gsα mutations, which include Arg201Cys, Arg201His, Glu227Arg, or Glu227His, are activating and are found only in cells of the abnormal tissues. Screening for hyperfunction of relevant endocrine glands and development of hypophosphatemia, which may be associated with elevated serum fibroblast growth factor 23 (FGF23) concentrations, is undertaken in MAS patients.

Acknowledgment

The author is grateful to the Medical Research Council (UK) for support and to Mrs. Tracey Walker for typing the manuscript.

FURTHER READING

American Thyroid Association Guidelines Task Force et al: Medullary thyroid cancer: management guidelines of the American Thyroid Association. Thyroid 19:565, 2009.

Christakis I et al: Clinical features, treatments, and outcomes of patients with thymic carcinoids and multiple endocrine neoplasia type 1 syndrome at MD Anderson Cancer Center. Horm Cancer 7:279, 2016.

Mulligan LM: RET revisited: Expanding the oncogenic portfolio. Nat Rev Cancer 14:173, 2014.

Newey PJ et al: Asymptomatic children with multiple endocrine neoplasia type 1 (*MEN1*) mutations may harbor non-functioning pancreatic neuroendocrine tumors. J Clin Endocrinol Metab 10: 3640, 2009.

Pilarski R et al: Cowden syndrome and the PTEN hamartoma tumor syndrome: Systematic review and revised diagnostic criteria. J Natl Cancer Inst 105:1607, 2013.

Salpea P, Stratakis CA: Carney complex and McCune Albright syndrome: An overview of clinical manifestations and human molecular genetics. Mol Cell Endocrinol 386:85, 2014.

Thakker RV: Multiple endocrine neoplasia type 1 (MEN1) and type 4 (MEN4). Mol Cell Endocrinol 5:2, 2014.

2756 Thakker RV et al: Clinical practice guidelines for multiple endocrine neoplasia type 1 (MEN1). J Clin Endocrinol Metab 97:2990, 2012.
Wells SA Jr et al: Revised American Thyroid Association guidelines for the management of medullary thyroid carcinoma. Thyroid 25:567, 2015.
Yates CJ, Newey PJ, Thakker RV: Challenges and controversies in management of pancreatic neuroendocrine tumours in patients with MEN1. Lancet Diabetes Endocrinol 3:895, 2015.

382 Autoimmune Polyendocrine Syndromes

Peter A. Gottlieb, Aaron W. Michels

Polyglandular deficiency syndromes have been given many different names, reflecting the wide spectrum of disorders that have been associated with these syndromes and the heterogeneity of their clinical presentations. The name used in this chapter for this group of disorders is *autoimmune polyendocrine syndrome* (APS). In general, these disorders are divided into two major categories, APS type 1 (APS-1) and APS type 2 (APS-2). Some groups have further subdivided APS-2 into APS type 3 (APS-3) and APS type 4 (APS-4) depending on the type of autoimmunity involved. For the most part, this additional classification does not clarify our understanding of disease pathogenesis or prevention of complications in individual patients. Importantly, there are many nonendocrine disease associations included in these syndromes, suggesting that although the underlying autoimmune disorder predominantly involves endocrine targets, it does not exclude other tissues. The disease associations found in APS-1 and APS-2 are summarized in Table 382-1. Understanding these syndromes and their disease manifestations can lead to early diagnosis and treatment of additional disorders in patients and their family members.

APS-1

APS-1 (Online Mendelian Inheritance in Man [OMIM] 240300) has also been called autoimmune polyendocrinopathy–candidiasis–ectodermal dystrophy (APECED). Mucocutaneous candidiasis, hypoparathyroidism, and Addison's disease form the three major components of this disorder. However, as summarized in Table 382-1, many other organ systems can be involved over time. APS-1 is rare, with fewer than 500 cases reported in the literature.

The classical form of APS-1 is an autosomal recessive disorder caused by mutations in the *AIRE* gene (autoimmune regulator gene) found on chromosome 21. This gene is most highly expressed in thymic medullary epithelial cells (mTECs) where it appears to control the expression of tissue-specific self-antigens (e.g., insulin). Deletion of this regulator leads to decreased expression of tissue-specific self-antigens and is hypothesized to allow autoreactive T cells to avoid central deletion, which normally occurs during T cell maturation in the thymus. The *AIRE* gene is also expressed in epithelial cells found in peripheral lymphoid organs, but its role in these extrathymic cells remains controversial. To date, over 100 mutations have been described in this gene, and there is a higher frequency within certain ethnic groups including Iranian Jews, Sardinians, Finns, Norwegians, and Irish. Recently, several autosomal dominant mutations have been identified, and are localized primarily in the PHD1 domain of the *AIRE* gene, rather than the CARD region where the autosomal recessive mutations have been found. Individuals with this non-classical form of APS-1 may have a later onset of symptoms, and less aggressive disease, without the full spectrum of autoimmune components being expressed.

Clinical Manifestations Classical APS-1 develops very early in life, often in infancy (Table 382-2). Chronic mucocutaneous candidiasis without signs of systemic disease is often the first manifestation. It affects the mouth and nails more frequently than the skin and esophagus. Chronic oral candidiasis can result in atrophic disease with areas suggestive of leukoplakia, which can pose a risk for future carcinoma. The etiology is associated with anticytokine autoantibodies (anti-IL-17A, IL-17F, and IL-22) related to T helper (T_H) 17 T cells and depressed production of these cytokines by peripheral blood mononuclear cells. Hypoparathyroidism

TABLE 382-1 Disease Associations with Autoimmune Polyendocrine Syndromes

AUTOIMMUNE POLYENDOCRINE SYNDROME TYPE 1	AUTOIMMUNE POLYENDOCRINE SYNDROME TYPE 2	OTHER AUTOIMMUNE POLYENDOCRINE DISORDERS
Endocrine	**Endocrine**	IPEX (immune dysfunction polyendocrinopathy X-linked)
Addison's disease	Addison's disease	Thymic tumors
Hypoparathyroidism	Type 1 diabetes	Anti-insulin receptor antibodies
Hypogonadism	*Graves' disease or autoimmune thyroiditis*	POEMS syndrome
Graves' disease or autoimmune thyroiditis	*Hypogonadism*	Insulin autoimmune syndrome (Hirata's syndrome)
Type 1 diabetes		Adult combined pituitary hormone deficiency (CPHD) with anti-Pit1 autoantibodies
		Kearns-Sayre syndrome
		DIDMOAD syndrome
Nonendocrine	**Nonendocrine**	Congenital rubella associated with thyroiditis and/or diabetes
Mucocutaneous candidiasis	Celiac disease, dermatitis herpetiformis	
Chronic active hepatitis	Pernicious anemia	
Pernicious anemia	Vitiligo	
Vitiligo	*Alopecia*	
Asplenism	*Myasthenia gravis*	
Ectodermal dysplasia	*IgA deficiency*	
Alopecia	*Parkinson's disease*	
Malabsorption syndromes	*Idiopathic thrombocytopenia*	
IgA deficiency		

Abbreviations: DIDMOAD, *d*iabetes *i*nsipidus, *d*iabetes *m*ellitus, progressive bilateral optic *a*trophy, and sensorineural *d*eafness; POEMS, *p*olyneuropathy, *o*rganomegaly, *e*ndocrinopathy, *M*-protein, and *s*kin changes.

Note: Italics denote less common disorders.

TABLE 382-2 Comparison of APS-1 and APS-2

APS-1	APS-2
Early onset: infancy	Later onset
Siblings often affected and at risk	Multigenerational
Equivalent sex distribution	Females > males affected
Monogenic: *AIRE* gene, chromosome 21, autosomal recessive	Polygenic: *HLA, MICA, PTNP22, CTLA4*
Not HLA associated for entire syndrome, some specific component risk	DR3/DR4 associated; other HLA class III gene associations noted
Autoantibodies to type 1 interferons and IL-17 and IL-22	No autoantibodies to cytokines
Autoantibodies to specific target organs	Autoantibodies to specific target organs
Asplenism	No defined immunodeficiency
Mucocutaneous candidiasis	Association with other nonendocrine immunologic disorders like myasthenia gravis and idiopathic thrombocytopenic purpura

Abbreviations: APS, autoimmune polyendocrine syndrome; IL, interleukin.

usually develops next, followed by adrenal insufficiency. The time from development of one component of the disorder to the next can be many years, and the order of disease appearance is variable.

Chronic candidiasis is nearly always present and is not very responsive to treatment. Hypoparathyroidism is found in >85% of cases, and Addison's disease is found in nearly 80%. Gonadal failure appears to affect women more than men (70% vs 25%, respectively), and hypoplasia of the dental enamel also occurs frequently (77% of patients). Other endocrine disorders that occur less frequently include type 1 diabetes (23%) and autoimmune thyroid disease (18%). Nonendocrine manifestations that present less frequently include alopecia (40%), vitiligo (26%), intestinal malabsorption (18%), pernicious anemia (31%), chronic active hepatitis (17%), and nail dystrophy. An unusual and debilitating manifestation of the disorder is the development of refractory diarrhea/obstipation that may be related to autoantibody-mediated destruction of enterochromaffin or enterochromaffin-like cells. The incidence rates for many of these disorders peak in the first or second decade of life, but the individual disease components continue to emerge over time. Therefore, prevalence rates may be higher than originally reported.

Diagnosis The diagnosis of APS-1 is usually made clinically when two of the three major component disorders are found in an individual patient. Siblings of individuals with APS-1 should be considered affected even if only one component disorder has been detected due to the known inheritance of the syndrome. Genetic analysis of the *AIRE* gene should be undertaken to identify mutations. Detection of anti–interferon α and anti–interferon omega antibodies can identify nearly 100% of cases with APS-1. The autoantibody arises independent of the type of *AIRE* gene mutation and is not found in other autoimmune disorders.

Diagnosis of each underlying disorder should be done based on their typical clinical presentations **(Table 382-3)**. Mucocutaneous candidiasis may present throughout the gastrointestinal tract, and it may be detected in the oral mucosa or from stool samples. Evaluation by a gastroenterologist to examine the esophagus for candidiasis or secondary stricture may be merited based on symptoms. Other gastrointestinal manifestations of APS-1, including malabsorption and obstipation, may also bring these young patients to the attention of gastroenterologists for first evaluation. Specific physical examination findings of hyperpigmentation, vitiligo, alopecia, tetany, and signs of hyper- or hypothyroidism should be considered as signs of development of component disorders.

The development of disease-specific autoantibody assays can help confirm disease and also detect risk for future disease. For example, where possible, detection of anti-cytokine antibodies to interleukin (IL) 17 and IL-22 would confirm the diagnosis of mucocutaneous candidiasis due to APS-1. The presence of anti-21-hydroxylase antibody or anti-17-hydroxylase antibody (which may be found more commonly in adrenal insufficiency associated with APS-1) would confirm the presence or risk for Addison's disease. Other autoantibodies found in type 1 diabetes (e.g., anti-GAD65), pernicious anemia, and other component conditions should be screened for on a regular basis (6- to 12-month intervals depending on the age of the subject).

Laboratory tests, including a complete metabolic panel, phosphorous and magnesium, thyroid-stimulating hormone (TSH), adrenocorticotropic hormone (ACTH; morning), hemoglobin A_{1c}, plasma vitamin B_{12} level, and complete blood count with peripheral smear looking for Howell-Jolly bodies (asplenism), should also be performed at these time points. Detection of abnormal physical findings or test results should prompt subsequent examinations of the relevant organ system (e.g., presence of Howell-Jolly bodies indicates need for ultrasound of spleen).

TABLE 382-3 Clinical Features and Recommended Follow-Up for APS-1 and APS-2

COMPONENT DISEASE	RECOMMENDED EVALUATION
APS-1	
Addison's disease	Sodium, potassium, ACTH, cortisol, 21- and 17-hydroxylase autoantibodies
Diarrhea	History
Ectodermal dysplasia	Physical examination
Hypoparathyroidism	Serum calcium, phosphate, PTH
Hepatitis	Liver function tests
Hypothyroidism/Graves' disease	TSH; thyroid peroxidase and/or thyroglobulin autoantibodies and anti-TSH receptor Ab
Male hypogonadism	FSH/LH, testosterone
Malabsorption	Physical examination, anti-IL-17 and anti-IL-22 autoantibodies
Mucocutaneous candidiasis	Physical examination, mucosal swab, stool samples
Obstipation	History
Ovarian failure	FSH/LH, estradiol
Pernicious anemia	CBC, vitamin B_{12} levels
Splenic atrophy	Blood smear for Howell-Jolly bodies; platelet count; ultrasound if positive
Type 1 diabetes	Glucose, hemoglobin A_{1c}, diabetes-associated autoantibodies (insulin, GAD65, IA-2, ZnT8)
APS-2	
Addison's disease	21-Hydroxylase autoantibodies, ACTH stimulation testing if positive
Alopecia	Physical examination
Autoimmune hyper- or hypothyroidism	TSH; thyroid peroxidase and/or thyroglobulin autoantibodies, anti-TSH receptor Ab
Celiac disease	Transglutaminase autoantibodies; small intestine biopsy if positive
Cerebellar ataxia	Dictated by signs and symptoms of disease
Chronic inflammatory demyelinating polyneuropathy	Dictated by signs and symptoms of disease
Hypophysitis	Dictated by signs and symptoms of disease, anti-Pit1 autoantibody
Idiopathic heart block	Dictated by signs and symptoms of disease
IgA deficiency	IgA level
Myasthenia gravis	Dictated by signs and symptoms of disease, antiacetylcholinesterase Ab
Myocarditis	Dictated by signs and symptoms of disease
Pernicious anemia	Anti–parietal cell autoantibodies
	CBC, vitamin B_{12} levels if positive
Serositis	Dictated by signs and symptoms of disease
Stiff man syndrome	Dictated by signs and symptoms of disease
Vitiligo	Physical examination, NALP-1 polymorphism

Abbreviations: Ab, antibody; ACTH, adrenocorticotropic hormone; APS, autoimmune polyendocrine syndrome; CBC, complete blood count; FSH, follicle-stimulating hormone; IL, interleukin; LH, luteinizing hormone; PTH, parathyroid hormone; TSH, thyroid-stimulating hormone.

TREATMENT

APS-1

Therapy of individual disease components is carried out as outlined in other relevant chapters. Replacement of deficient hormones (e.g., adrenal, pancreas, ovaries/testes) will treat most of the endocrinopathies noted. Several unique issues merit special emphasis.

Adrenal insufficiency can be masked by primary hypothyroidism by prolonging the half-life of cortisol. The caveat therefore is that replacement therapy with thyroid hormone can precipitate an adrenal crisis in an undiagnosed individual. Hence, all patients with hypothyroidism and the possibility of APS should be screened for adrenal insufficiency to allow treatment with glucocorticoids prior to the initiation of thyroid hormone replacement. Treatment of mucocutaneous candidiasis with ketoconazole in an individual with subclinical adrenal insufficiency may also precipitate adrenal crisis. Furthermore, mucocutaneous candidiasis may be difficult to eradicate entirely. Severe cases of disease involvement may require systemic immunomodulatory therapy, but this is not commonly needed.

APS-2 (OMIM 269200) is more common than APS-1 with a prevalence of 1–2 in 100,000. It has a gender bias and occurs more often in female patients with a ratio of at least 3:1 compared to male patients. In contrast to APS-1, APS-2 often has its onset in adulthood with a peak incidence between 20 and 60 years of age. It shows a familial, multigenerational heritage (Table 382-2). The presence of two or more of the following endocrine deficiencies in the same patient defines the presence of APS-2: primary adrenal insufficiency (Addison's disease; 50–70%), Graves' disease or autoimmune thyroiditis (15–69%), type 1 diabetes mellitus (T1D; 40–50%), and primary hypogonadism. Frequently associated autoimmune conditions include celiac disease (3–15%), myasthenia gravis, vitiligo, alopecia, serositis, and pernicious anemia. These conditions occur with increased frequency in affected patients but are also are found in their family members (Table 382-3).

Genetic Considerations The overwhelming risk factor for APS-2 has been localized to the genes in the human lymphocyte antigen (HLA) complex on chromosome 6. Primary adrenal insufficiency in APS-2, but not APS-1, is strongly associated with both HLA-DR3 and HLA-DR4. Other class I and class II genes and alleles, such as HLA-B8, HLA-DQ2 and HLA-DQ8, and HLA-DR subtype such as DRB1*04:04, appear to contribute to organ-specific disease susceptibility (Table 382-4). HLA-B8- and HLA-DR3-associated illnesses include selective IgA deficiency, juvenile dermatomyositis, dermatitis herpetiformis, alopecia, scleroderma, autoimmune thrombocytopenia purpura, hypophysitis, metaphyseal osteopenia, and serositis.

Several other immune genes have been proposed to be associated with Addison's disease and therefore with APS-2 (Table 382-3). The "5.1" allele of a major histocompatibility complex (MHC) gene is an atypical class I HLA molecule MIC-A. The MIC-A5.1 allele has a very strong association with Addison's disease that is not accounted for by linkage disequilibrium with DR3 or DR4. Its role is complicated because certain HLA class I genes can offset this effect. PTPN22 codes for a polymorphism in a protein tyrosine phosphatase, which acts on intracellular signaling pathways in both T and B lymphocytes. It has been implicated in T1D, Addison's disease, and other autoimmune conditions. CTLA4 is a receptor on the T cell surface that modulates the activation state of the cell as part of the signal 2 pathway (i.e., binding to CD80/86 on antigen presenting cells). Polymorphisms of this gene appear to cause downregulation of the cell surface expression of the receptor, leading to decreased T cell activation and proliferation. This appears to contribute to Addison's disease and potentially other components of APS-2. Allelic variants of the IL-2Rα are linked to development of T1D and autoimmune thyroid disease and could contribute to the phenotype of APS-2 in certain individuals.

Diagnosis When one of the component disorders is present, a second associated disorder occurs more commonly than in the general population (Table 382-3). There is controversy as to which tests to use and how often to screen individuals for disease. A strong family history of autoimmunity should raise suspicion in an individual with an initial component diagnosis. The development of a rarer form of autoimmunity, such as Addison's disease, should prompt more extensive screening for other linked disorders, as ~50% of Addison's disease patients develop another autoimmune diseases during their lifetime.

Circulating autoantibodies, as previously discussed, can precede the development of clinical disease by many years but would allow the clinician to follow the patient and identify the disease onset at its earliest time point (Tables 382-3 and 382-4). For each of the endocrine components of the disorder, appropriate autoantibody assays are listed and, if positive, should prompt physiologic testing to diagnose clinical or subclinical disease. For Addison's disease, antibodies to 21-hydroxylase antibodies are highly diagnostic for risk of adrenal insufficiency. However, individuals may take many years to develop overt symptoms of hypoadrenalism. Screening of 21-hydroxylase antibody–positive patients can be performed measuring morning ACTH and cortisol on a yearly basis. Rising ACTH values over time or low morning cortisol in association with signs or symptoms of adrenal insufficiency should prompt testing via the cosyntropin stimulation test (Chap. 379). T1D can be screened for by measuring autoantibodies directed against insulin, GAD65, IA-2, and ZnT8. Risk for progression to disease is based on the number of antibodies (≥2 islet autoantibodies with normal glucose tolerance is now defined as stage 1 of T1D as the lifetime risk for developing clinical symptoms is nearly 100%), and metabolic factors (impaired oral glucose tolerance test). National Institutes of Health–sponsored trial groups such as Type 1 Diabetes TrialNet are screening first- and second-degree family members for these autoantibodies and identifying prediabetic individuals who may qualify for intervention trials to change the course of the disease prior to onset. Efforts are now underway to screen the general population for T1D risk with islet autoantibodies.

Screening tests for thyroid disease can include anti–thyroid peroxidase (TPO) or anti-thyroglobulin autoantibodies or anti-TSH receptor antibodies for Graves' disease. Yearly measurements of TSH can then be used to follow these individuals. Celiac disease can be screened for using the anti–tissue transglutaminase (tTg) antibody test. For those <20 years of age, testing every 1–2 years should be performed, whereas less frequent testing is indicated after the age of 20 because the majority of individuals who develop celiac disease have the antibody

TABLE 382-4 APS-2 and Other Polyendocrine Disorder Associations

DISEASE	HLA ASSOCIATION	INITIATING FACTOR	MECHANISM	AUTOANTIGEN
Graves' Disease	DR3	Iodine Anti-CD52	Antibody	TSH receptor
Myasthenia gravis	DR3, DR7	Thymoma Penicillamine	Antibody	Acetylcholine receptor
Anti-insulin receptor	?	SLE or other autoimmune disease	Antibody	Insulin receptor
Hypoparathyroidism	?	?	Antibody	Cell surface inhibitor
Insulin autoimmune syndrome	DR4, DRB1*0406	Methimazole Sulfhydryl-containing drugs	Antibody	Insulin
Celiac disease	DQ2/DQ8	Gluten diet	T cell	Transglutaminase
Type 1 diabetes	DR3/DR4 DQ2/DQ8	? Congenital rubella	T cell	Insulin, GAD65, IA-2, ZnT8, IGRP
Addison's disease	DR3/DR4 DRB1*0404	Unknown	T cell	21-Hydroxylase P450-5cc
Thyroiditis	DR3/DQB1*0201 DQA1*0301	Iodine Interferon α	T cell	Thyroglobulin Thyroid peroxidase
Pernicious anemia	?	?	T cell	Intrinsic factor H+/K+ ATPase
Vitiligo	?	Melanoma Antigen Immunization	?	Melanocyte
Chromosome dysgenesis–trisomy 21 and Turner's syndrome	DQA1*0301	?	?	Thyroid, islet, transglutaminase
Hypophysitis	?	Pit-1, TDRD6	?	Pituitary, Pit-1

Abbreviations: APS, autoimmune polyendocrine syndrome; SLE, systemic lupus erythematosus; TSH, thyroid-stimulating hormone.

earlier in life. Positive tTg antibody test results should be confirmed on repeat testing, followed by small-bowel biopsy to document pathologic changes of celiac disease. Many patients have asymptomatic celiac disease that is nevertheless associated with osteopenia and impaired growth. If left untreated, symptomatic celiac disease has been reported to be associated with an increased risk of gastrointestinal malignancy, especially lymphoma, and osteoporosis later in life.

The knowledge of the particular disease associations should guide other autoantibody or laboratory testing. A complete history and physical examination should be performed every 1–3 years including CBC, metabolic panel, TSH, and vitamin B_{12} levels to screen for most of the possible abnormalities. More specific tests should be based on specific findings from the history and physical examination.

TREATMENT

APS-2

With the exception of Graves' disease, the management of each endocrine component of APS-2 involves hormone replacement and is covered in detail in the chapters on adrenal **(Chap. 379)**, thyroid **(Chap. 375)**, gonadal **(Chaps. 384 and 385)**, and parathyroid diseases **(Chap. 403)**. As noted for APS-1, adrenal insufficiency can be masked by primary hypothyroidism and should be considered and treated as discussed above. In patients with T1D, decreasing insulin requirements or hypoglycemia, without obvious secondary causes, may indicate the emergence of adrenal insufficiency. Hypocalcemia in APS-2 patients is more likely due to malabsorption, potentially from undiagnosed Celiac disease, than hypoparathyroidism.

Immunotherapy for autoimmune endocrine disease has been reserved for T1D, for the most part, reflecting the lifetime burden of the disease for the individual patient and society. Although several immunotherapies (e.g., modified anti-CD3, rituximab, abatacept, alefacept) can prolong the honeymoon phase of T1D, none has achieved long-term success. Active research using new approaches and combination therapy may change the treatment of this disease or other autoimmune conditions that share similar pathways. Furthermore, treatment of subclinical disease diagnosed by the presence of autoantibodies may provide a mechanism to preempt the development of overt disease and is the subject of active basic and clinical research.

IPEX

*I*mmune dysregulation, *p*olyendocrinopathy, *e*nteropathy, and *X*-linked disease (IPEX; OMIM 304790) is a rare X-linked recessive disorder. The disease onset is in infancy and is characterized by severe enteropathy, T1D, and skin disease, as well as variable association with several other autoimmune disorders. Many infants die within the first days of life, but the course is variable, with some children surviving for 12–15 years. Early onset of T1D, often at birth, is highly suggestive of the diagnosis because nearly 80% of IPEX patients develop T1D. Although treatment of the individual disorders can temporarily improve the situation, treatment of the underlying immune deficiency is required and includes immunosuppressive therapy generally followed by hematopoietic stem cell transplantation. Transplantation is the only life-saving form of therapy and can be fully curative by normalizing the imbalanced immune system found in this disorder.

IPEX is caused by mutations in the *FOXP3* gene, which is also mutated in the Scurfy mouse, an animal model that shares much of the phenotype of IPEX patients. The FOXP3 transcription factor is expressed in regulatory T cells designated CD4+CD25+FOXP3+ (Treg). Lack of this factor causes a profound deficiency of this Treg population and results in rampant autoimmunity due to the lack of peripheral tolerance normally provided by these cells. Certain mutations may lead to varying forms of expression of the full syndrome, and there are rare cases where the *FOXP3* gene is intact but other genes involved in this pathway (e.g., CD25, IL-2Rα) may be causative. Future therapy with autologous CD4+ T cells transfected with a functioning *FOXP3* gene may offer a better long-term outcome than has been seen in those treated with stem cell transplantation.

THYMIC TUMORS

Thymomas and thymic hyperplasia are associated with several autoimmune diseases, with the most common being myasthenia gravis (44%) and red cell aplasia (20%). Graves' disease, T1D, and Addison's disease may also be associated with thymic tumors. Patients with myasthenia gravis and thymoma may have unique anti–acetylcholine receptor autoantibodies. Most thymomas lack AIRE expression within the thymoma, and this could be a potential factor in the development of autoimmunity. In support of this concept, thymoma is the one other disease with "frequent" development of anticytokine antibodies and mucocutaneous candidiasis in adults. The majority of tumors are malignant, and temporary remissions of the autoimmune condition can occur with resection of the tumor.

ANTI-INSULIN RECEPTOR ANTIBODIES

This is a very rare disorder where severe insulin resistance (type B) is caused by the presence of anti-insulin receptor antibodies. It is associated with acanthosis nigricans, which can also be associated with other forms of less severe insulin resistance. About one-third of patients have an associated autoimmune illness such as systemic lupus erythematosus or Sjögren's syndrome. Therefore, the presence of antinuclear antibodies, elevated erythrocyte sedimentation rate, hyperglobulinemia, leukopenia, and hypocomplementemia may accompany the presentation. The presence of anti-insulin receptor autoantibodies leads to marked insulin resistance, requiring >100,000 units of insulin to be given daily with only partial control of hyperglycemia. Patients can also have severe hypoglycemia due to partial activation of the insulin receptor by the antibody. The course of the disease is variable, and several patients have had spontaneous remissions. A therapeutic approach that targets B lymphocytes, including rituximab, cyclophosphamide, and pulse steroids has been validated in follow-on case reports to induce remission of the disease.

INSULIN AUTOIMMUNE SYNDROME (HIRATA'S SYNDROME)

The insulin autoimmune syndrome, associated with Graves' disease and methimazole therapy (or other sulfhydryl-containing medications), is of particular interest due to a remarkably strong association with a specific HLA haplotype. Such patients with elevated titers of anti-insulin antibodies frequently present with hypoglycemia. In Japan, the disease is restricted to HLA-DR4-positive individuals with DRB1*04:06. Curiously, a recent report demonstrated that five out of six Caucasian patients taking lipoic acid (sulfhydryl group) who developed insulin autoimmune syndrome were primarily DRB1*04:03 (which is related to DRB1*04:06); the sixth was DRB1*04:06. In Hirata's syndrome the anti-insulin antibodies are often polyclonal. Discontinuation of the medication generally leads to resolution of the syndrome over time. There are very rare cases of insulin autoimmune syndrome not associated with sulfhydryl-containing medications that result in profound, life-threatening hypoglycemia. Treatment involves treating the underlying condition that causes anti-insulin antibodies, such as a B lymphocyte lymphoma (tend to have monoclonal insulin antibodies) or systemic lupus erythematosus. As hypoglycemia is profound when elevated titers of high affinity insulin antibodies bind secreted insulin and then release it into circulation, treatment begins with high dose glucocorticoids and potentially rituximab to target B lymphocytes.

POEMS SYNDROME

POEMS (*p*olyneuropathy, *o*rganomegaly, *e*ndocrinopathy, *M*-protein, and *s*kin changes; also known as Crow-Fukase syndrome; OMIM 192240) patients usually present with a progressive sensorimotor polyneuropathy, diabetes mellitus (50%), primary gonadal failure (70%), and a plasma cell dyscrasia with sclerotic bony lesions. Associated findings can be hepatosplenomegaly, lymphadenopathy, and hyperpigmentation. Patients often present in the fifth to sixth decade of life and have a median survival after diagnosis of <3 years. The syndrome is

assumed to be secondary to circulating immunoglobulins, but patients have excess vascular endothelial growth factor as well as elevated levels of other inflammatory cytokines such as IL1-β, IL-6, and tumor necrosis factor α. Patients have been treated with thalidomide, and more recently lenalidomide, leading to a decrease in vascular endothelial growth factor. Hyperglycemia responds to small, subcutaneous doses of insulin. The hypogonadism is due to primary gonadal disease with elevated plasma levels of follicle-stimulating hormone and luteinizing hormone. Temporary resolution of the features of POEMS, including normalization of blood glucose, may occur after radiotherapy for localized plasma cell lesions of bone or after chemotherapy, lenalidomide and dexamethasone, or autologous stem cell transplantation.

OTHER DISORDERS

Other diseases can exhibit polyendocrine deficiencies, including Kearns-Sayre syndrome, DIDMOAD syndrome (*d*iabetes *i*nsipidus, *d*iabetes *m*ellitus, progressive bilateral *o*ptic *a*trophy, and sensorineural *d*eafness; also termed Wolfram's syndrome), Down's syndrome or trisomy 21 (OMIM 190685), Turner's syndrome (monosomy X, 45,X0), and congenital rubella.

Kearns-Sayre syndrome (OMIM 530000) is a rare mitochondrial DNA disorder characterized by myopathic abnormalities leading to ophthalmoplegia and progressive weakness in association with several endocrine abnormalities, including hypoparathyroidism, primary gonadal failure, diabetes mellitus, and hypopituitarism. Crystalline mitochondrial inclusions are found in muscle biopsy specimens, and such inclusions have also been observed in the cerebellum. Antiparathyroid antibodies have not been described; however, antibodies to the anterior pituitary gland and striated muscle have been identified, and the disease may have autoimmune components. These mitochondrial DNA mutations occur sporadically and do not appear to be associated with a familial syndrome.

Wolfram's syndrome (OMIM 222300, chromosome 4; OMIM 598500, mitochondrial) is a rare autosomal recessive disease that is also called DIDMOAD. Neurologic and psychiatric disturbances are prominent in most patients and can cause severe disability. The disease is caused by defects in *Wolfram Syndrome 1* (WFS1) gene, which encodes a 100-kDa transmembrane protein that has been localized to the endoplasmic reticulum and is found in neuronal and neuroendocrine tissue. Its expression induces ion channel activity with a resultant increase in intracellular calcium and may play an important role in intracellular calcium homeostasis. Wolfram's syndrome appears to be a slowly progressive neurodegenerative process, and there is nonautoimmune selective destruction of the pancreatic beta cells. Diabetes mellitus with an onset in childhood is usually the first manifestation. Diabetes mellitus and optic atrophy are present in all reported cases, but expression of the other features is variable. Treatments targeting endoplasmic reticulum dysfunction are being tested and may be a bridge till gene therapy can be developed to treat the most severely affected cases.

Down's syndrome, or trisomy 21 (OMIM 190685), is associated with the development of T1D, thyroiditis, and celiac disease. Patients with Turner's syndrome also appear to be at increased risk for the development of thyroid disease and celiac disease. It is recommended to screen patients with trisomy 21 and Turner's syndrome for associated autoimmune diseases on a regular basis.

GLOBAL CONSIDERATIONS

Identification of these syndromes requires access to central laboratories with the ability to detect unique autoantibodies and to sequence the specific genes that may underlie these disorders. Early recognition of the clinical features of these disorders and timely referral and/or consultation with tertiary care centers to confirm the diagnosis and initiate therapy is important to improving outcomes. The *AIRE* recessive gene mutations found in APS-1 were originally described in high frequency in several populations including Finns, Iranian Jews, Sardinians, Norwegians, and Irish. Although individuals from many other countries have now been found to have these mutations, and the newly identified dominant *AIRE* gene mutations, understanding the frequency in the background population may raise the clinician's level of suspicion for these rare disorders. Hirata syndrome was originally reported in Japanese populations, but also may be found in other populations as noted.

FURTHER READING

Anderson MS, Su MA: AIRE expands: New roles in immune tolerance and beyond. Nat Rev Immunol 16:247, 2016.

Bruserud O et al: A longitudinal follow-up of autoimmune polyendocrine syndrome type 1. J Clin Endocrinol Metab 101:2975, 2016.

Oftedal BE et al: Dominant mutations in the autoimmune regulator *AIRE* are associated with common organ-specific autoimmune diseases. Immunity 42:1185, 2015.

Section 2 Sex- and Gender-Based Medicine

383 Disorders of Sex Development

John C. Achermann, J. Larry Jameson

Sex development begins in utero but continues into young adulthood with the achievement of sexual maturity and reproductive capability. The major determinants of sex development can be divided into three components: chromosomal sex, gonadal sex (sex determination), and phenotypic sex (sex differentiation) **(Fig. 383-1)**. Variations at each of these stages can result in disorders (or differences) of sex development (DSDs) **(Table 383-1)**. In the newborn period, ~1 in 4000 babies undergo investigation because of ambiguous (atypical) genitalia. Urgent assessment is indicated, because some causes such as congenital adrenal hyperplasia (CAH) can be associated with life-threatening adrenal crises. Support for the parents and clear communication about the diagnosis and management options are essential. The involvement of an experienced multidisciplinary team is important for counseling, planning appropriate investigations, and discussing long-term well-being. DSDs can also present at other ages and to a range of health professionals. Subtler forms of gonadal dysfunction (e.g., Klinefelter's syndrome [KS], Turner's syndrome [TS]) often are diagnosed later in life by internists. Because these conditions are associated with a variety of psychological, reproductive, and potential medical consequences, an open dialogue must be established between the patient and health care providers to ensure continuity and attention to these issues across the

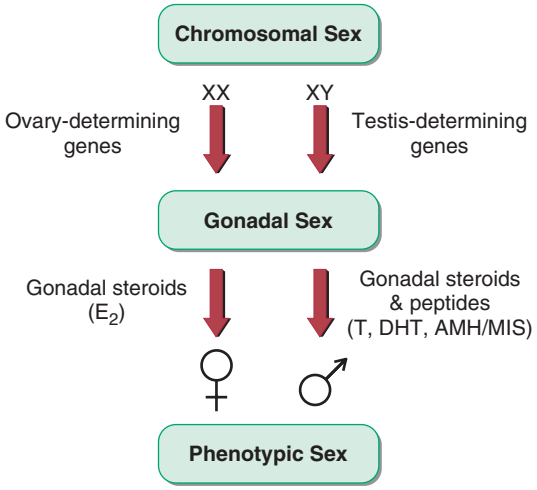


FIGURE 383-1 Sex development can be divided into three major components: chromosomal sex, gonadal sex, and phenotypic sex. DHT, dihydrotestosterone; MIS, müllerian-inhibiting substance also known as anti-müllerian hormone, AMH; T, testosterone.

TABLE 383-1 Classification of Disorders of Sex Development (DSDs)

SEX CHROMOSOME DSD	46,XY DSD (SEE TABLE 383-3)	46,XX DSD (SEE TABLE 383-4)
47,XXY (Klinefelter's syndrome and variants)	**Disorders of gonadal (testis) development**	**Disorders of gonadal (ovary) development**
45,X (Turner's syndrome and variants)	Complete or partial gonadal dysgenesis (e.g., *SRY, SOX9, SF1, WT1, DHH, GATA4/ZFPM2, MAP3K1*)	Gonadal dysgenesis
45,X/46,XY mosaicism (mixed gonadal dysgenesis)	Impaired fetal Leydig cell function (e.g., *SF1/NR5A1, CXorf6/MAMLD1, HHAT, SAMD9*)	Ovotesticular DSD
46,XX/46,XY (chimerism/mosaicism)	Ovotesticular DSD	Testicular DSD (e.g., *SRY*+, dup *SOX9, RSPO1, NR5A1*)
	Testis regression	**Androgen excess**
	Disorders in androgen synthesis or action	Fetal
	Disorders of androgen biosynthesis	3β-Hydroxysteroid dehydrogenase II (*HSD3β2*)
	LH receptor (*LHCGR*)	21-Hydroxylase (*CYP21A2*)
	Smith-Lemli-Opitz syndrome	P450 oxidoreductase (*POR*)
	Steroidogenic acute regulatory (*StAR*) protein	11β-Hydroxylase (*CYP11B1*)
	Cholesterol side-chain cleavage (*CYP11A1*)	Glucocorticoid receptor mutations
	3β-Hydroxysteroid dehydrogenase II (*HSD3B2*)	Fetoplacental
	17α-Hydroxylase/17,20-lyase (*CYP17A1*)	Aromatase deficiency (*CYP19*)
	P450 oxidoreductase (*POR*)	Oxidoreductase deficiency (*POR*)
	Cytochrome b5 (*CYB5A*)	Maternal
	17β-Hydroxysteroid dehydrogenase III (*HSD17B3*)	Maternal virilizing tumors (e.g., luteomas)
	5α-Reductase II (*SRD5A2*)	Androgenic drugs
	Aldo-keto reductase 1C2 (*AKR1C2*)	**Other**
	Disorders of androgen action	Syndromic associations (e.g., cloacal anomalies)
	Androgen insensitivity syndrome	Müllerian agenesis/hypoplasia (e.g., MRKH)
	Drugs and environmental modulators	Uterine abnormalities (e.g., MODY5)
	Other	Vaginal atresia (e.g., McKusick-Kaufman)
	Syndromic associations of male genital development	Labial adhesions
	Persistent müllerian duct syndrome	
	Vanishing testis syndrome	
	Isolated hypospadias	
	Congenital hypogonadotropic hypogonadism	
	Cryptorchidism	
	Environmental influences	

Abbreviations: MODY, maturity-onset diabetes of the young; MRKH, Mayer-Rokitansky-Kuster-Hauser syndrome.

Source: Modified from IA Hughes: Arch Dis Child 91:554, 2006.

life span **(Table 383-2)**. Support groups also have an important role to play for many of these conditions.

SEX DEVELOPMENT

Chromosomal sex, defined by a karyotype, describes the X and/or Y chromosome complement (46,XY; 46,XX) that is established at the time of fertilization. The presence of a normal Y chromosome determines that testis development will occur even in the presence of multiple X chromosomes (e.g., 47,XXY or 48,XXXY). The loss of an X chromosome impairs gonad development (45,X or 45,X/46,XY mosaicism). Fetuses with no X chromosome (45,Y) are not viable.

Gonadal sex refers to the histologic and functional characteristics of gonadal tissue as testis or ovary. The embryonic gonad is bipotential and can develop (from ~42 days after conception) into either a testis or an ovary, depending on which genes are expressed **(Fig. 383-2)**. Testis development is initiated by expression of the Y chromosome gene *SRY* (sex-determining region on the Y chromosome) that encodes an HMG box transcription factor. *SRY* is expressed transiently in cells destined to become Sertoli cells and serves as a pivotal switch to establish the testis lineage. Mutation of *SRY* prevents testis development in 46,XY individuals, whereas translocation of *SRY* in 46,XX individuals is sufficient to induce testis development and a male phenotype. Other

TABLE 383-2 Presentation of DSD at Different Stages of Life

PRESENTATION	FEATURES	PROFESSIONAL	EXAMPLES
Prenatal	Karyotype-phenotype discordance	Obstetrician; fetal medicine	Many
Neonatal	Atypical genitalia	Obstetrician; neonatal medicine	Many
	Salt losing crisis	Pediatrician	CAH (CYP21)
Childhood	Hernia	Surgeon	CAIS
	Androgenization	Endocrinologist	CAH (CYP21, CYP11B1)
	Poor growth	Pediatrician	Turner's, 45,X/46,XY
	Associated features	Oncologist/nephrologist	Wilms' tumor
Puberty	Androgenization	Endocrinologist	17β-HSD, 5α-reductase, SF1
	Absent puberty	Endocrinologist	Gonadal dysgenesis, CAH (CYP17), Turner's
Post-puberty	Amenorrhea	Gynecologist	CAIS
Adult	Infertility	Andrologist	Klinefelter's, 45,X/46,XY, SF1

Abbreviations: CAH, congenital adrenal hyperplasia; CAIS, complete androgen insensitivity syndrome; 17β-HSD, 17β-hydroxysteroid dehydrogenase deficiency; SF1, steroidogenic factor 1 (NR5A1).

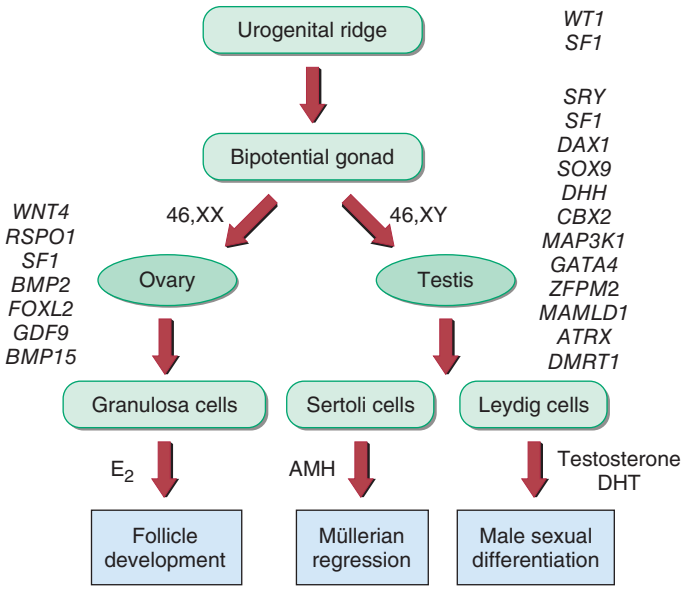


FIGURE 383-2 The genetic regulation of gonadal development. AMH, anti-müllerian hormone (müllerian-inhibiting substance); *ATRX*, α-thalassemia, mental retardation on the X; BMP2 and 15, bone morphogenic factors 2 and 15; *CBX2*, chromobox homologue 2; *DAX1*, dosage sensitive sex-reversal, adrenal hypoplasia congenita on the X chromosome, gene 1 (also known as NR0B1); *DHH*, desert hedgehog; DHT, dihydrotestosterone; *DMRT 1,2*, doublesex MAB3-related transcription factor 1,2; *FOXL2*, forkhead transcription factor L2; *GATA4*, GATA binding protein 4; GDF9, growth differentiation factor 9; *MAMLD1*, mastermind-like domain containing 1; *MAP3K1*, mitogen-activated protein kinase kinase kinase 1; *RSPO1*, R-spondin 1; *SF1*, steroidogenic factor 1 (also known as NR5A1); *SOX9*, *SRY*-related HMG-box gene 9; *SRY*, sex-determining region on the Y chromosome; *WNT4*, wingless-type MMTV integration site 4; *WT1*, Wilms' tumor–related gene 1; ZFPM2, zinc finger protein, multitype 2.

genes are necessary to continue testis development. *SOX9* (*SRY*-related HMG-box gene 9) is upregulated by *SRY* in the developing testis but is suppressed in the ovary. *WT1* (Wilms' tumor–related gene 1) acts early in the genetic pathway and regulates the transcription of several genes, including *SF1* (*NR5A1*), *DAX1* (*NR0B1*), and *AMH* (encoding müllerian-inhibiting substance [MIS]). *SF1* encodes steroidogenic factor 1, a nuclear receptor that functions in cooperation with other transcription factors to regulate a large array of adrenal and gonadal genes, including *SOX9* and many genes involved in steroidogenesis. *SF1* mutations causing loss of function are found in ~10% of XY patients with gonadal dysgenesis and impaired androgenization. In contrast, duplication of a related gene *DAX1* also impairs testis development, revealing the exquisite sensitivity of the testis-determining pathway to gene dosage effects. In addition to the genes mentioned above, at least 30 other genes are also involved in gonad development (Fig. 383-2). These genes encode an array of signaling molecules and paracrine growth factors in addition to transcription factors.

Although ovarian development once was considered a less active process, it is now clear that specific genes are expressed during the earliest stages of ovary development. Some of these factors may repress testis development (e.g., WNT4, R-spondin-1) (Fig. 383-2). Once the ovary has formed, additional factors are required for normal follicular development (e.g., follicle-stimulating hormone [FSH] receptor, *GDF9*). Steroidogenesis in the ovary requires the development of follicles that contain granulosa cells and theca cells surrounding the oocytes **(Chap. 385)**. Thus, there is relatively limited ovarian steroidogenesis until puberty.

Germ cells also develop in a sex dimorphic manner. In the developing ovary, primordial germ cells (PGCs) proliferate and enter meiosis, whereas they proliferate and then undergo mitotic arrest in the developing testis. PGC entry into meiosis is initiated by retinoic acid. The developing testis produces high levels of CYP26B1, an enzyme that degrades retinoic acid, preventing PGC entry into meiosis. Approximately 7 million germ cells are present in the fetal ovary in the second trimester, and 1 million remain at birth. Only 400 are ovulated during a woman's reproductive life span **(Chap. 385)**.

Phenotypic sex refers to the structures of the external and internal genitalia and secondary sex characteristics. The developing testis releases anti-müllerian hormone (AMH; also known as müllerian-inhibiting substance [MIS]) from Sertoli cells and testosterone from Leydig cells. AMH acts through specific receptors to cause regression of the müllerian structures from 60–80 days after conception. At ~60–140 days after conception, testosterone supports the development of wolffian structures, including the epididymides, vasa deferentia, and seminal vesicles. Testosterone is the precursor for dihydrotestosterone (DHT), a potent androgen that promotes development of the external genitalia, including the penis and scrotum (60–100 days, and thereafter) **(Fig. 383-3)**. The urogenital sinus develops into the prostate and prostatic urethra in the male and into the urethra and lower portion of the vagina in the female. The genital tubercle becomes the glans penis in the male and the clitoris in the female. The urogenital swellings form the scrotum or the labia majora, and the urethral folds fuse to form the shaft of the penis and the male urethra or the labia minora. In the female, wolffian ducts regress and the müllerian ducts form the fallopian tubes, uterus, and upper segment of the vagina. A female phenotype will develop in the absence of the gonad, but estrogen is needed for maturation of the uterus and breast at puberty.

The prenatal hormone environment likely influences aspects of gender identity and behavior. This is an area of ongoing research and is beyond the scope of this chapter.

DISORDERS OF CHROMOSOMAL SEX

Variations in sex chromosome number and structure can present as DSDs (e.g., 45,X/46,XY). KS (47,XXY) and TS (45,X) do not usually present with genital ambiguity but are associated with gonadal dysfunction **(Table 383-3)**.

KLINEFELTER'S SYNDROME (47,XXY)

Pathophysiology The classic form of KS (47,XXY) occurs after meiotic nondisjunction of the sex chromosomes during gametogenesis (40% during spermatogenesis, 60% during oogenesis). Mosaic forms of KS (46,XY/47,XXY) result from chromosomal mitotic nondisjunction within the zygote and occur in at least 10% of individuals with this condition. Other chromosomal variants of KS (e.g., 48,XXYY, 48,XXXY) are less common.

Clinical Features KS is characterized by small testes, infertility, gynecomastia, tall stature/increased leg length, and hypogonadism in phenotypic males. It has an incidence of at least 1 in 1000 men, but ~75% of cases are not diagnosed. Of those who are diagnosed, only 10% are identified prepubertally, usually because of small genitalia or cryptorchidism. Others are diagnosed after puberty, usually based on low androgens and/or gynecomastia. Developmental delay, speech difficulties, and poor motor skills may be features but are variable, especially in adolescence. Later in life, body habitus or infertility leads to the diagnosis. Testes are small and firm (median length 2.5 cm [4 mL volume]; almost always <3.5 cm [12 mL]) and typically seem inappropriately small for the degree of androgenization. Biopsies are not usually necessary but typically reveal seminiferous tubule hyalinization and azoospermia. Other clinical features of KS are listed in Table 383-3. Plasma concentrations of FSH and luteinizing hormone (LH) are increased in most adults with 47,XXY, and plasma testosterone is decreased (50–75%), reflecting primary gonadal insufficiency. Estradiol is often increased resulting in gynecomastia **(Chap. 384)**. Patients with mosaic forms of KS have less severe clinical features, have larger testes, and sometimes achieve spontaneous fertility.

TREATMENT

Klinefelter's Syndrome

Growth, endocrine function, and bone mineralization should be monitored, especially from adolescence. Educational and psychological support is important for many individuals with KS. Androgen supplementation improves virilization, libido, energy,

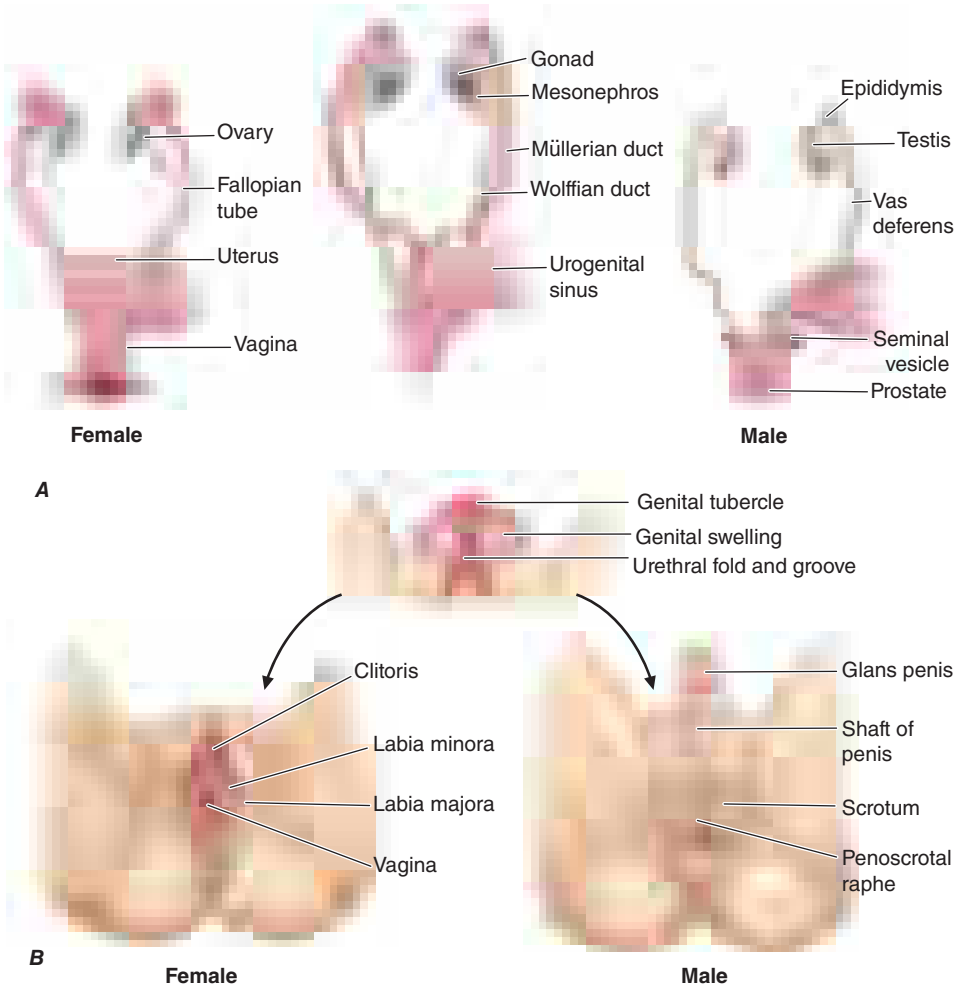


FIGURE 383-3 Sex development. *A.* Internal urogenital tract. ***B.*** External genitalia. *(After E Braunwald et al [eds]: Harrison's Principles of Internal Medicine, 15th ed. New York, McGraw-Hill, 2001.)*

TABLE 383-3 Clinical Features of Chromosomal Disorders of Sex Development (DSDs)

			GENITALIA		
DISORDER	**COMMON CHROMOSOMAL COMPLEMENT**	**GONAD**	**EXTERNAL**	**INTERNAL**	**BREAST DEVELOPMENT**
Klinefelter's syndrome	47,XXY or 46,XY/47,XXY	Hyalinized testes	Male	Male	Gynecomastia
	Clinical Features				
	Small testes, azoospermia, decreased facial and axillary hair, decreased libido, tall stature and increased leg length, decreased penile length, increased risk of breast tumors, thromboembolic disease, learning difficulties, speech delay and decreased verbal IQ, obesity, diabetes mellitus, metabolic syndrome, varicose veins, hypothyroidism, systemic lupus erythematosus, epilepsy			
Turner's syndrome	45,X or 45,X/46,XX	Streak gonad or immature ovary	Female	Hypoplastic female	Immature female
	Clinical Features				
	Infancy: lymphedema, web neck, shield chest, low-set hairline, cardiac defects and coarctation of the aorta, urinary tract malformations, and horseshoe kidney				
	Childhood: short stature, cubitus valgus, short neck, short fourth metacarpals, hypoplastic nails, micrognathia, scoliosis, otitis media and sensorineural hearing loss, ptosis and amblyopia, multiple nevi and keloid formation, autoimmune thyroid disease, visuospatial learning difficulties				
	Adulthood: pubertal failure and primary amenorrhea, hypertension, obesity, dyslipidemia, impaired glucose tolerance and insulin resistance, autoimmune thyroid disease, cardiovascular disease, aortic root dilation, osteoporosis, inflammatory bowel disease, chronic hepatic dysfunction, increased risk of colon cancer, hearing loss				
45,X/46,XY mosaicism	45,X/46,XY	Testis or streak gonad	Variable	Variable	Usually male
	Clinical Features				
	Short stature, increased risk of gonadal tumors, some Turner's syndrome features				
Ovotesticular DSD (true hermaphroditism)	46,XX/46,XY	Testis and ovary or ovotestis	Variable	Variable	Gynecomastia
	Clinical Features				
	Possible increased risk of gonadal tumors				

hypofibrinolysis, and bone mineralization in men with low testosterone levels but may occasionally worsen gynecomastia (Chap. 384). Gynecomastia can be treated by surgical reduction if it causes concern (Chap. 384). Fertility has been achieved by using in vitro fertilization in men with oligospermia or with intracytoplasmic sperm injection (ICSI) after retrieval of spermatozoa by testicular sperm extraction techniques. In specialized centers, successful spermatozoa retrieval using this technique is possible in >50% of men with nonmosaic KS. Results may be better in younger men. After ICSI and embryo transfer, successful pregnancies can be achieved in ~50% of these cases. The risk of transmitting chromosomal anomalies needs to be considered, although this outcome is much less common than originally predicted, and the role of preimplantation screening is debated. Long-term monitoring of men with KS is important given the increased risk of breast cancer, cardiovascular disease, metabolic syndrome, and autoimmune disorders. Because most men with KS are never diagnosed, it is important that all internists consider this diagnosis in men with these features who might be seeking medical advice for other conditions.

TURNER'S SYNDROME (GONADAL DYSGENESIS; 45,X)

Pathophysiology Approximately one-half of women with TS have a 45,X karyotype, about 20% have 45,X/46,XX mosaicism, and the remainder have structural abnormalities of the X chromosome such as X fragments, isochromosomes, or rings. The clinical features of TS result from haploinsufficiency of multiple X chromosomal genes (e.g., short stature homeobox, *SHOX*). However, imprinted genes also may be affected when the inherited X has different parental origins.

Clinical Features TS is characterized by bilateral streak gonads, primary amenorrhea, short stature, and other phenotypic features (Table 383-3). It affects ~1 in 2500 women and is diagnosed at different ages. Prenatally, a diagnosis of TS usually is made incidentally after chorionic villus sampling or amniocentesis for unrelated reasons such as advanced maternal age. Prenatal ultrasound findings include increased nuchal translucency. The postnatal diagnosis of TS should be considered in female neonates or infants with lymphedema, nuchal folds, low hairline, or left-sided cardiac defects and in girls with unexplained growth failure or pubertal delay. Although limited spontaneous pubertal development occurs in up to 30% of girls with TS (10%, 45,X; 30–40%, 45,X/46,XX) and ~2% have menarche, the vast majority of women with TS develop complete ovarian insufficiency. Therefore, this diagnosis should be considered in all women who present with primary or secondary amenorrhea and elevated gonadotropin levels.

TREATMENT

Turner's Syndrome

The management of girls and women with TS requires a multidisciplinary approach because many different organ systems can be affected. Detailed cardiac and renal evaluation should be performed at the time of diagnosis. Individuals with congenital heart defects (CHDs) (30%) (bicuspid aortic valve, 30–50%; coarctation of the aorta, 30%; aortic root dilation, 5%) require long-term follow-up by an experienced cardiologist, antibiotic prophylaxis for dental or surgical procedures, and serial magnetic resonance imaging (MRI) of aortic root dimensions, because progressive aortic root dilation is associated with increased risk of aortic dissection. Individuals found to have congenital renal and urinary tract malformations (30%) are at risk for urinary tract infections, hypertension, and nephrocalcinosis. Hypertension can occur independently of cardiac and renal malformations and should be monitored and treated as in other patients with essential hypertension. Clitoral enlargement or other evidence of virilization suggests the presence of covert Y chromosomal material and is associated with increased risk of gonadoblastoma. Regular assessment of thyroid function, weight, dentition, hearing, speech, vision, and educational issues should be performed during childhood. Otitis media and middle-ear disease are prevalent in childhood (50–85%), and sensorineural hearing loss becomes progressively common with age (70–90%). Autoimmune hypothyroidism (15–30%) can occur in childhood but has a mean age of onset in the third decade. Counseling about long-term growth and fertility issues should be provided. Patient support groups are active throughout the world and can play an invaluable role.

Short stature can be an issue for some girls because untreated final height rarely exceeds 150 cm in nonmosaic 45,X TS. Recombinant growth hormone stimulates growth rate in children with TS and is occasionally combined with low doses of the nonaromatizable anabolic steroid oxandrolone (up to 0.05 mg/kg per day) in an older child (>9 years). However, final height increments are often about 5–10 cm, and individualization of treatment response to regimens may be beneficial. Girls with evidence of ovarian insufficiency require estrogen replacement to induce breast and uterine development, support growth, and maintain bone mineralization. Most physicians now initiate low-dose estrogen therapy (one-tenth to one-eighth of the adult replacement dose) to induce puberty at an age-appropriate time (~11 years). Doses of estrogen are increased gradually to allow development over a 2- to 4-year period. Progestins are added later to regulate withdrawal bleeds. Some women with TS have achieved successful pregnancy after ovum donation and in vitro fertilization but the risks of cardiac complications are high, and expert counseling and management are needed. Long-term follow-up of women with TS involves careful surveillance of sex hormone replacement and reproductive function, bone mineralization, cardiac function and aortic root dimensions, blood pressure, weight and glucose tolerance, hepatic and lipid profiles, thyroid function, and hearing. This service is provided by a dedicated TS clinic in some centers.

45,X/46,XY MOSAICISM (MIXED GONADAL DYSGENESIS)

The phenotype of individuals with 45,X/46,XY mosaicism (sometimes called *mixed gonadal dysgenesis*) can vary considerably. Some have a predominantly female phenotype with somatic features of TS, streak gonads, and müllerian structures, and are managed as TS with a Y chromosome. Most 45,X/46,XY individuals have a male phenotype and testes, and the diagnosis is made incidentally after amniocentesis or during investigation of infertility. In practice, most newborns referred for assessment have atypical genitalia and variable somatic features. Management is complex and needs to be individualized. A female sex-of-rearing is often assigned if uterine structures are present, gonads are intraabdominal, and the phallus is very small. In such situations, gonadectomy usually is considered to prevent further androgen secretion at puberty and prevent risk of gonadoblastoma (up to 25%). Individuals raised as males usually have reconstructive surgery for hypospadias and removal of dysgenetic or streak gonads if the gonads cannot be brought down into the scrotum. Scrotal testes can be preserved but require regular examination for tumor development and sonography at the time of puberty. Biopsy for carcinoma in situ is recommended in adolescence, and testosterone supplementation may be required to support androgenization in puberty or if low testosterone is detected in adulthood. Height potential is usually reduced; some children receive recombinant growth hormone using TS protocols. Screening for cardiac, renal, and other TS features should be considered, and psychological support offered for the family and young person.

OVOTESTICULAR DSD

Ovotesticular DSD (formerly called *true hermaphroditism*) occurs when both an ovary and a testis—or when an ovotestis—are found in one individual. Most individuals with this diagnosis have a 46,XX karyotype, especially in sub-Saharan Africa, and present with ambiguous genitalia at birth or with breast development and phallic development at puberty. A 46,XX/46,XY chimeric karyotype is less common and has a variable phenotype.

DISORDERS OF GONADAL AND PHENOTYPIC SEX

Disorders of gonadal and phenotypic sex can result in reduced androgen production or action in individuals with a 46,XY karyotype (46,XY DSD), or excess androgen production in individuals with a 46,XX karyotype (46,XX DSD) (Table 383-1). These conditions cover a spectrum of phenotypes ranging from phenotypic females with a Y-chromosome to phenotypic males with a 46,XX karyotype to individuals with atypical genitalia. Karyotype is a useful starting investigation for diagnosis, but does not define an individual's gender.

46,XY DSD

Underandrogenization of the 46,XY fetus (formerly called *male pseudohermaphroditism*) reflects defects in androgen production or action. It can result from disorders of testis development, defects of androgen synthesis, or resistance to testosterone and DHT (Table 383-1).

Disorders of Testis Development • **TESTICULAR DYSGENESIS** *Pure* (or *complete*) *gonadal dysgenesis* (*Swyer's syndrome*) is associated with streak gonads, müllerian structures (due to insufficient AMH/MIS secretion), and a complete absence of androgenization. Phenotypic females with this condition often present because of absent pubertal development and are found to have a 46,XY karyotype. Serum sex steroids, AMH/MIS, and inhibin B are low, and LH and FSH are elevated. Patients with *partial gonadal dysgenesis* (*dysgenetic testes*) may produce enough MIS to regress the uterus and sufficient testosterone for partial androgenization, and therefore usually present in the newborn period with atypical genitalia. Gonadal dysgenesis can result from mutations or deletions of testis-promoting genes (*WT1*, *CBX2*, *SF1*, *SRY*, *SOX9*, *MAP3K1*, *DHH*, *GATA4/ZFPM2*, *ATRX*, *ARX*, *DMRT*) or duplication of chromosomal loci containing "antitestis" genes (e.g., *WNT4/RSPO1*, *DAX1*) **(Table 383-4)**. Among these, deletions or mutations of *SRY* and heterozygous mutations of *SF1* (*NR5A1*) appear to be most common but still account collectively for <25% of cases. Associated clinical features may be present, reflecting additional functional roles for these genes. For example, renal dysfunction occurs in patients with specific *WT1* mutations (Denys-Drash and Frasier's syndromes), primary adrenal failure occurs in some patients with *SF1* mutations, and severe cartilage abnormalities (campomelic dysplasia) are the predominant clinical feature of *SOX9* mutations. A family history of DSD, infertility, or early menopause is important because mutations in *SF1/NR5A1* can be inherited from a mother in a sex-limited dominant manner (which can mimic X-linked inheritance). In some situations, a woman may later develop primary ovarian insufficiency because of the effect of *SF1* on the ovary. Intraabdominal dysgenetic testes should be removed to prevent malignancy, and estrogens can be used to induce secondary sex characteristics and uterine development in 46,XY individuals raised as females, if person feels that female gender is appropriate for them. *Absent (vanishing) testis syndrome* (*bilateral anorchia*) reflects regression of the testis during development. The etiology is unknown, but the absence of müllerian structures indicates adequate secretion of AMH early in utero. Usually, androgenization of the external genitalia is normal. These individuals can be offered testicular prostheses and should receive androgen replacement in adolescence.

Disorders of Androgen Synthesis Defects in the pathway that regulates androgen synthesis **(Fig. 383-4)** cause underandrogenization of the 46,XY fetus (Table 383-1). Müllerian regression is unaffected because Sertoli cell function is preserved. These conditions can present with a spectrum of genital appearances, ranging from female-typical external genitalia or clitoromegaly in some individuals to penoscrotal hypospadias or a small phallus in others.

LH RECEPTOR Mutations in the LH receptor (LHCGR) cause Leydig cell hypoplasia and androgen deficiency, due to impaired actions of human chorionic gonadotropin in utero and LH late in gestation and during the neonatal period. As a result, testosterone and DHT synthesis are reduced.

STEROIDOGENIC ENZYME PATHWAYS Mutations in *steroidogenic acute regulatory protein* (*StAR*) and *CYP11A1* affect both adrenal and gonadal steroidogenesis (Fig. 383-4) **(Chap. 379)**. Affected individuals (46,XY) usually have severe early-onset salt-losing adrenal failure and a female phenotype, although later-onset milder variants have been reported. Defects in *3β-hydroxysteroid dehydrogenase type 2* (*HSD3β2*) also cause adrenal insufficiency in severe cases, but the accumulation of dehydroepiandrosterone (DHEA) has a mild androgenizing effect, resulting in ambiguous genitalia or hypospadias. Salt loss occurs in many but not all children. Patients with CAH due to *17α-hydroxylase (CYP17) deficiency* have variable underandrogenization and develop hypertension and hypokalemia due to the potent salt-retaining effects of corticosterone and 11-deoxycorticosterone. Patients with complete loss of 17α-hydroxylase function often present as phenotypic females who do not enter puberty and are found to have inguinal testes and hypertension in adolescence. Some mutations in *CYP17* selectively impair 17,20-lyase activity without altering 17α-hydroxylase activity, leading to underandrogenization without mineralocorticoid excess and hypertension. Disruption of the coenzyme, *cytochrome b5* (*CYB5A*), can present similarly, and methemoglobinemia is usually present. Mutations in *P450 oxidoreductase* (*POR*) affect multiple steroidogenic enzymes, leading to reduced androgen production and a biochemical pattern of apparent combined 21-hydroxylase and 17α-hydroxylase deficiency, sometimes with skeletal abnormalities (Antley-Bixler craniosynostosis). Defects in *17β-hydroxysteroid dehydrogenase type 3* (*HSD17β3*) and *5α-reductase type 2* (*SRD5A2*) interfere with the synthesis of testosterone and DHT, respectively. These conditions are characterized by minimal or absent androgenization in utero, but some phallic development can occur during adolescence due to the action of other enzyme isoforms. Individuals with 5α-reductase type 2 deficiency have normal wolffian structures and usually do not develop breast tissue. At puberty, the increase in testosterone induces muscle mass and other virilizing features despite DHT deficiency. Some individuals change gender from female to male at puberty. Thus, the management of this disorder requires expert support. DHT cream can improve prepubertal phallic growth in patients raised as male. Gonadectomy before adolescence and estrogen replacement at puberty can be considered in individuals raised as females who feel they have a female gender identity. Disruption of alternative pathways to fetal DHT production might also present with 46,XY DSD (*AKR1C2/AKR1C4*).

Disorders of Androgen Action • **ANDROGEN INSENSITIVITY SYNDROME** Mutations in the androgen receptor cause resistance to androgen (testosterone, DHT) action or the *androgen insensitivity syndrome* (*AIS*). AIS is a spectrum of disorders that affects at least 1 in 100,000 46,XY individuals. Because the androgen receptor is X-linked, only 46,XY offspring are affected if the mother is a carrier of a mutation. XY individuals with *complete AIS* (formerly called *testicular feminization syndrome*) have a female phenotype, normal breast development (due to aromatization of testosterone), a short vagina but no uterus (because MIS production is normal), scanty pubic and axillary hair, and a female gender identity and sex role behavior. Gonadotropins and testosterone levels can be low, normal, or elevated, depending on the degree of androgen resistance and the contribution of estradiol to feedback inhibition of the hypothalamic-pituitary-gonadal axis. AMH/MIS levels in childhood are normal or high. CAIS sometimes presents as inguinal hernias (containing testes) in childhood or more often with primary amenorrhea in late adolescence. Because there is a low risk of malignancy, gonadectomy has been recommended for girls diagnosed in childhood, with estrogen replacement prescribed at puberty. Alternatively, the gonads can be left in situ until breast development is complete and then removed to avoid tumor risk. Some adults with complete AIS decline gonadectomy, but should be counseled about the risk of malignancy, especially because early detection of premalignant changes by imaging or biomarkers is currently not possible. The use of graded dilators in adolescence is usually sufficient to dilate the vagina for sexual intercourse.

Partial AIS (Reifenstein's syndrome) results from androgen receptor mutations that maintain residual function. Patients often present in infancy with penoscrotal hypospadias and undescended testes and with gynecomastia at the time of puberty. Those individuals raised as

TABLE 383-4 Selected Genetic Causes of 46,XY Disorders of Sex Development (DSDs)

GENE	INHERITANCE	GONAD	UTERUS	EXTERNAL GENITALIA	ASSOCIATED FEATURES
Disorders of Testis Development					
WT1	AD	Dysgenetic testis	+/–	Female or ambiguous	Wilms' tumor, renal abnormalities, gonadal tumors (WAGR, Denys-Drash and Frasier's syndromes)
CBX2	AD	Ovary	+	Female	
SF1	AR/AD (SL)	Dysgenetic testis/Leydig dysfunction	+/–	Female or ambiguous	Primary adrenal failure; primary ovarian insufficiency in female (46,XX) relatives
SRY	Y	Dysgenetic testis or ovotestis	+/–	Female or ambiguous	
SOX9	AD	Dysgenetic testis or ovotestis	+/–	Female or ambiguous	Campomelic dysplasia
MAP3K1	AD (SL)	Dysgenetic testis	+/–	Female or ambiguous	
GATA4	AD	Dysgenetic testis	–	Female, ambiguous or male	Congenital heart disease
ZFPM2	AD	Dysgenetic testis		Female, ambiguous or male	
ARX	X	Dysgenetic testis	–	Male or ambiguous	Developmental delay; X-linked lissencephaly
SAMD9	AD	Dysgenetic testis/Leydig dysfunction	+	Female, ambiguous or male	Myelodysplasia, infection, growth restriction, adrenal hypoplasia, enteropathy
DHH	AR	Dysgenetic testis/Leydig dysfunction	+	Female	Minifascicular neuropathy
MAMLD1	X	Dysgenetic testis/Leydig dysfunction	–	Hypospadias	
DAX1	dupXp21	Dysgenetic testis	+/–	Female or ambiguous	
WNT4/RSPO1	dup1p35	Dysgenetic testis	+	Ambiguous	
Disorders of Androgen Synthesis					
LHR	AR	Testis	–	Female, ambiguous or micropenis	Leydig cell hypoplasia
DHCR7	AR	Testis	–	Variable	Smith-Lemli-Opitz syndrome: coarse facies, second-third toe syndactyly, failure to thrive, developmental delay, cardiac and visceral abnormalities
StAR	AR	Testis	–	Female or ambiguous	Congenital lipoid adrenal hyperplasia (primary adrenal failure)
CYP11A1	AR	Testis	–	Ambiguous	Primary adrenal failure
HSD3B2	AR	Testis	–	Ambiguous	CAH, primary adrenal failure ± salt loss, partial androgenization due to ↑ DHEA
CYP17	AR	Testis	–	Female or ambiguous	CAH, hypertension due to ↑ corticosterone and 11-deoxycorticosterone, except in isolated 17,20-lyase deficiency
CYB5A	AR	Testis	–	Ambiguous	Apparent isolated 17,20-lyase deficiency; methemoglobinemia
POR	AR	Testis	–	Ambiguous or male	Mixed features of 21-hydroxylase deficiency and 17α-hydroxylase/17,20-lyase deficiency, sometimes associated with Antley-Bixler craniosynostosis
HSD17B3	AR	Testis	–	Female or ambiguous	Partial androgenization at puberty, ↑ androstenedione-to-testosterone ratio
SRD5A2	AR	Testis	–	Ambiguous or micropenis	Partial androgenization at puberty, ↑ testosterone-to-dihydrotestosterone ratio
AKR1C2 (*AKR1C4*)	AR	Testis	–	Female or ambiguous	Decreased fetal DHT production
Disorders of Androgen Action					
Androgen receptor	X	Testis	–	Female, ambiguous, micropenis or normal male	Phenotypic spectrum from complete androgen insensitivity syndrome (female external genitalia) and partial androgen insensitivity (ambiguous) to normal male genitalia and infertility

Abbreviations: AD, autosomal dominant; *AKR1C2*, aldo-keto reductase family 1 member 2; AR, autosomal recessive; *ARX*, aristaless related homeobox, X-linked; CAH, congenital adrenal hyperplasia; *CBX2*, chromobox homologue 2; *CYB5A*, cytochrome b5 POR, P450 oxidoreductase; *CYP11A1*, P450 cholesterol side-chain cleavage; *CYP17*, 17α-hydroxylase and 17,20-lyase; *DAX1*, dosage sensitive sex-reversal, adrenal hypoplasia congenita on the X chromosome, gene 1; DHEA, dehydroepiandrosterone; *DHCR7*, sterol 7 δ reductase; *DHH*, desert hedgehog; *GATA4*, GATA binding protein 4; *HSD17B3,* 17β-hydroxysteroid dehydrogenase type 3; *HSD3B2*, 3β-hydroxysteroid dehydrogenase type 2; *LHR*, LH receptor; *MAP3K1*, mitogen-activated protein kinase kinase kinase 1; *SF1,* steroidogenic factor 1; SL, sex-limited; *SOX9, SRY*-related HMG-box gene 9; *SRD5A2*, 5α-reductase type 2; *SRY,* sex-related gene on the Y chromosome; *StAR*, steroidogenic acute regulatory protein; WAGR, Wilms' tumor, aniridia, genitourinary anomalies, and mental retardation; *WNT4*, wingless-type mouse mammary tumor virus integration site, 4; *WT1,* Wilms' tumor–related gene 1; ZFPM2, zinc finger protein, multitype 2.

males usually have hypospadias repair in childhood and may request breast reduction in adolescence. Some boys enter puberty spontaneously. High-dose testosterone has been given to support development if puberty does not progress, but long-term data are limited. Other patients present with clitoral enlargement and labial fusion and may be raised as females. The surgical and psychosexual management of these patients is complex and requires active involvement of the parents and the patient during the appropriate stages of development. *Azoospermia* and male-factor infertility also have been described in association with mild loss-of-function mutations in the androgen receptor.

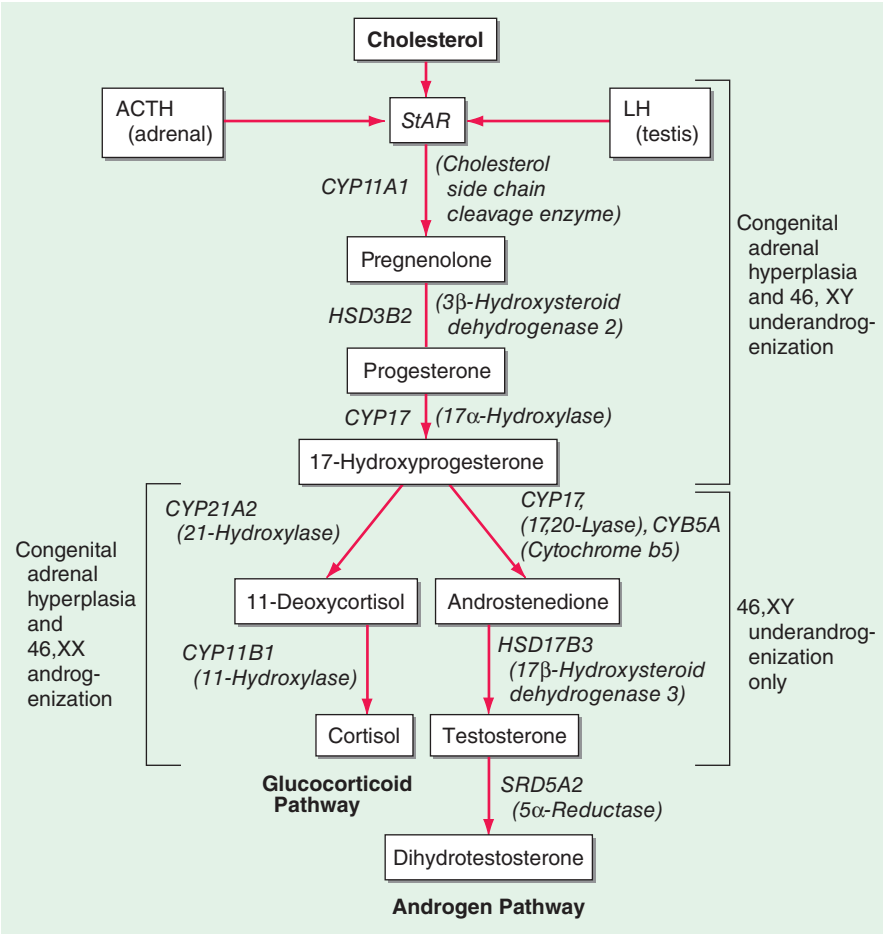


FIGURE 383-4 Simplified overview of glucocorticoid and androgen synthesis pathways. Defects in *CYP21A2* and *CYP11B1* shunt steroid precursors into the androgen pathway and cause androgenization of the 46,XX fetus. Testosterone is synthesized in the testicular Leydig cells and converted to dihydrotestosterone peripherally. Defects in enzymes involved in androgen synthesis result in underandrogenization of the 46,XY fetus. *StAR*, steroidogenic acute regulatory protein. *(After E Braunwald et al [eds]: Harrison's Principles of Internal Medicine, 15th ed. New York, McGraw-Hill, 2001.)*

OTHER DISORDERS AFFECTING 46,XY MALES

Persistent müllerian duct syndrome is the presence of a uterus in an otherwise phenotypic male. This condition can result from mutations in AMH or its receptor (AMHR2). The uterus may be removed, but only if damage to the vasa deferentia and blood supply can be avoided. *Isolated hypospadias* occurs in ~1 in 250 males. Most cases are idiopathic, although evidence of penoscrotal hypospadias, poor phallic development, and/or bilateral cryptorchidism requires investigation for an underlying DSD (e.g., partial gonadal dysgenesis, mild defect in testosterone action, or even severe forms of 46,XX CAH). Unilateral undescended testes (cryptorchidism) affect >3% of boys at birth. Orchidopexy should be considered if the testis has not descended by 6–9 months of age. Bilateral cryptorchidism occurs less frequently and should raise suspicion of gonadotropin deficiency or DSD. *Syndromic associations* and *intrauterine growth retardation* also occur relatively frequently in association with impaired testicular function or target tissue responsiveness, but the underlying etiology of many of these conditions is unknown.

46,XX DSD

Inappropriate androgenization of the 46,XX fetus (formerly called *female pseudohermaphroditism*) occurs when the gonad (ovary) contains androgen-secreting testicular tissue or after increased androgen exposure, which is usually adrenal in origin (Table 383-1).

46,XX Testicular/Ovotesticular DSD Testicular tissue can develop in 46,XX testicular DSD (46,XX males) after translocation of *SRY*, duplication of *SOX9*, or defects in *RSPO1* or *SF1/NR5A1* **(Table 383-5)**.

Increased Androgen Exposure • 21-HYDROXYLASE DEFICIENCY (CONGENITAL ADRENAL HYPERPLASIA) The *classic form* of 21-hydroxylase deficiency (21-OHD) is the most common cause of CAH **(Chap. 379)**. It has an incidence between 1 in 10,000 and 1 in 15,000 and is the most common cause of androgenization in chromosomal 46,XX females (Table 383-5). Affected individuals are homozygous or compound heterozygous for severe mutations in the enzyme 21-hydroxylase (*CYP21A2*). This mutation causes a block in adrenal glucocorticoid and mineralocorticoid synthesis, increasing 17-hydroxyprogesterone and shunting steroid precursors into the androgen synthesis pathway (Fig. 383-4). Glucocorticoid insufficiency causes a compensatory elevation of adrenocorticotropin (ACTH), resulting in adrenal hyperplasia and additional synthesis of steroid precursors proximal to the enzymatic block. Increased androgen synthesis in utero causes androgenization of the 46,XX fetus in the first trimester. Ambiguous genitalia are seen at birth, with varying degrees of clitoral enlargement and labial fusion. Excess androgen production causes gonadotropin-independent precocious puberty in males with 21-OHD.

The *salt-wasting* form of 21-OHD results from severe combined glucocorticoid and mineralocorticoid deficiency. A salt-wasting crisis usually manifests between 5 and 21 days of life and is a potentially life-threatening event that requires urgent fluid resuscitation and steroid treatment. Thus, a diagnosis of 21-OHD should be considered in any baby with atypical genitalia with bilateral nonpalpable gonads. Males (46,XY) with 21-OHD have no genital abnormalities at birth but are equally susceptible to adrenal insufficiency and salt-losing crises.

Females with the *classic simple virilizing* form of 21-OHD also present with genital ambiguity. They have impaired cortisol biosynthesis but do not develop salt loss. Patients with *nonclassic 21-OHD* produce normal amounts of cortisol and aldosterone but at the expense of producing excess androgens. Hirsutism (60%), oligomenorrhea (50%), and acne (30%) are the most common presenting features. This is one of the most common recessive disorders in humans, with an incidence as high as 1 in 100 to 500 in many populations and 1 in 27 in Ashkenazi Jews of Eastern European origin.

Biochemical features of acute salt-wasting 21-OHD are hyponatremia, hyperkalemia, hypoglycemia, inappropriately low cortisol and aldosterone, and elevated 17-hydroxyprogesterone, ACTH, and plasma renin activity. Presymptomatic diagnosis of classic 21-OHD is now made by neonatal screening tests for increased 17-hydroxyprogesterone in many centers. In most cases, 17-hydroxyprogesterone is markedly increased. In adults, ACTH stimulation (0.25 mg of cosyntropin IV) with assays for 17-hydroxyprogesterone at 0 and 30 min can be useful for detecting nonclassic 21-OHD and heterozygotes **(Chap. 379)**.

TREATMENT

Congenital Adrenal Hyperplasia

Acute salt-wasting crises require fluid resuscitation, IV hydrocortisone, and correction of hypoglycemia. Once the patient is stabilized, glucocorticoids must be given to correct the cortisol insufficiency and suppress ACTH stimulation, thereby preventing further virilization, rapid skeletal maturation, and the development of polycystic ovaries. Typically, hydrocortisone (10–15 mg/m^2 per day in three divided doses) is used in childhood with a goal of partially

TABLE 383-5 Selected Genetic Causes of 46,XX Disorders of Sex Development (DSDs)

GENE	INHERITANCE	GONAD	UTERUS	EXTERNAL GENITALIA	ASSOCIATED FEATURES
Testicular/Ovotesticular DSD					
SRY	Translocation	Testis or ovotestis	–	Male or ambiguous	
SOX9	dup17q24	Unknown	–	Male or ambiguous	
SF1 (codon 92)	AD	Testis or ovotestis	±	Male or ambiguous	
RSPO1	AR	Testis or ovotestis	±	Male or ambiguous	Palmar plantar hyperkeratosis, squamous cell skin carcinoma
WNT4	AR	Testis or ovotestis	–	Male or ambiguous	SERKAL syndrome (renal dysgenesis, adrenal and lung hypoplasia)
Increased Androgen Synthesis					
HSD3B2	AR	Ovary	+	Clitoromegaly	CAH, primary adrenal failure, mild androgenization due to ↑ DHEA
CYP21A2	AR	Ovary	+	Ambiguous	CAH, phenotypic spectrum from severe salt-losing forms associated with adrenal failure to simple virilizing forms with compensated adrenal function, ↑ 17-hydroxyprogesterone
POR	AR	Ovary	+	Ambiguous or female	Mixed features of 21-hydroxylase deficiency and 17α-hydroxylase/17,20-lyase deficiency, sometimes associated with Antley-Bixler craniosynostosis
CYP11B1	AR	Ovary	+	Ambiguous	CAH, hypertension due to ↑ 11-deoxycortisol and 11-deoxycorticosterone
CYP19	AR	Ovary	+	Ambiguous	Maternal virilization during pregnancy, absent breast development at puberty
Glucocorticoid receptor	AR	Ovary	+	Ambiguous	↑ ACTH, 17-hydroxyprogesterone and cortisol; failure of dexamethasone suppression

Abbreviations: ACTH, adrenocorticotropin; AR, autosomal recessive; CAH, congenital adrenal hyperplasia; *CYP11B1*, 11β-hydroxylase; *CYP19*, aromatase; *CYP21A2*, 21-hydroxylase; DHEA, dehydroepiandrosterone; *HSD3B2*, 3β-hydroxysteroid dehydrogenase type 2; *POR*, P450 oxidoreductase; *RSPO1*, R-spondin 1; *SF1,* steroidogenic factor 1; *SOX9*, *SRY*-related HMG-box gene 9; *SRY*, sex-related gene on the Y chromosome.

suppressing 17-hydroxyprogesterone. The aim of treatment is to use the lowest glucocorticoid dose that adequately suppresses adrenal androgen production without causing signs of glucocorticoid excess such as impaired growth and obesity. Salt-wasting conditions are treated with mineralocorticoid replacement. Infants usually need salt supplements up to the first year of life. Plasma renin activity and electrolytes are used to monitor mineralocorticoid replacement, remembering that normal ranges are age-dependent. Some patients with simple virilizing 21-OHD also benefit from mineralocorticoid supplements. Parents and patients should be educated about the need for increased doses of steroids during sickness, and patients should carry medic alert systems.

Steroid treatment for older adolescents and adults varies depending on lifestyle, age, and factors such as a desire to optimize fertility. Hydrocortisone remains a useful approach, but treatment with prednisolone at night may provide more complete ACTH suppression. Steroid doses should be adjusted to individual requirements because overtreatment can result in iatrogenic Cushing's-like features, including weight gain, insulin resistance, hypertension, and osteopenia. Because it is long acting, dexamethasone given at night is useful for ACTH suppression but is often associated with more side effects, making hydrocortisone or prednisolone preferable for most patients. Androstenedione and testosterone may be useful measurements of long-term control, with less fluctuation than 17-hydroxyprogesterone. Mineralocorticoid requirements often decrease in adulthood, and doses should be reassessed and reduced to avoid hypertension in adults. In very severe cases, adrenalectomy has been advocated but incurs the risks of surgery and total adrenal insufficiency. Newer approaches to treatment under investigation include more physiological cortisol replacement strategies and specifically targeting androgen excess.

Girls with significant genital androgenization due to classic 21-OHD usually undergo vaginal reconstruction and sometimes clitoral reduction (maintaining the glans and nerve supply), but the optimal timing of these procedures is debated, as is the need for the individual to be able to consent. There is a higher threshold for undertaking clitoral surgery in some centers because long-term sensation and ability to achieve orgasm can be affected, but the long-term results of newer techniques are not yet known. Full information about all options should be provided, with appropriate support. Good endocrine control to reduce testosterone levels is also important. If surgery is performed in infancy, surgical revision or regular vaginal dilatation may be needed in adolescence or adulthood, and long-term psychological support and psychosexual counseling may be appropriate. Women with 21-OHD frequently develop polycystic ovaries and have reduced fertility, especially when control is poor. Fecundity is achieved in 60–90% of women with good metabolic control, but ovulation induction (or even adrenalectomy) may be required. Dexamethasone should be avoided in pregnancy. Men with poorly controlled 21-OHD may develop testicular adrenal rests and are at risk for reduced fertility. Prenatal treatment of 21-OHD by the administration of dexamethasone to mothers is still under evaluation. Treatment must be started early in pregnancy, but has the risk that both affected and nonaffected fetuses are exposed. The long-term effects of prenatal dexamethasone exposure on fetal development are still under evaluation, and current guidelines recommend full informed consent before treatment, ideally in a study protocol that allows long-term follow-up of all children treated. Newer techniques such as cell-free fetal DNA testing and early genotyping may potentially reduce treatment of nonaffected fetuses.

The treatment of other forms of CAH includes mineralocorticoid and glucocorticoid replacement for salt-losing conditions (e.g., *StAR*, *CYP11A1*, *HSD3β2*), suppression of ACTH drive with glucocorticoids in disorders associated with hypertension (e.g., *CYP17*, *CYP11B1*), and appropriate sex hormone replacement in adolescence and adulthood, when necessary.

OTHER CAUSES Increased androgen synthesis can also occur in CAH due to defects in *POR*, *11β-hydroxylase* (*CYP11B1*), and *3β-hydroxysteroid dehydrogenase type 2* (*HSD3B2*) and with mutations in the genes encoding *aromatase* (*CYP19*) and the glucocorticoid receptor. Increased androgen exposure in utero can occur with maternal virilizing tumors and with ingestion of androgenic compounds.

OTHER DISORDERS AFFECTING 46,XX FEMALES

Congenital absence of the vagina occurs in association with *müllerian agenesis* or *hypoplasia* as part of the Mayer-Rokitansky-Kuster-Hauser (MRKH) syndrome (rarely caused by *WNT4* mutations). This diagnosis should be considered in otherwise phenotypically normal females with primary amenorrhea. Associated features include renal (agenesis) and cervical spinal abnormalities.

GLOBAL CONSIDERATIONS

The approach to a child or adolescent with ambiguous genitalia or another DSD requires cultural sensitivity, as the concepts of sex and gender vary widely. Rare genetic DSDs can occur more frequently in specific populations (e.g., *5α-reductase type 2* in the Dominican Republic). Different forms of CAH also show ethnic and geographic variability. In many countries, appropriate biochemical tests may not be readily available, and access to appropriate forms of treatment and support may be limited.

FURTHER READING

ACHERMANN JC et al: Disorders of sex development: Effect of molecular diagnostics. Nat Rev Endocrinol 11:478, 2015.

AHMED SF et al: Society for Endocrinology UK guidance on the initial evaluation of an infant or an adolescent with a suspected disorder of sex development (Revised 2015). Clin Endocrinol (Oxf) 84:771, 2016.

BACKELJAUW PF et al: Proceedings from the Turner Resource Network Symposium: The crossroads of health care research and health care delivery. Am J Med Genet 167:1962, 2015.

COOLS M et al: Gonadal pathology and tumor risk in relation to clinical characteristics in patients with 45,X/46,XY mosaicism. J Clin Endocrinol Metab 96:E1171, 2011.

DAVIS SM et al: Testis Development and Fertility Potential in Boys with Klinefelter Syndrome. Endocrinol Metab Clin North Am 44:843, 2015.

GRAVHOLT CH et al: Clinical practice guidelines for the care of girls and women with Turner syndrome: Proceedings from the 2016 Cincinnati International Turner Syndrome Meeting. Eur J Endocrinol 177:G1, 2017.

HUGHES IA et al: Androgen insensitivity syndrome. Lancet 380:1490, 2012.

LEE PA et al: Global disorders of sex development update since 2006: Perceptions, approach and care. Horm Res Paediatr 85:158, 2016.

TURCU AF, AUCHUS RJ: The next 150 years of congenital adrenal hyperplasia. J Steroid Biochem Mol Biol 153:63, 2015.

WARNE GL, RAZA J: Disorders of sex development (DSDs), their presentation and management in different cultures. Rev Endocr Metab Disord 9:227, 2008.

384 Disorders of the Testes and Male Reproductive System

Shalender Bhasin, J. Larry Jameson

The male reproductive system regulates sex differentiation, androgenization, and the hormonal changes that accompany puberty, ultimately leading to spermatogenesis and fertility. Under the control of the pituitary hormones—luteinizing hormone (LH) and follicle-stimulating hormone (FSH)—the Leydig cells of the testes produce testosterone and germ cells are nurtured by Sertoli cells to divide, differentiate, and mature into sperm. During embryonic development, testosterone and dihydrotestosterone (DHT) induce the wolffian duct and virilization of the external genitalia. During puberty, testosterone promotes somatic growth and the development of secondary sex characteristics. In the adult, testosterone is necessary for spermatogenesis, libido and normal sexual function, and maintenance of muscle and bone mass. This chapter focuses on the physiology of the testes and disorders associated with decreased androgen production, which may be caused by gonadotropin deficiency or by primary testis dysfunction. A variety of testosterone formulations now allow more physiologic androgen replacement. Infertility occurs in ~5% of men and is increasingly amenable to treatment by hormone replacement or by using sperm transfer techniques. **For further discussion of sexual dysfunction, disorders of the prostate, and testicular cancer, see Chaps. 390, 83, 84, respectively.**

DEVELOPMENT AND STRUCTURE OF THE TESTIS

The fetal testis develops from the undifferentiated gonad after expression of a genetic cascade that is initiated by the SRY (sex-related gene on the Y chromosome) **(Chap. 383)**. SRY induces differentiation of Sertoli cells, which surround germ cells and, together with peritubular myoid cells, form testis cords that will later develop into seminiferous tubules. Fetal Leydig cells and endothelial cells migrate into the gonad from the adjacent mesonephros but may also arise from interstitial cells that reside between testis cords. Fetal Leydig cells atrophy after birth and do not contribute to the origin of adult Leydig cells, which originate from undifferentiated progenitor cells that appear in the testis after birth and acquire full steroidogenic function during puberty. Testosterone produced by the fetal Leydig cells supports the growth and differentiation of wolffian duct structures that develop into the epididymis, vas deferens, and seminal vesicles. Testosterone is also converted to DHT (see below), which induces formation of the prostate and the external male genitalia, including the penis, urethra, and scrotum. Testicular descent through the inguinal canal is controlled in part by Leydig cell production of insulin-like factor 3 (INSL3), which acts via a receptor termed *Great* (*G* protein–coupled *re*ceptor *a*ffecting *t*estis descent). Sertoli cells produce müllerian inhibiting substance (MIS), which causes regression of the müllerian structures, including the fallopian tube, uterus, and upper segment of the vagina.

NORMAL MALE PUBERTAL DEVELOPMENT

Puberty commonly refers to the maturation of the reproductive axis and the development of secondary sex characteristics. In addition to reproductive hormones, it requires a coordinated response of multiple hormonal systems including metabolic signals (e.g., leptin), as well as the adrenal and growth hormone (GH) axes **(Fig. 384-1)**. The development of secondary sex characteristics is initiated by *adrenarche*, which usually occurs between 6 and 8 years of age when the adrenal gland begins to produce greater amounts of androgens from the zona reticularis, the principal site of dehydroepiandrosterone (DHEA) production. The sex maturation process is greatly accelerated by the activation of the hypothalamic-pituitary axis and the production of gonadotropin-releasing hormone (GnRH). The GnRH pulse generator in the hypothalamus is active during fetal life and early infancy, but is restrained until the early stages of puberty by a neuroendocrine brake imposed by the inhibitory actions of glutamate and γ–amino butyric acid (GABA)

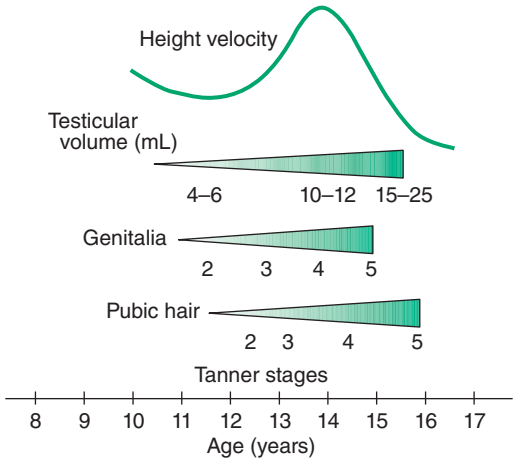


FIGURE 384-1 Pubertal events in males. Sexual maturity ratings for genitalia and pubic hair and divided into five stages. *(From WA Marshall, JM Tanner: Variations in the pattern of pubertal changes in boys. Arch Dis Child 45:13, 1970.)*

in the mediobasal hypothalamus, and neuropeptide Y. Although the pathways that initiate reactivation of the GnRH pulse generator at the onset of puberty remain incompletely understood, mounting evidence supports involvement of GPR54, a G protein–coupled receptor that binds an endogenous ligand, kisspeptin. Individuals with mutations of GPR54 fail to enter puberty, and experiments in primates demonstrate that infusion of the ligand is sufficient to induce premature puberty. Kisspeptin signaling plays an important role in mediating the feedback action of sex steroids on gonadotropin secretion and in regulating the tempo of sexual maturation at puberty. Leptin, a hormone produced by adipose cells, plays a permissive role in the resurgence of GnRH secretion at the onset of puberty, as leptin-deficient individuals also fail to enter puberty (Chap. 394). Adipocyte hormone leptin, gut hormone ghrelin, neuropeptide Y, and kisspeptin integrate the signals originating in energy stores and metabolic tissues with mechanisms that control onset of puberty through regulation of GnRH secretion. Energy deficit and excess, and metabolic stress are associated with disturbed reproductive maturation and timing of puberty.

The early stages of puberty are characterized by nocturnal surges of LH and FSH. Growth of the testes is usually the first clinical sign of puberty, reflecting an increase in seminiferous tubule volume. Increasing levels of testosterone deepen the voice and stimulate muscle growth. Conversion of testosterone to DHT leads to growth of the external genitalia and pubic hair. DHT also stimulates prostate and facial hair growth and initiates recession of the temporal hairline. The growth spurt occurs at a testicular volume of about 10–12 mL. GH increases early in puberty and is stimulated in part by the rise in gonadal steroids. GH increases the level of insulin-like growth factor 1 (IGF-1), which enhances linear bone growth. The prolonged pubertal exposure to gonadal steroids (mainly estradiol) ultimately induces epiphyseal closure and limits further bone growth.

REGULATION OF TESTICULAR FUNCTION

REGULATION OF THE HYPOTHALAMIC-PITUITARY-TESTIS AXIS IN ADULT MAN

Hypothalamic GnRH regulates the production of the pituitary gonadotropins, LH and FSH (Fig. 384-2). GnRH is released in discrete pulses approximately every 2 h, resulting in corresponding pulses of LH and FSH. These dynamic hormone pulses account in part for the wide variations in LH and testosterone, even within the same individual. LH acts primarily on the Leydig cell to stimulate testosterone synthesis. The regulatory control of androgen synthesis is modulated by dynamic integration of the feedforward elements exerted on the testis by LH and FSH, and the feedback exerted by testosterone and estrogen on both the hypothalamus and the pituitary. FSH acts on the Sertoli cell to regulate spermatogenesis and the production of Sertoli products such as inhibin B, which acts to selectively suppress pituitary FSH. Despite these somewhat distinct Leydig and Sertoli cell–regulated pathways, testis function is integrated at several levels: GnRH regulates both gonadotropins; spermatogenesis requires high levels of testosterone; numerous paracrine interactions between Leydig and Sertoli cells are necessary for normal testis function.

THE LEYDIG CELL: ANDROGEN SYNTHESIS

LH binds to its seven transmembrane, G protein–coupled receptor to activate the cyclic AMP pathway. Stimulation of the LH receptor induces *st*eroid *a*cute *r*egulatory (StAR) protein, along with several steroidogenic enzymes involved in androgen synthesis. LH receptor mutations cause Leydig cell hypoplasia or agenesis, underscoring the importance of this pathway for Leydig cell development and function. The rate-limiting process in testosterone synthesis is the transport of intracellular cholesterol by the StAR protein to the inner mitochondrial membrane. Mutations of the StAR protein are associated with Congenital Lipoid Adrenal Hyperplasia, a rare form of congenital adrenal hyperplasia (CAH) characterized by very low adrenal and gonadal steroids. Peripheral benzodiazepine receptor, a mitochondrial cholesterol-binding protein, is also an acute regulator of Leydig cell steroidogenesis. The major enzymatic steps involved in testosterone synthesis

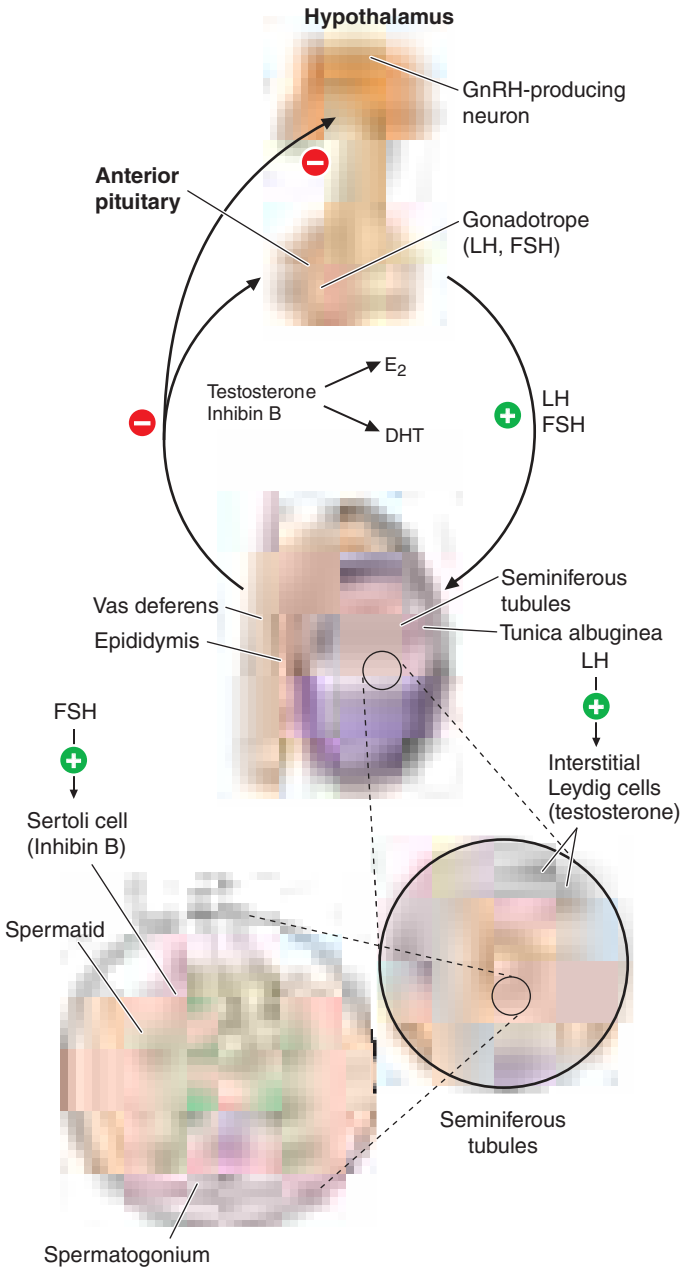


FIGURE 384-2 Hypothalamic pituitary gonadotropin axis, structure of testis, seminiferous tubule. E_2, 17β-estradiol; DHT, dihydrotestosterone.

are summarized in Fig. 384-3. After cholesterol transport into the mitochondrion, the formation of pregnenolone by CYP11A1 (side chain cleavage enzyme) is a limiting enzymatic step. The 17α-hydroxylase and the 17,20-lyase reactions are catalyzed by a single enzyme, CYP17; posttranslational modification (phosphorylation) of this enzyme and the presence of specific enzyme cofactors, such as cytochrome B, confer 17,20-lyase activity selectively in the testis and zona reticularis of the adrenal gland. Abiraterone is a dual inhibitor of 17 α-hydroxylase and 17,20-lyase activities, which play an important role in androgen synthesis in castration-resistant prostate cancers. Testosterone can be converted to the more potent DHT by a family of steroid 5α-reductase enzymes, or it can be aromatized to estradiol by CYP19 (aromatase). At least two isoforms of steroid 5α-reductase, SRD5A1 and SRD5A2, have been described; all known patients with 5α-reductase deficiency have had mutations in SRD5A2, the predominant form in the prostate and the skin. Finasteride predominantly inhibits SRD5A2, while dutasteride is a dual inhibitor of both SRD5A1 and SRD5A2. DHT can also be derived directly through the backdoor pathway in which 17 α-hydroxyprogesterone is converted through a series of 5 α and 3 α reductions to androsterone and eventually to DHT. Recent reports of

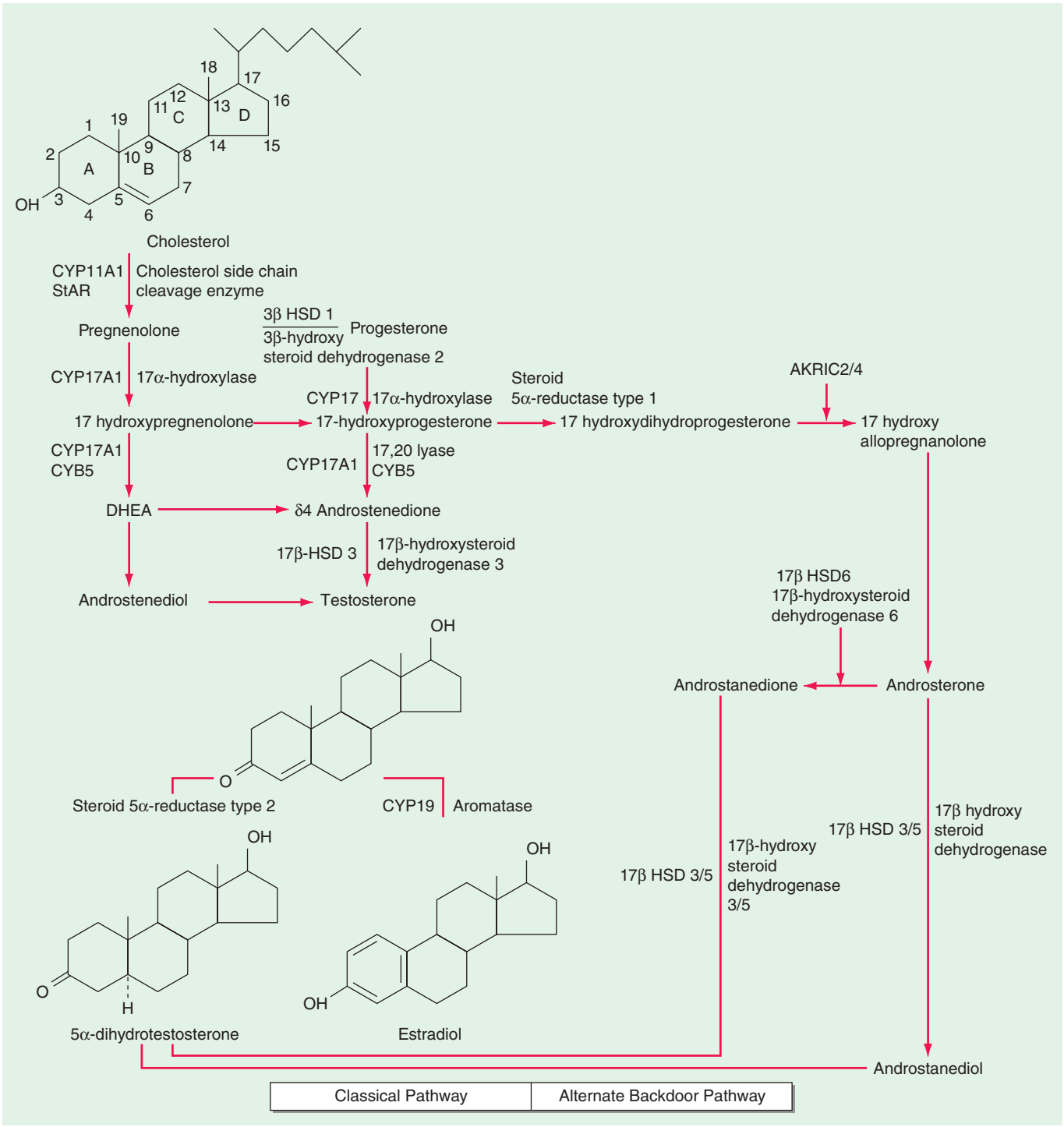


FIGURE 384-3 **The biochemical pathway** in the conversion of 27-carbon sterol cholesterol to androgens and estrogens.

mutations in AKR1C2/4 genes in undervirilized 46, XY individuals suggest that the backdoor pathway for DHT formation, which was originally described in the tammar wallaby, is active in the human fetal testis.

Testosterone Transport and Metabolism In males, 95% of circulating testosterone is derived from testicular production (3–10 mg/d). Direct secretion of testosterone by the adrenal and the peripheral conversion of androstenedione to testosterone collectively account for another 0.5 mg/d of testosterone. Only a small amount of DHT (70 μg/d) is secreted directly by the testis; most circulating DHT is derived from peripheral conversion of testosterone. Most of the daily production of estradiol (~45 μg/d) in men is derived from aromatase-mediated peripheral conversion of testosterone and androstenedione.

Circulating testosterone is bound predominantly to sex hormone–binding globulin (SHBG) and albumin (Fig. 384-4), and to a lesser extent to cortisol binding globulin (CBG), and orosomucoid. SHBG binds testosterone with much greater affinity than albumin, CBP, and orosomucoid. Only 1.0–4.0% of testosterone is unbound. According to the "free hormone" hypothesis, only the unbound fraction is biologically active. The term "bioavailable testosterone" refers to unbound testosterone plus testosterone bound loosely to albumin, and reflects the concept that albumin-bound testosterone can dissociate at the capillary level, especially in tissues with long transit time, such as the liver and the brain. SHBG-bound testosterone also may be internalized through endocytic pits by binding to a protein called megalin. SHBG concentrations are decreased by androgens, obesity, diabetes mellitus, hypothyroidism, nephrotic syndrome, and genetic factors. Conversely,

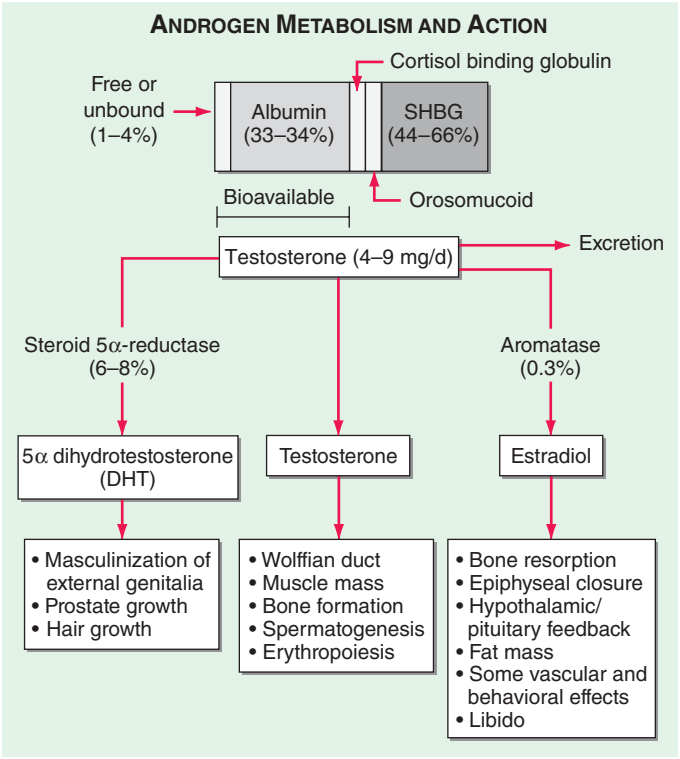


FIGURE 384-4 Androgen metabolism and actions. SHBG, sex hormone–binding globulin.

estrogen administration, hyperthyroidism, many chronic inflammatory illnesses, infections such as HIV or hepatitis B and C, aging, and the use of some anticonvulsants are associated with high SHBG concentrations.

Testosterone is metabolized predominantly in the liver, although some degradation occurs in peripheral tissues, particularly the prostate and the skin. In the liver, testosterone is converted by a series of enzymatic steps that involve 5α- and 5β-reductases, 3α- and 3β-hydroxysteroid dehydrogenases, and 17β-hydroxysteroid dehydrogenase into androsterone, etiocholanolone, DHT, and 3-α-androstanediol. These compounds undergo glucuronidation or sulfation before being excreted by the kidneys.

Mechanism of Androgen Action Testosterone exerts some of its biologic effects by binding to androgen receptor (AR), either directly or after its conversion to DHT by the steroid 5-α reductase. The actions of testosterone on the Wolffian structures, skeletal muscle, erythropoiesis, and bone in men do not require its obligatory conversion to DHT. However, the conversion of testosterone to DHT is necessary for the masculinization of the urogenital sinus and genital tubercle. Aromatization of testosterone to estradiol mediates additional effects of testosterone on the bone resorption, epiphyseal closure, sexual desire, vascular endothelium, and fat. DHT can also be converted in some tissues by 3-keto reductase/3β-hydroxysteroid dehydrogenase enzymes to 5α-androstane-3β,17β-diol, which is a high-affinity ligand and agonist of estrogen receptor ER β.

The AR is structurally related to the nuclear receptors for estrogen, glucocorticoids, and progesterone **(Chap. 370)**. The AR is encoded by a gene on the long arm of the X chromosome and has a molecular mass of about 110 kDa. A polymorphic region in the amino terminus of the receptor, which contains a variable number of glutamine repeats, modifies the transcriptional activity of the receptor. The AR protein is distributed in both the cytoplasm and the nucleus. The ligand binding to the AR induces conformational changes that allow the recruitment and assembly of tissue-specific cofactors, and causes it to translocate into the nucleus, where it binds to specific androgen response elements in the DNA or other transcription factors already bound to DNA. Thus, the AR is a ligand-regulated transcription factor that regulates the expression of androgen-dependent genes in a tissue-specific manner. Some androgen effects, such as those on the smooth muscle, may be mediated by nongenomic AR signal transduction pathways. Testosterone binds to AR with half the affinity of DHT. The DHT-AR complex also has greater thermostability and a slower dissociation rate than the testosterone-AR complex. However, the molecular basis for selective testosterone versus DHT actions remains incompletely explained.

THE SEMINIFEROUS TUBULES: SPERMATOGENESIS

The seminiferous tubules are convoluted, closed loops with both ends emptying into the rete testis, a network of progressively larger efferent ducts that ultimately form the epididymis (Fig. 384-2). The seminiferous tubules total about 600 m in length and comprise about two-thirds of testis volume. The walls of the tubules are formed by polarized Sertoli cells that are apposed to peritubular myoid cells. Tight junctions between Sertoli cells create the blood-testis barrier. Germ cells comprise the majority of the seminiferous epithelium (~60%) and are intimately embedded within the cytoplasmic extensions of the Sertoli cells, which function as "nurse cells." Germ cells progress through characteristic stages of mitotic and meiotic divisions. A pool of type A spermatogonia serve as stem cells capable of self-renewal. Primary spermatocytes are derived from type B spermatogonia and undergo meiosis before progressing to spermatids that undergo spermiogenesis (a differentiation process involving chromatin condensation, acquisition of an acrosome, elongation of cytoplasm, and formation of a tail) and are released from Sertoli cells as mature spermatozoa. The complete differentiation process into mature sperm requires 74 days. Peristaltic-type action by peritubular myoid cells transports sperm into the efferent ducts. The spermatozoa spend an additional 21 days in the epididymis, where they undergo further maturation and capacitation. The normal adult testes produce >100 million sperm per day.

Naturally occurring mutations in FSHβ or in the FSH receptor confirm an important, but not essential, role for this pathway in spermatogenesis. Females with mutations in FSHβ or the FSH receptor are hypogonadal and infertile because ovarian follicles do not mature; males with these mutations exhibit variable degrees of reduced spermatogenesis, presumably because of impaired Sertoli cell function. Because Sertoli cells produce inhibin B, an inhibitor of FSH, seminiferous tubule damage (e.g., by radiation) causes a selective increase of FSH. Testosterone reaches very high concentrations locally in the testis and is essential for spermatogenesis. The cooperative actions of FSH and testosterone are important in the progression of meiosis and spermiation. FSH and testosterone regulate germ cell survival via the intrinsic and the extrinsic apoptotic mechanisms. FSH may also play an important role in supporting spermatogonia. Gonadotropin-regulated testicular RNA helicase (GRTH/DDX25), a testis-specific gonadotropin/androgen-regulated RNA helicase, is present in germ cells and Leydig cells and may be an important factor in the paracrine regulation of germ cell development. Several cytokines and growth factors are also involved in the regulation of spermatogenesis by paracrine and autocrine mechanisms. A number of knockout mouse models exhibit impaired germ cell development or spermatogenesis, presaging possible mutations associated with male infertility.

The human Y chromosome contains a small pseudoautosomal region that can recombine with homologous regions of the X chromosome. Most of the Y chromosome does not recombine with the X chromosome and is referred to as the male-specific region of the Y (MSY). The MSY contains 156 transcription units that encode for 26 proteins, including nine families of Y-specific multicopy genes; many of these Y-specific genes are also testis-specific and necessary for spermatogenesis. Microdeletions in several nonoverlapping subregions of the Y chromosome—AZFa, AZFb, AZFc and AZFd, which contain many spermatogenic genes (e.g., RNA-binding motif, *RBM*; deleted in azoospermia, *DAZ*)—are associated with oligospermia or azoospermia. Approximately 15% of infertile men with azoospermia and about 6% of men with severe oligozoospermia harbor a Y microdeletion. Microdeletions of the AZFa and AZFb subregions are typically associated with Sertoli cell only or maturation arrest histology, azoospermia, and poor prognosis for sperm retrieval. In contrast, AZFc subregion microdeletions are typically associated with oligozoospermia and higher success rates for sperm retrieval. Microdeletion involving the DAZ gene in the

AZFc region is one of the commonest Y chromosome microdeletions in infertile men. A partial deletion of the AZFc region called the *gr/gr deletion* is associated with infertility among Caucasian men in Europe and the Western Pacific region.

TREATMENT

Male Factor Infertility

Treatment options for male factor infertility have expanded greatly in recent years. Secondary hypogonadism is highly amenable to treatment with pulsatile GnRH or gonadotropins (see below). Assisted reproductive technologies, such as the in vitro fertilization (IVF) and intracytoplasmic sperm injection (ICSI), have provided new opportunities for patients with primary testicular failure and disorders of sperm transport. Choice of initial treatment options depends on sperm concentration and motility. Expectant management should be attempted initially in men with mild male factor infertility (sperm count of 15–20 × 10^6/mL and normal motility). Moderate male factor infertility (10–15 x 10^6/mL and 20–40% motility) should begin with intrauterine insemination alone or in combination with treatment of the female partner with clomiphene or gonadotropins, but it may require IVF with or without ICSI. For men with a severe defect (sperm count of <10 × 10^6/mL, 10% motility), IVF with ICSI or donor sperm has become the treatment of choice. Yq microdeletions will be transmitted through ICSI from the affected father to his male offspring if sperm carrying Yq microdeletion is used.

CLINICAL AND LABORATORY EVALUATION OF MALE REPRODUCTIVE FUNCTION

HISTORY AND PHYSICAL EXAMINATION

The history should focus on developmental stages such as puberty and growth spurts, as well as androgen-dependent events such as early morning erections, frequency and intensity of sexual thoughts, and frequency of masturbation or intercourse. Although libido and the overall frequency of sexual acts are decreased in androgen-deficient men, young hypogonadal men can achieve erections in response to visual erotic stimuli. Men with acquired androgen deficiency often report decreased energy and low mood.

The physical examination should focus on secondary sex characteristics such as hair growth, gynecomastia, testicular volume, prostate, and height and body proportions. *Eunuchoid proportions* are defined as an arm span >2 cm greater than height and suggest that androgen deficiency occurred before epiphyseal fusion. Hair growth in the face, axilla, chest, and pubic regions is androgen-dependent; however, changes may not be noticeable unless androgen deficiency is severe and prolonged. Ethnicity also influences the intensity of hair growth **(Chap. 387)**. Testicular volume is best assessed by using a Prader orchidometer. Testes range from 3.5 to 5.5 cm in length, which corresponds to a volume of 12–25 mL. Advanced age does not influence testicular size, although the consistency becomes less firm. Asian men generally have smaller testes than western Europeans, independent of differences in body size. Because of its possible role in infertility, the presence of varicocele should be sought by palpation while the patient is standing; it is more common on the left side. Patients with Klinefelter syndrome have markedly reduced testicular volumes (1–2 mL). In congenital hypogonadotropic hypogonadism, testicular volumes provide a good index for the degree of gonadotropin deficiency and the likelihood of response to therapy.

GONADOTROPIN AND INHIBIN MEASUREMENTS

LH and FSH are measured using two-site immunoradiometric, immunofluorometric, or chemiluminescent assays, which have very low cross-reactivity with other pituitary glycoprotein hormones and human chorionic gonadotropin (hCG) and have sufficient sensitivity to measure the low levels present in patients with hypogonadotropic hypogonadism. In men with a low testosterone level, an LH level can distinguish primary (high LH) versus secondary (low or inappropriately normal LH) hypogonadism. An elevated LH level indicates a primary defect at the testicular level, whereas a low or inappropriately normal LH level suggests a defect at the hypothalamic-pituitary level. LH pulses occur about every 1–3 h in normal men. Thus, gonadotropin levels fluctuate, and samples should be pooled or repeated when results are equivocal. FSH is less pulsatile than LH because it has a longer half-life. Selective increase in FSH suggests damage to the seminiferous tubules. Inhibin B, a Sertoli cell product that suppresses FSH, is reduced with seminiferous tubule damage. Inhibin B is a dimer with α-β_B subunits and is measured by two-site immunoassays.

GnRH Stimulation Testing The GnRH test is performed by measuring LH and FSH concentrations at baseline and at 30 and 60 min after intravenous administration of 100 μg of GnRH. A minimally acceptable response is a twofold LH increase and a 50% FSH increase. In the prepubertal period or with severe GnRH deficiency, the gonadotrope may not respond to a single bolus of GnRH because it has not been primed by endogenous hypothalamic GnRH; in these patients, GnRH responsiveness may be restored by chronic, pulsatile GnRH administration. With the availability of sensitive and specific LH assays, GnRH stimulation testing is used rarely.

TESTOSTERONE ASSAYS

Total Testosterone Total testosterone includes both unbound and protein-bound testosterone and is measured by radioimmunoassays, immunometric assays, or liquid chromatography tandem mass spectrometry (LC-MS/MS). LC-MS/MS involves extraction of serum by organic solvents, separation of testosterone from other steroids by high-performance liquid chromatography and mass spectrometry, and quantitation of unique testosterone fragments by mass spectrometry. LC-MS/MS provides accurate and sensitive measurements of testosterone levels even in the low range and has emerged as the method of choice for testosterone measurement. The use of laboratories that have been certified by the Centers for Disease Control's Hormone Standardization Program for Testosterone (HoST) can ensure that testosterone measurements are accurate and calibrated to an international standard. A single fasting morning sample provides a good approximation of the average testosterone concentration with the realization that testosterone levels fluctuate in response to pulsatile LH. Testosterone is generally lower in the late afternoon and is reduced by acute illness. The harmonized normal range for total testosterone, measured using LC-MS/MS in nonobese populations of European and American men, 19–39 years, is 264–916 ng/dL. This harmonized reference range can be applied to values from laboratories that are certified by the CDC's Hormone Standardization Program for Testosterone.

Alterations in SHBG levels due to aging, obesity, diabetes mellitus, hyperthyroidism, some types of medications, chronic illness, or on a congenital basis, can affect total testosterone levels. Heritable factors contribute substantially to the population level variation in testosterone levels and genome wide association studies have revealed polymorphisms in SHBG gene as important contributors to variation in testosterone levels.

Measurement of Unbound Testosterone Levels Most circulating testosterone is bound to SHBG and to albumin; only 1.0–4% of circulating testosterone is unbound, or "free." Free testosterone should ideally be measured by equilibrium dialysis under standardized conditions using an accurate and reliable assay for total testosterone. The unbound testosterone concentration also can be calculated from total testosterone, SHBG, and albumin concentrations. Recent research has shown that testosterone binding to SHBG is a multi-step process that involves complex allosteric interactions between the two binding sites within the SHBG dimer; a novel ensemble allosteric model of testosterone's binding to SHBG dimers provides good estimates of free testosterone concentrations. The previous law-of-mass action equations based on linear models of testosterone binding to SHBG used assumptions that have been shown to be erroneous. Tracer analogue methods are relatively inexpensive and convenient, but they are inaccurate. *Bioavailable testosterone* refers to unbound testosterone plus testosterone

that is loosely bound to albumin; it can be determined by the ammonium sulfate precipitation method. However, the measurements of bioavailable testosterone using the ammonium sulfate precipitation are technically challenging, susceptible to imprecision, and are not recommended.

hCG Stimulation Test The hCG stimulation test is performed by administering a single injection of 1500–4000 IU of hCG intramuscularly and measuring testosterone levels at baseline and 24, 48, 72, and 120 h after hCG injection. An alternative regimen involves three injections of 1500 units of hCG on successive days and measuring testosterone levels 24 h after the last dose. An acceptable response to hCG is a doubling of the testosterone concentration in adult men. In prepubertal boys, an increase in testosterone to >150 ng/dL indicates the presence of testicular tissue. No response may indicate an absence of testicular tissue or marked impairment of Leydig cell function. Measurement of MIS, a Sertoli cell product, is also used to detect the presence of testes in prepubertal boys with cryptorchidism.

SEMEN ANALYSIS

Semen analysis is the most important step in the evaluation of male infertility. Samples are collected by masturbation following a period of abstinence for 2–3 days. Semen volumes and sperm concentrations vary considerably among fertile men, and several samples may be needed before concluding that the results are abnormal. Analysis should be performed within an hour of collection. Using semen samples from over 4500 men in 14 countries, whose partners had a time-to-pregnancy of <12 months, WHO has generated the following one-sided reference limits for semen parameters: semen volume, 1.5 mL; total sperm number, 39 million per ejaculate; sperm concentration, 15 million per mL; vitality, 58% live; progressive motility, 32%; total (progressive + non-progressive) motility, 40%; morphologically normal forms, 4.0%. Some men with low sperm counts are nevertheless fertile. A variety of tests for sperm function can be performed in specialized laboratories, but these add relatively little to the treatment options.

TESTICULAR BIOPSY

Testicular biopsy is useful in some patients with oligospermia or azoospermia as an aid in diagnosis and indication for the feasibility of treatment. Using fine-needle aspiration biopsy is performed under local anesthesia to aspirate tissue for histology. Alternatively, open biopsies can be performed under local or general anesthesia when more tissue is required. A normal biopsy in an azoospermic man with a normal FSH level suggests obstruction of the vas deferens, which may be correctable surgically. Biopsies are also used to harvest sperm for ICSI and to classify disorders such as hypospermatogenesis (all stages present but in reduced numbers), germ cell arrest (usually at primary spermatocyte stage), and Sertoli cell–only syndrome (absent germ cells) or hyalinization (sclerosis with absent cellular elements).

Testing for Y Chromosome Microdeletions Y chromosome microdeletions are detected by extracting DNA from peripheral blood leukocytes and using polymerase chain reaction (PCR) amplification using primers for some 300 sequence-tagged sites on the Y chromosome, followed by gel electrophoresis to determine whether the DNA sequences corresponding to the selected Y chromosome markers are present. However, because these ~300 Y chromosome markers account for only a small fraction of the 23 million base pairs on the Y chromosome, a negative test does not exclude microdeletions in other subregions of the Y chromosome.

DISORDERS OF SEXUAL DIFFERENTIATION

See Chap. 383.

DISORDERS OF PUBERTY

The onset and tempo of puberty varies greatly in the general population and is affected by genetic, nutritional, and environmental factors. Although a substantial fraction of the variance in the timing of puberty is explained by heritable factors, the genes involved remain unknown.

TABLE 384-1 Causes of Precocious or Delayed Puberty in Boys

- I. Precocious puberty
  - A. Gonadotropin-dependent
    - 1. Idiopathic
    - 2. Hypothalamic hamartoma or other lesions
    - 3. CNS tumor or inflammatory state
  - B. Gonadotropin-independent
    - 1. Congenital adrenal hyperplasia
    - 2. hCG-secreting tumor
    - 3. McCune-Albright syndrome
    - 4. Activating LH receptor mutation
    - 5. Exogenous androgens
    - 6. Androgen producing tumors of the adrenal or the testis
- II. Delayed puberty
  - A. Constitutional delay of growth and puberty
  - B. Systemic disorders
    - 1. Chronic disease
    - 2. Malnutrition
    - 3. Anorexia nervosa
  - C. CNS tumors and their treatment (radiotherapy and surgery)
  - D. Hypothalamic-pituitary causes of pubertal failure (low gonadotropins)
    - 1. Congenital disorders (Table 384-2)
    - 2. Acquired disorders
      - a. Pituitary tumors
      - b. Hyperprolactinemia
      - c. Infiltrative disorders, such as hemachromatosis
  - E. Gonadal causes of pubertal failure (elevated gonadotropins)
    - 1. Klinefelter syndrome
    - 2. Bilateral undescended testes
    - 3. Orchitis
    - 4. Chemotherapy or radiotherapy
    - 5. Anorchia
  - F. Androgen insensitivity

Abbreviations: CNS, central nervous system; GnRH, gonadotropin-releasing hormone; hCG, human chronic gonadotropin; LH, luteinizing hormone.

PRECOCIOUS PUBERTY

Puberty in boys aged <9 years is considered precocious. *Isosexual precocity* refers to premature sexual development consistent with phenotypic sex and includes features such as the development of facial hair and phallic growth. Isosexual precocity is divided into gonadotropin-dependent and gonadotropin-independent causes of androgen excess (Table 384-1). *Heterosexual precocity* refers to the premature development of estrogenic features in boys, such as breast development.

Gonadotropin-Dependent Precocious Puberty This disorder, called *central precocious puberty* (CPP), is less common in boys than in girls. It is caused by premature activation of the GnRH pulse generator, sometimes because of central nervous system (CNS) lesions such as hypothalamic hamartomas, but it is often idiopathic. CPP is characterized by gonadotropin levels that are inappropriately elevated for age. Because pituitary priming has occurred, GnRH elicits LH and FSH responses typical of those seen in puberty or in adults. MRI should be performed to exclude a mass, structural defect, infection, or inflammatory process. Mutations in kisspeptin, kisspeptin receptor, and MKRN3, an imprinted gene encoding makorin RING-finder protein 3, which is expressed only from the paternally inherited allele, have been associated with CPP. Loss of function mutations in MKRN3 remove the brake that restrains pulsatile GnRH, resulting in precocious puberty.

Gonadotropin-Independent Precocious Puberty Androgens from the testis or the adrenal are increased but gonadotropins are low. This group of disorders includes hCG-secreting tumors; CAH; sex steroid–producing tumors of the testis, adrenal, and ovary; accidental or deliberate exogenous sex steroid administration; hypothyroidism; and activating mutations of the LH receptor or $G_s\alpha$ subunit.

Familial Male-Limited Precocious Puberty Also called *testotoxicosis*, familial male-limited precocious puberty is an autosomal dominant disorder caused by activating mutations in the LH receptor, leading to constitutive stimulation of the cyclic AMP pathway and testosterone production. Clinical features include premature androgenization in boys, growth acceleration in early childhood, and advanced bone age followed by premature epiphyseal fusion. Testosterone is elevated, and LH is suppressed. Treatment options include inhibitors of testosterone synthesis (e.g., ketoconazole, medroxyprogesterone acetate), AR antagonists (e.g., flutamide and bicalutamide), and aromatase inhibitors (e.g., anastrozole).

MCCUNE-ALBRIGHT SYNDROME This is a sporadic disorder caused by somatic (postzygotic) activating mutations in the $G_s\alpha$ subunit that links G protein–coupled receptors to intracellular signaling pathways (Chap. 405). The mutations impair the guanosine triphosphatase activity of the $G_s\alpha$ protein, leading to constitutive activation of adenylyl cyclase. Like activating LH receptor mutations, this stimulates testosterone production and causes gonadotropin-independent precocious puberty. In addition to sexual precocity, affected individuals may have autonomy in the adrenals, pituitary, and thyroid glands. *Café au lait* spots are characteristic skin lesions that reflect the onset of the somatic mutations in melanocytes during embryonic development. Polyostotic fibrous dysplasia is caused by activation of the parathyroid hormone receptor pathway in bone. Treatment is similar to that in patients with activating LH receptor mutations. Bisphosphonates have been used to treat bone lesions.

CONGENITAL ADRENAL HYPERPLASIA Boys with CAH who are not well controlled with glucocorticoid suppression of adrenocorticotropic hormone (ACTH) can develop premature virilization because of excessive androgen production by the adrenal gland (Chaps. 379 and 383). LH is low, and the testes are small. Adrenal rests may develop within the testis of poorly controlled patients with CAH because of chronic ACTH stimulation; adrenal rests do not require surgical removal and regress with effective glucocorticoid therapy. Some children with CAH may develop gonadotropin-dependent precocious puberty with early maturation of the hypothalamic-pituitary-gonadal axis, elevated gonadotropins, and testicular growth.

Heterosexual Sexual Precocity Breast enlargement in prepubertal boys can result from familial aromatase excess, estrogen-producing tumors in the adrenal gland, Sertoli cell tumors in the testis, marijuana smoking, or exogenous estrogens or androgens. Occasionally, germ cell tumors that secrete hCG can be associated with breast enlargement due to excessive stimulation of estrogen production (see "Gynecomastia," below).

APPROACH TO THE PATIENT

Precocious Puberty

After verification of precocious development, serum testosterone, LH and FSH levels should be measured to determine whether gonadotropins are increased in relation to chronologic age (gonadotropin-dependent) or whether sex steroid secretion is occurring independent of LH and FSH (gonadotropin-independent). In children with gonadotropin-dependent precocious puberty, CNS lesions should be excluded by history, neurologic examination, and MRI scan of the head. If organic causes are not found, one is left with the diagnosis of idiopathic central precocity. Patients with high testosterone but suppressed LH concentrations have gonadotropin-independent sexual precocity; in these patients, DHEA sulfate (DHEAS) and 17α-hydroxyprogesterone should be measured. High levels of testosterone and 17α-hydroxyprogesterone suggest the possibility of CAH due to 21α-hydroxylase or 11β-hydroxylase deficiency. If testosterone and DHEAS are elevated, adrenal tumors should be excluded by obtaining a CT scan of the adrenal glands. Patients with elevated testosterone but without increased 17α-hydroxyprogesterone or DHEAS should undergo careful evaluation of the testis by palpation and ultrasound to exclude a Leydig cell neoplasm. Activating mutations of the LH receptor should be considered in children with gonadotropin-independent precocious puberty in whom CAH, androgen abuse, and adrenal and testicular neoplasms have been excluded.

TREATMENT

Precocious Puberty

In patients with a known cause (e.g., a CNS lesion or a testicular tumor), therapy should be directed towards the underlying disorder. In patients with idiopathic CPP, long-acting GnRH analogues can be used to suppress gonadotropins and decrease testosterone, halt early pubertal development, delay accelerated bone maturation, prevent early epiphyseal closure, promote final height gain, and mitigate the psychosocial consequences of early pubertal development without causing osteoporosis. The treatment is most effective for increasing final adult height if it is initiated before age 6. Puberty resumes after discontinuation of the GnRH analogue. Counseling is an important aspect of the overall treatment strategy.

In children with gonadotropin-independent precocious puberty, inhibitors of steroidogenesis, such as ketoconazole, AR antagonists, and aromatase inhibitors have been used empirically. Long-term treatment with spironolactone (a weak androgen antagonist) and ketoconazole has been reported to normalize growth rate and bone maturation and to improve predicted height in small, nonrandomized trials in boys with familial male-limited precocious puberty. Aromatase inhibitors, such as testolactone and letrozole, have been used as adjuncts to antiandrogen therapy for children with familial male-limited precocious puberty, CAH, and McCune-Albright syndrome. More potent novel inhibitors of testosterone synthesis, such as abiraterone, have not been evaluated in boys with gonadotropin-independent precocious puberty.

DELAYED PUBERTY

Puberty is delayed in boys if it has not ensued by age 14, an age that is 2–2.5 standard deviations above the mean for healthy children. Delayed puberty is more common in boys than in girls. There are four main categories of delayed puberty: (1) constitutional delay of growth and puberty (~60% of cases); (2) functional hypogonadotropic hypogonadism caused by systemic illness or malnutrition (~20% of cases); (3) hypogonadotropic hypogonadism caused by genetic or acquired defects in the hypothalamic-pituitary region (~10% of cases); and (4) hypergonadotropic hypogonadism secondary to primary gonadal failure (~15% of cases) (Table 384-1). The constitutional delay of growth and puberty clusters in families displays an autosomal dominant pattern of inheritance, and has been linked in some families with a locus on pericentromeric region of chromosome 2. Functional hypogonadotropic hypogonadism is more common in girls than in boys. Permanent causes of hypogonadotropic or hypergonadotropic hypogonadism are identified in <25% of boys with delayed puberty.

APPROACH TO THE PATIENT

Delayed Puberty

History of systemic illness, eating disorders, excessive exercise, social and psychological problems, and abnormal patterns of linear growth during childhood should be verified. Boys with pubertal delay may have accompanying emotional and physical immaturity relative to their peers, which can be a source of anxiety. Physical examination should focus on height; arm span; weight; visual fields; and secondary sex characteristics, including hair growth, testicular volume, phallic size, and scrotal reddening and thinning. Testicular size >2.5 cm generally indicates that the child has entered puberty.

The main diagnostic challenge is to distinguish those with constitutional delay, who will progress through puberty at a later age,

from those with an underlying pathologic process. Constitutional delay should be suspected when there is a family history and when there are delayed bone age and short stature. Pituitary priming by pulsatile GnRH is required before LH and FSH are synthesized and secreted normally. Thus, blunted responses to exogenous GnRH can be seen in patients with constitutional delay, GnRH deficiency, or pituitary disorders. On the other hand, low-normal basal gonadotropin levels or a normal response to exogenous GnRH is consistent with an early stage of puberty, which is often heralded by nocturnal GnRH secretion. Thus, constitutional delay is a diagnosis of exclusion that requires ongoing evaluation until the onset of puberty and the growth spurt.

TREATMENT

Delayed Puberty

If therapy is considered appropriate, it can begin with 25–50 mg testosterone enanthate or testosterone cypionate every 2 weeks, or by using a 2.5-mg testosterone patch or 25-mg testosterone gel. Because aromatization of testosterone to estrogen is obligatory for mediating androgen effects on epiphyseal fusion, concomitant treatment with aromatase inhibitors may allow attainment of greater final adult height. Testosterone treatment should be interrupted after 6 months to determine if endogenous LH and FSH secretion have ensued. Other causes of delayed puberty should be considered when there are associated clinical features or when boys do not enter puberty spontaneously after a year of observation or treatment.

Reassurance without hormonal treatment is appropriate for many individuals with presumed constitutional delay of puberty. However, the impact of delayed growth and pubertal progression on a child's social relationships and school performance should be weighed. The boys with constitutional delay of puberty are less likely to achieve their full genetic height potential and have reduced total body bone mass as adults, mainly due to narrow limb bones and vertebrae as a result of impaired periosteal expansion during puberty. Furthermore, the time of onset of puberty is negatively associated with bone mineral content and density in boys at skeletal maturity. Judicious use of androgen therapy in carefully selected boys with constitutional delay can induce pubertal induction and progression, and promote short-term growth without compromising final height, and when administered with an aromatase inhibitor, it may improve final height.

DISORDERS OF THE MALE REPRODUCTIVE AXIS DURING ADULTHOOD

HYPOGONADOTROPIC HYPOGONADISM

Because LH and FSH are trophic hormones for the testes, impaired secretion of these pituitary gonadotropins results in secondary hypogonadism, which is characterized by low testosterone in the setting of low or inappropriately normal LH and FSH. Those with the most severe gonadotropin deficiency have complete absence of pubertal development, sexual infantilism, and, in some cases, hypospadias and undescended testes. Patients with partial gonadotropin deficiency have delayed or arrested sex development. The 24-h LH secretory profiles are heterogeneous in patients with hypogonadotropic hypogonadism, reflecting variable abnormalities of LH pulse frequency or amplitude. In severe cases, basal LH is low and there are no LH pulses. A smaller subset of patients has low-amplitude LH pulses or markedly reduced pulse frequency. Occasionally, only sleep-entrained LH pulses occur, reminiscent of the pattern seen in the early stages of puberty. Hypogonadotropic hypogonadism can be classified into congenital and acquired disorders. Congenital disorders most commonly involve GnRH deficiency, which leads to gonadotropin deficiency. Acquired disorders are much more common than congenital disorders and may result from a variety of sellar mass lesions or infiltrative diseases of the hypothalamus or pituitary, or due to the effects of drugs, nutritional or psychiatric disorders, or systemic diseases.

Congenital Disorders Associated with Gonadotropin Deficiency Congenital hypogonadotropic hypogonadism is a heterogeneous group of disorders characterized by decreased gonadotropin secretion and testicular dysfunction either due to impaired function of the GnRH pulse generator or the gonadotrope. The disorders characterized by GnRH deficiency represent a family of oligogenic disorders whose phenotype spans a wide spectrum. Some individuals with GnRH deficiency may suffer from complete absence of pubertal development, while others may manifest varying degrees of gonadotropin deficiency and pubertal delay, and a subset that carries the same mutations as their affected family members may even have normal reproductive function. In ~10% of men with idiopathic hypogonadotropic hypogonadism (IHH), reversal of gonadotropin deficiency may occur in adult life after sex steroid therapy. Also, a small fraction of men with IHH may present with androgen deficiency and infertility in adult life after having gone through apparently normal pubertal development. Nutritional, emotional, or metabolic stress may unmask gonadotropin deficiency and reproductive dysfunction (e.g., hypothalamic amenorrhea) in some patients who harbor mutations in the candidate genes but who previously had normal reproductive function. The clinical phenotype may include isolated anosmia or hyposmia. Oligogenicity, and gene-gene and gene-environment interactions may contribute to variations in clinical phenotype.

Mutations in a number of genes involved in the development and migration of GnRH neurons, or in the regulation of GnRH secretion have been linked to GnRH deficiency, although the genetic defect remains elusive in nearly two thirds of cases. Familial hypogonadotropic hypogonadism can be transmitted as an X-linked (20%), autosomal recessive (30%), or autosomal dominant (50%) trait. Some individuals with IHH have sporadic mutations in the same genes that cause inherited forms of the disorder. The genetic defects associated with GnRH deficiency can ben conveniently classified as anosmic (Kallmann syndrome) or normosmic **(Table 384-2)**, although the occurrence of both anosmic and normosmic forms of GnRH deficiency in the same families suggests commonality of pathophysiologic mechanisms. *Kallmann syndrome*, the anosmic form of GnRH deficiency, can result from mutations in one or more genes associated with olfactory bulb morphogenesis and the migration of GnRH neurons from their origin in the region of the olfactory placode, along the scaffold established by the olfactory nerves, through the cribriform plate into their final location into the pre-optic region of the hypothalamus. Thus, mutations in *KAL1*, genes involved in fibroblast growth factor (FGF) signaling (*FGF8*, *FGFR1*, *FGF17*, *IL17RD*, *DUSP6*, *SPRY4*, and *FLRT3*), *NELF*, genes involved in *PROK* signaling (*PROK2* and *PROK2R*), *WDR11*, *SEMA3*, *HS6ST1*, *CHD7*, and *FEZF1* have been described in patients with Kallmann syndrome. An X-linked form of IHH is caused by mutations in the *KAL1* gene, which encodes anosmin, a protein that mediates the migration of neural progenitors of the olfactory bulb and GnRH-producing neurons. These individuals have GnRH deficiency and variable combinations of anosmia or hyposmia, renal defects, and neurologic abnormalities including mirror movements. Proteins such as those involved in FGF and prokineticin signaling, and *KAL1*, which account for the great majority of Kallmann syndrome cases, interact with heparin sulfate glycosominoglycan compounds within the extracellular matrix in promoting GnRH neuronal migration. Mutations in the *FGFR1* gene cause an autosomal dominant form of hypogonadotropic hypogonadism that clinically resembles Kallmann syndrome; mutations in its putative ligand, the *FGF8* gene product have also been associated with IHH. Craniofacial tissues and olfactory ensheathing cells also play important roles in neurogenesis and migration of the GnRH neurons, and additional proteins that regulate these cell types may also be involved in the pathogenesis of Kallmann syndrome. The co-occurrence of tooth anomalies, cleft palate, craniofacial anomalies, pigmentation, and other neurological defects in patients with Kallmann Syndrome suggest that the syndrome may be a part of the spectrum of neurocristopathies.

TABLE 384-2 Causes of Congenital Hypogonadotropic Hypogonadism

GENE	LOCUS	INHERITANCE	ASSOCIATED FEATURES
A. Hypogonadotropic Hypogonadism due to GnRH Deficiency			
A1. GnRH Deficiency Associated with Hyposmia or Anosmia			
KAL1	Xp22	X-linked	Anosmia, renal agenesis, synkinesia, cleft lip/palate, oculomotor/visuospatial defects, gut malformations
NELF	9q34.3	AR	Anosmia, hypogonadotropic hypogonadism
FGF8	10q24	AR	Anosmia (some patients may be normosmic), skeletal abnormalities
FGFR1	8p11-p12	AD	Anosmia, cleft lip/palate, synkinesia, syndactyly
PROK2	3p21	AR	Anosmia/ sleep dysregulation
PROK2R	20p12.3	AR	Variable
CHD7	8q12.1		Anosmia, other features of CHARGE syndrome
FEZ1	8p22	AR	Anosmia, olfactory bulb aplasia
WDR11	10q26	AD	Anosmia
SOX10	22q13		Deafness
SEMA3A	7q21		Anosmia; some persons with mutations are normal
HS6ST1	2q14	complex	Anosmia
A2. GnRH Deficiency with Normal Sense of Smell			
GNRHR	4q21	AR	None
GnRH1	8p21	AR	None
KISS1R	19p13	AR	None
TAC3	12q13	AR	Microphallus, cryptorchidism, reversal of GnRH deficiency
TAC3R	4q25	AR	Microphallus, cryptorchidism, reversal of GnRH deficiency
LEPR	1p31	AR	Obesity
LEP	7q31	AR	Obesity
B. Hypogonadotropic Hypogonadism not due to GnRH Deficiency			
PC1	5q15-21	AR	Obesity, diabetes mellitus, ACTH deficiency
HESX1	3p21	AR AD	Septooptic dysplasia, CPHD Isolated GH insufficiency
LHX3	9q34	AR	CPHD (ACTH spared), cervical spine rigidity
PROP1	5q35	AR	CPHD (ACTH usually spared)
FSHβ	11p13	AR	↑ LH
LHβ	19q13	AR	↑ FSH
SF1 (NR5A1)	9p33	AD/AR	Primary adrenal failure, XY sex reversal

Abbreviations: ACTH, adrenocorticotropic hormone; AD, autosomal dominant; AR, autosomal recessive; CHARGE syndrome: eye coloboma, heart defects, choanal atresia, growth and developmental retardation, genitourinary anomalies, ear anomalies; CPHD, combined pituitary hormone deficiency; DAX1, *d*osage-sensitive sex-reversal, *a*drenal hypoplasia congenita, *X*-chromosome; FGFR1, fibroblast growth factor receptor 1; FSHβ, follicle-stimulating hormone β-subunit; GNRHR, gonadotropin-releasing hormone receptor; GPR54, G protein–coupled receptor 54; HESX1, homeo box gene expressed in embryonic stem cells 1; KAL1, Kallmann Syndrome Interval Gene 1, also known as anosmin 1; LEP, leptin; LEPR, leptin receptor; LHX3, LIM homeobox gene 3; LHβ, luteinizing hormone β-subunit; NELF, nasal embryonic LHRH factor; PC1, prohormone convertase 1; PROK2, prokineticin 2; PROP1, Prophet of Pit 1; SF1, steroidogenic factor 1.

Normosmic GnRH deficiency results from defects in pulsatile GnRH secretion, its regulation, or its action on the gonadotrope and has been associated with mutations in GnRHR, GNRH1, KISS1R, TAC3, TACR3, NROB1 (DAX1). Some mutations, such as those in PROK2, PROKR2, and CHD7 have been associated with both anosmic as well as normosmic form of IHH. *GnRH receptor* mutations, the most frequent identifiable cause of normosmic IHH, account for ~40% of autosomal recessive and 10% of sporadic cases of hypogonadotropic hypogonadism. These patients have decreased LH response to exogenous GnRH. Some receptor mutations alter GnRH binding affinity, allowing apparently normal responses to pharmacologic doses of exogenous GnRH, whereas other mutations may alter signal transduction downstream of hormone binding. Mutations of the *GnRH1* gene have also been reported in patients with hypogonadotropic hypogonadism, although they are rare. G protein–coupled receptor *KISS1R* (GPR54) and its cognate ligand, kisspeptin (*KISS1*), are important regulators of sexual maturation in primates. Recessive mutations in GPR54 cause gonadotropin deficiency without anosmia. Patients retain responsiveness to exogenous GnRH, suggesting an abnormality in the neural pathways controlling GnRH release. The genes encoding neurokinin B (TAC3), which is involved in preferential activation of GnRH release in early development, and its receptor (TAC3R) have been implicated in some families with normosmic IHH. X-linked hypogonadotropic hypogonadism also occurs in *adrenal hypoplasia congenita*, a disorder caused by mutations in the *DAX1* gene, which encodes a nuclear receptor in the adrenal gland and reproductive axis. Adrenal hypoplasia congenita is characterized by absent development of the adult zone of the adrenal cortex, leading to neonatal adrenal insufficiency. Puberty usually does not occur or is arrested, reflecting variable degrees of gonadotropin deficiency. Although sexual differentiation is normal, some patients have testicular dysgenesis and impaired spermatogenesis despite gonadotropin replacement. Less commonly, adrenal hypoplasia congenita, sex reversal, and hypogonadotropic hypogonadism can be caused by mutations of steroidogenic factor 1 (SF1). Rarely, recessive mutations in the *LHβ* or *FSHβ* genes have been described in patients with selective deficiencies of these gonadotropins.

A number of homeodomain transcription factors are involved in the development and differentiation of the specialized hormone-producing cells within the pituitary gland (Table 384-2). Patients with mutations of *PROP1* have combined pituitary hormone deficiency that includes GH, prolactin (PRL) thyroid-stimulating hormone (TSH), LH, and FSH, but not ACTH. *LHX3* mutations cause combined pituitary hormone deficiency in association with cervical spine rigidity. *HESX1* mutations cause septooptic dysplasia and combined pituitary hormone deficiency. Mutations of *ARNT1*, inherited as an autosomal recessive disorder, are associated with diabetes insipidus, ACTH deficiency, GH, LH, FSH deficiency, anterior pituitary hypoplasia, hypoplastic frontal and temporal lobes, thin corpus callosum, prominent forehead, and retrognathia. Patients with *SOX2* mutations can have gonadotropin deficiency, variable deficiencies of TSH and ACTH, pituitary hypoplasia, microphthalmia, and intellectual disability.

Prader-Willi syndrome is characterized by obesity, hypotonic musculature, mental retardation, hypogonadism, short stature, and small hands and feet. Prader-Willi syndrome is a genomic imprinting disorder caused by deletions of the proximal portion of paternally derived chromosome 15q11-15q13 region, which contains a bipartite imprinting center; uniparental disomy of the maternal alleles; or mutations of the genes/loci involved in imprinting (Chap. 456). *Laurence-Moon syndrome* is an autosomal recessive disorder characterized by obesity, hypogonadism, mental retardation, polydactyly, and retinitis pigmentosa. Recessive mutations of leptin, or its receptor, cause severe obesity and pubertal arrest, apparently because of hypothalamic GnRH deficiency (Chap. 394).

Acquired Hypogonadotropic Disorders • **SEVERE ILLNESS, STRESS, MALNUTRITION, AND EXERCISE** These may cause reversible gonadotropin deficiency. Although gonadotropin deficiency and reproductive dysfunction are well documented in these conditions in women, men exhibit similar but less-pronounced responses. Unlike women, most male runners and other endurance athletes have normal gonadotropin and sex steroid levels, despite low body fat and frequent intensive exercise. Testosterone levels fall at the onset of illness and recover during recuperation. The magnitude of gonadotropin suppression generally correlates with the severity of illness. Although hypogonadotropic hypogonadism is the most common cause of androgen deficiency in patients with acute illness, some have elevated levels of LH and FSH, which suggest primary gonadal dysfunction. The pathophysiology of reproductive dysfunction during acute illness is

unknown but likely involves a combination of cytokine and/or glucocorticoid effects. There is a high frequency of low testosterone levels in patients with chronic illnesses such as HIV infection, end-stage renal disease, chronic obstructive lung disease, and many types of cancer and in patients receiving glucocorticoids. About 20% of HIV-infected men with low testosterone levels have elevated LH and FSH levels; these patients presumably have primary testicular dysfunction. The remaining 80% have either normal or low LH and FSH levels; these men have a central hypothalamic-pituitary defect or a dual defect involving both the testis and the hypothalamic-pituitary centers. Muscle wasting is common in chronic diseases associated with hypogonadism, which also leads to debility, poor quality of life, and adverse outcome of disease. There is great interest in exploring strategies that can reverse androgen deficiency or attenuate the sarcopenia associated with chronic illness.

Men using opioids for relief of cancer or noncancerous pain or because of addiction often have suppressed testosterone and LH levels and high prevalence of sexual dysfunction and osteoporosis; the degree of suppression is dose-related and particularly severe with long acting opioids such as methadone. Opioids suppress GnRH secretion and alter the sensitivity to feedback inhibition by gonadal steroids. Men who are heavy users of marijuana have decreased testosterone secretion and sperm production. The mechanism of marijuana-induced hypogonadism is decreased GnRH secretion. Gynecomastia observed in marijuana users can also be caused by plant estrogens in crude preparations. Androgen deprivation therapy in men with prostate cancer has been associated with increased risk of bone fractures, diabetes mellitus, cardiovascular events, fatigue, sexual dysfunction, tender gynecomastia, and poor quality of life.

OBESITY In men with mild to moderate obesity, SHBG levels decrease in proportion to the degree of obesity, resulting in lower total testosterone levels. However, free testosterone levels usually remain within the normal range. SHBG production in the liver is inhibited by hepatic lipids, and by TNF-α and interleukin-1, but it is not affected by insulin. Thus, the low SHBG levels seen in obesity and diabetes are likely the result of low grade inflammation and the increased amount of hepatic lipids rather than high insulin levels. Estradiol levels are higher in obese men compared to healthy, nonobese controls, because of aromatization of testosterone to estradiol in adipose tissue. Weight loss is associated with reversal of these abnormalities including an increase in total and free testosterone levels and a decrease in estradiol levels. A subset of obese men with moderate to severe obesity may have a defect in the hypothalamic-pituitary axis as suggested by low free testosterone in the absence of elevated gonadotropins. Weight gain in adult men can accelerate the rate of age-related decline in testosterone levels.

HYPERPROLACTINEMIA (See also Chap. 373) Elevated PRL levels are associated with hypogonadotropic hypogonadism. PRL inhibits hypothalamic GnRH secretion either directly or through modulation of tuberoinfundibular dopaminergic pathways. A PRL-secreting tumor may also destroy the surrounding gonadotropes by invasion or compression of the pituitary stalk. Treatment with dopamine agonists reverses gonadotropin deficiency, although there may be a delay relative to PRL suppression.

SELLAR MASS LESIONS Neoplastic and nonneoplastic lesions in the hypothalamus or pituitary can directly or indirectly affect gonadotrope function. In adults, pituitary adenomas constitute the largest category of space-occupying lesions affecting gonadotropin and other pituitary hormone production. Pituitary adenomas that extend into the suprasellar region can impair GnRH secretion and mildly increase PRL secretion (usually <50 μg/L) because of impaired tonic inhibition by dopaminergic pathways. These tumors that cause hyperprolactinemia by stalk compression should be distinguished from prolactinomas, which typically are associated with higher PRL levels. The presence of diabetes insipidus suggests the possibility of a craniopharyngioma, infiltrative disorder, or other hypothalamic lesions (Chap. 374).

HEMOCHROMATOSIS (See also Chap. 407) Both the pituitary and testis can be affected by excessive iron deposition. However, the pituitary defect is the predominant lesion in most patients with hemochromatosis and hypogonadism. The diagnosis of hemochromatosis is suggested by the association of characteristic skin discoloration, hepatic enlargement or dysfunction, diabetes mellitus, arthritis, cardiac conduction defects, and hypogonadism.

PRIMARY TESTICULAR CAUSES OF HYPOGONADISM

Common causes of primary testicular dysfunction include Klinefelter syndrome, uncorrected cryptorchidism, cancer chemotherapy, radiation to the testes, trauma, torsion, infectious orchitis, HIV infection, anorchia syndrome, and myotonic dystrophy. Primary testicular disorders may be associated with impaired spermatogenesis, decreased androgen production, or both. **See Chap. 383 for disorders of testis development, androgen synthesis, and androgen action.**

Klinefelter Syndrome (See also Chap. 383) Klinefelter syndrome is the most common chromosomal disorder associated with testicular dysfunction and male infertility. It occurs in about 1 in 600 live-born males. Azoospermia is the rule in men with Klinefelter syndrome who have the 47,XXY karyotype; however, men with mosaicism may have germ cells, especially at a younger age. The clinical phenotype of Klinefelter syndrome can be variable, possibly because of mosaicism, polymorphisms in AR gene, the parental origin of the X chromosome, X-linked copy number variations, gene-dosage effects in conjunction with X chromosome inactivation, variable testosterone levels, or other genetic factors. Testicular histology shows hyalinization of seminiferous tubules and absence of spermatogenesis. Although their function is impaired, the number of Leydig cells appears to increase. Testosterone is decreased and estradiol is increased, leading to clinical features of undervirilization and gynecomastia. Men with Klinefelter syndrome are at increased risk of systemic lupus erythematosus, Sjögren's syndrome, breast cancer, diabetes mellitus, osteoporosis, non-Hodgkin's lymphoma, and some types of lung cancer, and reduced risk of prostate cancer. Periodic mammography for breast cancer surveillance is recommended for men with Klinefelter syndrome. Fertility can be achieved by intracytoplasmic injection of sperm retrieved surgically from the testes of men with Klinefelter syndrome, including some men with nonmosaic form of Klinefelter syndrome. The karyotypes 48,XXXY and 49,XXXXY are associated with a more severe phenotype, increased risk of congenital malformations, and lower intelligence than 47,XXY individuals.

Cryptorchidism Cryptorchidism occurs when there is incomplete descent of the testis from the abdominal cavity into the scrotum. About 3% of full-term and 30% of premature male infants have at least one undescended testis at birth, but descent is usually complete by the first few weeks of life. The incidence of cryptorchidism is <1% by 9 months of age. Androgens regulate predominantly the inguinoscrotal descent of the testes through degeneration of the cranio-suspensory ligament and a shortening of the gubernaculums, respectively. Mutations in INSL3 and leucine-rich repeat family of G-protein-coupled receptor 8 (LGR8), which regulate transabdominal portion of testicular descent, have been found in some patients with cryptorchidism.

Cryptorchidism is associated with increased risk of malignancy, infertility, inguinal hernia, and torsion. Unilateral cryptorchidism, even when corrected before puberty, is associated with decreased sperm count, possibly reflecting unrecognized damage to the fully descended testis or other genetic factors. Epidemiologic, clinical, and molecular evidence supports the idea that cryptorchidism, hypospadias, impaired spermatogenesis, and testicular cancer may be causally related to common genetic and environment perturbations, and are components of the testicular dysgenesis syndrome.

Acquired Testicular Defects *Viral orchitis* may be caused by the mumps virus, echovirus, lymphocytic choriomeningitis virus, and group B arboviruses. Orchitis occurs in as many as one-fourth of adult men with mumps; the orchitis is unilateral in about two-thirds and bilateral in the remainder. Orchitis usually develops a few days after the onset of parotitis but may precede it. The testis may return to normal size and function or undergo atrophy. Semen analysis returns

to normal for three-fourths of men with unilateral involvement but normal for only one-third of men with bilateral orchitis. *Trauma*, including testicular torsion, can also cause secondary atrophy of the testes. The exposed position of the testes in the scrotum renders them susceptible to both thermal and physical trauma, particularly in men with hazardous occupations.

The testes are sensitive to *radiation damage*. Doses >200 mGy (20 rad) are associated with increased FSH and LH levels and damage to the spermatogonia. After ~800 mGy (80 rad), oligospermia or azoospermia develops, and higher doses may obliterate the germinal epithelium. Permanent androgen deficiency in adult men is uncommon after therapeutic radiation; however, most boys given direct testicular radiation therapy for acute lymphoblastic leukemia have permanently low testosterone levels. Sperm banking should be considered before patients undergo radiation treatment or chemotherapy.

Drugs interfere with testicular function by several mechanisms, including inhibition of testosterone synthesis (e.g., ketoconazole), blockade of androgen action (e.g., spironolactone), increased estrogen (e.g., marijuana), or direct inhibition of spermatogenesis (e.g., chemotherapy).

Combination chemotherapy for acute leukemia, Hodgkin's disease, and testicular and other cancers may impair Leydig cell function and cause infertility. The degree of gonadal dysfunction depends on the type of chemotherapeutic agent and the dose and duration of therapy. Because of high response rates and the young age of these men, infertility and androgen deficiency have emerged as important long-term complications of cancer chemotherapy. Cyclophosphamide and combination regimens containing procarbazine are particularly toxic to germ cells. Thus, 90% of men with Hodgkin's lymphoma receiving MOPP (mechlorethamine, oncovin, procarbazine, prednisone) therapy develop azoospermia or extreme oligozoospermia; newer regimens that do not include procarbazine, such as ABVD (adriamycin, bleomycin, vinblastine, dacarbazine), are less toxic to germ cells.

Alcohol, when consumed in excess for prolonged periods, decreases testosterone, independent of liver disease or malnutrition. Elevated estradiol and decreased testosterone levels may occur in men taking digitalis.

The occupational and recreational history should be carefully evaluated in all men with infertility because of the toxic effects of many *chemical agents* on spermatogenesis. Known environmental hazards include pesticides (e.g., vinclozolin, dicofol, atrazine), sewage contaminants (e.g., ethinyl estradiol in birth control pills, surfactants such as octylphenol, nonyphenol), plasticizers (e.g., pthalates), flame retardants (PCBs, polybrominated diphenol ethers), industrial pollutants (e.g., heavy metals cadmium and lead, dioxins, polycyclic aromatic hydrocarbons), microwaves and ultrasound. In some populations, sperm density is said to have declined by as much as 40% in the past 50 years. Environmental estrogens or antiandrogens may be partly responsible.

Testicular failure also occurs as a part of *polyglandular autoimmune insufficiency* **(Chap. 381)**. Sperm antibodies can cause isolated male infertility. In some instances, these antibodies are secondary phenomena resulting from duct obstruction or vasectomy. Granulomatous diseases can affect the testes, and testicular atrophy occurs in 10–20% of men with lepromatous leprosy because of direct tissue invasion by the mycobacteria. The tubules are involved initially, followed by endarteritis and destruction of Leydig cells.

Systemic disease can cause primary testis dysfunction in addition to suppressing gonadotropin production. In cirrhosis, a combined testicular and pituitary abnormality leads to decreased testosterone production independent of the direct toxic effects of ethanol. Impaired hepatic extraction of adrenal androstenedione leads to extraglandular conversion to estrone and estradiol, which partially suppresses LH. Testicular atrophy and gynecomastia are present in approximately one-half of men with cirrhosis. In chronic renal failure, androgen synthesis and sperm production decrease despite elevated gonadotropins. The elevated LH level is due to reduced clearance, but it does not restore normal testosterone production. About one-fourth of men with renal failure have hyperprolactinemia. Improvement in testosterone production with hemodialysis is incomplete, but successful renal transplantation may return testicular function to normal. Testicular atrophy is present in one-third of men with sickle cell anemia. The defect may be at either the testicular or the hypothalamic-pituitary level. Sperm density can decrease temporarily after acute febrile illness in the absence of a change in testosterone production. Infertility in men with celiac disease is associated with a hormonal pattern typical of androgen resistance, namely elevated testosterone and LH levels.

Neurologic diseases associated with altered testicular function include myotonic dystrophy, spinobulbar muscular atrophy, and paraplegia. In myotonic dystrophy, small testes may be associated with impairment of both spermatogenesis and Leydig cell function. Spinobulbar muscular atrophy is caused by an expansion of the glutamine repeat sequences in the amino-terminal region of the AR; this expansion impairs function of the AR, but it is unclear how the alteration is related to the neurologic manifestations. Men with spinobulbar muscular atrophy often have undervirilization and infertility as a late manifestation. Spinal cord injury that causes paraplegia is often associated with low testosterone levels and may cause persistent defects in spermatogenesis; some patients retain the capacity for penile erection and ejaculation.

ANDROGEN INSENSITIVITY SYNDROMES

Mutations in the AR cause resistance to the action of testosterone and DHT. These X-linked mutations are associated with variable degrees of defective male phenotypic development and undervirilization **(Chap. 383)**. Although not technically hormone-insensitivity syndromes, two genetic disorders impair testosterone conversion to active sex steroids. Mutations in the *SRD5A2* gene, which encodes 5α-reductase type 2, prevent the conversion of testosterone to DHT, which is necessary for the normal development of the male external genitalia. Mutations in the *CYP19* gene, which encodes aromatase, prevent testosterone conversion to estradiol. Males with *CYP19* mutations have delayed epiphyseal fusion, tall stature, eunuchoid proportions, visceral adiposity, and osteoporosis, consistent with evidence from an estrogen receptor–deficient individual that these testosterone actions are mediated via estrogen.

GYNECOMASTIA

Gynecomastia refers to enlargement of the male breast. It is caused by excess estrogen action and is usually the result of an increased estrogen/androgen ratio. True gynecomastia is associated with glandular breast tissue that is >4 cm in diameter and often tender. Glandular tissue enlargement should be distinguished from excess adipose tissue: glandular tissue is firmer and contains fibrous-like cords. Gynecomastia occurs as a normal physiologic phenomenon in the newborn (due to transplacental transfer of maternal and placental estrogens), during puberty (high estrogen to androgen ratio in early stages of puberty), and with aging (increased fat tissue and increased aromatase activity along with the age-related decline in testosterone levels), but it can also result from pathologic conditions associated with androgen deficiency or estrogen excess. The prevalence of gynecomastia increases with age and body mass index (BMI), likely because of increased aromatase activity in adipose tissue. Medications that alter androgen metabolism or action may also cause gynecomastia. The relative risk of breast cancer is increased in men with gynecomastia, although the absolute risk is relatively small.

PATHOLOGIC GYNECOMASTIA

Any cause of *androgen deficiency* can lead to gynecomastia, reflecting an increased estrogen/androgen ratio, as estrogen synthesis still occurs by aromatization of residual adrenal and gonadal androgens. Gynecomastia is a characteristic feature of Klinefelter syndrome **(Chap. 383)**. *Androgen insensitivity* disorders also cause gynecomastia. *Excess estrogen production* may be caused by tumors, including Sertoli cell tumors in isolation or in association with Peutz-Jegher syndrome or Carney complex. Tumors that produce hCG, including some testicular tumors, stimulate Leydig cell estrogen synthesis. *Increased conversion of androgens to estrogens* can be a result of increased availability of substrate (androstenedione) for extraglandular estrogen formation

(CAH, hyperthyroidism, and most feminizing adrenal tumors) or to diminished catabolism of androstenedione (liver disease) so that estrogen precursors are shunted to aromatase in peripheral sites. Obesity is associated with increased aromatization of androgen precursors to estrogens. Extraglandular aromatase activity can also be increased in tumors of the liver or adrenal gland or rarely as an inherited disorder. Several families with *increased peripheral aromatase activity* inherited as an autosomal dominant or as an X-linked disorder have been described. In some families with this disorder, an inversion in chromosome 15q21.2-3 causes the CYP19 gene to be activated by the regulatory elements of contiguous genes resulting in excessive estrogen production in the fat and other extragonadal tissues. The familial aromatase excess syndrome due to CYP19 mutation or chromosomal rearrangement is characterized by pre- or peripubertal onset of gynecomastia, advanced bone age, short adult height due to premature epiphyseal closure, and hypogonadotropic hypogonadism. *Drugs* can cause gynecomastia by acting directly as estrogenic substances (e.g., oral contraceptives, phytoestrogens, digitalis), inhibiting androgen synthesis (e.g., ketoconazole), or action (e.g., spironolactone); for many drugs, such as cimetidine, imatinib, or some antiretroviral drugs for HIV, the precise mechanism is unknown.

Because up to two-thirds of pubertal boys and about half of hospitalized men have palpable glandular tissue that is benign, detailed investigation or intervention is not indicated in all men presenting with gynecomastia **(Fig. 384-5)**. In addition to the extent of gynecomastia, recent onset, rapid growth, tender tissue, and occurrence in a lean subject should prompt more extensive evaluation. This should include a careful drug history, measurement and examination of the testes, assessment of virilization, evaluation of liver function, and hormonal measurements including testosterone, estradiol, and androstenedione, LH, and hCG. A karyotype should be obtained in men with very small testes to exclude Klinefelter syndrome. In spite of extensive evaluation, the etiology is established in fewer than one-half of patients.

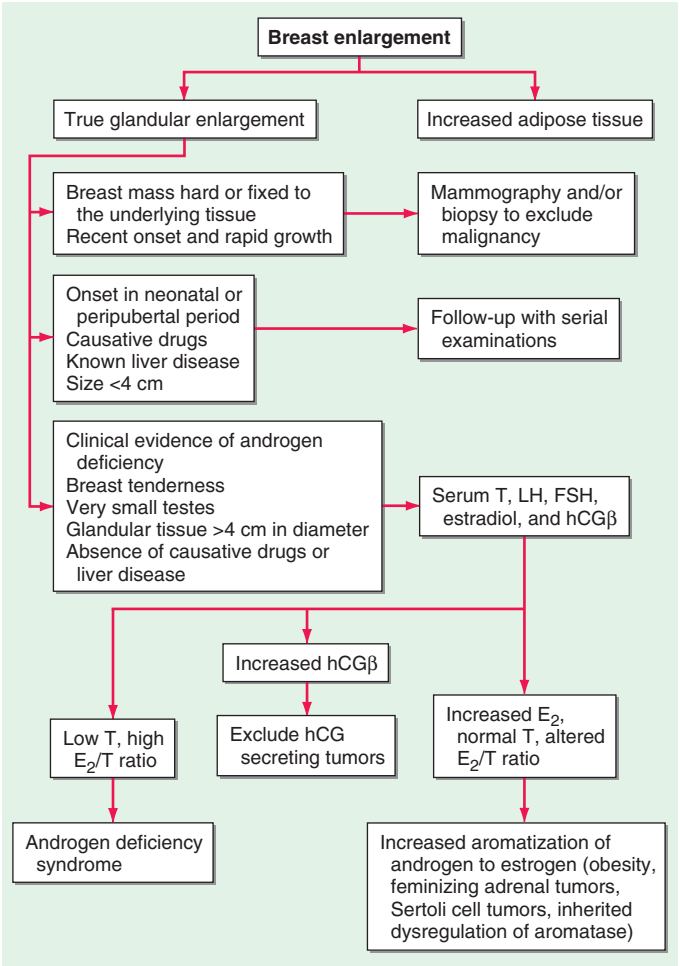


FIGURE 384-5 Evaluation of gynecomastia. E_2, 17β-estradiol; FSH, follicle-stimulating hormones; hCGβ, human chorionic gonadotropin β; LH, luteinizing hormone; T, testosterone.

TREATMENT

Gynecomastia

When the primary cause can be identified and corrected shortly after the onset of gynecomastia, breast enlargement usually subsides over several months. However, if gynecomastia is of long duration, surgery is the most effective therapy. Indications for surgery include severe psychological and/or cosmetic problems, continued growth or tenderness, or suspected malignancy. In patients who have painful gynecomastia and in whom surgery cannot be performed, treatment with antiestrogens such as tamoxifen (20 mg/d) can reduce pain and breast tissue size in over half the patients. Estrogen receptor antagonists, tamoxifen and raloxifen, have been reported in small trials to reduce breast size in men with pubertal gynecomastia, although complete regression of breast enlargement is unusual with the use of estrogen receptor antagonists. Aromatase inhibitors can be effective in the early proliferative phase of the disorder. However, in a randomized trial in men with established gynecomastia, anastrozole proved no more effective than placebo in reducing breast size. Tamoxifen is effective in prevention and treatment of breast enlargement and breast pain in men with prostate cancer who are receiving anti-androgen therapy.

AGING-RELATED CHANGES IN MALE REPRODUCTIVE FUNCTION

A number of cross-sectional and longitudinal studies (e.g., The Baltimore Longitudinal Study of Aging, the Framingham Heart Study, the Massachusetts Male Aging Study, and the European Male Aging Study [EMAS]) have established that testosterone concentrations decrease with advancing age. This age-related decline starts in the third decade of life and progresses slowly; the rate of decline in testosterone concentrations is greater in obese men, in men with chronic illness, and in those taking medications. Because SHBG concentrations are higher in older men than in younger men, free or bioavailable testosterone concentrations decline with aging to a greater extent than total testosterone concentrations. The age-related decline in testosterone is due to defects at all levels of the hypothalamic-pituitary-testicular axis: pulsatile GnRH secretion is attenuated, LH response to GnRH is reduced, and testicular response to LH is impaired. However, the gradual rise of LH with aging suggests that testis dysfunction is the main cause of declining androgen levels. The term *andropause* has been used to denote age-related decline in testosterone concentrations; this term is a misnomer because there is no discrete time when testosterone concentrations decline abruptly.

Several epidemiologic studies, such as the Framingham Heart Study, the EMAS, and the Study of Osteoporotic Fractures in Men (MrOS) that used mass spectrometry for measuring testosterone levels have reported ~10% prevalence of low testosterone levels in middle-aged and older men; the prevalence of unequivocally low testosterone and sexual symptoms in men aged 40–70 years in the EMAS was 2.1%, and increased with age from 0.1% for men aged 40–49 years of age to 5.1% for those aged 70–79 years. The age-related decline in testosterone should be distinguished from classical hypogonadism due to diseases of the testes, the pituitary, and the hypothalamus. Low total and bioavailable testosterone concentrations have been associated with decreased appendicular skeletal muscle mass and strength, decreased self-reported physical function, higher visceral fat mass, insulin resistance, and increased risk of coronary artery disease and mortality. An analysis of signs and symptoms in older men in the EMAS revealed a syndromic association of sexual symptoms with total testosterone levels below 320 ng/dL and free testosterone levels below 64 pg/mL in community-dwelling older men.

A series of placebo-controlled testosterone trials have provided important information about the efficacy of testosterone in improving

outcomes in older men. Testosterone replacement in older men, aged ≥65, with sexual symptoms improved sexual activity, sexual desire, and erectile function more than placebo. Testosterone replacement did not improve fatigue or cognitive function, and had only a small effect on mood and mobility. Among older men with low testosterone and age-associated memory impairment, testosterone replacement did not improve memory or other measures of cognition relative to placebo. Testosterone replacement was associated with significantly greater increase in vertebral as well as femoral volumetric bone mineral density and estimated bone strength relative to placebo. Testosterone replacement was associated with a greater increase in hemoglobin levels and corrected anemia in a greater proportion of men who had unexplained anemia of aging. Testosterone administration was associated with a significantly greater increase in coronary artery noncalcified plaque volume, as measured by coronary artery computerized tomography angiography. Neither the testosterone trials nor a randomized trial of the effects of testosterone on atherosclerosis progression in aging men (TEAAM Trial) with low or low normal testosterone levels found significant differences between testosterone and placebo arms in the rates of change in either the coronary artery calcium scores or the common carotid artery intima-media thickness. Neither of the trials were long enough or large enough to determine the effects of testosterone replacement therapy on prostate or major adverse cardiovascular events. In systematic reviews of randomized controlled trials, testosterone therapy of healthy older men with low or low-normal testosterone levels was associated with greater increments in lean body mass, grip strength, and self-reported physical function than that associated with placebo. Testosterone therapy has not been shown to improve clinical depression, fracture risk, progression to dementia, progression from prediabetes to diabetes, or response to phosphodiesterase inhibitors in older men.

The long-term risks of testosterone therapy remain largely unknown. While there is no evidence that testosterone causes prostate cancer, there is concern that testosterone therapy might cause subclinical prostate cancers to grow. Testosterone therapy is associated with increased risk of detection of prostate events.

The data relating cardiovascular disease (CVD) and venous thromboembolic (VTE) risk with the use of testosterone supplementation in men with low testosterone levels and hypogonadal symptoms are few and inconclusive. The relationship of testosterone and cardiovascular events in cross-sectional and prospective cohort studies has been inconsistent. A small number of epidemiologic studies have reported an inverse relationship between testosterone concentrations and common carotid artery intima-media thickness. Low testosterone level has been associated with increased risk of all-cause mortality, especially cardiovascular mortality. It is possible that testosterone is a marker of health; older men with multiple co-morbid conditions who are at increased risk of death may have low testosterone levels as a result of comorbid conditions.

Most meta-analyses have not shown a statistically significant association between testosterone and cardiovascular events, major adverse cardiovascular events, or deaths. No adequately powered randomized trials have been conducted to determine the effects of testosterone replacement in major adverse cardiovascular events. Thus, there are insufficient data to establish a causal link between testosterone therapy and CV events.

Population screening of all older men for low testosterone levels is not recommended, and testing should be restricted to men who have symptoms or signs attributable to androgen deficiency. Testosterone therapy is not recommended for all older men with low testosterone levels. In older men with significant symptoms of androgen deficiency who have unequivocally low testosterone levels, testosterone therapy may be considered on an individualized basis and should be instituted after careful discussion of the risks and benefits (see "Testosterone Replacement," below).

Testicular morphology, semen production, and fertility are maintained up to a very old age in men. Although concern has been expressed about age-related increases in germ cell mutations and impairment of DNA repair mechanisms, there is no clear evidence that the frequency of chromosomal aneuploidy is increased in the sperm of older men. However, the incidence of autosomal dominant diseases, such as achondroplasia, polyposis coli, Marfan syndrome, and Apert syndrome, increases in the offspring of men who are advanced in age, consistent with transmission of sporadic missense mutations. Advanced paternal age may be associated with increased rates of de novo mutations, which may contribute to an increased risk of neurodevelopmental diseases such as schizophrenia and autism. The somatic mutations in male germ cells that enhance the proliferation of germ cells could lead to within-testis expansion of mutant clonal lines, thus favoring the propagation of germ cells carrying these pathogenic mutations, and increasing the risk of mutations in the offspring of older fathers (the "selfish spermatogonial selection" hypothesis).

APPROACH TO THE PATIENT

Androgen Deficiency

Hypogonadism is often characterized by decreased sex drive, reduced frequency of sexual activity, inability to maintain erections, reduced beard growth, loss of muscle mass, decreased testicular size, and gynecomastia. Erectile dysfunction and androgen deficiency are two distinct clinical disorders that can co-exist in middle-aged and older men. Less than 10% of patients with erectile dysfunction have testosterone deficiency. Thus, it is useful to evaluate men presenting with erectile dysfunction for androgen deficiency. Except when extreme, these clinical features of androgen deficiency may be difficult to distinguish from changes that occur with normal aging. Moreover, androgen deficiency may develop gradually. When symptoms or clinical features suggest possible androgen deficiency, the laboratory evaluation is initiated by the measurement of total testosterone, preferably in the morning using a reliable assay, such as liquid chromatography tandem mass spectrometry (LC-MS/MS) that has been calibrated to an international testosterone standard **(Fig. 384-6)**. A consistently low total testosterone level >264 ng/dL measured by an LC-MS/MS assay in a Centers for Disease Control (CDC)—certified laboratory, in association with symptoms, is evidence of testosterone deficiency. An early-morning testosterone level >400 ng/dL makes the diagnosis of androgen deficiency unlikely. In men with testosterone levels between 200 and 400 ng/dL, the total testosterone level should be repeated and a free testosterone level should be measured. In older men and in patients with other clinical states that are associated with alterations in SHBG levels, a direct measurement of free testosterone level by equilibrium dialysis can be useful in unmasking testosterone deficiency.

When androgen deficiency has been confirmed by the consistently low testosterone concentrations, LH should be measured to classify the patient as having primary (high LH) or secondary (low or inappropriately normal LH) hypogonadism. An elevated LH level indicates that the defect is at the testicular level. Common causes of primary testicular failure include Klinefelter syndrome, HIV infection, uncorrected cryptorchidism, cancer chemotherapeutic agents, radiation, surgical orchiectomy, or prior infectious orchitis. Unless causes of primary testicular failure are known, a karyotype should be performed in men with low testosterone and elevated LH to diagnose Klinefelter syndrome. Men who have a low testosterone but "inappropriately normal" or low LH levels have secondary hypogonadism; their defect resides at the hypothalamic-pituitary level. Common causes of acquired secondary hypogonadism include space-occupying lesions of the sella, hyperprolactinemia, chronic illness, hemochromatosis, excessive exercise, and the use of anabolic-androgenic steroids, opiates, marijuana, glucocorticoids, and alcohol. Measurement of PRL and MRI scan of the hypothalamic-pituitary region can help exclude the presence of a space-occupying lesion. Patients in whom known causes of hypogonadotropic hypogonadism have been excluded are classified as having IHH. It is not unusual for congenital causes of hypogonadotropic hypogonadism, such as Kallmann syndrome, to be diagnosed in young adults.

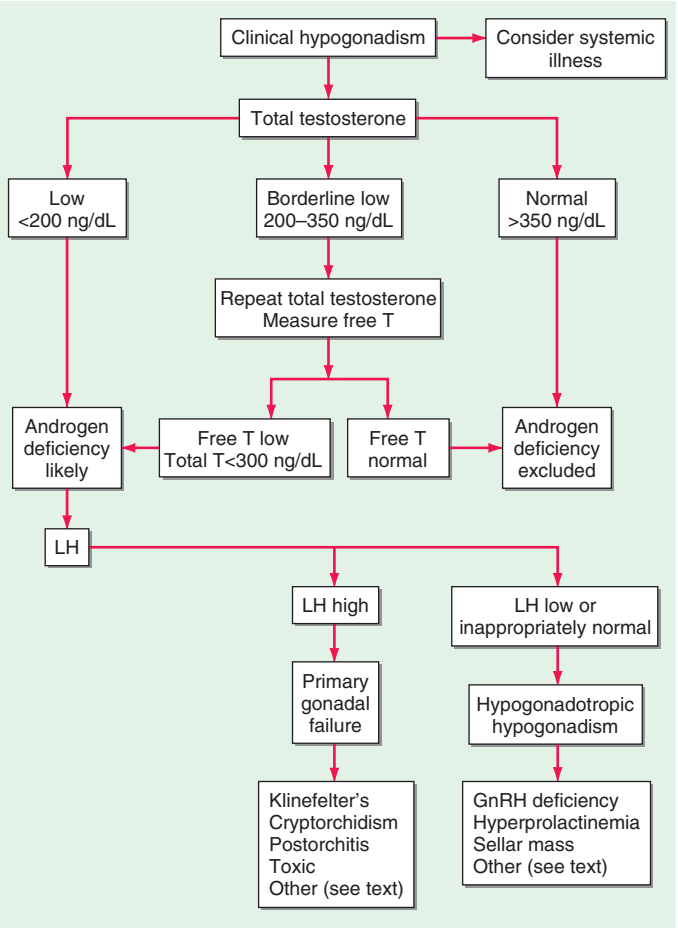


FIGURE 384-6 Evaluation of hypogonadism. GnRH, gonadotropin-releasing hormone; LH, luteinizing hormone; T, testosterone.

TREATMENT

Androgen Deficiency

GONADOTROPINS

Gonadotropin therapy is used to establish or restore fertility in patients with gonadotropin deficiency of any cause. Several gonadotropin preparations are available. Human menopausal gonadotropin (hMG; purified from the urine of postmenopausal women) contains 75 IU FSH and 75 IU LH per vial. hCG (purified from the urine of pregnant women) has little FSH activity and resembles LH in its ability to stimulate testosterone production by Leydig cells. Recombinant LH is also available. Treatment is usually begun with hCG alone, and hMG is added later to promote the FSH-dependent stages of spermatid development. Recombinant human FSH (hFSH) is available and is indistinguishable from purified urinary hFSH in its biologic activity and pharmacokinetics in vitro and in vivo, although the mature β subunit of recombinant hFSH has seven fewer amino acids. Recombinant hFSH is available in ampoules containing 75 IU (~7.5 μg FSH), which accounts for >99% of protein content. Once spermatogenesis is restored using combined FSH and LH therapy, hCG alone is often sufficient to maintain spermatogenesis.

Although a variety of treatment regimens are used, 1000 IU of hCG or recombinant human LH (rhLH) administered intramuscularly three times weekly is a reasonable starting dose. Testosterone levels should be measured 6–8 weeks later and 48–72 h after the hCG or rhLH injection; the hCG/rhLH dose should be adjusted to achieve testosterone levels in the mid-normal range. Sperm counts should be monitored on a monthly basis. It may take several months for spermatogenesis to be restored; therefore, it is important to forewarn patients about the potential length and expense of the treatment and to provide conservative estimates of success rates. If testosterone levels are in the mid-normal range but the sperm concentrations are low after 6 months of therapy with hCG alone, FSH should be added. This can be done by using hMG, highly purified urinary hFSH, or recombinant hFSH. The selection of FSH dose is empirical. A common practice is to start with the addition of 75 IU FSH three times each week in conjunction with the hCG/rhLH injections. If sperm densities are still low after 3 months of combined treatment, the FSH dose should be increased to 150 IU. Occasionally, it may take ≥18–24 months for spermatogenesis to be restored.

The two best predictors of success using gonadotropin therapy in hypogonadotropic men are testicular volume at presentation and time of onset of gonadotropin deficiency. In general, men with testicular volumes >8 mL have better response rates than those who have testicular volumes <4 mL. Patients who become hypogonadotropic after puberty experience higher success rates than those who have never undergone pubertal changes. Spermatogenesis can usually be reinitiated by hCG alone, with high rates of success for men with postpubertal onset of hypogonadotropism. The presence of a primary testicular abnormality, such as cryptorchidism, will attenuate testicular response to gonadotropin therapy. Prior androgen therapy does not preclude subsequent response to gonadotropin therapy, although some studies suggest that it may attenuate response to subsequent gonadotropin therapy.

TESTOSTERONE REPLACEMENT

Androgen therapy is indicated to restore testosterone levels to normal to correct features of androgen deficiency in men with organic hypogonadism due to known diseases of the testes, pituitary, and the hypothalamus. Testosterone replacement induces secondary sex characteristics, improves libido and overall sexual activity; increases lean muscle mass, hemoglobin and hematocrit, and bone mineral density, and decreases fat mass. The benefits of testosterone replacement therapy have only been proven in men who have documented symptomatic androgen deficiency, as demonstrated by testosterone levels that are well below the lower limit of normal.

Testosterone is available in a variety of formulations with distinct pharmacokinetics **(Table 384-3)**. Testosterone serves as a prohormone and is converted to 17β-estradiol by aromatase and to 5α-dihydrotestosterone by steroid 5α-reductase. Therefore, when evaluating testosterone formulations, it is important to consider whether the formulation being used can achieve physiologic estradiol and DHT concentrations, in addition to normal testosterone concentrations. The current recommendation is to restore testosterone levels to the mid-normal range.

Oral Derivatives of Testosterone Testosterone is well-absorbed after oral administration but is quickly degraded during the first pass through the liver. Therefore, it is difficult to achieve sustained blood levels of testosterone after oral administration of crystalline testosterone. 17α-Alkylated derivatives of testosterone (e.g., 17α-methyl testosterone, oxandrolone, fluoxymesterone) are relatively resistant to hepatic degradation and can be administered orally; however, because of the potential for hepatotoxicity, including cholestatic jaundice, peliosis, and hepatoma, these formulations should not be used for testosterone replacement. Hereditary angioedema due to C1 esterase deficiency is the only exception to this general recommendation; in this condition, oral 17α-alkylated androgens are useful because they stimulate hepatic synthesis of the C1 esterase inhibitor.

Injectable Forms of Testosterone The esterification of testosterone at the 17β-hydroxy position makes the molecule hydrophobic and extends its duration of action. The slow release of testosterone ester from an oily depot in the muscle accounts for its extended duration of action. The longer the side chain, the greater the hydrophobicity of the ester and longer the duration of action. Thus, testosterone enanthate, cypionate, and undecanoate with longer side chains have longer duration of action than testosterone propionate. Within 24 h after intramuscular administration of 200 mg testosterone enanthate or cypionate, testosterone levels rise into the high-normal

TABLE 384-3 Clinical Pharmacology of Some Testosterone Formulations

FORMULATION	REGIMEN	PHARMACOKINETIC PROFILE	DHT AND E2	ADVANTAGES	DISADVANTAGES
T enanthate or cypionate	150–200 mg IM q 2 wk or 75–100 mg/wk	After a single IM injection, serum T levels rise into the supraphysiological range, then decline gradually into the low normal or the hypogonadal range by the end of the dosing interval	DHT and E2 levels rise in proportion to the increase in T levels; T:DHT and T:E2 ratios do not change	Corrects symptoms of androgen deficiency; relatively inexpensive, if self-administered; flexibility of dosing	Requires IM injection; peaks and valleys in serum T levels
Topical testosterone gels and axillary testosterone solution	Available in sachets, tubes and pumps	When used in appropriate doses, these topical formulations restore serum T and E2 levels to the physiological male range	Serum DHT levels and DHT to T ratio are higher in hypogonadal men treated with the transdermal gels than in healthy eugonadal men	Corrects symptoms of androgen deficiency, ease of application, good skin tolerability	Potential of transfer to a female partner or child by direct skin-to-skin contact; skin irritation in a small proportion of treated men; moderately high DHT levels; considerable interindividual and intra-indivudal variation in on-treatment testosterone levels
Transdermal testosterone patch	1 or 2 patches, designed to nominally deliver 4–8 mg T over 24 h applied daily on nonpressure areas	Restores serum T, DHT, and E2 levels to the physiological male range	T:DHT and T:E2 levels are in the physiological male range	Ease of application, corrects symptoms of androgen deficiency	Serum T levels in some androgen-deficient men may be in the low-normal range; these men may need application of 2 patches daily; skin irritation at the application site occurs frequently in many patients
Buccal, bioadhesive, T tablets	30 mg controlled release, bioadhesive tablets bid	Absorbed from the buccal mucosa	Normalizes serum T and DHT levels in hypogonadal men	Corrects symptoms of androgen deficiency	Gum-related adverse events in 16% of treated men
T pellets	Several pellets implanted sc; dose and regimen vary with formulation	Serum T peaks at 1 mo and then is sustained in normal range for 3–4 mo, depending on formulation	T:DHT and T:E2 ratios do not change	Corrects symptoms of androgen deficiency	Requires surgical incision for insertions; pellets may extrude spontaneously
17-α-methyl T	This 17-α-alkylated compound should **not** be used because of potential for liver toxicity.	Orally active			Clinical responses are variable; potential for liver toxicity; should **not** be used for treatment of androgen deficiency
Oral T undecanoate*	40–80 mg po bid or tid with meals	When administered in oleic acid, T undecanoate is absorbed through the lymphatics, bypassing the portal system; considerable variability in the same individual on different days and among individuals	High DHT to T ratio	Convenience of oral administration	Not approved in the US;variable clinical responses, variable serum T levels, high DHT:T ratio
Injectable long-acting T undecanoate in oil[1]	US regimen 750 mg IM, followed by 750 mg at 4 wk, and 750 mg every 10 weeks	When administered at the recommended dose, serum T levels are maintained in the normal range in a majority of treated men	DHT and E2 levels rise in proportion to the increase in T levels; T:DHT and T:E2 ratios do not change	Corrects symptoms of androgen deficiency; requires infrequent administration.	Requires IM injection of a large volume; cough reported immediately after injection in a small number of men
Testosterone-in-adhesive matrix patch*	2 × 60 cm^2 patches delivering ~4.8 mg of T/d	Restores serum T, DHT and E_2 to the physiological range	T:DHT and T:E_2 are in the physiological range.	Lasts 2 d	Some skin irritation
Intranasal Testosterone	2 actuations of the metered dose pump (11 mg) applied into the nostrils three times daily	Restores T into the normal male range	T:DHT and T:E2 ratio in the physiologic range		Requires 3x daily application; nasal irritation, epistaxis, nasophyringitis

*These formulations are not approved for clinical use in the United States, but are available outside the United States in many countries. Physicians in those countries where these formulations are available should follow the approved drug regimens.

Abbreviations: DHT, dihydrotestosterone; E2, estradiol; T, testosterone.

or supraphysiologic range and then gradually decline into the hypogonadal range over the next 2 weeks. A bimonthly regimen of testosterone enanthate or cypionate therefore results in peaks and troughs in testosterone levels that may be accompanied by changes in a patient's mood, sexual desire, and energy level; weekly administration of testosterone enanthate or cypionate can reduce these variations in testosterone levels during the dosing interval. The kinetics of testosterone enanthate and cypionate are similar. Estradiol and DHT levels are normal if testosterone replacement is physiologic.

A long-acting testosterone undecanoate in oil, administered at an initial priming dose of 750 mg intramuscularly followed by a second dose of 750 mg 4 weeks later, and then at a maintenance dose of 750 mg every 10 weeks, maintains serum testosterone, estradiol, and DHT in the normal male range and corrects symptoms of androgen deficiency in a majority of treated men. However, its relative drawback is the large injection volume and cough in a small proportion of patients.

Transdermal Testosterone Patch The nongenital testosterone patch, when applied in an appropriate dose, can normalize

testosterone, DHT, and estradiol levels 4–12 h after application. Sexual function and well-being are restored in androgen-deficient men treated with the nongenital patch. One 4-mg patch may not be sufficient to increase testosterone into the mid-normal male range in all hypogonadal men; many patients may need two 4-mg patches daily to achieve the targeted testosterone concentrations. The use of testosterone patches may be associated with skin irritation in some individuals.

Testosterone Gel Several transdermal testosterone gels, Androgel, Testim, Fortesta, and Axiron, and some generic versions, when applied topically to the skin in appropriate doses (Table 384-3), can maintain total and free testosterone concentrations in the normal range in hypogonadal men. The current recommendations are to begin with an initial FDA-recommended dose and adjust the dose based on testosterone levels. The advantages of the testosterone gel include the ease of application. A major concern is the potential for inadvertent transfer of the gel to a sexual partner or to children who may come in close contact with the patient. The ratio of DHT to testosterone concentrations is higher in men treated with the testosterone gel than in healthy men. Also, there is considerable intra- and inter-individual variation in serum testosterone levels in men treated with the transdermal gel due to variations in transdermal absorption and plasma clearance of testosterone. Therefore, monitoring of serum testosterone levels and multiple dose adjustments may be required to achieve and maintain testosterone levels in the target range.

Buccal Adhesive Testosterone A buccal testosterone tablet, which adheres to the buccal mucosa and releases testosterone as it is slowly dissolved, has been approved. After twice-daily application of 30-mg tablets, serum testosterone levels are maintained within the normal male range in a majority of treated hypogonadal men. The adverse effects include buccal ulceration and gum problems in a few subjects. The effects of food and brushing on absorption have not been studied in detail.

Pellets of crystalline testosterone can be inserted in the subcutaneous tissue through a small skin incision. Testosterone is released by surface erosion of the implant and absorbed into the systemic circulation, and testosterone levels can be maintained in the normal range for 3–4 months. Potential drawbacks include incising the skin for insertion and removal, and spontaneous extrusions and fibrosis at the site of the implant.

Testosterone Formulations Not Available in the United States Testosterone undecanoate, when administered orally in oleic acid, is absorbed preferentially through the lymphatics into the systemic circulation and is spared the first-pass degradation in the liver. Doses of 40–80 mg orally, two or three times daily, are typically used. However, the clinical responses are variable and suboptimal. DHT-to-testosterone ratios are higher in hypogonadal men treated with oral testosterone undecanoate, as compared to eugonadal men.

An intranasal testosterone gel is now available as a metered dose pump and is administered typically at a starting dose of 11 mg testosterone in the form of 2 pump actuations, one in each nostril three times daily. Formulation-specific adverse effects include rhinorrhea, nasal discomfort, epistaxis, nasopharyngitis, and nasal scab.

Novel Androgen Formulations A number of androgen formulations with better pharmacokinetics or more selective activity profiles are under development. Initial clinical trials have demonstrated the feasibility of administering testosterone by the sublingual, oral, or buccal routes. Long-acting biodegradable microsphere formulations have also been investigated. 7α-Methyl-19-nortestosterone is an androgen that cannot be 5α-reduced; therefore, compared to testosterone, it has relatively greater agonist activity in muscle and gonadotropin suppression but lesser activity on the prostate.

Selective Androgen Receptor Modulators (SARMs) are a class of AR ligands that bind the AR and display tissue-selective actions. A number of nonsteroidal SARMs that act as agonists on the muscle and bone and which spare the prostate to varying degrees have advanced to phase III human trials. Nonsteroidal SARMs do not serve as substrates for either the steroid 5-α reductase or the CYP19aromatase. SARM binding to AR induces specific conformational changes in the AR protein, which then modulates protein-protein interactions between AR and its coregulators, resulting in tissue-specific regulation of gene expression. SARMs that are strong agonists for the muscle, bone, and sexual function, and antagonists for the prostate may be valuable in treating men with prostate cancer, who are receiving androgen deprivation therapy.

Pharmacologic Uses of Androgens Androgens and selective AR modulators are being evaluated as anabolic therapies for functional limitations associated with aging and chronic illness. Testosterone supplementation increases skeletal muscle mass, maximal voluntary strength, and muscle power in healthy men, hypogonadal men, older men with low testosterone levels, HIV-infected men with weight loss, and men receiving glucocorticoids. These anabolic effects of testosterone are related to testosterone dose and circulating concentrations. Systematic reviews have confirmed that testosterone therapy of HIV-infected men with weight loss promotes improvements in body weight, lean body mass, muscle strength, and depression indices, leading to the recommendation that testosterone be considered as an adjunctive therapy in HIV-infected men who are experiencing unexplained weight loss and who have low testosterone levels. It is unknown whether testosterone therapy of older men with functional limitations is safe and effective in improving physical function, vitality, and health-related quality of life, and reducing disability. Concerns about potential adverse effects of testosterone on prostate and cardiovascular event rates have encouraged the development of selective AR modulators that are preferentially anabolic and spare the prostate.

Testosterone administration induces hypertrophy of both type 1 and 2 fibers and increases satellite cell (muscle progenitor cells) and myonuclear number. Androgens promote the differentiation of mesenchymal, multipotent progenitor cells into the myogenic lineage and inhibit their differentiation into the adipogenic lineage. Testosterone binding to AR promotes the association of liganded AR with β-catenin and its translocation into the nucleus where it binds TCF-4 and activates Wnt-target genes, including follistatin, which blocks signaling through the TGFβ pathway, thereby promoting myogenic differentiation of muscle progenitor cells. Testosterone may have additional effects on satellite cell replication and muscle protein synthesis, which may contribute to an increase in skeletal muscle mass.

Other indications for androgen therapy are in selected patients with anemia due to bone marrow failure (an indication largely supplanted by erythropoietin) or for hereditary angioedema.

Male Hormonal Contraception Based on Combined Administration of Testosterone and Gonadotropin Inhibitors Supraphysiologic doses of testosterone (200 mg testosterone enanthate weekly) suppress LH and FSH secretion and induce azoospermia in 50% of Caucasian men and >95% of Chinese men. The WHO-supported multicenter efficacy trials have demonstrated that suppression of spermatogenesis to azoospermia or severe oligozoozspermia (<3 million/mL) by administration of supraphysiologic doses of testosterone enanthate to men results in highly effective contraception. Because of concern about long-term adverse effects of supraphysiologic testosterone doses, regimens that combine other gonadotropin inhibitors, such as GnRH antagonists and progestins with replacement doses of testosterone, have been investigated. Regimens containing an androgen plus a progestin such as depo medroxyprogesterone acetate, etonogestrelm, or norethisterone enanthate have been highly effective in inducing azoospermia or severe oligozoospermia (sperm density <1 million/mL) in nearly 99% of treated men over a 1-year period. The combined regimens of testosterone plus a progestin have been associated with weight gain, acne, mood changes including depressed mood, libido changes, and decreased plasma high-density lipoprotein (HDL) cholesterol and their long-term safety has not been demonstrated. One such trial of a combined regimen of testosterone undecanoate plus norethisterone enanthate was stopped early due to adverse

events. Selective AR modulators that are more potent inhibitors of gonadotropins than testosterone and spare the prostate hold promise for their contraceptive potential.

Recommended Regimens for Androgen Replacement Testosterone esters are administered typically at doses of 75–100 mg intramuscularly every week, or 150–200 mg every 2 weeks. Testosterone undecanoate is administered at an initial dose of 750 mg followed 4 weeks later by a second injection of 750 mg and then 750 mg every 10 weeks. Testosterone gels are typically applied over a covered area of skin at initial doses that vary with the formulation. The patients should wash their hands after gel application and keep the area of gel application covered with clothing to minimize the risk of gel transfer to another person. One or two 4-mg nongenital testosterone patches are applied daily over the skin of the back, thigh, or upper arm away from pressure areas. Bioadhesive buccal testosterone tablets at a dose of 30 mg are applied twice daily on the buccal mucosa. Intranasal testosterone is administered as a spray in each nostril 3 times a day (33 mg/day).

Establishing Efficacy of Testosterone Replacement Therapy Because a clinically useful marker of androgen action is not available, correction of symptoms, induction and maintenance of secondary sex characteristics, and restoration of testosterone levels into the mid-normal range remain the goal of therapy. Measurements of LH and FSH are not useful in assessing the adequacy of testosterone replacement. Testosterone should be measured 3 months after initiating therapy to assess adequacy of therapy. There is substantial inter-individual variability in serum testosterone levels, especially with transdermal gels, presumably due to genetic differences in testosterone clearance and substantial variation in transdermal absorption. In patients who are treated with testosterone enanthate or cypionate, testosterone levels should be 350–600 ng/dL 1 week after the injection. If testosterone levels are outside this range, adjustments should be made either in the dose or in the interval between injections. In men on transdermal patch, gel, or buccal testosterone therapy, testosterone levels should be in the mid-normal range (400–750 ng/dL) 4–12 h after application. If testosterone levels are outside this range, the dose should be adjusted. Multiple dose adjustments are often necessary to achieve testosterone levels in the desired therapeutic range.

Restoration of sexual function, induction and maintenance of secondary sex characteristics, well-being, and maintenance of muscle and bone health are important objectives of testosterone replacement therapy. The patient should be asked about sexual desire and activity, the presence of early morning erections, and the ability to achieve and maintain erections adequate for sexual intercourse. The hair growth in response to androgen replacement is variable and depends on ethnicity. Hypogonadal men with prepubertal onset of androgen deficiency who begin testosterone therapy in their late twenties or thirties may find it difficult to adjust to their newly found sexuality and may benefit from counseling. If the patient has a sexual partner, the partner should be included in counseling because of the dramatic physical and sexual changes that occur with androgen treatment.

Contraindications for Androgen Administration Testosterone administration is contraindicated in men with prostate or breast cancer **(Table 384-4)**. Testosterone therapy should not be administered without further urologic evaluation to men with a palpable prostate nodule or induration, or prostate-specific antigen >3 ng/mL, or with severe lower urinary tract symptoms (American Urological Association lower urinary tract symptom score >19). Testosterone replacement should not be administered to men with baseline hematocrit ≥50%, severe untreated obstructive sleep apnea, uncontrolled or poorly controlled congestive heart failure, or to men with myocardial infarction, stroke, or acute coronary syndrome in the preceding 3 months.

Monitoring Potential Adverse Experiences The clinical effectiveness and safety of testosterone replacement therapy should be assessed 3–6 months after initiating testosterone therapy and annually thereafter **(Table 384-5)**. Potential adverse effects include acne, oiliness of skin, erythrocytosis, breast tenderness and enlargement, leg edema, and increased risk of detection of prostate events. In addition, there may be formulation-specific adverse effects such as skin irritation with transdermal patch; risk of gel transfer to a sexual partner with testosterone gels; buccal ulceration and gum problems with buccal testosterone; pain and mood fluctuation with injectable testosterone esters; cough and injection site pain with long-acting testosterone undecanoate; and, nasal irritation, epistaxis, and nasal scab with intranasal formulation.

TABLE 384-4 Conditions in Which Testosterone Administration Is Associated with an Increased Risk of Adverse Outcomes

Conditions in which testosterone administration is associated with very high risk of serious adverse outcomes:
Metastatic prostate cancer
Breast cancer
Conditions in which testosterone administration is associated with moderate to high risk of adverse outcomes:
Undiagnosed prostate nodule or induration
PSA > 3
Erythrocytosis (hematocrit >50%)
Severe lower urinary tract symptoms associated with benign prostatic hypertrophy as indicated by American Urological Association/International prostate symptom score >19
Uncontrolled or poorly controlled congestive heart failure
Myocardial infarction, stroke, or acute coronary syndrome in the preceding 3 months

Abbreviation: PSA, prostate-specific antigen.

Source: Reproduced from the Endocrine Society Guideline for Testosterone Therapy of Androgen Deficiency Syndromes in Men (Bhasin S et al: J Clin Endocrinol Metab 95:2536, 2010).

Hemoglobin Levels Administration of testosterone to androgen-deficient men is typically associated with a ~3% increase in hemoglobin levels, due to increased erythropoiesis, stimulation of erythropoietin, suppression of hepcidin, and increased iron availability for erythropoiesis. The magnitude of hemoglobin increase during testosterone therapy is greater in older men than younger men, and in men who have sleep apnea, a significant smoking history, or chronic obstructive lung disease, or who live at high altitude. The frequency of erythrocytosis is higher in hypogonadal men treated with injectable testosterone esters than in those treated with transdermal formulations, presumably due to the higher testosterone dose delivered by the typical regimens of testosterone esters. Erythrocytosis is the most frequent adverse event reported in testosterone trials in middle-aged and older men and is also the most frequent cause of treatment discontinuation in these trials. If hematocrit rises above 54%, testosterone therapy should be stopped until hematocrit has fallen to <50%. After evaluation of the patient for hypoxia and sleep apnea, testosterone therapy may be reinitiated at a lower dose.

Prostate and Serum PSA Levels Testosterone replacement therapy increases prostate volume to the size seen in age-matched controls but does not increase prostate volume beyond that expected for age. There is no evidence that testosterone therapy causes prostate cancer. However, androgen administration can exacerbate preexisting metastatic prostate cancer. Many older men harbor microscopic foci of cancer in their prostates. It is not known whether long-term testosterone administration will induce these microscopic foci to grow into clinically significant cancers.

PSA levels are lower in testosterone-deficient men and are restored to normal after testosterone replacement. There is considerable test-retest variability in PSA measurements. Increments in PSA levels after testosterone supplementation in androgen-deficient men are generally <0.5 ng/mL, and increments >1.0 ng/mL over a 3–6-month period are unusual. The 90% confidence interval for the change in PSA values in men with benign prostatic hypertrophy, measured

TABLE 384-5 Monitoring Men Receiving Testosterone Therapy

1. Evaluate the patient 3–6 months after treatment initiation and then annually to assess whether symptoms have responded to treatment and whether the patient is suffering from any adverse effects.
2. Monitor testosterone level 3–6 months after initiation of testosterone therapy:
   - Therapy should aim to raise serum testosterone level into the mid-normal range.
     - Injectable testosterone enanthate or cypionate: Measure serum testosterone level midway between injections. If testosterone is >600 ng/dL (20.9 nmol/L) or <350 ng/dL (12.2 nmol/L), adjust dose or frequency.
     - Transdermal patches: Assess testosterone level 3–12 h after application of the patch; adjust dose to achieve testosterone level in the midnormal range.
     - Buccal testosterone bioadhesive tablet: Assess level immediately before application of fresh system.
     - Transdermal gels and solution: Assess testosterone level 2–12 h after patient has been on treatment for at least 2 weeks; adjust dose to achieve serum testosterone level in the midnormal range.
     - Testosterone pellets: Measure testosterone levels at the end of the dosing interval. Adjust the number of pellets and/or the dosing interval to achieve serum testosterone levels in the normal range.
     - Injectable testosterone undecanoate: Measure serum testosterone level just prior to each subsequent injection and adjust the dosing interval to maintain serum testosterone in mid-normal range.
3. Check hematocrit at baseline, at 3–6 months, and then annually. If hematocrit is >54%, stop therapy until hematocrit decreases to a safe level; evaluate the patient for hypoxia and sleep apnea; reinitiate therapy with a reduced dose.
4. Measure bone mineral density of lumbar spine and/or femoral neck after 1–2 yr of testosterone therapy in hypogonadal men with osteoporosis or low trauma fracture, consistent with regional standard of care.
5. In men aged ≥40 years with baseline PSA >0.6 ng/mL, perform digital rectal examination and check PSA level before initiating treatment, at 3–6 months, and then in accordance with guidelines for prostate cancer screening depending on the age and race of the patient.
6. Obtain urological consultation if there is:
   - An increase in serum PSA concentration >1.4 ng/mL within any 12-month period of testosterone treatment.
   - A PSA velocity of >0.4 ng/mL·yr using the PSA level after 6 months of testosterone administration as the reference (only applicable if PSA data are available for a period exceeding 2 yr).
   - Detection of a prostatic abnormality on digital rectal examination.
   - An AUA/ IPSS prostate symptom score of >19 along with an increase in IPSS score of ≥5 points above baseline.
7. Evaluate formulation-specific adverse effects at each visit:
   - Buccal testosterone tablets*: Inquire about alterations in taste and examine the gums and oral mucosa for irritation.
   - Injectable testosterone esters (enanthate, cypionate, and undecanoate): Ask about fluctuations in mood or libido, and rarely cough after injections.
   - Testosterone patches: Look for skin reaction at the application site.
   - Testosterone gels: Advise patients to cover the application sites with a shirt and to wash the skin with soap and water before having skin-to-skin contact, because testosterone gels leave a testosterone residue on the skin that can be transferred to a woman or child who might come in close contact. Serum testosterone levels are maintained when the application site is washed 4–6 h after application of the testosterone gel.
   - Testosterone pellets: Look for signs of infection, fibrosis, or pellet extrusion.
     - Intranasal testosterone: Look for signs of nasal irritation or scab

*Not approved for clinical use in the United States.

Abbreviations: AUA/IPSS, American Urological Association International Prostate Symptom Score; PSA, prostate-specific antigen.

Source: Reproduced with permission from the Endocrine Society Guideline for Testosterone Therapy of Androgen Deficiency Syndromes in Adult Men (Bhasin S et al: J Clin Endocrinol Metab 95:2536, 2010).

3–6 months apart, is 1.4 ng/mL. Therefore, the Endocrine Society expert panel suggested that an increase in PSA >1.4 ng/mL in any one year after starting testosterone therapy, if confirmed, should lead to urologic evaluation. PSA velocity criterion can be used for patients who have sequential PSA measurements for >2 years; a change of >0.40 ng/mL per year merits closer urologic follow-up.

Cardiovascular Risk As discussed above, there is insufficient evidence to determine whether testosterone replacement therapy increases the risk of major adverse cardiovascular events in hypogonadal men. A large prospective randomized trial is being planned to determine the effects of testosterone replacement therapy on major adverse cardiovascular events in middle-aged and older men with low testosterone levels and symptoms of androgen deficiency.

Androgen Abuse by Athletes and Recreational Bodybuilders The illicit use of androgenic-anabolic steroids (AAS) to enhance athletic performance first surfaced in the 1950s among powerlifters and spread rapidly to other sports, professional as well as high school athletes, and recreational bodybuilders. In the early 1980s, the use of AAS spread beyond the athletic community into the general population, and now, as many as 3 million Americans—most of them men—have likely used these compounds. Most AAS users are not athletes, but rather recreational weightlifters, who use these drugs to look lean and more muscular. The most commonly used AAS include testosterone esters, nandrolone, stanozolol, methandienone, and methenolol. AAS users generally use increasing doses of multiple steroids in a practice known as stacking.

The adverse effects of long-term AAS abuse remain poorly understood. Most of the information about the adverse effects of AAS has emerged from case reports, uncontrolled studies, or from clinical trials that used replacement doses of testosterone. The adverse event data from clinical trials using physiologic replacement doses of testosterone have been extrapolated unjustifiably to AAS users who may administer 10–100 times the replacement doses of testosterone over many years, to support the claim that AAS use is safe. A substantial fraction of androgenic steroid users also use other drugs that are perceived to be muscle-building or performance-enhancing, such as growth hormone; erythropoiesis stimulating agents; insulin; stimulants such as amphetamine, clenbuterol, cocaine, ephedrine, and thyroxine; and drugs perceived to reduce adverse effects such as hCG, aromatase inhibitors, or estrogen antagonists. The adverse events associated with AAS use may be due to AAS themselves, concomitant use of other drugs, high-risk behaviors, and host characteristics that may render these individuals more susceptible to AAS use or to other high risk behaviors.

The high rates of premature mortality and morbidities observed in AAS users are alarming. One Finnish study reported 4.6 times the risk of death among elite power lifters than in age-matched men from the general population. The causes of death among power lifters included suicides, myocardial infarction, and hepatic coma. A retrospective review of patient records in Sweden also reported higher standardized mortality ratios for AAS users than for nonusers. Increased death rates among AAS users include suicide, homicide, and accidents. The median age of death among AAS users—24 years—is even lower than that for heroin or amphetamine users.

Four categories of adverse events associated with AAS abuse are of particular concern: cardiovascular events, psychiatric, prolonged suppression of the hypothalamic-pituitary-testicular axis, and potential neurotoxicity. Numerous reports of premature cardiac death among young AAS users raise concerns about the adverse cardiovascular effects of AAS. High doses of AAS may induce proatherogenic dyslipidemia, accelerate atherogenesis, increase thrombosis risk via effects on clotting factors and platelets, and induce vasospasm through their effects on vascular nitric oxide. Recent studies of AAS users using tissue Doppler and strain imaging, and magnetic resonance imaging have reported diastolic and systolic dysfunction, including significantly lower early and late diastolic tissue velocities, reduced E/A ratio, and reduced peak systolic strain in AAS users than in nonusers. Power athletes using AAS often have short QT intervals but increased QT dispersion, which may predispose them to ventricular arrhythmias. Long-term AAS use may be associated with myocardial hypertrophy and fibrosis.

Myocardial tissue of power lifters using AAS has been shown to be infiltrated with fibrous tissue and fat droplets. The finding of AR on myocardial cells suggests that AAS might be directly toxic to myocardial cells. Studies of long-term AAS users using computerized tomography angiography have revealed accelerated atherogenesis.

Unlike replacement doses of testosterone, which are associated with only a small decrease in HDL cholesterol and little or no effect on total cholesterol, LDL cholesterol and triglyceride levels, supraphysiologic doses of testosterone and orally administered, 17-α-alylated, nonaromatizable AAS are associated with marked reductions in HDL cholesterol and increases in LDL cholesterol.

Some AAS users develop hypomanic and manic symptoms (irritability, aggressiveness, reckless behavior, and occasional psychotic symptoms, sometimes associated with violence) during AAS exposure, and major depression (sometimes associated with suicidality) during AAS withdrawal. Users may also be susceptible to other forms of illicit drug use, which may be potentiated or exacerbated by AAS.

Long-term AAS use suppresses LH and FSH secretion and inhibits endogenous testosterone production and spermatogenesis. Men, who have used AAS for more than a few months, experience marked suppression of the hypothalamic-pituitary-testicular (HPT) axis after stopping AAS that may be associated with sexual dysfunction, fatigue, infertility, depressed mood, and even suicidality. In some long-term AAS users, HPT suppression may last more than a year, and in a few individuals, recovery of the HPT axis may be incomplete or may never occur. The symptoms of androgen deficiency caused by androgen withdrawal may cause some men to revert back to using AAS, leading to continued use and AAS dependence. As many as 30% of AAS users develop a syndrome of AAS dependence, characterized by long-term AAS use despite adverse medical and psychiatric effects. AAS withdrawal hypogonadism has emerged as an important cause of androgen deficiency accounting for a substantial fraction of testosterone prescriptions in many men's health clinics.

Supraphysiologic doses of testosterone may also impair insulin sensitivity. Orally administered androgens also have been associated with insulin resistance and diabetes.

Unsafe injection practices, high-risk behaviors, and increased rates of incarceration render AAS users at increased risk of HIV, and hepatitis B and C. In one survey, nearly 1 in 10 gay men had injected AAS or other substances, and AAS users were more likely to report high-risk unprotected anal sex than other men.

Elevated liver enzymes, cholestatic jaundice, hepatic neoplasms, and peliosis hepatis have been reported with oral, 17-α-alkylated AAS. AAS use may cause muscle hypertrophy without compensatory adaptations in tendons, ligaments, and joints, thus increasing the risk of tendon and joint injuries. Upper extremity tendon ruptures are observed almost exclusively among weightlifters who use AAS. AAS use is associated with acne, baldness, as well as increased body hair.

The suspicion of AAS use should be raised by the increased hemoglobin and hematocrit, suppressed LH and FSH and testosterone levels, low high-density lipoproteins cholesterol, and low testicular volume and sperm density in a person who looks highly muscular. In most AAS users seeking medical attention, direct nonjudgmental questioning is sufficient to uncover AAS use and formal testing for AAS usually is not needed. However, if needed, accredited laboratories use gas chromatography-mass spectrometry or liquid chromatography-mass spectrometry to detect anabolic steroid abuse. In recent years, the availability of high-resolution mass spectrometry and tandem mass spectrometry has further improved the sensitivity of detecting androgen abuse. Illicit testosterone use is detected generally by the application of the measurement of urinary testosterone to epitestosterone ratio and further confirmed by the use of the ^{13}C:^{12}C ratio in testosterone by the use of isotope ratio combustion mass spectrometry. Exogenous testosterone administration increases urinary testosterone glucuronide excretion and consequently the testosterone to epitestosterone ratio. Ratios above 4 suggest exogenous testosterone use but can also reflect genetic variation. Genetic variations in the uridine diphospho-glucuronyl transferase 2B17 (*UGT2B17*), the major enzyme for testosterone glucuronidation, affect testosterone to epitestosterone ratio. Synthetic testosterone has a lower ^{13}C:^{12}C ratio than endogenously produced testosterone and these differences in ^{13}C:^{12}C ratio can be detected by isotope ratio combustion mass spectrometry, which is used to confirm exogenous testosterone use in individuals with a high testosterone to epitestosterone ratio.

FURTHER READING

Abreu AP et al: Central precocious puberty caused by mutations in the imprinted gene *MKRN3*. N Engl J Med 368:26, 2013.

Balasubramanian R et al: Human GnRH deficiency: A unique disease model to unravel the ontogeny of GnRH neurons. Neuroendocrinology 92:81, 2010.

Basaria S et al: Effects of testosterone administration for 3 years on subclinical atherosclerosis progression in older men with low or low-normal testosterone levels: A randomized clinical trial. JAMA 314:570, 2015.

Bhasin S et al: Testosterone therapy in men with androgen deficiency syndromes: An endocrine society clinical practice guideline. J Clin Endocrinol Metab 95:2536, 2010.

Bhasin S et al: Effect of testosterone supplementation with and without a dual 5α-reductase inhibitor on fat-free mass in men with suppressed testosterone production: a randomized controlled trial. JAMA 307:931, 2012.

Budoff MJ et al: Testosterone treatment and coronary artery plaque volume in older men with low testosterone. JAMA 317:708, 2017.

Finkelstein JS et al: Gonadal steroids and body composition, strength, and sexual function in men. N Engl J Med 369:1011, 2013.

Forni PE, Wray S: GnRH, anosmia and hypogonadotropic hypogonadism—Where are we? Front Neuroendocrinol 36:165, 2015.

Pope HG Jr et al: Adverse health consequences of performance-enhancing drugs: An Endocrine Society scientific statement. Endocr Rev 35:341, 2014.

Snyder PJ et al: Effects of testosterone treatment in older men. N Engl J Med 74:611, 2016.

Spitzer M et al: Risks and benefits of testosterone therapy in older men. Nat Rev Endocrinol 9:414, 2013.

Travison TG et al: Harmonized reference ranges for circulating testosterone levels in men of four cohort studies in the USA and Europe. J Clin Endocrinol Metab 2017 in press.

Zakharov MN et al: A multi-step, dynamic allosteric model of testosterone's binding to sex hormone binding globulin. Mol Cell Endocrinol 399:190, 2015.

385 Disorders of the Female Reproductive System

Janet E. Hall

The female reproductive system regulates the hormonal changes responsible for puberty and adult reproductive function. Normal reproductive function in women requires the dynamic integration of hormonal signals from the hypothalamus, pituitary, and ovary, resulting in repetitive cycles of follicle development, ovulation, and preparation of the endometrial lining of the uterus for implantation should conception occur. It is critical to understand pubertal development in normal girls (and boys) as a yardstick for identifying precocious and delayed puberty.

For further discussion of related topics, see the following chapters: amenorrhea and pelvic pain (Chap. 386), infertility and contraception (Chap. 389), menopause (Chap. 388), disorders of sex development (Chap. 383), and disorders of the male reproductive system (Chap. 384).

DEVELOPMENT OF THE OVARY AND EARLY FOLLICULAR GROWTH

The ovary orchestrates the development and release of a mature oocyte and secretes hormones (e.g., estrogen, progesterone, inhibins A and B, relaxin) that play critical roles in a variety of target tissues, including breast, bone, and uterus, in addition to the hypothalamus and pituitary. To achieve these functions in repeated monthly cycles, the ovary undergoes some of the most dynamic changes of any organ in the body. Primordial germ cells can be identified by the third week of gestation, and their migration to the genital ridge is complete by 6 weeks of gestation. Germ cells persist within the genital ridge, are then referred to as *oogonia*, and are essential for induction of ovarian development. In patients with 45,X Turner syndrome, primordial germ cells proliferate and migrate to the genital ridge, but do not persist as their survival requires the presence of pregranulosa cells that are dependent on the presence of both X chromosomes. (Chap. 383).

The germ cell population expands, and starting at ~8 weeks of gestation, oogonia begin to enter prophase of the first meiotic division and become primary oocytes. This allows the oocyte to be surrounded by a single layer of flattened granulosa cells to form a primordial follicle (Fig. 385-1). Granulosa cells are derived from mesonephric cells that invade the ovary early in its development, pushing the germ cells to the periphery. Although there is evidence that both oocyte-like cells and follicle-like structures can form from embryonic stem cells in culture, there is, as yet, no clear evidence that this occurs in vivo and thus, the ovary appears to contain a nonrenewable pool of germ cells. Through the combined processes of mitosis, meiosis, and atresia, the population of oogonia reaches its maximum of 6–7 million by 20 weeks of gestation, after which there is a progressive loss of both oogonia and primordial follicles through the process of atresia. It appears that entry into meiosis provides some degree of protection from programmed cell death. At birth, oogonia are no longer present in the ovary, and only 1–2 million germ cells remain in the form of primordial follicles (Fig. 385-2). The oocyte persists in prophase of the first meiotic division until just before ovulation, when meiosis resumes.

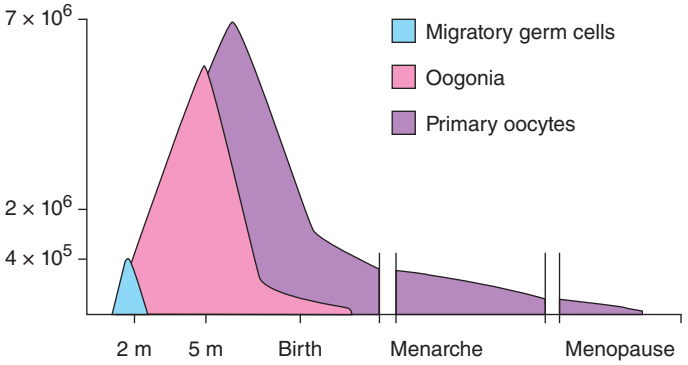


FIGURE 385-2 Ovarian germ cell number is maximal at mid-gestation and decreases precipitously thereafter.

The quiescent primordial follicles are recruited to further growth and differentiation through a highly regulated process that limits the size of the developing cohort to ensure that folliculogenesis can continue throughout the reproductive life span. This initial recruitment of primordial follicles to form primary follicles (Fig. 385-1) is characterized by growth of the oocyte and the transition from squamous to cuboidal granulosa cells. The theca interna cells that surround the developing follicle begin to form as the primary follicle grows. Acquisition of a zona pellucida by the oocyte and the presence of several layers of surrounding cuboidal granulosa cells mark the development of secondary follicles. It is at this stage that granulosa cells develop follicle-stimulating hormone (FSH), estradiol, and androgen receptors and communicate with one another through the development of gap junctions.

Bidirectional signaling between the germ cells and the somatic cells in the ovary is a necessary component underlying the maturation of the oocyte and the capacity for hormone secretion. For example, oocyte-derived growth differentiation factor 9 (GDF-9) and bone morphogenic protein-15 (BMP-15), also known as GDF-9b, are required for migration of pregranulosa and pretheca cells to the outer surface of the developing follicle and, hence, initial follicle formation. GDF-9 is also required for formation of secondary follicles, as are granulosa cell–derived KIT ligand (KITL) and the forkhead transcription factor (FOXL2). A significant number of genes have been identified that are required for development of the normal complement of oogonia in the ovary, initial follicle development and resistance to follicle loss; all are candidates for premature ovarian insufficiency (POI) and mutations in >50 genes have been identified in patients with POI with even more that have been associated with an earlier age at natural menopause.

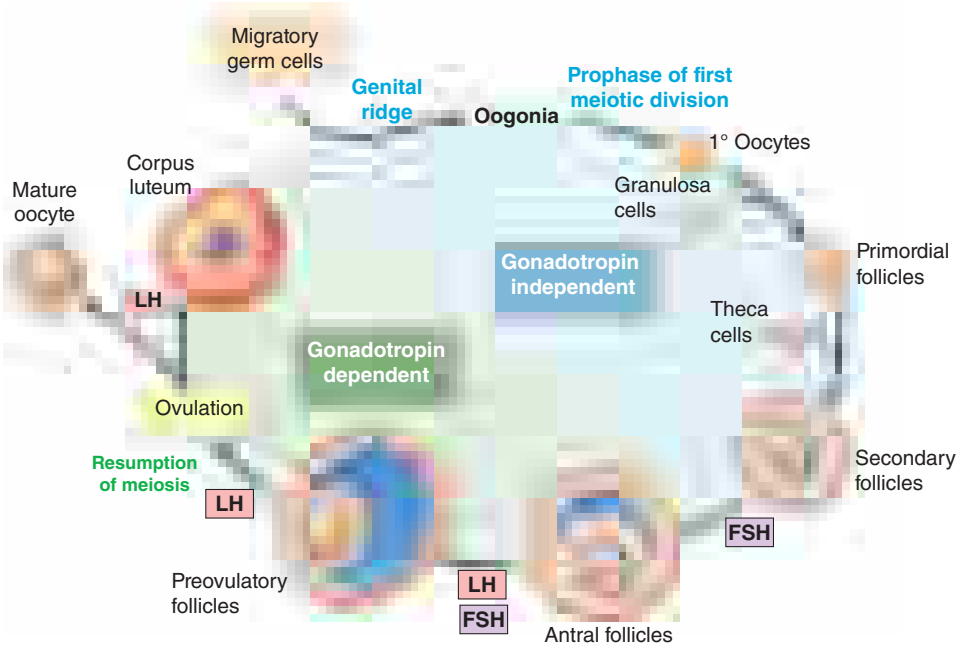


FIGURE 385-1 Stages of ovarian development from the arrival of the migratory germ cells at the genital ridge through gonadotropin-independent and gonadotropin-dependent phases that ultimately result in ovulation of a mature oocyte. FSH, follicle-stimulating hormone; LH, luteinizing hormone.

DEVELOPMENT OF A MATURE FOLLICLE

The early stages of follicle growth are primarily driven by intraovarian factors; after initial recruitment, development to the secondary follicle stage may take close to a year. Further maturation to the preovulatory stage, including the resumption of meiosis in the oocyte, requires the combined stimulus of FSH and luteinizing hormone (LH) (Fig. 385-1). Recruitment of secondary follicles from the resting follicle pool requires the direct action of FSH, whereas anti-müllerian hormone (AMH) produced from small growing follicles, restrains this effect of FSH controlling the number of follicles entering the actively growing pool. Accumulation of follicular fluid between

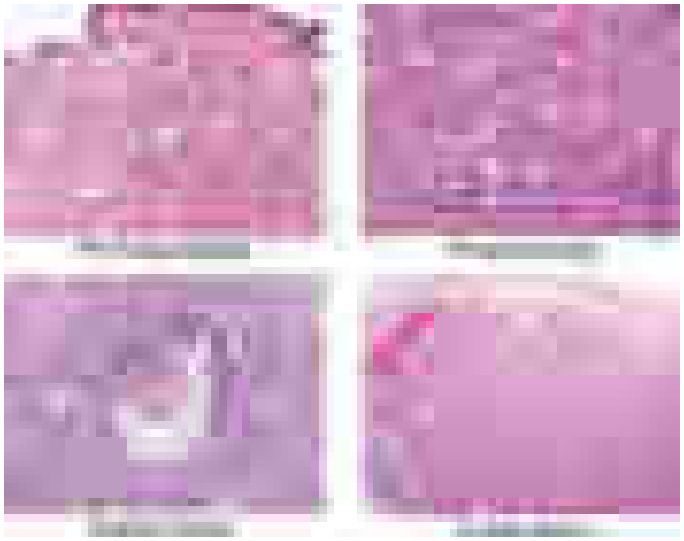


FIGURE 385-3 Development of ovarian follicles. The Graafian follicle is also known as a tertiary or preovulatory follicle. *(Courtesy of J.H. Eichhorn and D. Roberts, Massachusetts General Hospital; with permission.)*

the layers of granulosa cells creates an antrum that divides the granulosa cells into two functionally distinct groups: mural cells that line the follicle wall and cumulus cells that surround the oocyte **(Fig. 385-3)**. In addition to its role in normal development of the müllerian system, the WNT signaling pathway is required for normal antral follicle development and may also play a role in ovarian steroidogenesis. Recruitment to the small antral stage generally occurs over several cycles with further growth to follicle sizes of >4–7 mm in several waves during a single cycle. A single dominant follicle emerges from the growing follicle pool within the first 5–7 days after the onset of menses while the majority of follicles fall off their growth trajectory and become atretic. Autocrine actions of activin and BMP-6, derived from the granulosa cells, and paracrine actions of GDF-9, BMP-15, BMP-6, and Gpr149, derived from the oocyte, are involved in granulosa cell proliferation and modulation of FSH responsiveness. Differential exposure to these factors, and to vascular endothelial growth factor (VEGF), can attenuate vascular density and permeability, likely explaining the mechanism whereby a given follicle is selected for continued growth to the preovulatory stage. The dominant follicle can be distinguished by its size, evidence of granulosa cell proliferation, large number of FSH receptors, high aromatase activity, and elevated concentrations of estradiol and inhibin A in follicular fluid. In addition, secretion of estradiol and inhibin from the dominant follicle inhibits FSH and the growth of other follicles.

The dominant follicle undergoes rapid expansion during the 5–6 days prior to ovulation, reflecting granulosa cell proliferation and accumulation of follicular fluid. FSH induces LH receptors on the granulosa cells, and the preovulatory, or Graafian, follicle moves to the outer ovarian surface in preparation for ovulation. The LH surge triggers the resumption of meiosis, the suppression of granulosa cell proliferation, and the induction of cyclooxygenase 2 (COX-2), prostaglandins, the progesterone receptor (PR), and the epidermal growth factor (EGF)-like growth factors amphiregulin, epiregulin, betacellulin, and neuroregulin 1, all of which are required for ovulation. Ovulation requires production of extracellular matrix leading to expansion of the cumulus cell population that surrounds the oocyte and the controlled expulsion of the egg and follicular fluid. Both progesterone and prostaglandins (induced by the ovulatory stimulus) are essential for this process as are members of the matrix metalloproteinase family. After ovulation, luteinization of theca and granulosa cells is induced by LH in conjunction with the acquisition of a rich vascular network in response to VEGF and basic fibroblast growth factor (FGF). Traditional regulators of central reproductive control, gonadotropin-releasing hormone (GnRH) and its receptor (GnRHR), as well as kisspeptin, are also produced in the ovary and may be involved in corpus luteum function.

REGULATION OF OVARIAN FUNCTION

HYPOTHALAMIC AND PITUITARY SECRETION

GnRH neurons derive from cells in the olfactory placode and to a lesser extent, the neural crest. They migrate along the scaffold of the olfactory neurons across the cribiform plate to the hypothalamus where they separate from the olfactory neurons. Studies in GnRH-deficient patients who fail to undergo puberty have provided insights into genes that control the ontogeny and function of GnRH neurons **(Fig. 385-4)**. *KAL1*, *FGF8/FGFR1*, *PROK2/PROKR2*, *NSMF*, *HS6SD1*, and *CDH7*, among others **(Chap. 384)**, have been implicated in the migration of GnRH neurons to the hypothalamus. Approximately 7000 GnRH neurons, scattered throughout the medial basal hypothalamus, establish contacts with capillaries of the pituitary portal system in the median eminence. GnRH is secreted into the pituitary portal system in discrete pulses to stimulate synthesis and secretion of LH and FSH from pituitary gonadotropes, which comprise ~10% of cells in the pituitary **(Chap. 371)**. Functional connections of GnRH neurons with the portal system are established by the end of the first trimester, coinciding with the production of pituitary gonadotropins. Thus, like the ovary, the hypothalamic and pituitary components of the reproductive system are present before birth. However, the high levels of estradiol and progesterone produced by the placenta suppress hypothalamic-pituitary stimulation of ovarian hormonal secretion in the fetus.

After birth and the loss of placenta-derived steroids, gonadotropin levels rise. FSH levels are much higher in girls than in boys. This rise in FSH results in circulating estradiol and increased inhibin B, but without terminal follicle maturation or ovulation . Studies that have identified mutations in *TAC3*, which encodes neurokinin B, and its receptor, *TAC3R*, in patients with GnRH deficiency indicate that both are involved in control of GnRH secretion and may be particularly important at this early stage of development. By 12–20 months of age, the reproductive axis is again suppressed, and a period of relative quiescence persists until puberty **(Fig. 385-5)**. At the onset of puberty, pulsatile GnRH secretion induces pituitary gonadotropin production. In the early stages of puberty, LH and FSH secretion are apparent only during sleep, but as puberty develops, pulsatile gonadotropin secretion occurs throughout the day and night.

The mechanisms responsible for the childhood quiescence and pubertal reactivation of the reproductive axis remain incompletely

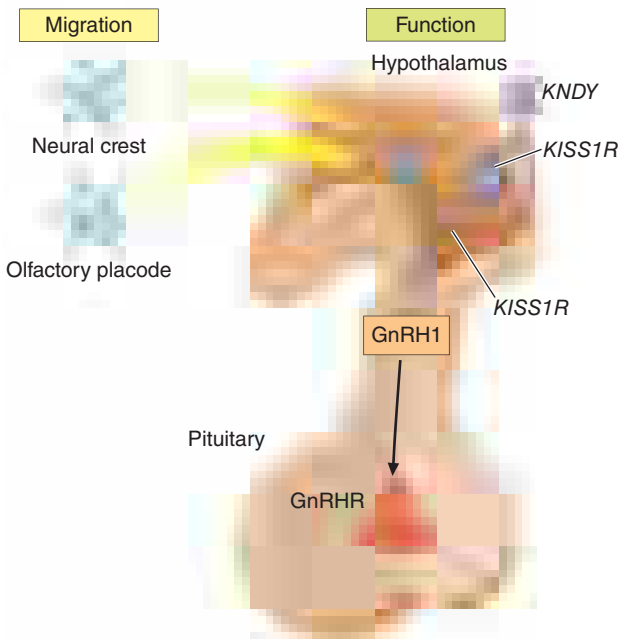


FIGURE 385-4 Genetic studies in patients with congenital forms of hypogonadotropic hypogonadism have expanded our understanding of the development and migration of gonadotropin-releasing hormone (GnRH) neurons from the olfactory placode and neural crest to the hypothalamus as well as the upstream regulation of GnRH secretion.

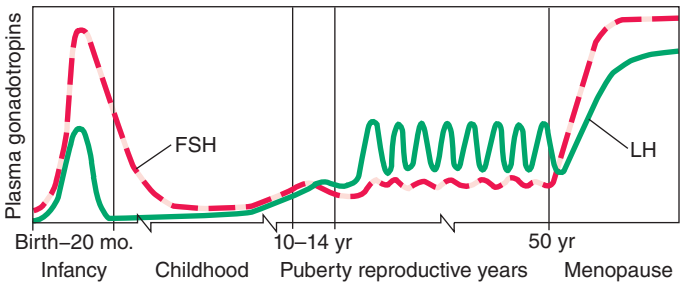


FIGURE 385-5 Follicle-stimulating hormone (FSH) and luteinizing hormone (LH) are increased during the neonatal years but go through a period of childhood quiescence before increasing again during puberty. Gonadotropin levels are cyclic during the reproductive years and increase dramatically with the loss of negative feedback that accompanies menopause.

understood. GnRH neurons in the hypothalamus respond to both excitatory and inhibitory factors. Increased sensitivity to the inhibitory influence of gonadal steroids has long been implicated in the inhibition of GnRH secretion during childhood but has not been definitively established in the human. Metabolic signals, including adipocyte-derived leptin, play a permissive role in reproductive function **(Chap. 394)**. Studies of patients with isolated GnRH deficiency reveal that mutations in the G protein–coupled receptor 54 (*GPR54*) gene (now known as *KISS1R*) preclude the onset of puberty. The ligand for this receptor is derived from the parent peptide, kisspeptin-1 (*KISS1*), and is a powerful stimulant for GnRH release. A potential role for kisspeptin in the onset of puberty has been suggested by upregulation of *KISS1* and *KISS1R* transcripts in the hypothalamus at the time of puberty. *TAC3*, which stimulates GnRH secretion through kisspeptin signaling, and dynorphin (*Dyn*), which plays an inhibitory role in GnRH control, are frequently co-expressed with *KISS1* in KNDy neurons of the median eminence that project to GnRH neurons. This system is intimately involved in both estrogen and progesterone negative feedback regulation of GnRH secretion.

RFamide-Related peptides (RFRPs) are the mammalian orthologues of gonadotropin inhibitory hormone (GnIH) which was initially discovered in the quail. While RFRP-1 and RFRP-3 neurons send axonal projections to GnRH neurons in humans, and RFRPs are secreted into the pituitary portal system, further studies are required to determine their potential physiologic role in the human.

OVARIAN STEROIDS

Ovarian steroid-producing cells do not store hormones but produce them in response to LH and FSH during the normal menstrual cycle. The sequence of steps and the enzymes involved in the synthesis of steroid hormones are similar in the ovary, adrenal, and testis. However, the enzymes required to catalyze specific steps are compartmentalized and may not be abundant or even present in all cell types. Within the developing ovarian follicle, estrogen synthesis from cholesterol requires close integration between theca and granulosa cells—sometimes called the *two-cell model for steroidogenesis* **(Fig. 385-6)**. FSH receptors are confined to the granulosa cells, whereas LH receptors are restricted to the theca cells until the late stages of follicular development, when they are also found on granulosa cells. The theca cells surrounding the follicle are highly vascularized and use cholesterol, derived primarily from circulating lipoproteins, as the starting point for the synthesis of androstenedione and testosterone under the control of LH. These steroid precursors cross the basal lamina to the granulosa cells, which receive no direct blood supply. The mural granulosa cells are particularly rich in aromatase and, under the control of FSH, produce estradiol, the primary steroid secreted from the follicular phase ovary and the most potent estrogen. Theca cell–produced androstenedione and, to a lesser extent, testosterone are also secreted into peripheral blood, where they can be converted to dihydrotestosterone in skin and to estrogens in adipose tissue. The hilar interstitial cells of the ovary are functionally similar to Leydig cells and are also capable of secreting androgens.

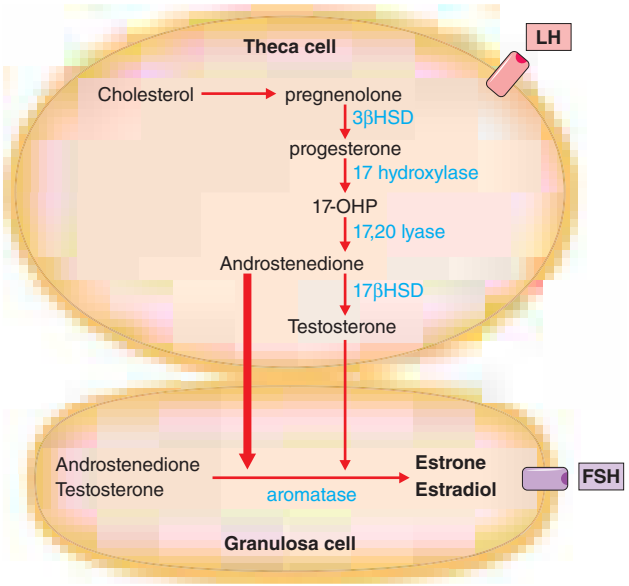


FIGURE 385-6 Estrogen production in the ovary requires the cooperative function of the theca and granulosa cells under the control of luteinizing hormone (LH) and follicle-stimulating hormone (FSH). HSD, hydroxysteroid dehydrogenase; OHP, hydroxyprogesterone.

Stromal cells proliferate in response to androgens (as in polycystic ovarian syndrome [PCOS]), but do not secrete androgens. However, high levels of androgens may be produced by luteinized theca cells in women with hyperthecosis.

Development of the rich capillary network following rupture of the follicle at the time of ovulation makes it possible for large molecules such as low-density lipoprotein (LDL) to reach the luteinized granulosa and theca lutein cells. As in the follicle, both cell types are required for steroidogenesis in the corpus luteum. The luteinized granulosa cells are the main source of progesterone production, whereas the smaller theca lutein cells produce 17-hydroxyprogesterone and androgenic substrates for aromatization to estradiol by the luteinized granulosa cells. Production of estrogen metabolites by the corpus luteum plays a significant role in maintenance of the vascularization required for its function. LH is critical for formation and maintenance of corpus luteum structure and function. LH and human chorionic gonadotropin (hCG) bind to a common receptor; thus, in conception cycles, hCG rescues the declining function of the corpus luteum, maintaining steroid and peptide secretion for the first 10 weeks of pregnancy. HCG is commonly used for luteal phase support in the treatment of infertility.

Steroid Hormone Actions Both estrogen and progesterone play critical roles in the expression of secondary sexual characteristics in women **(Chap. 370)**. Estrogen promotes development of the ductule system in the breast, whereas progesterone is responsible for glandular development. In the reproductive tract, estrogens create a receptive environment for fertilization and support pregnancy and parturition through carefully coordinated changes in the endometrium, thickening of the vaginal mucosa, thinning of the cervical mucus, and uterine growth and contractions. Progesterone induces secretory activity in the estrogen-primed endometrium, increases the viscosity of cervical mucus, and inhibits uterine contractions. Both gonadal steroids play critical roles in negative and positive feedback of gonadotropin secretion. Progesterone also increases basal body temperature and has therefore been used clinically as a marker of ovulation.

The vast majority of circulating estrogens and androgens are carried in the blood bound to carrier proteins, which restrain their free diffusion into cells and prolong their clearance, serving as a reservoir. High-affinity binding proteins include sex hormone–binding globulin (SHBG), which binds androgens with somewhat greater affinity than

estrogens, and corticosteroid-binding globulin (CBG), which also binds progesterone. Modulations in binding protein levels by insulin, androgens, and estrogens contribute to high bioavailable testosterone levels in PCOS and to high circulating total estrogen and progesterone levels during pregnancy.

Estrogens act primarily through binding to the nuclear receptors, estrogen receptor (ER) α and β. Transcriptional coactivators and co-repressors modulate ER action **(Chap. 370)**. Both ER subtypes are present in the hypothalamus, pituitary, ovary, and reproductive tract. Although ERα and β exhibit some functional redundancy, there is also a high degree of specificity, particularly in expression within cell types. For example, ERα functions in ovarian theca cells, whereas ERβ is critical for granulosa cell function. There is also evidence for membrane-initiated signaling by estrogen. Similar signaling mechanisms pertain for progesterone with evidence of transcriptional regulation through PR A and B protein isoforms, as well as rapid membrane signaling.

OVARIAN PEPTIDES

Inhibin was initially isolated from gonadal fluids based on its ability to selectively inhibit FSH secretion from pituitary cells. Inhibin is a heterodimer composed of an α subunit and a βA or βB subunit to form inhibin A or inhibin B, both of which are secreted from the ovary. Activin is a homodimer of inhibin β subunits with the capacity to stimulate the synthesis and secretion of FSH. Inhibins and activins are members of the transforming growth factor β (TGF-β) superfamily of growth and differentiation factors. During the purification of inhibin, follistatin, an unrelated monomeric protein that inhibits FSH secretion, was discovered. Within the pituitary, follistatin inhibits FSH secretion indirectly by binding and neutralizing activin.

Inhibin B is constitutively secreted from the granulosa cells of small antral follicles and its serum levels increase in conjunction with granulosa cell proliferation during recruitment of secondary follicles under the control of FSH. In addition to its role as a marker of decreasing ovarian reserve during reproductive aging, inhibin B is an important inhibitor of FSH, independent of estradiol, during the menstrual cycle. Inhibin A is present in both granulosa and theca cells and is secreted by the dominant follicle. Inhibin A is also present in luteinized granulosa cells and is a major secretory product of the corpus luteum. Synthesis and secretion of inhibin A are directly controlled by FSH and LH. Although activin is also secreted from the ovary, the excess of follistatin in serum, combined with its nearly irreversible binding of activin, make it unlikely that ovarian activin plays an endocrine role in FSH regulation. However, there is evidence that activin plays an autocrine/paracrine role in the ovary, in addition to its intra-pituitary role in modulation of FSH production.

AMH (also known as müllerian-inhibiting substance) is important in ovarian biology in addition to the function from which it derived its name (i.e., promotion of the degeneration of the müllerian system during embryogenesis in the male). AMH is produced by granulosa cells from small follicles and is a marker of ovarian reserve with advantages over inhibin B because of its relative stability across the menstrual cycle. AMH inhibits the recruitment of primordial follicles into the follicle pool and counters FSH stimulation of aromatase expression. AMH is increased in polycystic ovarian syndrome in conjunction with the abundance of small follicles in this disorder.

Gonadotropin Surge Attenuating Factor (GnSAF) is an ovarian factor that attenuates GnRH-induced gonadotropin secretion. Its role is not yet fully understood, but there is an inverse relationship between GnSAF and follicle size suggesting that its primary role involves the early stages of follicle development rather than curtailing the gonadotropin surge as its name implies.

Relaxin is produced primarily by the theca lutein cells of the corpus luteum. Both relaxin and its receptor, RXFP1, are highly expressed in the uterus during the peri-implantation period in the marmoset and its primary role appears to be in promoting decidualization and vascularization of the endometrium prior to implantation. Relaxin was named for its ability to suppress myometrial contractility in pigs and rodents, but it does not appear to exert this activity in women.

HORMONAL INTEGRATION OF THE NORMAL MENSTRUAL CYCLE

The sequence of changes responsible for mature reproductive function is coordinated through a series of negative and positive feedback loops that alter pulsatile GnRH secretion, the pituitary response to GnRH, and the relative secretion of LH and FSH from the gonadotrope. The frequency and amplitude of pulsatile GnRH secretion differentially modulate the synthesis and secretion of LH and FSH. Slow GnRH pulse frequencies favor FSH synthesis whereas increased GnRH pulse frequency and amplitude favor LH synthesis. Activin is produced in both pituitary gonadotropes and folliculostellate cells and stimulates the synthesis and secretion of FSH through autocrine-paracrine mechanisms that are modulated by follistatin. Inhibins function as potent antagonists of activins through sequestration of the activin receptors. Although inhibin is expressed in the pituitary, gonadal inhibin is the principal source of feedback inhibition of FSH.

For the majority of the cycle, the reproductive system functions in a classic endocrine negative feedback mode. Estradiol and progesterone inhibit GnRH secretion, acting through kisspeptin and dynorphin in the KNDy neurons, and the inhibins act at the pituitary to selectively inhibit FSH synthesis and secretion **(Fig. 385-7)**. Estradiol also contributes to negative feedback at the pituitary with an effect that is greater for FSH than LH. This tightly regulated negative feedback control of FSH is critical for development of the single mature oocyte that characterizes normal reproductive function in women. In addition to these negative feedback controls, the menstrual cycle is uniquely dependent on estrogen-induced positive feedback to produce an LH surge that is essential for ovulation of a mature follicle. Estrogen negative feedback in women occurs primarily at the hypothalamus with a small pituitary contribution, whereas estrogen positive feedback occurs at the pituitary in women with upregulation of GnRH signaling. In women, hypothalamic GnRH secretion plays a permissive role in generating the preovulatory gonadotropin surge, a mechanism that differs significantly from that in rodents and other species that rely on seasonal and circadian cues, in which a surge of GnRH also occurs.

THE FOLLICULAR PHASE

The follicular phase is characterized by recruitment of a cohort of secondary follicles and the ultimate selection of a dominant preovulatory follicle **(Fig. 385-8)**. The follicular phase begins, by convention, on the first day of menses. However, follicle recruitment is initiated

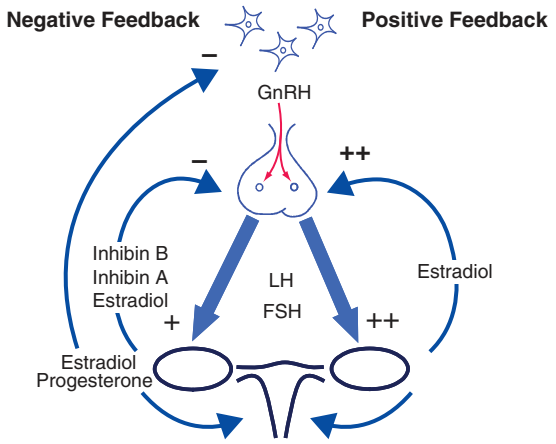


FIGURE 385-7 The reproductive system in women is critically dependent on both negative feedback of gonadal steroids and inhibin to modulate follicle-stimulating hormone (FSH) secretion and on estrogen positive feedback to generate the preovulatory luteinizing hormone (LH) surge. GnRH, gonadotropin-releasing hormone.

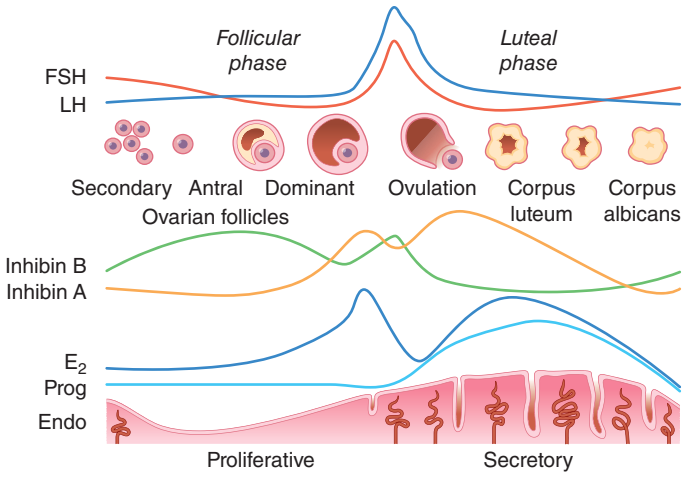


FIGURE 385-8 Relationship between gonadotropins, follicle development, gonadal secretion, and endometrial changes during the normal menstrual cycle. E_2, estradiol; Endo, endometrium; FSH, follicle-stimulating hormone; LH, luteinizing hormone; Prog, progesterone.

by the rise in FSH that begins in the late luteal phase of the previous cycle in conjunction with the loss of negative feedback of gonadal steroids and likely inhibin A. The fact that a 20–30% increase in FSH is adequate for follicular recruitment speaks to the marked sensitivity of the resting follicle pool to FSH. The resultant granulosa cell proliferation is responsible for increasing early follicular phase levels of inhibin B. Inhibin B in conjunction with rising levels of estradiol and inhibin A, restrain FSH secretion during this critical period such that only a single follicle matures in the vast majority of cycles. The increased risk of multiple gestation associated with the increased levels of FSH characteristic of advanced maternal age, or with exogenous gonadotropin administration in the treatment of infertility, attests to the importance of negative feedback regulation of FSH. With further growth of the dominant follicle, estradiol and inhibin A increase and the follicle acquires LH receptors. Increasing levels of estradiol are responsible for proliferative changes in the endometrium. The exponential rise in estradiol results in positive feedback on the pituitary, leading to the generation of an LH surge (and a smaller FSH surge), thereby triggering ovulation and luteinization of granulosa and theca cells.

THE LUTEAL PHASE

The luteal phase begins with the formation of the corpus luteum from the ruptured follicle (Fig. 385-8). Progesterone and inhibin A are produced from the luteinized granulosa cells, which continue to aromatize theca-derived androgen precursors, producing estradiol. The combined actions of estrogen and progesterone are responsible for the secretory changes in the endometrium that are necessary for implantation. The corpus luteum is supported by LH but has a finite life span because of diminished sensitivity to LH. The demise of the corpus luteum results in a progressive decline in hormonal support of the endometrium. Inflammation or local hypoxia and ischemia result in vascular changes in the endometrium, leading to the release of cytokines, cell death, and shedding of the endometrium.

If conception occurs, hCG produced by the trophoblast binds to LH receptors on the corpus luteum, maintaining steroid hormone production and preventing involution of the corpus luteum until its hormonal function is taken over by the placenta 6–10 weeks after conception.

CLINICAL ASSESSMENT OF OVARIAN FUNCTION

Menstrual bleeding should become regular within 2–4 years of menarche, although anovulatory and irregular cycles are common before that. For the remainder of adult reproductive life, the cycle length counted from the first day of menses to the day preceding subsequent menses is ~28 days, with a range of 25–35 days. However, cycle-to-cycle variability for an individual woman is ±2 days. Luteal phase length is relatively constant between 12 and 14 days in normal cycles; thus, the major variability in cycle length is due to variations in follicular phase length. The duration of menstrual bleeding in ovulatory cycles varies between 4 and 6 days. There is a gradual shortening of cycle length with age such that women aged >35 years have cycles that are shorter than during their younger reproductive years. Anovulatory cycles increase as women approach menopause, and bleeding patterns may be erratic.

Women who report regular monthly bleeding with cycles that do not vary by >4 days generally have ovulatory cycles, but several other clinical signs can be used to assess the likelihood of ovulation. Some women experience *mittelschmerz*, described as midcycle pelvic discomfort that is thought to be caused by the rapid expansion of the dominant follicle at the time of ovulation. A constellation of premenstrual moliminal symptoms such as bloating, breast tenderness, and food cravings often occur several days before menses in ovulatory cycles, but their absence cannot be used as evidence of anovulation. Methods that can be used to determine whether ovulation is likely include a serum progesterone level >5 ng/mL ~7 days before expected menses, an increase in basal body temperature of 0.24°C (>0.5°F) in the second half of the cycle due to the thermoregulatory effect of progesterone, or the detection of the urinary LH surge using ovulation predictor kits. Because ovulation occurs ~36 h after the LH surge, urinary LH can be helpful in timing intercourse to coincide with ovulation.

Ultrasound can be used to detect the growth of the fluid-filled antrum of the developing follicle and to assess endometrial proliferation in response to increasing estradiol levels in the follicular phase. It can also be used to provide evidence of ovulation by documenting collapse of the dominant follicle and/or the presence of a corpus luteum as well as the characteristic echogenicity of the secretory endometrium of the luteal phase.

PUBERTY

NORMAL PUBERTAL DEVELOPMENT IN GIRLS

The first menstrual period (*menarche*) occurs relatively late in the series of developmental milestones that characterize normal pubertal development **(Table 385-1)**. Menarche is preceded by the appearance of pubic and then axillary hair (*adrenarche*) as a result of maturation of the zona reticularis in the adrenal gland and increased adrenal androgen secretion, particularly dehydroepiandrosterone (DHEA). The triggers for adrenarche remain unknown but may involve increases in body mass index, as well as in utero and neonatal factors. Menarche is also preceded by breast development (*thelarche*). The breast is exquisitely sensitive to the very low levels of estrogen that result from peripheral conversion of adrenal androgens and the low levels of estrogen secreted from the ovary early in pubertal maturation. Breast development precedes the appearance of pubic and axillary hair in ~60% of girls. The interval between the onset of breast development and menarche is ~2 years. There has been a gradual decline in the age of menarche over the past century, attributed in large part to improvement in nutrition, and there is a relationship between adiposity and earlier sexual maturation in girls. In the United States, menarche occurs at an average age of 12.5 years (Table 385-1).

TABLE 385-1 Mean Age (Years) of Pubertal Milestones in Girls

	ONSET OF BREAST/ PUBIC HAIR DEVELOPMENT	AGE OF PEAK HEIGHT VELOCITY	MENARCHE	FINAL BREAST/ PUBIC HAIR DEVELOPMENT	ADULT HEIGHT
White	10.2	11.9	12.6	14.3	17.1
Black	9.6	11.5	12	13.6	16.5

Source: From FM Biro et al: J Pediatr 148:234, 2006.

Much of the variation in the timing of puberty is due to genetic factors. Heritability estimates from twin studies range between 50 and 80%. Adrenarche and thelarche occur ~1 year earlier in black compared with white girls, although the difference in the timing of menarche is less pronounced. Genome-wide association studies have identified over a hundred genes associated with pubertal timing in boys and girls attesting to the high degree of coordination of this reproductive and growth milestone. These findings include genes involved in GnRH secretion (e.g., *TACR3,* and the maternally imprinted gene, *MKRN3*, that has been associated with familial precocious puberty), pituitary development and function (e.g., *POU1F1*), hormone synthesis and bioactivity (e.g., *STARD4, ESR1, RXRG*), gonadal feedback (e.g., *INHBA, ESR1*), and energy homeostasis and growth including *LIN28B*, a sentinel puberty gene, which is a potent regulator of microRNA processing.

Other important hormonal changes also occur in conjunction with puberty. Growth hormone (GH) levels increase early in puberty, stimulated in part by the pubertal increase in estrogen secretion. GH increases insulin-like growth factor-I (IGF-I), which enhances linear growth. The growth spurt is generally less pronounced in girls than in boys, with a peak growth velocity of ~7 cm/year. Linear growth is ultimately limited by closure of epiphyses in the long bones as a result of prolonged exposure to estrogen. Puberty is also associated with mild insulin resistance.

DISORDERS OF PUBERTY

The differential diagnosis of precocious and delayed puberty is similar in boys **(Chap. 384)** and girls. However, there are differences in the timing of normal puberty and differences in the relative frequency of specific disorders in girls compared with boys.

Precocious Puberty Traditionally, precocious puberty has been defined as the development of secondary sexual characteristics before the age of 8 in girls based on data from Marshall and Tanner in British girls studied in the 1960s. More recent studies led to recommendations that girls be evaluated for precocious puberty if breast development or pubic hair is present at <7 years of age for white girls or <6 years for black girls; however, these guidelines have not been widely accepted in favor of careful follow-up in girls presenting at <8 years.

Precocious puberty in girls is most often centrally mediated **(Table 385-2)**, resulting from early activation of the hypothalamic-pituitary-ovarian axis. It is characterized by pulsatile LH secretion (which is initially associated with deep sleep) and an enhanced LH and FSH response to exogenous GnRH or a GnRH agonist (two- to threefold stimulation) **(Table 385-3)**. True precocity is marked by advancement in bone age of >2 standard deviations, a recent history of growth acceleration, and progression of secondary sexual characteristics. In girls, centrally mediated precocious puberty (CPP) is idiopathic in ~85% of cases; however, neurogenic causes must be considered. Activating mutations in *KISS* and *KISS1R* have been found in a small number of patients with CPP, and loss of function mutations in *MKRN3* have been reported in familial CPP. However, the frequency of these mutations is insufficient to justify their use in routine clinical testing. GnRH agonists that induce pituitary desensitization are the mainstay of treatment to prevent premature epiphyseal closure and preserve adult height, as well as to manage psychosocial repercussions of precocious puberty.

Peripherally mediated precocious puberty does not involve activation of the hypothalamic-pituitary-ovarian axis and is characterized by suppressed gonadotropins in the presence of elevated estradiol. Management of peripheral precocious puberty involves treating the underlying disorder (Table 385-2) and limiting the effects of gonadal steroids using aromatase inhibitors, inhibitors of steroidogenesis, and ER blockers. It is important to be aware that central precocious puberty can also develop in girls whose precocity was initially peripherally mediated, as in McCune-Albright syndrome and congenital adrenal hyperplasia.

Incomplete and intermittent forms of precocious puberty may also occur. For example, premature breast development may occur in girls before the age of 2 years, with no further progression and without significant advancement in bone age, estrogen production, or compromised height. Premature adrenarche can also occur in the absence of progressive pubertal development, but it must be distinguished from late-onset congenital adrenal hyperplasia and androgen-secreting tumors, in which case it may be termed *heterosexual precocity*. Premature

TABLE 385-2 Differential Diagnosis of Precocious Puberty

CENTRAL (GnRH DEPENDENT)	PERIPHERAL (GnRH INDEPENDENT)
Idiopathic	Congenital adrenal hyperplasia
CNS tumors	Estrogen-producing tumors
Hamartomas	Adrenal tumors
Astrocytomas	Ovarian tumors
Adenomyomas	Gonadotropin/hCG-producing tumors
Gliomas	Exogenous exposure to estrogen or androgen or lavender or tea-tree oil
Germinomas	McCune-Albright syndrome
CNS infection	Aromatase excess syndrome
Head trauma	
Iatrogenic	
Radiation	
Chemotherapy	
Surgical	
CNS malformation	
Arachnoid or suprasellar cysts	
Septo-optic dysplasia	
Hydrocephalus	

Abbreviations: CNS, central nervous system; GnRH, gonadotropin-releasing hormone; hCG, human chorionic gonadotropin.

TABLE 385-3 Evaluation of Precocious and Delayed Puberty

	PRECOCIOUS	DELAYED
Initial Screening Tests		
History and physical	×	×
Assessment of growth velocity	×	×
Bone age	×	×
LH, FSH	×	×
Estradiol, testosterone	×	×
DHEAS	×	×
17-Hydroxyprogesterone	×	
TSH, T_4	×	×
Complete blood count		×
Sedimentation rate, C-reactive protein		×
Electrolytes, renal function		×
Liver enzymes		×
IGF-I, IGFBP-3		×
Urinalysis		×
Secondary Tests		
Pelvic ultrasound	×	×
Cranial MRI	×	×
β-hCG	×	
GnRH/agonist stimulation test	×	×
ACTH stimulation test	×	
Inflammatory bowel disease panel	×	×
Celiac disease panel		×
Prolactin		×
Karyotype		×

Abbreviations: ACTH, adrenocorticotropic hormone; DHEAS, dehydroepiandrosterone sulfate; FSH, follicle-stimulating hormone; hCG, human chorionic gonadotropin; IGF-I, insulin-like growth factor-I; IGFBP-3, IGF-binding protein 3; LH, luteinizing hormone; MRI, magnetic resonance imaging; TSH, thyroid-stimulating hormone; T_4, thyroxine.

adrenarche may be associated with obesity, hyperinsulinemia, and the subsequent predisposition to PCOS.

Delayed Puberty Delayed puberty **(Table 385-4)** is defined as the absence of secondary sexual characteristics by age 13 in girls. The diagnostic considerations are very similar to those for primary amenorrhea **(Chap. 386)**. Between 25 and 40% of delayed puberty in girls is of ovarian origin, with Turner's syndrome accounting for the majority of such patients. Delayed puberty may occur in the setting of systemic illnesses, including celiac disease and chronic renal disease, and endocrinopathies such as diabetes and hypothyroidism. In addition, girls appear to be particularly susceptible to the adverse effects of decreased energy balance resulting from exercise, dieting, and/or eating disorders and thus, functional hypothalamic amenorrhea (HA) can present with primary amenorrhea. Together these reversible conditions account for ~25% of delayed puberty in girls. Congenital hypogonadotropic hypogonadism in girls or boys can be caused by mutations in several different genes or combinations of genes (Fig. 385-4, **Chap. 384**, Table 385-2). Approximately 50% of girls with congenital hypogonadotropic hypogonadism, with or without anosmia, have a history of some degree of breast development, and 10% report one to two episodes of vaginal bleeding. Family studies suggest that genes identified in association with absent puberty may also cause delayed puberty, and recent reports have further suggested that a genetic susceptibility to environmental stresses such as diet and exercise may account for at least some cases of functional HA, including in girls who present with primary amenorrhea. Although neuroanatomic causes of delayed puberty are considerably less common in girls than in boys, it is always important to rule these out in the setting of hypogonadotropic hypogonadism.

TABLE 385-4 Differential Diagnosis of Delayed Puberty

Hypergonadotropic
Ovarian
Turner's syndrome
Gonadal dysgenesis
Chemotherapy/radiation therapy
Galactosemia
Autoimmune oophoritis
Congenital lipoid hyperplasia
Steroidogenic enzyme abnormalities
17α-Hydroxylase deficiency
Aromatase deficiency
Gonadotropin/receptor mutations
*FSH*β, *LHR*, *FSHR*
Androgen resistance syndrome
Hypogonadotropic
Genetic
Hypothalamic syndromes
Leptin/leptin receptor
HESX1 (septo-optic dysplasia)
PC1 (prohormone convertase)
IHH and Kallmann's syndrome
KAL1, FGF8, FGFR1, NSMF, PROK2, PROKR2, SEM3A, HS6ST1, WDR11, CHD7
KISS1, KISS1R, TAC3, TAC3R, GnRH1, GnRHR, and others
Abnormalities of pituitary development/function
PROP1
CNS tumors/infiltrative disorders
Craniopharyngioma
Astrocytoma, germinoma, glioma
Prolactinomas, other pituitary tumors
Histiocytosis X
Chemotherapy/radiation
Functional
Chronic diseases
Malnutrition
Excessive exercise
Eating disorders

Abbreviations: CHD7, chromodomain-helicase-DNA-binding protein 7; CNS, central nervous system; *FGF8*, fibroblast growth factor 8; *FGFR1*, fibroblast growth factor 1 receptor; *FSH*β, follicle-stimulating hormone β chain; *FSHR*, FSH receptor; *GNRHR*, gonadotropin-releasing hormone receptor; *HESX1*, homeobox, embryonic stem cell expressed 1; *HS6ST1*, heparin sulfate 6-O sulfotransferase 1; IHH, idiopathic hypogonadotropic hypogonadism; KAL, Kallmann; *KISS1*, kisspeptin 1; *KISSR1*, KISS1 receptor; *LHR*, luteinizing hormone receptor; *NSMF*, NMDA receptor synaptonuclear signaling and neuronal migration factor; *PROK2*, prokineticin 2; *PROKR2* prokineticin receptor 2; *PROP1*, prophet of Pit1, paired-like homeodomain transcription factor *SEMA3A*, semaphorin-3A; *WDR11*, WD repeat-containing protein 11.

FURTHER READING

Devoto L et al: The significance of estradiol metabolites in human corpus luteum physiology. Steroids 123:50, 2017.

Francois CM et al: A novel action of follicle-stimulating hormone in the ovary promotes estradiol production without inducing excessive follicular growth before puberty. Sci Rep 7:46222, 2017.

George JT et al: Effect of gonadotropin-inhibitory hormone on luteinizing hormone secretion in humans. Clin Endocrinol (Oxf) 86:731, 2017.

Kaplowitz P, Bloch C: The section on endocrinology. Evaluation and referral of children with signs of early puberty. Pediatrics 137:e20153732, 2016.

Macedo DB et al: Sexual precocity—Genetic bases of central precocious puberty and automonous gonadal activation. Endocr Dev 29:50, 2016.

Marshall SA et al: The role of relaxin in normal and abnormal uterine function during the menstrual cycle and early pregnancy. Reprod Sci 24:342, 2017.

Perry JR et al: Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. Nature 514:92, 2014.

Skorupskaite K et al: Interactions between neurokinin B and kisspeptin in mediating estrogen feedback in healthy women. J Clin Endocrinol Metab 101:4628, 2016.

Stamou MI et al: Discovering genes essential to the hypothalamic regulation of human reproduction using a human disease model: Adjusting to life in the "-omics" era. Endocr Rev 36:603, 2015.

Tsutsui K, Ubuka T: GnIH control of feeding and reproductive behaviors. Front Endocrinol (Lausanne) 7:170, 2016.

Tucker EJ et al: Premature ovarian insufficiency: New perspectives on genetic cause and phenotypic spectrum. Endocr Rev 37:609, 2016.

Willemsen RH, Dunger DB: Normal variation in pubertal timing: Genetic determinants in relation to growth and adiposity. Endocr Dev 29:17, 2016.

Zhu J et al: A shared genetic basis for self-limited delayed puberty and idiopathic hypergonadotropic hypogonadism. J Clin Endocrinol Metab 100:E646, 2015.

386 Menstrual Disorders and Pelvic Pain

Janet E. Hall

Menstrual dysfunction can signal an underlying abnormality that may have long-term health consequences. Although frequent or prolonged bleeding usually prompts a woman to seek medical attention, infrequent or absent bleeding may seem less troubling and the patient may not bring it to the attention of the physician. Thus, a focused menstrual history is a critical part of every encounter with a female patient.

Pelvic pain is a common complaint that may relate to an abnormality of the reproductive organs but also may be of gastrointestinal, urinary tract, or musculoskeletal origin. Depending on its cause, pelvic pain may require urgent surgical attention. Recent guidelines no longer recommend routine pelvic examination in asymptomatic, average-risk women other than periodic cervical cancer screening. However, pelvic examination is an important part of the evaluation of amenorrhea, abnormal uterine bleeding, and pelvic pain.

MENSTRUAL DISORDERS

DEFINITION AND PREVALENCE

Amenorrhea refers to the absence of menstrual periods. Amenorrhea is classified as *primary* if menstrual bleeding has never occurred in the absence of hormonal treatment or *secondary* if menstrual periods cease for 3–6 months. Primary amenorrhea is a rare disorder that occurs in <1% of the female population. However, between 3 and 5% of women experience at least 3 months of secondary amenorrhea in any specific year. There is no evidence that race or ethnicity influences the prevalence of amenorrhea. However, because of the importance of adequate nutrition for normal reproductive function, both the age at menarche and the prevalence of secondary amenorrhea vary significantly in different parts of the world.

Oligomenorrhea is defined as a cycle length >35 days or <10 menses per year. Both the frequency and the amount of vaginal bleeding are irregular in oligomenorrhea, and moliminal symptoms (premenstrual breast tenderness, food cravings, mood lability), suggestive of ovulation, are variably present. Anovulation can also present with intermenstrual intervals <24 days or vaginal bleeding for >7 days. Frequent or heavy irregular bleeding is termed *dysfunctional uterine bleeding* if anatomic uterine and outflow tract lesions or a bleeding diathesis have been excluded. Oligo- or anovulation are most frequently associated with polycystic ovarian syndrome (PCOS).

Primary Amenorrhea The absence of menarche (the first menstrual period) by age 16 has been used traditionally to define primary amenorrhea. However, other factors, such as growth, secondary sexual characteristics, and the presence of cyclic pelvic pain, also influence the age at which primary amenorrhea should be investigated. Recent studies suggest that puberty is occurring at an earlier age, particularly in obese girls. However, it is important to note that these data reflect earlier breast development alone with minimal change in the age of menarche. Thus, an evaluation for amenorrhea should be initiated by age 15 or 16 in the presence of normal growth and secondary sexual characteristics; age 13 in the absence of secondary sexual characteristics or if height is less than the third percentile; age 12 or 13 in the presence of breast development and cyclic pelvic pain; or within 2 years of breast development if menarche, has not occurred.

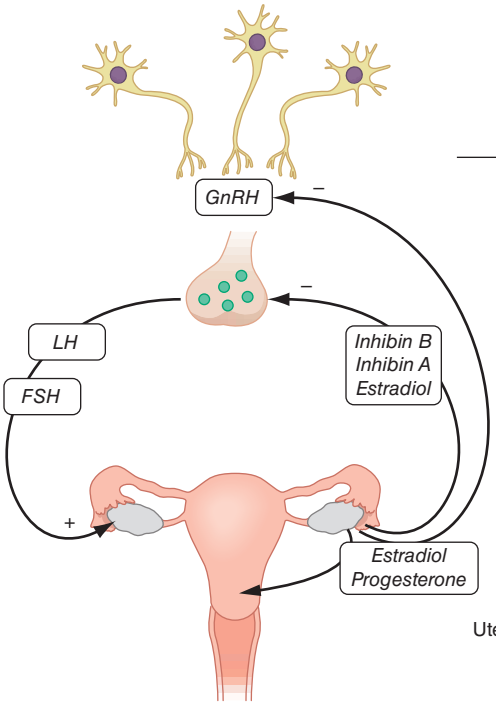
Secondary Amenorrhea or Oligomenorrhea Irregular cycles are relatively common for up to 3 years after menarche and for 1–2 years before the final menstrual period. In the intervening years, menstrual cycle length is ~28 days, with an intermenstrual interval normally ranging between 25 and 35 days. Cycle-to-cycle variability in an individual woman who is ovulating consistently is generally +/− 2 days. Pregnancy is the most common cause of amenorrhea and should be excluded early in any evaluation of menstrual irregularity. However, many women occasionally miss a single period. Three months of secondary amenorrhea, or 6 months in women with previously irregular cycles, should prompt an evaluation, as should a history of intermenstrual intervals >35 or <21 days or bleeding that persists for >7 days.

DIAGNOSIS

Pregnancy is the most common cause of amenorrhea, and must be excluded in all cases, regardless of patient history. Evaluation of menstrual dysfunction depends on understanding the interrelationships between the four critical components of the reproductive tract: (1) the hypothalamus, (2) the pituitary, (3) the ovaries, and (4) the uterus and outflow tract (**Fig. 386-1**; **Chap. 385**). This system is maintained by complex negative and positive feedback loops involving the ovarian steroids (estradiol and progesterone) and peptides (inhibin B and inhibin A) and the hypothalamic (gonadotropin-releasing hormone [GnRH]) and pituitary (follicle-stimulating hormone [FSH] and luteinizing hormone [LH]) components of this system (Fig. 386-1).

Disorders of menstrual function can be thought of in two main categories: disorders of the uterus and outflow tract and disorders of ovulation. Many of the conditions that cause primary amenorrhea are congenital but go unrecognized until the time of normal puberty (e.g., genetic, chromosomal, and anatomic abnormalities). All causes of secondary amenorrhea also can cause primary amenorrhea.

Disorders of the Uterus or Outflow Tract Abnormalities of the uterus and outflow tract typically present as primary amenorrhea. In patients with normal pubertal development and a blind vagina, the differential diagnosis includes *obstruction* by a transverse vaginal septum or imperforate hymen; *müllerian agenesis* (Mayer-Rokitansky-Kuster-Hauser syndrome), which can be caused by mutations in the *WNT4* gene; and *androgen insensitivity syndrome* (AIS), which is an X-linked recessive disorder that accounts for ~10% of all cases of primary amenorrhea **(Chap. 384)**. Patients with AIS have a 46,XY karyotype, but because of the lack of androgen receptor responsiveness, those with complete AIS lack features of androgenization and have female external genitalia. The absence of pubic and axillary hair



	Primary	Secondary
Hypothalamus	27%	36%
Pituitary	2%	15%
PCOS	7%	30%
Ovary	43%	12%
Uterus/outflow tract	19%	7%

FIGURE 386-1 Role of the hypothalamic-pituitary-gonadal axis in the etiology of amenorrhea. Gonadotropin-releasing hormone (GnRH) secretion from the hypothalamus stimulates follicle-stimulating hormone (FSH) and luteinizing hormone (LH) secretion from the pituitary to induce ovarian folliculogenesis and steroidogenesis. Ovarian secretion of estradiol and progesterone controls the shedding of the endometrium, resulting in menses, and, in combination with the inhibins, provides feedback regulation of the hypothalamus and pituitary to control secretion of FSH and LH. The prevalence of amenorrhea resulting from abnormalities at each level of the reproductive system (hypothalamus, pituitary, ovary, uterus, and outflow tract) varies depending on whether amenorrhea is primary or secondary. PCOS, polycystic ovarian syndrome.

distinguishes them clinically from patients with müllerian agenesis, as does a testosterone level in the male range. The rare patient with 5α reductase type 2 enzyme deficiency has a similar presentation, but undergoes virilization at the time of puberty. *Asherman's syndrome* presents as secondary amenorrhea or hypomenorrhea and results from partial or complete obliteration of the uterine cavity by adhesions that prevent normal growth and shedding of the endometrium. Curettage performed for pregnancy complications accounts for >90% of cases; genital tuberculosis is an important cause in regions where it is endemic.

TREATMENT

Disorders of the Uterus or Outflow Tract

Obstruction of the outflow tract requires surgical correction. It is important that this be performed as soon as the diagnosis is made as the risk of endometriosis is increased with retrograde menstrual flow. *Müllerian agenesis* may require surgical intervention to allow sexual intercourse, although vaginal dilatation is adequate in some patients. Because ovarian function is normal, assisted reproductive techniques can be used with a surrogate carrier. *Androgen resistance syndrome* requires gonadectomy because there is risk of gonadoblastoma in the dysgenetic gonads, although surgery is generally delayed until after breast development and the pubertal growth spurt. Estrogen replacement is indicated after gonadectomy, and vaginal dilatation may be required to allow sexual intercourse.

Disorders of Ovulation Once uterus and outflow tract abnormalities have been excluded, other causes of amenorrhea involve disorders of ovulation. The differential diagnosis is based on the results of initial tests, including a pregnancy test, an FSH level (to determine whether the cause is likely to be ovarian or central), and assessment of hyperandrogenism **(Fig. 386-2)**.

HYPOGONADOTROPIC HYPOGONADISM Low estrogen levels in combination with normal or low levels of LH and FSH are seen with anatomic, genetic, or functional abnormalities that interfere with hypothalamic GnRH secretion or normal pituitary responsiveness to GnRH. Although relatively uncommon, tumors and infiltrative diseases should be considered in the differential diagnosis of hypogonadotropic hypogonadism **(Chap. 373)**. These disorders may present with primary or secondary amenorrhea. They may occur in association with other features suggestive of hypothalamic or pituitary dysfunction, such as short stature, diabetes insipidus, galactorrhea, and headache. Hypogonadotropic hypogonadism also may be seen after cranial irradiation. In the postpartum period, amenorrhea occurs normally in association with breast feeding, but may also be caused by pituitary necrosis (Sheehan's syndrome) or lymphocytic hypophysitis. Because reproductive dysfunction is commonly associated with hyperprolactinemia from neuroanatomic lesions or medications, prolactin should be measured in all patients with hypogonadotropic hypogonadism **(Chap. 373)**.

Isolated hypogonadotropic hypogonadism (IHH) occurs in women, although it is three times more common in men. IHH generally presents with primary amenorrhea, although 50% have some degree of breast development, and ~10% report one to two menses. IHH is associated with anosmia in half of women (termed Kallmann's syndrome). Genetic causes of IHH have been identified in ~50% of patients **(Chaps. 384 and 385)**.

Functional hypothalamic amenorrhea (HA) is a diagnosis of exclusion of other causes of hypogonadotropic hypogonadism including chronic diseases (type 1 diabetes, celiac disease, hyperthyroidism, Cushing Syndrome) and use of opioids, glucocorticoids or psychotropic medications that increase prolactin levels. Functional HA is most commonly associated with conditions causing a mismatch between energy expenditure and energy intake and/or significant stress. Variants in genes associated with IHH may increase susceptibility to these environmental inputs, accounting in part for the clinical variability in this disorder. Metabolic and stress signaling is transduced to the

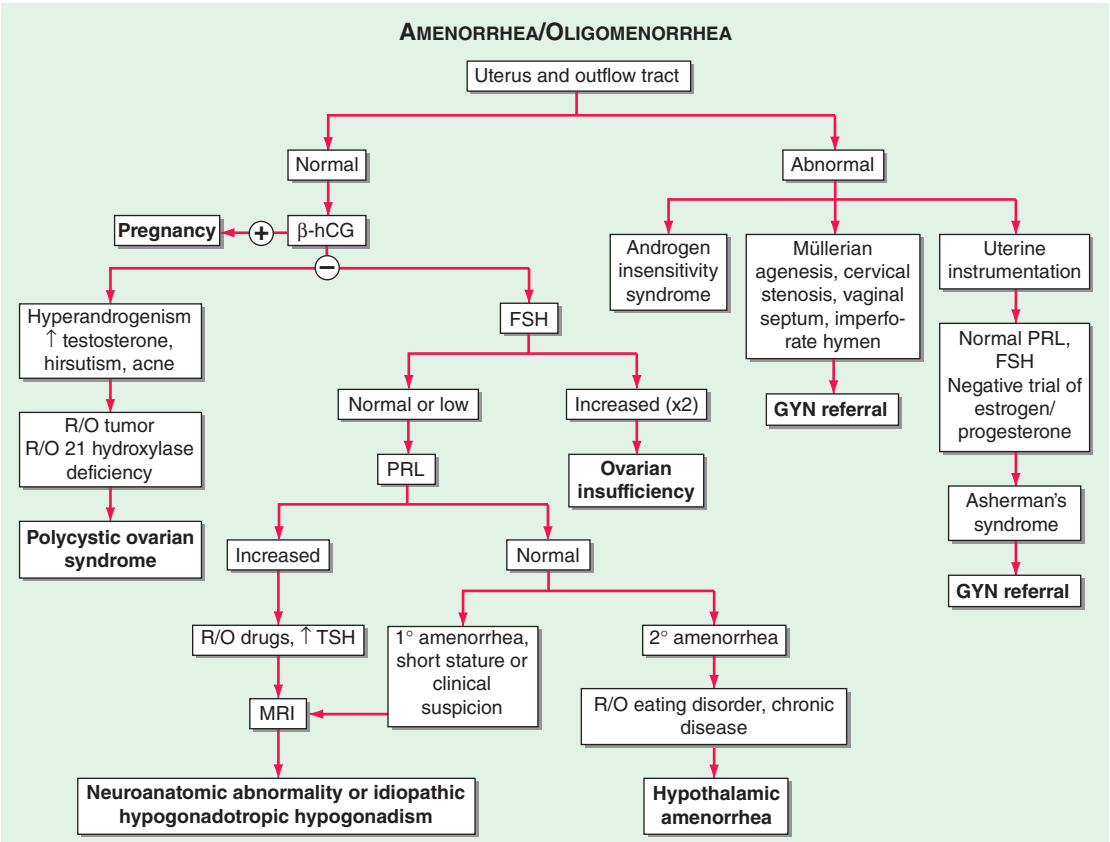


FIGURE 386-2 Algorithm for evaluation of amenorrhea. β-hCG, human chorionic gonadotropin; FSH, follicle-stimulating hormone; GYN, gynecologist; MRI, magnetic resonance imaging; PRL, prolactin; R/O, rule out; TSH, thyroid-stimulating hormone.

reproductive axis, at least in part, through leptin signaling from the periphery and via hypothalamic kisspeptin control of GnRH. The diagnosis of HA generally can be made on the basis of a careful history, a physical examination, and the demonstration of low levels of gonadotropins and normal prolactin levels. Eating disorders and chronic disease must be specifically excluded. An atypical history, headache, signs of other hypothalamic dysfunction, or hyperprolactinemia, even if mild, necessitates cranial magnetic resonance imaging (MRI) to exclude a neuroanatomic cause. Up to 10% of women with HA may have some features of PCOS (increased ovarian volume, higher anti-mullerian hormone [AMH] levels, and slightly elevated androgen levels).

HYPERGONADOTROPIC HYPOGONADISM Ovarian failure is considered premature when it occurs in women <40 years old and accounts for ~10% of secondary amenorrhea. *Primary ovarian insufficiency* (POI) has replaced the terms *premature menopause* and *premature ovarian failure* in recognition of the continuum of impaired ovarian function encompassed by this disorder. Ovarian insufficiency is associated with the loss of negative-feedback restraint on the hypothalamus and pituitary, resulting in increased FSH and LH levels. FSH is a better marker of ovarian failure because of loss of negative feedback effects of both estradiol and the inhibins and because its levels are less variable than those of LH. AMH levels will also be low in patients with POI, but are more frequently used in management of infertility. As with natural menopause, POI may wax and wane, and serial measurements may be necessary to establish the diagnosis.

Once the diagnosis of POI has been established, further evaluation is indicated because of other health problems that may be associated with POI. Although POI is most commonly of unknown cause, it also occurs in association with a variety of chromosomal abnormalities (most often Turner's syndrome), autoimmune polyglandular failure syndromes, and other rare disorders. Radiotherapy and chemotherapy may reduce ovarian reserve, with effects on both the oocytes and the supporting granulosa cells. New approaches, including ovarian preservation, are being developed to support long-term fertility choices in women of reproductive age prior to oncologic treatment. The recognition that early ovarian failure occurs in premutation carriers of the fragile X syndrome is important because of the increased risk of severe mental retardation in male children with *FMR1* mutations. Thus, follow-up testing should include a karyotype in all POI patients, serum anticortical and 21-hydroxylase antibodies (specific but not sensitive for subsequent adrenal insufficiency), thyroid function and thyroid peroxidase antibodies, *FMR1* premutation screening, and assessment of bone mineral density. Ovarian biopsy is of no diagnostic or prognostic value. Although the number of genetic causes POI is increasing, routine testing for mutations other than *FMR1* is currently not recommended.

Hypergonadotropic hypogonadism occurs rarely in other disorders, such as mutations in the FSH or LH receptors. Aromatase deficiency and 17α-hydroxylase deficiency are associated with decreased estrogen and elevated gonadotropins and with hyperandrogenism and hypertension, respectively. Gonadotropin-secreting tumors in women of reproductive age generally present with high, rather than low, estrogen levels and cause ovarian hyperstimulation or dysfunctional bleeding.

TREATMENT

Hypo- and Hypergonadotropic Causes of Amenorrhea

Amenorrhea almost always is associated with chronically low levels of estrogen, whether it is caused by hypogonadotropic hypogonadism or ovarian insufficiency. Development of secondary sexual characteristics requires gradual titration of estradiol replacement with eventual addition of progestin. Hormone replacement with either low-dose estrogen/progesterone regimens or oral contraceptive pills is recommended until the usual age of menopause for bone and cardiovascular protection. In women with functional HA or anorexia nervosa, hormone replacement alone may not be sufficient to restore or maintain bone density. Patients with hypogonadotropic hypogonadism who are interested in fertility require treatment with both exogenous FSH and LH or pulsatile GnRH. Patients with ovarian failure can consider oocyte donation, which has a high rate of success in this population, although its use in women with Turner's syndrome is limited by significant maternal cardiovascular risk.

POLYCYSTIC OVARIAN SYNDROME PCOS is diagnosed based on a combination of clinical or biochemical evidence of hyperandrogenism, amenorrhea or oligomenorrhea, and the ultrasound appearance of polycystic ovaries. Approximately half of patients with PCOS are obese, and abnormalities in insulin dynamics are common, as is metabolic syndrome. Symptoms generally begin shortly after menarche and are slowly progressive. Lean oligo-ovulatory patients with PCOS generally have high LH levels in the presence of normal to low levels of FSH and estradiol, although these may be suppressed by undernutrition or stress. The LH/FSH ratio is less pronounced in obese patients in whom insulin resistance is a more prominent feature.

TREATMENT

Polycystic Ovarian Syndrome

A major abnormality in patients with PCOS is the failure of regular, predictable ovulation. Thus, these patients are at risk for the development of dysfunctional bleeding and endometrial hyperplasia associated with unopposed estrogen exposure. Endometrial protection can be achieved with the use of oral contraceptives or progestins (medroxyprogesterone acetate, 5–10 mg, or prometrium, 200 mg daily for 10–14 days of each month). Oral contraceptives are also useful for management of hyperandrogenic symptoms, as are spironolactone and cyproterone acetate (not available in the United States), which function as weak androgen receptor blockers. Management of the associated metabolic syndrome may be appropriate for some patients (Chap. 401). For patients interested in fertility, weight control is a critical first step. Clomiphene citrate is highly effective as a first-line treatment, as is the aromatase inhibitor letrozole. Exogenous gonadotropins can be used by experienced practitioners; a diagnosis of polycystic ovaries increases the risk of hyperstimulation, even in women with regular, ovulatory menstrual cycles. Metformin is frequently used in patients with PCOS, and is appropriate as an adjunct with diet and exercise for obese women with PCOS, or for treatment of diabetes or impaired glucose tolerance, as in non-PCOS patients. However, metformin is not recommended for endometrial protection or treatment of hyperandrogenic symptoms, infertility, pregnancy loss or prevention of gestational diabetes.

PELVIC PAIN

The mechanisms that cause pelvic pain are similar to those that cause abdominal pain (Chap. 12) and include inflammation of the parietal peritoneum, obstruction of hollow viscera, vascular disturbances, and pain originating in the abdominal wall. Pelvic pain may reflect pelvic disease per se but also may reflect extrapelvic disorders that refer pain to the pelvis. In up to 60% of cases, pelvic pain can be attributed to gastrointestinal problems, including appendicitis, cholecystitis, infections, intestinal obstruction, diverticulitis, and inflammatory bowel disease. Urinary tract and musculoskeletal disorders are also common causes of pelvic pain.

APPROACH TO THE PATIENT

Pelvic Pain

As with all types of abdominal pain, the first priority is to identify life-threatening conditions (shock, peritoneal signs) that may require emergent surgical management. The possibility of pregnancy should be identified as soon as possible by menstrual history and/or testing. A thorough history that includes the type, location, radiation,

TABLE 386-1 Causes of Pelvic Pain

	ACUTE	CHRONIC
Cyclic pelvic pain		Mittelschmerz Dysmenorrhea Endometriosis
Noncyclic pelvic pain	Pelvic inflammatory disease Ruptured or hemorrhagic ovarian cyst, endometrioma, or ovarian torsion Ectopic pregnancy Endometritis Acute growth or degeneration of uterine myoma Threatened abortion	Pelvic congestion syndrome Adhesions and retroversion of the uterus Pelvic malignancy Vulvodynia Chronic pelvic inflammatory disease Tuberculous salpingitis History of sexual abuse

and status with respect to increasing or decreasing severity can help identify the cause of acute pelvic pain. Specific associations with vaginal bleeding, sexual activity, defecation, urination, movement, or eating should be specifically sought. Determination of whether the pain is acute versus chronic and cyclic versus noncyclic will direct further investigation **(Table 386-1)**. However, disorders that cause cyclic pain occasionally may cause noncyclic pain, and the converse is also true.

ACUTE PELVIC PAIN

Pelvic inflammatory disease (PID) most commonly presents with bilateral lower abdominal pain. It is generally of recent onset and is exacerbated by intercourse or jarring movements. Fever is present in about half of these patients; abnormal uterine bleeding occurs in about one-third. New vaginal discharge, urethritis, and chills may be present but are less specific signs. With public health efforts to control sexually transmitted diseases, the rate and severity of PID have declined in the United States and Europe; however, this is not the case in the developing world. Subclinical PID with its attendant risks of infertility and ectopic pregnancy remains a significant problem worldwide. Public health and professional organizations recommend annual testing for *C. trachomatis* in all sexually active women <25 and both *C. trachomatis* and *N. gonorrhea* in all women at increased risk. *Adnexal pathology* can present acutely and may be due to rupture, bleeding or torsion of cysts, or, much less commonly, the fallopian tubes. Neoplasms of the ovary or fallopian tube are much less common causes. Fever may be present with ovarian torsion. *Ectopic pregnancy* is associated with right- or left-sided lower abdominal pain, with clinical signs generally appearing 6–8 weeks after the last normal menstrual period. Vaginal bleeding occurs in ~50% of cases. Orthostatic signs and fever may be present. Risk factors include the presence of known tubal disease, previous ectopic pregnancies, a history of infertility, diethylstilbestrol (DES) exposure of the mother in utero, or a history of pelvic infections. Rupture of the fallopian tube remains a life-threatening emergency; the incidence depends on access to care but is ~18% in developed countries. *Threatened abortion* may also present with amenorrhea, abdominal pain, and vaginal bleeding. Although more common than ectopic pregnancy, it is rarely associated with systemic signs. *Uterine pathology* includes endometritis and, less frequently, degenerating leiomyomas (fibroids). Endometritis often is associated with vaginal bleeding and systemic signs of infection. It occurs in the setting of sexually transmitted infections, uterine instrumentation, or postpartum infection.

A sensitive pregnancy test, complete blood count with differential, urinalysis, tests for chlamydial and gonococcal infections, and abdominal ultrasound aid in making the diagnosis and directing further management.

TREATMENT

Acute Pelvic Pain

Treatment of acute pelvic pain depends on the suspected etiology but may require surgical or gynecologic intervention. Conservative management is an important consideration for ovarian cysts, if torsion is not suspected, to avoid unnecessary pelvic surgery and the subsequent risk of infertility due to adhesions. Surgical treatment may be required for ectopic pregnancies; however, women presenting with unruptured ectopic pregnancies may be appropriate for treatment with methotrexate, which is effective in ~90% of cases when multiple doses are used.

TREATMENT

Chronic Pelvic Pain

Some women experience discomfort at the time of ovulation (*mittelschmerz*). The pain can be quite intense but is generally of short duration. The mechanism is thought to involve rapid expansion of the dominant follicle, although it also may be caused by peritoneal irritation by follicular fluid released at the time of ovulation.

DYSMENORRHEA

Dysmenorrhea refers to the crampy lower abdominal midline discomfort that begins with the onset of menstrual bleeding and gradually decreases over the next 12–72 h. It may be associated with nausea, diarrhea, fatigue, and headache and occurs in 60–93% of adolescents, beginning with the establishment of regular ovulatory cycles. Its prevalence decreases after pregnancy and with the use of oral contraceptives.

Primary dysmenorrhea results, in a majority of cases, from hormone-dependent prostaglandin (PG)-pathway mechanisms that cause intense uterine contractions, decreased blood flow, and increased peripheral nerve hypersensitivity, resulting in pain. However, variability in response to COX inhibitors suggests that PG-independent pathways, such as platelet activating factor, may also mediate inflammation.

Secondary dysmenorrhea is caused by underlying pelvic pathology. *Endometriosis* results from the presence of endometrial glands and stroma outside the uterus. These deposits of ectopic endometrium respond to hormonal stimulation and cause dysmenorrhea, which begins several days before menses. Endometriosis also may be associated with painful intercourse, painful bowel movements, and tender nodules in the uterosacral ligament. Fibrosis and adhesions can produce lateral displacement of the cervix, which is a useful sign on speculum examination. Transvaginal pelvic ultrasound is part of the initial workup and may detect an endometrioma within the ovary, rectovaginal or bladder nodules, or ureteral involvement. The CA125 level may be increased, but it has low negative predictive value. Definitive diagnosis requires laparoscopy. Symptomatology does not always predict the extent of endometriosis. The prevalence is lower in black and Hispanic women than in Caucasians and Asians. *Other secondary causes* of dysmenorrhea include adenomyosis, a condition caused by the presence of ectopic endometrial glands and stroma within the myometrium. Cervical stenosis, which may result from trauma, infection, or surgery also may cause pain associated with menses. Pelvic congestion syndrome is associated with pelvic varicosities with low blood flow.

TREATMENT

Dysmenorrhea

Local application of heat is of some benefit. Exercise, sexual activity, a vegetarian diet, use of vitamins D, B_1, B_6, and E and fish oil, acupuncture, and yoga have all been suggested to be of benefit but studies are not adequate to provide recommendations. However, nonsteroidal anti-inflammatory drugs (NSAIDs) are very effective and provide >80% sustained response rates. Ibuprofen, naproxen, ketoprofen, mefanamic acid, and nimesulide are all superior to

placebo. Treatment should be started a day before expected menses and continued for 2–3 days. Oral contraceptives also reduce symptoms of dysmenorrhea. The use of tocolytics, antiphosphodiesterase inhibitors, and magnesium has been suggested, but there are insufficient data to recommend them. Failure of response to NSAIDs and/or oral contraceptives is suggestive of a pelvic disorder such as *endometriosis*, and diagnostic laparoscopy should be considered to guide further treatment.

FURTHER READING

Bloomfield H et al: Screening pelvic examinations in asymptomatic average risk adult women. WA-ESP Project #09-009; 2013.

Bouilly J et al: Identification of multiple gene mutations accounts for the new genetic architecture of ovarian insufficiency. J Clin Endocrinol Metab 101:4541, 2016.

Brunham RC et al: Pelvic inflammatory disease. N Engl J Med 372:2039, 2015.

Fourman LR, Fazeli PK: Neuroendocrine causes of amenorrhea—An update. J Clin Endocrinol Metab 100:812, 2015.

Ju H et al: The prevalence and risk factors of dysmenorrhea. Epidem Rev 36:104, 2014.

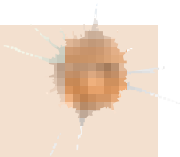
Morley LC et al: On behalf of the Royal College of Obstetricians and Gynecologists. Metformin therapy for the management of infertility in women with polycystic ovary syndrome. Scientific Impact Paper No. 13. BJOG 124:e306, 2017.

Oladosu FA et al: Nonsteroidal anti-inflammatory drug resistance in dysmenorrhea: Epidemiology, causes, and treatment. Am J Obstet Gynecol Sept 6, 2017 [Epub ahead of print].

Vercellini P et al: Endometriosis: Pathogenesis and treatment. Nat Rev Endocrinol 10:261, 2014.

387 Hirsutism

David A. Ehrmann



Hirsutism, which is defined as androgen-dependent excessive male-pattern hair growth, affects ~10% of women. Hirsutism is most often idiopathic or the consequence of androgen excess associated with polycystic ovary syndrome (PCOS). Less frequently, it may result from adrenal androgen overproduction as occurs in nonclassic congenital adrenal hyperplasia (CAH) **(Table 387-1)**. Rarely, it is a harbinger of a serious underlying condition. Cutaneous manifestations commonly associated with hirsutism include acne and male-pattern balding (androgenic alopecia). *Virilization* refers to a condition in which androgen levels are sufficiently high to cause additional signs and symptoms, such as deepening of the voice, breast atrophy, increased muscle bulk, clitoromegaly, and increased libido. Virilization may be due to benign hyperplasia of ovarian theca and stroma cells (e.g., *hyperthecosis*); it may also result from an ovarian or adrenal neoplasm.

TABLE 387-1 Causes of Hirsutism

Gonadal hyperandrogenism
Ovarian hyperandrogenism
Polycystic ovary syndrome/functional ovarian hyperandrogenism
Ovarian steroidogenic blocks
Syndromes of extreme insulin resistance
Ovarian neoplasms
Hyperthecosis
Adrenal hyperandrogenism
Premature adrenarche
Functional adrenal hyperandrogenism
Congenital adrenal hyperplasia (nonclassic and classic)
Abnormal cortisol action/metabolism
Adrenal neoplasms
Other endocrine disorders
Cushing's syndrome
Hyperprolactinemia
Acromegaly
Peripheral androgen overproduction
Obesity
Idiopathic
Pregnancy-related hyperandrogenism
Hyperreactio luteinalis
Thecoma of pregnancy
Drugs
Androgens
Oral contraceptives containing androgenic progestins
Minoxidil
Phenytoin
Diazoxide
Cyclosporine
Valproic Acid
True hermaphroditism

HAIR FOLLICLE GROWTH AND DIFFERENTIATION

Hair can be categorized as either *vellus* (fine, soft, and not pigmented) or *terminal* (long, coarse, and pigmented). The number of hair follicles does not change over an individual's lifetime, but the follicle size and type of hair can change in response to numerous factors, particularly androgens. Androgens are necessary for terminal hair and sebaceous gland development and mediate differentiation of pilosebaceous units (PSUs) into either a terminal hair follicle or a sebaceous gland. In the former case, androgens transform the vellus hair into a terminal hair; in the latter case, the sebaceous component proliferates and the hair remains vellus.

There are three phases in the cycle of hair growth: (1) *anagen* (growth phase), (2) *catagen* (involution phase), and (3) *telogen* (rest phase). Depending on the body site, hormonal regulation may play an important role in the hair growth cycle. For example, the eyebrows, eyelashes, and vellus hairs are androgen-insensitive, whereas the axillary and pubic areas are sensitive to low levels of androgens. Hair growth on the face, chest, upper abdomen, and back requires higher levels of androgens and is therefore more characteristic of the pattern typically seen in men. Androgen excess in women can lead to increased hair growth in most androgen-sensitive sites except in the scalp region, where hair loss occurs because androgens cause scalp hairs to spend less time in the anagen phase.

Although androgen excess underlies most cases of hirsutism, there is only a modest correlation between androgen levels and the quantity of hair growth. This is due to the fact that hair growth from the follicle also depends on local growth factors, and there is variability in end organ (PSU) sensitivity. Genetic factors and ethnic background also influence hair growth. In general, dark-haired individuals tend to be more hirsute than blond or fair individuals. Asians and Native Americans have relatively sparse hair in regions sensitive to high androgen levels, whereas people of Mediterranean descent are more hirsute.

CLINICAL ASSESSMENT

Historic elements relevant to the assessment of hirsutism include the age at onset and rate of progression of hair growth and associated symptoms or signs (e.g., menstrual irregularity and acne). Depending on the cause, excess hair growth typically is first noted during the second and third decades of life. The growth is usually slow but progressive. Sudden development and rapid progression of hirsutism suggest the possibility of an androgen-secreting neoplasm, in which case virilization also may be present.

The age at onset of menstrual cycles (menarche) and the pattern of the menstrual cycle should be ascertained; irregular cycles from the time of menarche onward are more likely to result from ovarian rather than adrenal androgen excess. Associated symptoms such

as galactorrhea should prompt evaluation for hyperprolactinemia (Chap. 373) and possibly hypothyroidism (Chap. 375). Hypertension, striae, easy bruising, centripetal weight gain, and weakness suggest hypercortisolism (Cushing's syndrome; Chap. 379). Rarely, patients with growth hormone excess (i.e., acromegaly) present with hirsutism. Use of medications such as phenytoin, minoxidil, and cyclosporine may be associated with androgen-independent excess hair growth (i.e., hypertrichosis). A family history of infertility and/or hirsutism may indicate disorders such as nonclassic CAH (Chap. 379).

Physical examination should include measurement of height and weight and calculation of body mass index (BMI). A BMI >25 kg/m^2 is indicative of excess weight for height, and values >30 kg/m^2 are often seen in association with hirsutism, probably the result of increased conversion of androgen precursors to testosterone. Notation should be made of blood pressure, as adrenal causes may be associated with hypertension. Cutaneous signs sometimes associated with androgen excess and insulin resistance include acanthosis nigricans and skin tags.

An objective clinical assessment of hair distribution and quantity is central to the evaluation in any woman presenting with concerns about excessive hair growth. This assessment permits the distinction between hirsutism and hypertrichosis and provides a baseline reference point to gauge the response to treatment. A simple and commonly used method to grade hair growth is the modified scale of Ferriman and Gallwey (Fig. 387-1), in which each of nine androgen-sensitive sites is graded from 0 to 4. Approximately 95% of white women have a score <8 on this scale; thus, it is normal for most women to have some hair growth in androgen-sensitive sites. Scores >8 suggest excess androgen-mediated hair growth, a finding that should be assessed further by means of hormonal evaluation (see below). In racial/ethnic groups that are less likely to manifest hirsutism (e.g., Asian women), additional cutaneous evidence of androgen excess should be sought, including pustular acne and thinning scalp hair.

HORMONAL EVALUATION

Androgens are secreted by the ovaries and adrenal glands in response to their respective tropic hormones: luteinizing hormone (LH) and adrenocorticotropic hormone (ACTH). Testosterone is the principal circulating steroid involved in the etiology of hirsutism; other steroids that may contribute to the development of hirsutism include androstenedione, dehydroepiandrosterone (DHEA) and its sulfated form (DHEAS). The ovaries and adrenal glands normally contribute about equally to testosterone production. Approximately half of the total testosterone originates from direct glandular secretion, and the remainder is derived from the peripheral conversion of androstenedione and DHEA (Chap. 384).

Although it is the most important circulating androgen, testosterone is in effect the penultimate androgen in mediating hirsutism; it is converted to the more potent dihydrotestosterone (DHT) by the enzyme 5α-reductase, which is located in the PSU. DHT has a higher affinity for, and slower dissociation from, the androgen receptor. The local production of DHT allows it to serve as the primary mediator of androgen action at the level of the pilosebaceous unit. There are two isoenzymes of 5α-reductase: type 2 is found in the prostate gland and in hair follicles, and type 1 is found primarily in sebaceous glands.

One approach to testing for hyperandrogenemia is depicted in Fig. 387-2. In addition to measuring blood levels of testosterone and DHEAS, it is important to measure the level of free (or unbound) testosterone. The fraction of testosterone that is not bound to its carrier protein, sex hormone–binding globulin (SHBG), is biologically available for conversion to DHT and binding to androgen receptors. Hyperinsulinemia and/or androgen excess decrease hepatic production of SHBG, resulting in levels of total testosterone within the high-normal range, whereas the unbound hormone is elevated more substantially. Although there is a decline in ovarian testosterone production after menopause, ovarian estrogen production decreases to an even greater extent, and the concentration of SHBG is reduced. Consequently, there is an increase in the relative proportion of unbound testosterone, and it may exacerbate hirsutism after menopause.

A baseline plasma total testosterone level >12 nmol/L (>3.5 ng/mL) usually indicates a virilizing tumor, whereas a level >7 nmol/L (>2 ng/mL) is suggestive of tumor but may also be observed in women with hyperthecosis. A basal DHEAS level >18.5 μmol/L (>7000 μg/L) suggests an adrenal tumor. Although DHEAS has been proposed as a "marker" of predominant adrenal androgen excess, it is not unusual to find modest elevations in DHEAS among women with PCOS. Computed tomography (CT) or magnetic resonance imaging (MRI) should be used to localize an adrenal mass, and ultrasound usually suffices to identify an ovarian mass if clinical evaluation and hormonal levels suggest these possibilities.

PCOS is the most common cause of ovarian androgen excess (Chap. 385). An increased ratio of LH to follicle-stimulating hormone is characteristic in carefully studied patients with PCOS. However, because of the pulsatile nature of gonadotropin secretion, this finding may be absent in up to half of women with PCOS. Transvaginal ultrasound classically shows enlarged ovaries and increased stroma in women with PCOS. However, cystic ovaries also may be found in women without clinical or laboratory features of PCOS. Although usually limited to a research setting, a gonadotropin-releasing hormone agonist test can be used to make a specific diagnosis of ovarian hyperandrogenism. A peak 17-hydroxyprogesterone level ≥7.8 nmol/L (≥2.6 μg/L) after the administration of 100 μg nafarelin (or 10 μg/kg leuprolide) subcutaneously is virtually diagnostic of ovarian hyperandrogenism.

Because adrenal androgens are readily suppressed by low doses of glucocorticoids, the dexamethasone androgen-suppression test may broadly distinguish ovarian from adrenal androgen overproduction. A blood sample is obtained before and after the administration of dexamethasone (0.5 mg orally every 6 h for 4 days). An adrenal source is suggested by suppression of unbound testosterone into the normal range; incomplete suppression suggests ovarian androgen excess. An overnight 1-mg dexamethasone suppression test, with measurement of 8:00 A.M. serum cortisol, is useful when there is clinical suspicion of Cushing's syndrome (Chap. 379).

Nonclassic CAH is most commonly due to 21-hydroxylase deficiency but also can be caused by autosomal recessive defects in other steroidogenic enzymes necessary for adrenal corticosteroid synthesis (Chap. 379). Because of the enzyme defect, the adrenal gland cannot secrete glucocorticoids (especially cortisol) efficiently. This results in diminished negative feedback inhibition of ACTH, leading to compensatory adrenal hyperplasia and the accumulation of steroid precursors that subsequently are converted to androgen. Deficiency of 21-hydroxylase can be reliably excluded by determining a morning 17-hydroxyprogesterone level <6 nmol/L (<2 μg/L) (drawn in the follicular phase). Alternatively, 21-hydroxylase deficiency can be diagnosed by measurement of 17-hydroxyprogesterone 1 h after the administration of 250 μg of synthetic ACTH (cosyntropin) intravenously.

TREATMENT

Hirsutism

Treatment of hirsutism may be accomplished pharmacologically or by mechanical means of hair removal. Nonpharmacologic treatments should be considered in all patients either as the only treatment or as an adjunct to drug therapy.

Nonpharmacologic treatments include (1) bleaching, (2) depilatory (removal from the skin surface) such as shaving and chemical treatments, and (3) epilatory (removal of the hair including the root) such as plucking, waxing, electrolysis, laser and intense pulsed light (IPL). Despite perceptions to the contrary, shaving does not increase the rate or density of hair growth. Chemical depilatory treatments may be useful for mild hirsutism that affects only limited skin areas, though they can cause skin irritation. Wax treatment removes hair temporarily but is uncomfortable. Electrolysis is effective for more permanent hair removal, particularly in the hands of a skilled electrologist. Laser and IPL are used to treat large areas of pigmented, terminal hair. Light of specific wavelength, duration, and energy is absorbed by melanin in the hair shaft and follicle leading to

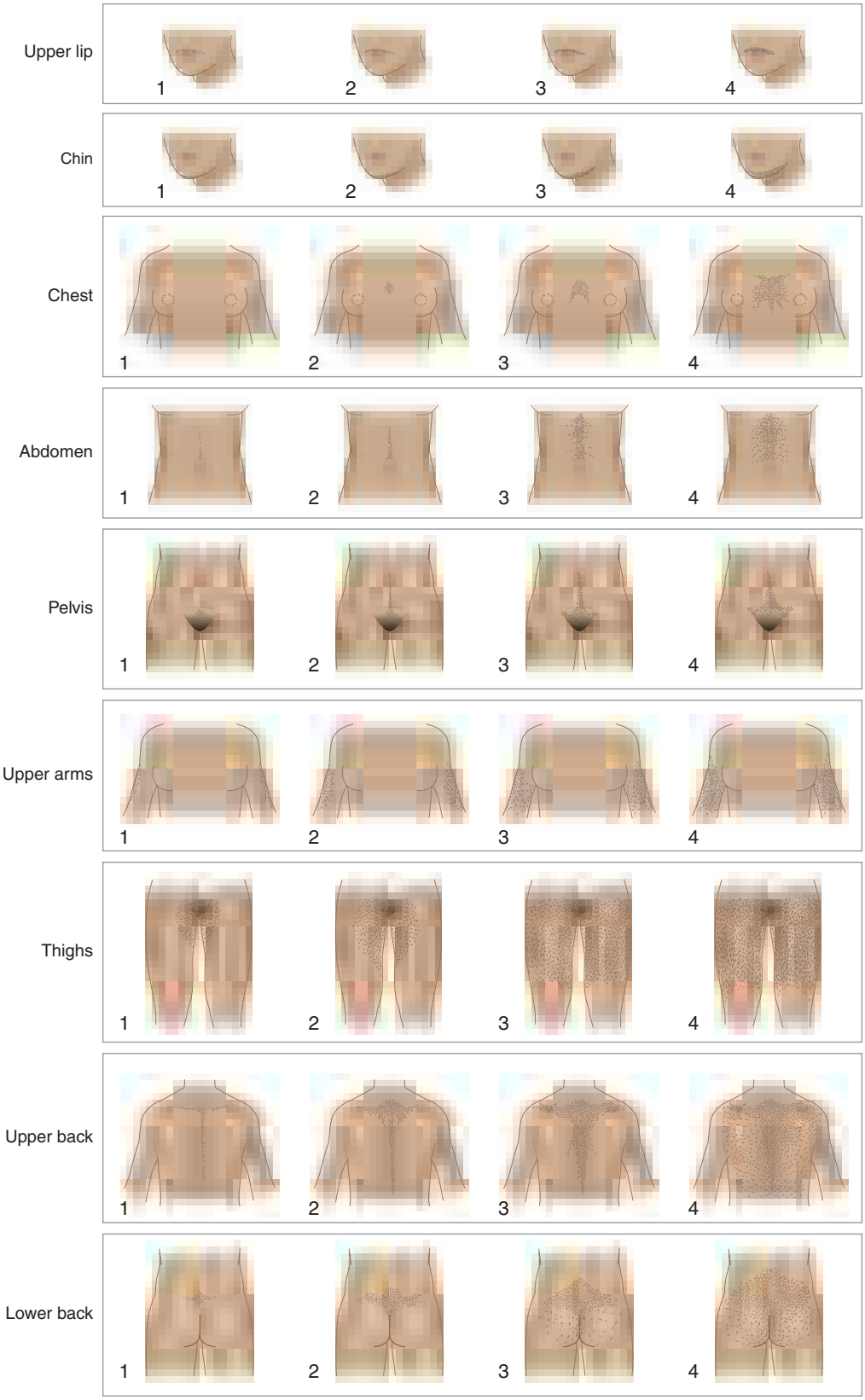


FIGURE 387-1 Hirsutism scoring scale of Ferriman and Gallwey. The nine body areas that have androgen-sensitive areas are graded from 0 (no terminal hair) to 4 (frankly virile) to obtain a total score. A normal hirsutism score is <8. *(Modified from DA Ehrmann et al: Hyperandrogenism, hirsutism, and polycystic ovary syndrome, in LJ DeGroot and JL Jameson [eds], Endocrinology, 5th ed. Philadelphia, Saunders, 2006; with permission.)*

photothermolysis. Properly delivered, this treatment delays hair regrowth and causes permanent hair removal in many patients.

Pharmacologic therapy is directed at interrupting one or more of the steps in the pathway of androgen synthesis and action: (1) suppression of adrenal and/or ovarian androgen production, (2) enhancement of androgen-binding to plasma-binding proteins, particularly SHBG, (3) impairment of the peripheral conversion of androgen precursors to active androgen, and (4) inhibition of androgen action at the target tissue level. Attenuation of hair growth is typically not evident until 4–6 months after initiation of medical treatment and in most cases leads to only a modest reduction in hair growth.

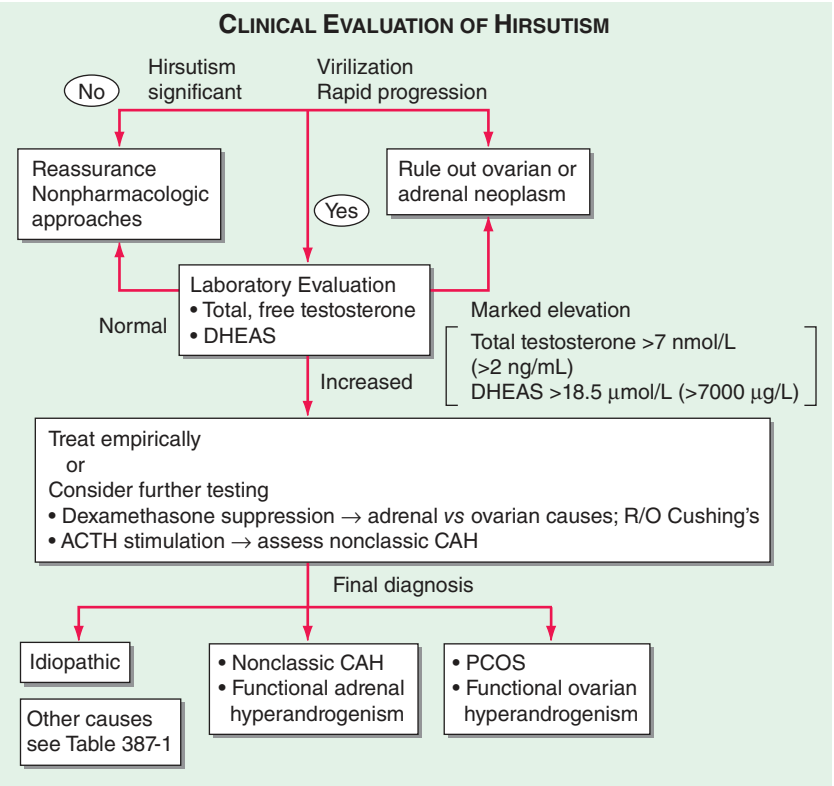


FIGURE 387-2 Algorithm for the evaluation and differential diagnosis of hirsutism. ACTH, adrenocorticotropic hormone; CAH, congenital adrenal hyperplasia; DHEAS, sulfated form of dehydroepiandrosterone; PCOS, polycystic ovarian syndrome.

Combination estrogen-progestin therapy in the form of an oral contraceptive is usually the first-line endocrine treatment for hirsutism and acne, after cosmetic and dermatologic management. The estrogenic component of most oral contraceptives currently in use is either ethinyl estradiol or mestranol. The suppression of LH leads to reduced production of ovarian androgens. The reduced androgen levels also result in a dose-related increase in SHBG, thus lowering the fraction of unbound plasma testosterone. Estrogens also have a direct, dose-dependent suppressive effect on sebaceous cell function.

The choice of a specific oral contraceptive should be predicated on the progestational component, as progestins vary in their suppressive effect on SHBG levels and in their androgenic potential. Ethynodiol diacetate has relatively low androgenic potential, whereas progestins such as norgestrel and levonorgestrel are particularly androgenic, as judged from their attenuation of the estrogen-induced increase in SHBG. Norgestimate exemplifies the newer generation of progestins that are virtually nonandrogenic. Drospirenone, an analogue of spironolactone that has both antimineralocorticoid and antiandrogenic activities, has been approved for use as a progestational agent in combination with ethinyl estradiol.

Oral contraceptives are contraindicated in women with a history of thromboembolic disease and women with increased risk of breast or other estrogen-dependent cancers (Chap. 388). There is a relative contraindication to the use of oral contraceptives in smokers and those with hypertension or a history of migraine headaches. In most trials, estrogen-progestin therapy alone improves the extent of acne by a maximum of 50–70%. The effect on hair growth may not be evident for 6 months, and the maximum effect may require 9–12 months owing to the length of the hair growth cycle. Improvements in hirsutism are typically in the range of 20%, but there may be an arrest of further progression of hair growth.

Because oral contraceptives are efficacious and have fewer side effects, they are recommended over glucocorticoids as first-line treatment of hirsutism in CAH. If the response to oral contraceptives is inadequate, glucocorticoids may be used. The lowest effective dose of glucocorticoid should be used (e.g., dexamethasone [0.2–0.5 mg] or prednisone [5–10 mg]) taken at bedtime to achieve maximal suppression by inhibiting the nocturnal surge of ACTH.

Cyproterone acetate is the prototypic antiandrogen. It acts mainly by competitive inhibition of the binding of testosterone and DHT to the androgen receptor. In addition, it may enhance the metabolic clearance of testosterone by inducing hepatic enzymes. Although not available for use in the United States, cyproterone acetate is widely used in Canada, Mexico, and Europe. Cyproterone (50–100 mg) is given on days 1–15 and ethinyl estradiol (50 µg) is given on days 5–26 of the menstrual cycle. Side effects include irregular uterine bleeding, nausea, headache, fatigue, weight gain, and decreased libido.

Spironolactone, which usually is used as a mineralocorticoid antagonist, is also a weak antiandrogen. It is almost as effective as cyproterone acetate when used at high enough doses (100–200 mg daily). Patients should be monitored intermittently for hyperkalemia or hypotension, though these side effects are uncommon. Pregnancy should be avoided because of the risk of feminization of a male fetus. Spironolactone can also cause menstrual irregularity. It often is used in combination with an oral contraceptive, which suppresses ovarian androgen production and helps prevent pregnancy.

Flutamide is a potent nonsteroidal antiandrogen that is effective in treating hirsutism, but concerns about the induction of hepatocellular dysfunction preclude its use. Finasteride is a competitive inhibitor of 5α-reductase type 2. Beneficial effects on hirsutism have been reported, but the predominance of 5α-reductase type 1 in the PSU appears to account for its limited efficacy. Finasteride would also be expected to impair sexual differentiation in a male fetus, and it should not be used in women who may become pregnant.

Eflornithine cream (Vaniqa) has been approved as a novel treatment for unwanted facial hair in women, but long-term efficacy remains to be established. It can cause skin irritation under exaggerated conditions of use. Ultimately, the choice of any specific agent(s) must be tailored to the unique needs of the patient being treated. As noted previously, pharmacologic treatments for hirsutism should be used in conjunction with nonpharmacologic approaches. It is also helpful to review the pattern of female hair distribution in the normal population to dispel unrealistic expectations.

FURTHER READING

Brown DL et al: Ovarian stromal hyperthecosis: Sonographic features and histologic associations. J Ultrasound Med 28:587, 2009.

Haak CS et al: Hair removal in hirsute women with normal testosterone levels: A randomized controlled trial of long-pulsed diode laser vs. intense pulsed light. Br J Dermatol 163:1007, 2010.

McCartney CR, Marshall JC: Polycystic Ovary Syndrome. N Engl J Med 375:1398, 2016.

Rosenfield RL, Ehrmann DA: The Pathogenesis of Polycystic Ovary Syndrome (PCOS): The Hypothesis of PCOS as Functional Ovarian Hyperandrogenism Revisited. Endocr Rev 37:467, 2016.

van Zuuren EJ et al: Interventions for hirsutism (excluding laser and photoepilation therapy alone). Cochrane Database Syst Rev CD010334, 2015.

Zimmerman Y et al: The effect of combined oral contraception on testosterone levels in healthy women: A systematic review and meta-analysis. Human Reprod Update 20:76, 2013.

388 Menopause and Postmenopausal Hormone Therapy

JoAnn E. Manson, Shari S. Bassuk

Menopause is the permanent cessation of menstruation due to loss of ovarian follicular function. It is diagnosed retrospectively after 12 months of amenorrhea. The average age at menopause is 51 years among U.S. women. *Perimenopause* refers to the time period preceding menopause, when fertility wanes and menstrual cycle irregularity increases, until the first year after cessation of menses. The onset of perimenopause precedes the final menses by 2–8 years, with a mean duration of 4 years. Smoking accelerates the menopausal transition by 2 years.

Although the peri- and postmenopausal transitions share many symptoms, the physiology and clinical management of the two differ. Low-dose oral contraceptives have become a therapeutic mainstay in perimenopause, whereas postmenopausal hormone therapy (HT) has been a common method of symptom alleviation after menstruation ceases.

PERIMENOPAUSE

PHYSIOLOGY

Ovarian mass and fertility decline sharply after age 35 and even more precipitously during perimenopause; depletion of primary follicles, a process that begins before birth, occurs steadily until menopause **(Chap. 385)**. In perimenopause, intermenstrual intervals shorten significantly (typically by 3 days) as a result of an accelerated follicular phase. Follicle-stimulating hormone (FSH) levels rise because of altered folliculogenesis and reduced inhibin secretion. In contrast to the consistently high FSH and low estradiol levels seen in menopause, perimenopause is characterized by "irregularly irregular" hormone levels. The propensity for anovulatory cycles can produce a hyperestrogenic, hypoprogestagenic environment that may account for the increased incidence of endometrial hyperplasia or carcinoma, uterine polyps, and leiomyoma observed among women of perimenopausal age. Mean serum levels of selected ovarian and pituitary hormones during the menopausal transition are shown in **Fig. 388-1**. With transition into menopause, estradiol levels fall markedly, whereas estrone levels are relatively preserved, a pattern reflecting peripheral aromatization of adrenal and ovarian androgens. Levels of FSH increase more than those of luteinizing hormone, presumably because of the loss of inhibin as well as estrogen feedback.

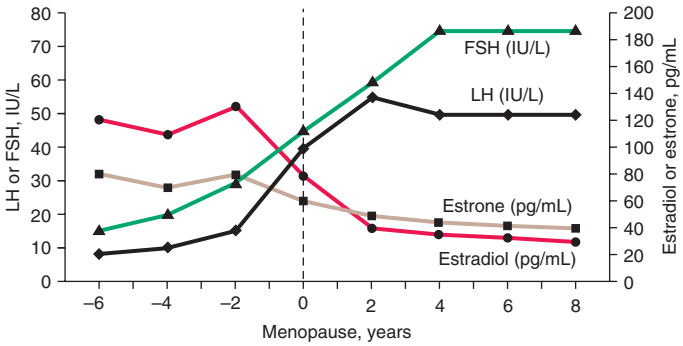


FIGURE 388-1 Mean serum levels of ovarian and pituitary hormones during the menopausal transition. FSH, follicle-stimulating hormone; LH, luteinizing hormone. *(From JL Shifren, I Schiff: J Womens Health Gend Based Med 9 Suppl 1: S3, 2000. Reproduced with permission.)*

DIAGNOSTIC TESTS

The Stages of Reproductive Aging Workshop +10 (STRAW+10) classification provides a comprehensive framework for the clinical assessment of ovarian aging. As shown in **Fig. 388-2**, menstrual cycle characteristics are the principal criteria for characterizing the menopausal transition, with biomarker measures as supportive criteria. Because of their extreme intraindividual variability, FSH and estradiol levels are imperfect diagnostic indicators of perimenopause in menstruating women. However, a consistently low FSH level in the early follicular phase (days 2–5) of the menstrual cycle does not support a diagnosis of perimenopause, while levels >25 IU/L in a random blood sample are characteristic of the late menopause transition. FSH measurement can also aid in assessing fertility; levels of <20 IU/L, 20 to <30 IU/L, and ≥30 IU/L measured on day 3 of the cycle indicate a good, fair, and poor likelihood of achieving pregnancy, respectively. Antimüllerian hormone and inhibin B may also be useful for assessing reproductive aging.

SYMPTOMS

Determining whether symptoms that develop in midlife are due to ovarian senescence or to other age-related changes is difficult. There is strong evidence that the menopausal transition can cause hot flashes, night sweats, irregular bleeding, and vaginal dryness, and there is moderate evidence that it can cause sleep disturbances in some women. There is inconclusive or insufficient evidence that ovarian aging is a major cause of mood swings, depression, impaired memory or concentration, somatic symptoms, urinary incontinence, or sexual dysfunction. In one U.S. study, nearly 60% of women reported hot flashes in the 2 years before their final menses. Symptom intensity, duration, frequency, and effects on quality of life are highly variable.

TREATMENT

Perimenopause

PERIMENOPAUSAL THERAPY

For women with irregular or heavy menses or hormone-related symptoms that impair quality of life, low-dose combined oral contraceptives are a staple of therapy. Static doses of estrogen and progestin (e.g., 20 μg of ethinyl estradiol and 1 mg of norethindrone acetate daily for 21 days each month) can eliminate vasomotor symptoms and restore regular cyclicity. Oral contraceptives provide other benefits, including protection against ovarian and endometrial cancers and increased bone density, although it is not clear whether use during perimenopause decreases fracture risk later in life. Moreover, the contraceptive benefit is important, given that the unintentional pregnancy rate among women in their forties rivals that of adolescents. Contraindications to oral contraceptive use include cigarette smoking, liver disease, a history of thromboembolism or cardiovascular disease, breast cancer, or unexplained vaginal bleeding. Progestin-only formulations (e.g., 0.35 mg of norethindrone daily) or medroxyprogesterone (Depo-Provera) injections (e.g.,

Menarche

FMP (0)

Stage	–5	–4	–3b	–3a	–2	–1	+1a	+1b	+1c	+2
Terminology	**Reproductive**				**Menopausal transition**		**Postmenopause**			
	Early	Peak	Late		Early	Late	Early			Late
					Perimenopause					
Duration	*Variable*				*Variable*	1–3 years	2 years (1+1)		3–6 years	*Remaining lifespan*
Principal criteria										
Menstrual cycle	Variable to regular	Regular	Regular	Subtle changes in flow/ length	*Variable Length* Persistent ≥7-day difference in length of consecutive cycles	Interval of amenorrhea of ≥60 days				
Supportive criteria										
Endocrine FSH AMH Inhibin B			 Low Low	 Variable* Low Low	 ↑ Variable* Low Low	 ↑ >25 IU/L** Low Low	 ↑ Variable Low Low		 Stabilizes Very low Very low	
Antral follicle count			Low	Low	Low	Low	Very low		Very low	
Descriptive characteristics										
Symptoms						Vasomotor symptoms *Likely*	Vasomotor symptoms *Most likely*			*Increasing* symptoms of urogenital atrophy

*Blood draw on cycle days 2–5 ↑ = elevated.
**Approximate expected level based on assays using current international pituitary standard.

FIGURE 388-2 The Stages of Reproductive Aging Workshop +10 (STRAW +10) staging system for reproductive aging in women. AMH, antimüllerian hormone; FSH, follicle-stimulating hormone. *(From SD Harlow et al: Menopause 14:387, 2012. Reproduced with permission.)*

150 mg IM every 3 months) may provide an alternative for the treatment of perimenopausal menorrhagia in women who smoke or have cardiovascular risk factors. Although progestins neither regularize cycles nor reduce the number of bleeding days, they reduce the volume of menstrual flow.

Nonhormonal strategies to reduce menstrual flow include the use of nonsteroidal anti-inflammatory agents such as mefenamic acid (an initial dose of 500 mg at the start of menses, then 250 mg qid for 2–3 days) or, when medical approaches fail, endometrial ablation. It should be noted that menorrhagia requires an evaluation to rule out uterine disorders. Transvaginal ultrasound with saline enhancement is useful for detecting leiomyomata or polyps, and endometrial aspiration can identify hyperplastic changes.

TRANSITION TO MENOPAUSE

For sexually active women using contraceptive hormones to alleviate perimenopausal symptoms, the question of when and if to switch to HT must be individualized. Doses of estrogen and progestogen (either synthetic progestins or natural forms of progesterone) in HT are lower than those in oral contraceptives and have not been documented to prevent pregnancy. Although a 1-year absence of spontaneous menses reliably indicates ovulation cessation, it is not possible to assess the natural menstrual pattern while a woman is taking an oral contraceptive. Women willing to switch to a barrier method of contraception should do so; if menses occur spontaneously, oral contraceptive use can be resumed. The average age of final menses among relatives can serve as a guide for when to initiate this process, which can be repeated yearly until menopause has occurred.

MENOPAUSE AND POSTMENOPAUSAL HT

One of the most complex health care decisions facing women is whether to use postmenopausal HT. Once prescribed primarily to relieve vasomotor symptoms, HT has been promoted as a strategy to forestall various disorders that accelerate after menopause, including osteoporosis and cardiovascular disease. In 2000, nearly 40% of postmenopausal women aged 50–74 in the United States had used HT. This widespread use occurred despite the paucity of conclusive data, until recently, on the health consequences of such therapy. Although many women rely on their health care providers for a definitive answer to the question of whether to use postmenopausal hormones, balancing the benefits and risks for an individual patient is challenging.

Although observational studies suggest that HT prevents cardiovascular and other chronic diseases, the apparent benefits may result at least in part from differences between women who opt to take postmenopausal hormones and women who do not. Those choosing HT tend to be healthier, have greater access to medical care, are more compliant with prescribed treatments, and maintain a more health-promoting lifestyle. Randomized trials, which eliminate these confounding factors, have not consistently confirmed the benefits found in observational studies. Indeed, the largest HT trial to date, the Women's Health Initiative (WHI), which examined more than 27,000 postmenopausal women aged 50–79 (mean age, 63) for an average of 5–7 years, was stopped early because of an overall unfavorable benefit-risk ratio in the estrogen-progestin arm and an excess risk of stroke that was not offset by a reduced risk of coronary heart disease (CHD) in the estrogen-only arm.

The following summary offers a decision-making guide based on a synthesis of currently available evidence. Prevention of cardiovascular disease is eliminated from the equation due to lack of evidence for such benefits in randomized clinical trials.

BENEFITS AND RISKS OF POSTMENOPAUSAL HT

See Table 388-1.

Definite Benefits • **SYMPTOMS OF MENOPAUSE** Compelling evidence, including data from randomized clinical trials, indicates that estrogen therapy is highly effective for controlling vasomotor and

TABLE 388-1 Benefits and Risks of Postmenopausal Hormone Therapy in the Overall Study Population of Women aged 50–79 Years in the Intervention Phase of the Women's Health Initiative (WHI) Estrogen-Progestin and Estrogen-Alone Trials[a]

		ESTROGEN-PROGESTIN		ESTROGEN ALONE	
OUTCOME	EFFECT	RELATIVE BENEFIT OR RISK	ABSOLUTE BENEFIT OR RISK[b]	RELATIVE BENEFIT OR RISK	ABSOLUTE BENEFIT OR RISK[b]
Definite Benefits					
Symptoms of menopause	Definite improvement	↓65–90% decreased risk[c]		↓65–90% decreased risk[c]	
Osteoporosis	Definite increase in bone mineral density and decrease in fracture risk	↓33% decreased risk for hip fracture	6 fewer cases (11 vs. 17) of hip fracture	↓33% decreased risk for hip fracture	6 fewer cases (13 vs. 19) of hip fracture
Definite Risks[h]					
Endometrial cancer	Definite increase in risk with estrogen alone (see below for estrogen-progestin)	See below	See below		4.6 excess cases (observational studies)
Pulmonary embolism	Definite increase in risk	↑98% increased risk	9 excess cases (18 vs. 9)	↑35% increased risk (n.s.)	4 excess cases (14 vs. 10)
Deep-vein thrombosis	Definite increase in risk	↑87% increased risk	11.5 excess cases (25 vs. 14)	↑48% increased risk	7.5 excess cases (23 vs. 15)
Breast cancer	Definite increase in risk with long-term use (≥5 years) of estrogen-progestin	↑24% increased risk	8.5 excess cases (43 vs. 35)	↓21% decreased risk (n.s.)	7 fewer cases (28 vs. 35)
Gallbladder disease	Definite increase in risk	↑57% increased risk	47 excess cases (131 vs. 84)	↑55% increased risk	58 excess cases (164 vs. 106)
Probable or Uncertain Risks and Benefits[h]					
Coronary heart disease[d]	Probable increase in risk among older women and women many years past menopause; possible decrease in risk or no effect in younger or recently menopausal women[e]	↑18% increased risk (n.s.)	6 excess cases (41 vs. 35)	No increase in risk	No difference in risk
Myocardial infarction	Significant interaction by age group for estrogen alone, with reduced risk in younger—but not older—women (*p* for trend by age = 0.02)	↑24% increased risk (n.s.)	6 excess cases (35 vs. 29)	No increase in risk[e]	No difference in risk[e]
Stroke	Probable increase in risk	↑37% increased risk	9 excess cases (33 vs. 24)	↑35% increased risk	11 excess cases (45 vs. 34)
Ovarian cancer	Probable increase in risk with long-term use (≥5 years)	↑41% increased risk (n.s.)	1 excess case (5 vs. 4)	Not available	Not available
Endometrial cancer	Probable decrease in risk with estrogen-progestin during long-term follow-up (see above for estrogen alone)	↓33% decreased risk[f]	3 fewer cases (7 vs. 10)	See above	See above
Urinary incontinence	Probable increase in risk	↑49% increased risk	549 excess cases (1661 vs. 1112)	↑61% increased risk	852 excess cases (2255 vs. 1403)
Colorectal cancer	Probable decrease in risk with estrogen-progestin; possible increase in risk in older women with estrogen alone (*p* for trend by age = 0.02 for estrogen alone)	↓38% decreased risk	6.5 fewer cases (10 vs. 17)	No increase or decrease in risk[e]	No difference in risk[e]
Type 2 diabetes	Probable decrease in risk	↓19% decreased risk	16 fewer cases (72 vs. 88)	↓14% decreased risk	21 fewer cases (134 vs. 155)
Dementia (age ≥65)	Increase in risk in older women (but inconsistent data from observational studies and randomized trials)	↑101% increased risk	23 excess cases (46 vs. 23)	↑47% increased risk (n.s.)	15 excess cases (44 vs. 29)
Total mortality	Possible increase in risk among older women and women many years past menopause; possible decrease in risk or no effect in younger or recently menopausal women (*p* for trend by age <0.05 for both trials combined)	No increase in risk	No difference in risk	No increase in risk[e]	No difference in risk[e]
Global index[g]	Probable increase in risk or no effect among older women and women many years past menopause; possible decrease in risk or no effect in younger or recently menopausal women (*p* for trend by age = 0.02 for estrogen alone)	↑12% increased risk	20.5 excess cases (189 vs. 168)	No increase in risk[e]	No difference in risk[e]

[a]The estrogen-progestin arm of the WHI assessed 5.6 years of conjugated equine estrogens (0.625 mg/d) plus medroxyprogesterone acetate (2.5 mg/d) versus placebo. The estrogen-alone arm of the WHI assessed 7.1 years of conjugated equine estrogens (0.625 mg/d) versus placebo. [b]Number of cases per 10, 000 women per year. [c]The WHI was not designed to assess the effect of HT on menopausal symptoms. Data from other randomized trials suggest that HT reduces risk for menopausal symptoms by 65–90%. [d]*Coronary heart disease* is defined as nonfatal myocardial infarction or coronary death. [e]There was a significant interaction by age; that is, the association between HT and the specified outcome was different in younger women and older women. [f]This is the risk reduction that was observed during a cumulative 13-year follow-up period (5.6 years of treatment plus 8.2 years of postintervention observation). [g]The *global index* is a composite outcome representing the first event for each participant from among the following: coronary heart disease, stroke, pulmonary embolism, breast cancer, colorectal cancer, endometrial cancer (estrogen-progestin arm only), hip fracture, and death. Because participants can experience more than one type of event, the global index cannot be derived by a simple summing of the component events. [h]Includes some outcomes where results were divergent between the estrogen-progestin arm and the estrogen-alone arm.

Abbreviation: n.s., not statistically significant.

Source: Data from JE Manson et al: JAMA 310:1353, 2013.

genitourinary symptoms. Alternative approaches, including the use of antidepressants (such as paroxetine, 7.5 mg/d; or venlafaxine, 75–150 mg/d), gamma-aminobutyric acid analogues (such as gabapentin, 900–2400 mg/d [dose divided 3 times per day]; or pregabalin, 150–300 mg/d [dose divided twice per day]), or clonidine (0.1 mg/d), may also alleviate vasomotor symptoms, although they are less effective than HT. Paroxetine is the only nonhormonal drug approved by the U.S. Food and Drug Administration for treatment of vasomotor symptoms. Bazedoxifene, an estrogen agonist/antagonist, in combination with conjugated estrogens has also received approval for this use. Cognitive behavioral therapy and clinical hypnosis have been shown in randomized trials to help with vasomotor symptom management. Weight loss, mindfulness-based stress reduction, stellate ganglion block, and the consumption of S-equol soy derivatives are also promising strategies, although more trials are needed. For genitourinary syndrome of menopause, the efficacy of low-dose vaginal estrogen is similar to that of oral or transdermal estrogen; oral ospemifene or vaginal prasterone are additional options.

OSTEOPOROSIS (See also Chap. 404)

Bone density By reducing bone turnover and resorption rates, estrogen slows the aging-related bone loss experienced by most postmenopausal women. More than 50 randomized trials have demonstrated that postmenopausal estrogen therapy, with or without a progestogen, rapidly increases bone mineral density at the spine by 4–6% and at the hip by 2–3% and that those increases are maintained during treatment.

Fractures Data from observational studies indicate a 50–80% lower risk of vertebral fracture and a 25–30% lower risk of hip, wrist, and other peripheral fractures among current estrogen users; addition of a progestogen does not appear to modify this benefit. In the WHI, 5–7 years of either combined estrogen-progestin or estrogen-only therapy was associated with a 33% reduction in hip fractures and 25–30% fewer total fractures among a population unselected for osteoporosis. Bisphosphonates (such as alendronate, 10 mg/d or 70 mg once per week; risedronate, 5 mg/d or 35 mg once per week; ibandronate, 2.5 mg/d or 150 mg once per month or 3 mg every 3 months IV; or zoledronic acid 5 mg once per year IV) and denosumab (60 mg twice per year SC) increase bone mass density by reducing bone resorption and have been shown in randomized trials to decrease fracture rates. Other treatment options include bazedoxifene in combination with conjugated estrogens; the selective estrogen receptor modulator (SERM) raloxifene (60 mg/d); and parathyroid hormone (teriparatide, 20 µg/d SC). Unlike estrogen, these alternative therapies do not appear to have adverse effects on the endometrium or breast. Increased weight-bearing and resistance exercise; adequate calcium intake (1000–1200 mg/d through diet or supplements in two or three divided doses); and adequate vitamin D intake (600–1000 IU/d) may also reduce the risk of osteoporosis-related fractures. According to a 2011 report by the Institute of Medicine (now the National Academy of Medicine), 25-hydroxyvitamin D blood levels of ≥50 nmol/L are sufficient for bone-density maintenance and fracture prevention. The Fracture Risk Assessment (FRAX®) score, an algorithm that combines an individual's bone-density score with age and other risk factors to predict her 10-year risk of hip and major osteoporotic fracture, may be of use in guiding decisions about pharmacologic treatment (see *www.shef.ac.uk/FRAX/*).

Definite Risks

• **ENDOMETRIAL CANCER (WITH ESTROGEN ALONE)** A combined analysis of 30 observational studies found a tripling of endometrial cancer risk among short-term users (1–5 years) of unopposed estrogen and a nearly tenfold increased risk among long-term users (≥10 years). These findings are supported by results from the randomized Postmenopausal Estrogen/Progestin Interventions (PEPI) trial, in which 24% of women assigned to unopposed estrogen for 3 years developed atypical endometrial hyperplasia—a premalignant lesion—as opposed to only 1% of women assigned to placebo. Use of a progestogen, which opposes the effects of estrogen on the endometrium, eliminates these risks and may even reduce risk (see later).

VENOUS THROMBOEMBOLISM A meta-analysis of observational studies found that current oral estrogen use was associated with a 2.5-fold increase in risk of venous thromboembolism in postmenopausal women. A meta-analysis of randomized trials, including the WHI, found a 2.1-fold increase in risk. Results from the WHI indicate a nearly twofold increase in risk of pulmonary embolism and deep-vein thrombosis with estrogen-progestin and a 35–50% increase in these risks with estrogen-only therapy. Transdermal estrogen, taken alone or with certain progestogens (micronized progesterone or pregnane derivatives), appears to be a safer alternative with respect to thrombotic risk.

BREAST CANCER (WITH ESTROGEN-PROGESTIN) An increased risk of breast cancer has been found among current or recent estrogen users in observational studies; this risk is directly related to duration of use. In a meta-analysis of 51 case-control and cohort studies, short-term use (<5 years) of postmenopausal HT did not appreciably elevate breast cancer incidence, whereas long-term use (≥5 years) was associated with a 35% increase in risk. In contrast to findings for endometrial cancer, combined estrogen-progestin regimens appear to increase breast cancer risk more than estrogen alone. Data from randomized trials also indicate that estrogen-progestin raises breast cancer risk. In the WHI, women assigned to receive combination hormones for an average of 5.6 years were 24% more likely to develop breast cancer than women assigned to placebo, but 7.1 years of estrogen-only therapy did not increase risk. Indeed, the WHI showed a trend toward a reduction in breast cancer risk with estrogen alone, although it is unclear whether this finding would pertain to formulations of estrogen other than conjugated equine estrogens or to treatment durations of >7 years. In the Heart and Estrogen/Progestin Replacement Study (HERS), 4 years of combination therapy was associated with a 27% increase in breast cancer risk. Although the latter finding was not statistically significant, the totality of evidence strongly implicates estrogen-progestin therapy in breast carcinogenesis.

Some observational data suggest that the length of the interval between menopause onset and HT initiation may influence the association between such therapy and breast cancer risk, with a "gap time" of <3–5 years conferring a higher HT-associated breast cancer risk. (This pattern of findings contrasts with that for CHD, as discussed later in this chapter.) However, this association remains inconclusive and may be a spurious finding attributable to higher rates of screening mammography and thus earlier cancer detection in HT users than in nonusers, especially in early menopause. Indeed, in the WHI trial, hazard ratios for HT and breast cancer risk did not differ among women 50–59, those 60–69, and those 70–79 years of age at trial entry. (There was insufficient power to examine finer age categories.) Additional research is needed to clarify the issue.

GALLBLADDER DISEASE Large observational studies report a two- to threefold increased risk of gallstones or cholecystectomy among postmenopausal women taking oral estrogen. In the WHI, women randomized to estrogen-progestin or estrogen alone were ~55% more likely to develop gallbladder disease than those assigned to placebo. Risks were also increased in HERS. Transdermal HT might be a safer alternative, but further research is needed.

Probable or Uncertain Risks and Benefits

• **CORONARY HEART DISEASE/STROKE** Until recently, HT had been enthusiastically recommended as a possible cardioprotective agent. In the past three decades, multiple observational studies suggested, in the aggregate, that estrogen use leads to a 35–50% reduction in CHD incidence among postmenopausal women. The biologic plausibility of such an association is supported by data from randomized trials demonstrating that exogenous estrogen lowers plasma low-density lipoprotein (LDL) cholesterol levels and raises high-density lipoprotein (HDL) cholesterol levels by 10–15%. Administration of estrogen also favorably affects lipoprotein(a) levels, LDL oxidation, endothelial vascular function, fibrinogen, and plasminogen activator inhibitor 1. However, estrogen therapy has unfavorable effects on other biomarkers of cardiovascular risk: it boosts triglyceride levels; promotes coagulation via factor VII, prothrombin fragments 1 and 2, and fibrinopeptide A elevations; and raises levels of the inflammatory marker C-reactive protein.

Randomized trials of estrogen or combined estrogen-progestin in women with preexisting cardiovascular disease have not confirmed the benefits reported in observational studies. In HERS (a secondary-prevention trial designed to test the efficacy and safety of estrogen-

progestin therapy with regard to clinical cardiovascular outcomes), the 4-year incidence of coronary death and nonfatal myocardial infarction was similar in the active-treatment and placebo groups, and a 50% increase in risk of coronary events was noted during the first year among participants assigned to the active-treatment group. Although it is possible that progestin may mitigate estrogen's benefits, the Estrogen Replacement and Atherosclerosis (ERA) trial indicated that angiographically determined progression of coronary atherosclerosis was unaffected by either opposed or unopposed estrogen treatment. Moreover, no cardiovascular benefit was found in the Papworth Hormone Replacement Therapy Atherosclerosis Study, a trial of transdermal estradiol with and without norethindrone; the Women's Estrogen for Stroke Trial (WEST), a trial of oral 17β-estradiol; or the Estrogen in the Prevention of Reinfarction Trial (ESPRIT), a trial of oral estradiol valerate. Thus, in clinical trials, HT has not proved effective for the secondary prevention of cardiovascular disease in postmenopausal women.

Primary-prevention trials also suggest an early increase in cardiovascular risk and an absence of cardioprotection with postmenopausal HT. In the WHI, women assigned to 5.6 years of estrogen-progestin therapy were 18% more likely to develop CHD (defined in primary analyses as nonfatal myocardial infarction or coronary death) than those assigned to placebo, although this risk elevation was not statistically significant. However, during the trial's first year, there was a significant 80% increase in risk, which diminished in subsequent years (p for trend by time = 0.03). In the estrogen-only arm of the WHI, no overall effect on CHD was observed during the 7.1 years of the trial or in any specific year of follow-up. This pattern of results was similar to that for the outcome of total myocardial infarction.

However, a closer look at available data suggests that timing of initiation of HT may critically influence the association between such therapy and CHD. Estrogen may slow early stages of atherosclerosis but have adverse effects on advanced atherosclerotic lesions. It has been hypothesized that the prothrombotic and proinflammatory effects of estrogen manifest themselves predominantly among women with subclinical lesions who initiate HT well after the menopausal transition, whereas women with less arterial damage who start HT early in menopause may derive cardiovascular benefit because they have not yet developed advanced lesions. Data from experiments in nonhuman primates and from some recent randomized trials in humans support this concept. Conjugated estrogens had no effect on the extent of coronary artery plaque in cynomolgus monkeys assigned to receive estrogen alone or combined with progestin starting 2 years (~6 years in human terms) after oophorectomy and well after the establishment of atherosclerosis. However, administration of exogenous hormones immediately after oophorectomy, during the early stages of atherosclerosis, reduced the extent of plaque by 70%. In the Early versus Late Intervention Trial with Estradiol (ELITE), a 6-year trial among 643 healthy postmenopausal women that was designed to test whether effects of estrogen on the development and progression of atherosclerosis depend on age at initiation of therapy, oral 17β-estradiol administered with or without vaginal micronized progesterone significantly slowed carotid atherosclerotic progression in women within 6 years of menopause onset (mean age, 55.4 years) but not in women more than 10 years past menopause onset (mean age, 65.4 years) (p, interaction=0.007). On the other hand, in the Kronos Early Estrogen Prevention Study (KEEPS), a 4-year trial among 729 healthy postmenopausal women within 3 years of menopause onset at trial entry (mean age, 53 years), neither oral conjugated estrogens nor transdermal estradiol, administered with oral micronized progesterone, affected carotid atherosclerotic progression. However, the low prevalence of this endpoint in the overall study population may have curtailed power to detect a treatment difference.

Lending further credence to the timing hypothesis are results of subgroup analyses of data from observational studies and large clinical trials. For example, among women who entered the WHI trial with a relatively favorable cholesterol profile, estrogen with or without progestin led to a 40% lower risk of incident CHD. Among women who entered with a worse cholesterol profile, therapy resulted in a 73% higher risk (p for interaction = 0.02). The presence or absence of the metabolic syndrome **(Chap. 401)** also strongly influenced the relation between HT and incident CHD. Among women with the metabolic syndrome, HT more than doubled CHD risk, whereas no association was observed among women without the syndrome. Moreover, although there was no association between estrogen-only therapy and CHD in the WHI trial cohort as a whole, such therapy was associated with a CHD risk reduction of 40% among participants aged 50–59; in contrast, a risk reduction of only 5% was observed among those aged 60–69, and a risk increase of 9% was found among those aged 70–79 (p for trend by age = 0.08). For the outcome of total myocardial infarction, estrogen alone was associated with a borderline-significant 45% reduction and a nonsignificant 24% increase in risk among the youngest and oldest women, respectively (p for trend by age = 0.02). Estrogen was also associated with lower levels of coronary artery calcified plaque in the younger age group. Although age did not have a similar effect in the estrogen-progestin arm of the WHI, CHD risks increased with years since menopause (p for trend = 0.08), with a significantly elevated risk among women who were ≥20 years past menopause. For the outcome of total myocardial infarction, estrogen-progestin was associated with a 9% risk reduction among women <10 years past menopause as opposed to a 16% increase in risk among women 10–19 years past menopause and a twofold increase in risk among women >20 years past menopause (p for trend = 0.01). In the large observational Nurses' Health Study, women who chose to start HT within 4 years of menopause experienced a lower risk of CHD than did nonusers, whereas those who began therapy ≥10 years after menopause appeared to receive little coronary benefit. Observational studies include a high proportion of women who begin HT within 3–4 years of menopause, whereas clinical trials include a high proportion of women ≥12 years past menopause; this difference helps to reconcile some of the apparent discrepancies between the two types of studies.

For the outcome of stroke, WHI participants assigned to estrogen-progestin or estrogen alone were ~35% more likely to suffer a stroke than those assigned to placebo. Whether or not age at initiation of HT influences stroke risk is not well understood. In the WHI and the Nurses' Health Study, HT was associated with an excess risk of stroke in all age groups. Further research is needed on age, time since menopause, and other individual characteristics (including biomarkers) that predict increases or decreases in cardiovascular risk associated with exogenous HT. Furthermore, it remains uncertain whether different doses, formulations, or routes of administration of HT will produce different cardiovascular effects.

COLORECTAL CANCER Observational studies have suggested that HT reduces risks of colon and rectal cancer, although the estimated magnitudes of the relative benefits have ranged from 8 to 34% in various meta-analyses. In the WHI (the sole trial to examine the issue), estrogen-progestin was associated with a significant 38% reduction in colorectal cancer over a 5.6-year period, although no benefit was seen with 7 years of estrogen-only therapy. However, a modifying effect of age was observed, with a doubling of risk with HT in women aged 70–79 but no risk elevation in younger women (p for trend by age = 0.02).

COGNITIVE DECLINE AND DEMENTIA A meta-analysis of 10 case-control and two cohort studies suggested that postmenopausal HT is associated with a 34% decreased risk of dementia. Subsequent randomized trials (including the WHI), however, have failed to demonstrate any benefit of estrogen or estrogen-progestin therapy on the progression of mild to moderate Alzheimer's disease and/or have indicated a potential adverse effect of HT on the incidence of dementia, at least in women ≥65 years of age. Among women randomized to HT (as opposed to placebo) at age 50–55 in the WHI, no effect on cognition was observed during the postintervention phase. Determining whether timing of initiation of HT influences cognitive outcomes will require further study.

OVARIAN CANCER AND OTHER DISORDERS On the basis of limited observational and randomized data, it has been hypothesized that HT increases the risk of ovarian cancer and reduces the risk of type 2 diabetes mellitus. Results from the WHI support these hypotheses. The WHI also found that HT use was associated with an increased risk of urinary incontinence and that estrogen-progestin was associated with increased rates of lung cancer mortality.

ENDOMETRIAL CANCER (WITH ESTROGEN-PROGESTIN) In the WHI, use of estrogen-progestin was associated with a nonsignificant 17% reduction in risk of endometrial cancer. A significant reduction in risk emerged during the postintervention period (see later).

ALL-CAUSE MORTALITY In the overall WHI cohort, estrogen with or without progestin was not associated with all-cause mortality. However, there was a trend toward reduced mortality in younger women, particularly with estrogen alone. For women aged 50–59, 60–69, and 70–79 years, relative risks (RRs) associated with estrogen-only therapy were 0.70, 1.01, and 1.21, respectively (p for trend = 0.04).

OVERALL BENEFIT-RISK PROFILE Estrogen-progestin was associated with an unfavorable benefit-risk profile (excluding relief from menopausal symptoms) as measured by a "global index"—a composite outcome including CHD, stroke, pulmonary embolism, breast cancer, colorectal cancer, endometrial cancer, hip fracture, and death (Table 388-1)—in the WHI cohort as a whole, and this association did not vary by 10-year age group. Estrogen-only therapy was associated with a neutral benefit-risk profile in the WHI cohort as a whole. However, there was a significant trend toward a more favorable benefit-risk profile among younger women and a less favorable profile among older women, with RRs of 0.84, 0.99, and 1.17 for women aged 50–59, 60–69, and 70–79 years, respectively (p for trend by age = 0.02). The balance of benefits and risks of estrogen with and without progestin among women aged 50-59 is shown in Fig. 388-3.

CHANGES IN HEALTH STATUS AFTER DISCONTINUATION OF HT In the WHI, many but not all risks and benefits associated with active use of HT dissipated within 5–7 years after discontinuation of therapy. For estrogen-progestin, an elevated risk of breast cancer persisted (RR = 1.28 [95% confidence interval, 1.11–1.48]) during a median cumulative 13-year follow-up period (5.6 years of treatment plus 8.2 years of postintervention observation), but most cardiovascular disease risks became neutral. A reduction in hip fracture risk persisted (RR = 0.81 [0.68–0.97]), and a significant reduction in endometrial cancer risk emerged (RR = 0.67 [0.49–0.91]). For estrogen alone, the reduction in breast cancer risk became statistically significant (RR = 0.79 [0.65–0.97]) during a median cumulative 13-year follow-up period (6.8 years of treatment plus 6.6 years of postintervention observation), and significant differences by age group persisted for total myocardial infarction and the global index, with more favorable results for younger women.

APPROACH TO THE PATIENT

Postmenopausal HT

The rational use of postmenopausal HT requires balancing the potential benefits and risks. Figure 388-4 provides one approach to decision-making. The clinician should first determine whether the patient has moderate to severe menopausal symptoms—the

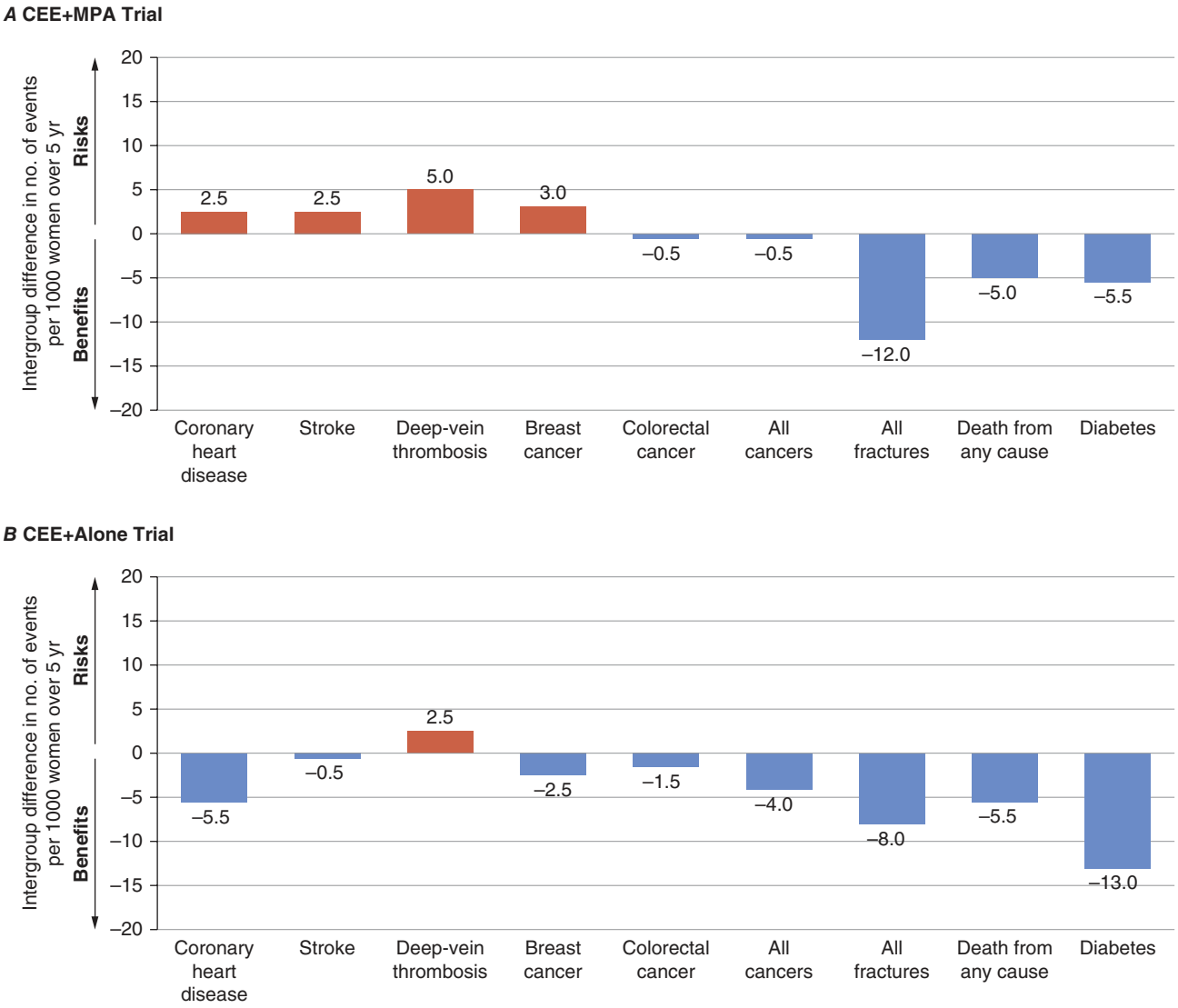


FIGURE 388-3 Benefits and risks of the two hormone therapy formulations evaluated in the Women's Health Initiative, in women aged 50–59 years. Results are shown for the two formulations, conjugated equine estrogens (CEE) alone or in combination with medroxyprogesterone acetate (MPA). Risks and benefits are expressed as the difference in number of events (number in the HT group minus the number in the placebo group) per 1000 women over 5 years. *(Data are from JE Manson et al: JAMA 310:1353, 2013.) (Graphic display is from JE Manson, AM Kaunitz: N Engl J Med 374:803, 2016 and is reproduced with permission.)*

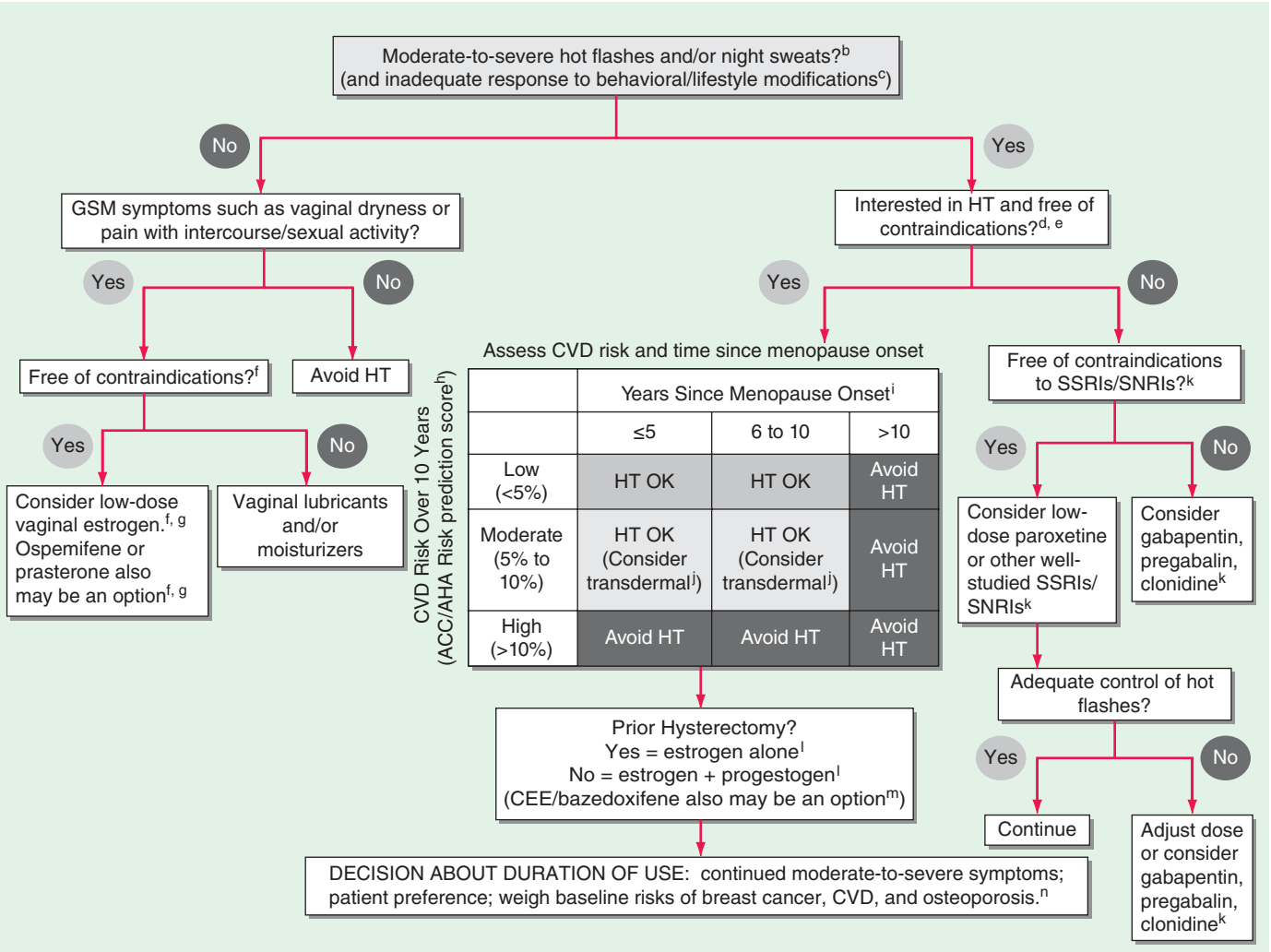


FIGURE 388-4 Algorithm for menopausal symptom management.[a] The algorithm was developed in collaboration with the North American Menopause Society and is available in a free mobile app called MenoPro (dual mode for clinicians and patients). ACC, American College of Cardiology; AHA, American Heart Association; CEE, conjugated equine estrogens; CVD, cardiovascular disease; GSM, genitourinary syndrome of menopause; HT, hormone therapy; SNRIs, serotonin-norepinephrine reuptake inhibitors; SSRIs, selective serotonin reuptake inhibitors. *(Adapted from JE Manson et al: Menopause 22:247, 2015. Used with permission.)*

[a]Algorithm applies to women with menopausal symptoms who are aged ≥45 years and to women who have had removal of both ovaries, regardless of age. Women aged <45 years or those with uncertain menopausal status may need additional clinical evaluation before applying this algorithm. [b]Women who are at high risk of osteoporotic fracture but are unable to tolerate alternative preventive medications may also be reasonable candidates for systemic HT even if they do not have moderate to severe vasomotor symptoms. [c]Patients should try lifestyle modifications for at least 3 months before using this algorithm. (A patient handout with suggested lifestyle modifications can be found at *http://www.menopause.org/docs/for-women/mnflashes.pdf*.) [d]Reassess each step at least once every 6–12 months (assuming the patient's continued preference for HT) or if the patient's health status changes. [e]See text for contraindications to systemic HT. [f]Contraindications to low-dose vaginal estrogen include unexplained vaginal bleeding, and breast cancer, endometrial cancer, or other estrogen-dependent cancer. Contraindications to ospemifene and prasterone are similar to those for low-dose vaginal estrogen, and contraindications for ospemifene additionally include venous or arterial thromboembolic disease, severe liver disease, and use of estrogens or estrogen agonists-antagonists. MenoPro, a free mobile app from the North American Menopause Society, is available for further guidance on use of these medications for treatment of GSM symptoms. [g]Regularly updated tables of all U.S. and Canadian formulations approved by regulatory authorities for the treatment of GSM symptoms are available from the North American Menopause Society (*http://www.menopause.org/docs/default-source/professional/nams-ht-tables.pdf*). [h]Ten-year risk of cardiovascular disease, as assessed by ACC/AHA risk prediction score (DC Goff Jr et al: 2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation 129[Suppl 2]:149, 2014.) [i]Women >10 years past menopause are not good candidates for initiation (first use) of HT. [j]Consider avoiding oral HT. Transdermal HT may be a preferable option because it has a less adverse effect on clotting factors, triglyceride levels, and inflammation factors than oral HT. [k]MenoPro, a free mobile app from the North American Menopause Society, is available for further guidance on use of these nonhormonal medications (as well as hormonal options) for treatment of vasomotor symptoms of menopause. [l]Regularly updated tables of all U.S. and Canadian formulations approved by regulatory authorities for the treatment of menopausal vasomotor symptoms are available from the North American Menopause Society (*http://www.menopause.org/docs/default-source/professional/nams-ht-tables.pdf*). In the United States, the most commonly prescribed oral estrogens for systemic treatment of vasomotor symptoms are 17β-estradiol (1.0 mg/d, 0.5 mg/d, and other doses) and conjugated equine estrogens (CEE, 0.625 mg/d, 0.3 mg/d, and other doses). The most commonly prescribed transdermal estrogen products are 17β-estradiol skin patches (0.0375 mg/d, 0.05 mg/d, and other doses). The most commonly prescribed progestogens are medroxyprogesterone acetate (MPA, 2.5 mg/d, 5 mg/d, and 10 mg/d) and micronized progesterone (100 mg/d and 200 mg/d). Also available are oral estrogen-progestin combinations, such as oral CEE and MPA, oral 17β-estradiol or ethinyl estradiol with norethindrone acetate, and other options. [m]CEE/bazedoxifene may be an option for women with a uterus, especially those with concerns about breast tenderness, breast density, or uterine bleeding. Contraindications to CEE/bazedoxifene are similar to those for systemic HT. [n]See text for additional discussion.

primary indication for initiation of systemic HT. Systemic HT may also be used to prevent osteoporosis in women at high risk of fracture who cannot tolerate alternative osteoporosis therapies. (Vaginal estrogen or other medications may be used to treat genitourinary syndrome of menopause in the absence of vasomotor symptoms.) The benefits and risks of such therapy should be reviewed with the patient, giving more emphasis to absolute than to relative measures of effect and pointing out uncertainties in clinical knowledge where

relevant. Because chronic disease rates generally increase with age, absolute risks tend to be greater in older women, even when RRs remain similar. Potential side effects—especially vaginal bleeding that may result from use of the combined estrogen-progestogen formulations recommended for women with an intact uterus—should be noted. The patient's own preference regarding therapy should be elicited and factored into the decision. Contraindications should be assessed routinely and include unexplained vaginal bleeding; liver dysfunction or disease; venous thromboembolism; known blood clotting disorder or thrombophilia (transdermal estrogen may be an option); untreated hypertension; history of endometrial cancer (except stage 1 without deep invasion), breast cancer or other estrogen-dependent cancer; and history of CHD, stroke, or transient ischemic attack. Relative contraindications to systemic HT include an elevated risk of breast cancer (e.g., women who have one or more first-degree relatives with breast cancer, susceptibility genes such as *BRCA1* or *BRCA2*, a personal history of cellular atypia detected by breast biopsy [see also Breast Cancer Risk Score at *http://www.cancer.gov/bcrisktool/*]); hypertriglyceridemia (>400 mg/dL); and active gallbladder disease (transdermal estrogen may be an option in the latter two cases). Primary prevention of heart disease should not be viewed as an expected benefit of HT, and an increase in the risk of stroke as well as a small early increase in the risk of coronary artery disease should be considered. Nevertheless, such therapy may be appropriate if the noncoronary benefits of treatment clearly outweigh the risks. A woman who suffers an acute coronary event or stroke while taking HT should discontinue therapy immediately.

Short-term use (<5 years for estrogen-progestogen and <7 years for estrogen alone) is appropriate for relief of menopausal symptoms among women without contraindications to such use. However, such therapy should be avoided by women with an elevated baseline risk of future cardiovascular events. Women who have contraindications for or are opposed to HT may derive benefit from the use of certain antidepressants (including venlafaxine, fluoxetine, or paroxetine), gabapentin or pregabalin, or clonidine, and, for genitourinary symptoms, intravaginal estrogen creams or devices, ospemifene, or prasterone.

Long-term use (≥5 years for estrogen-progestogen and ≥7 years for estrogen alone) is more problematic because a heightened risk of breast cancer must be factored into the decision, especially for estrogen-progestogen. Reasonable candidates for such use include the small percentage of postmenopausal women who have persistent severe vasomotor symptoms along with an increased risk of osteoporosis (e.g., those with osteopenia, a personal or family history of nontraumatic fracture, or a weight below 125 lbs), who also have no personal or family history of breast cancer in a first-degree relative or other contraindications, and who have a strong personal preference for therapy. Poor candidates are women with elevated cardiovascular risk, those at increased risk of breast cancer, and those at low risk of osteoporosis. Even for reasonable candidates, strategies to minimize dose and duration of use should be employed. For example, women using HT to relieve intense vasomotor symptoms in early postmenopause should consider discontinuing therapy within 5 years, resuming it only if such symptoms persist. Because of the role of progestogens in increasing breast cancer risk, regimens that employ cyclic rather than continuous progestogen exposure as well as formulations other than medroxyprogesterone acetate should be considered if treatment is extended. For prevention of osteoporosis, alternative therapies such as bisphosphonates or SERMs should be considered. Research on alternative progestogens and androgen-containing preparations has been limited, particularly with respect to long-term safety. Additional research on the effects of these agents on cardiovascular disease, glucose tolerance, and breast cancer will be of particular interest.

In addition to HT, lifestyle choices such as smoking abstention, adequate physical activity, and a healthy diet can play a role in controlling symptoms and preventing chronic disease. An expanding array of pharmacologic options (e.g., bisphosphonates, SERMs, and other agents for osteoporosis; cholesterol-lowering or antihypertensive agents for cardiovascular disease) should also reduce the widespread reliance on hormone use. However, short-term HT may still benefit some women.

FURTHER READING

Bassuk SS, Manson JE: Menopausal hormone therapy and cardiovascular disease risk: Utility of biomarkers and clinical factors for risk stratification. Clin Chem 60:68, 2014.

Canonico M et al: Hormone replacement therapy and risk of venous thromboembolism in postmenopausal women: Systematic review and meta-analysis. BMJ 336:1227, 2008.

Chlebowski RT et al: Oestrogen plus progestin and lung cancer in postmenopausal women (Women's Health Initiative trial): A post-hoc analysis of a randomised controlled trial. Lancet 374:1243, 2009.

Kaunitz AM, Manson JE: Management of menopausal symptoms. Obstet Gynecol 126:859, 2015.

Manson JE, Bassuk SS: *Hot Flashes, Hormones & Your Health*. New York: McGraw-Hill, 2007.

Manson JE et al: Menopausal hormone therapy and health outcomes during the intervention and extended poststopping phases of the Women's Health Initiative randomized trials. JAMA 310:1353, 2013.

Manson JE et al: Algorithm and mobile app for menopausal symptom management and hormonal/non-hormonal therapy decision making: A clinical decision-support tool from the North American Menopause Society. Menopause 22:247, 2015.

North American Menopause Society: Nonhormonal management of menopause-associated vasomotor symptoms: 2015 position statement of the North American Menopause Society. Menopause 22:1155, 2015.

North American Menopause Society: The 2017 hormone therapy position statement of the North American Menopause Society. Menopause 24:728, 2017.

Rossouw JE et al: Postmenopausal hormone therapy and risk of cardiovascular disease by age and years since menopause. JAMA 297:1465, 2007.

389 Infertility and Contraception

Janet E. Hall

INFERTILITY

DEFINITION AND PREVALENCE

Infertility is the inability to conceive after 12 months of unprotected sexual intercourse or after 6 months in women ≥35. This revised definition is based on data indicating that 50% of apparently normal couples will conceive within 3 months, 75–82% within 6 months, and 85–92% within 12 months, but recognizes the age-related decrease in fertility. In the United States, the overall rate of infertility in married women aged 15–44 is 6.7% based on the recent National Survey of Family Growth. The infertility rate has remained relatively stable over the past 30 years in most countries. However, the proportion of couples without children has risen, reflecting both higher numbers of couples in childbearing years and a trend to delay childbearing. This trend has important implications because of the age-related decrease in fecundability, the ability to conceive and carrying a baby to term; the incidence of primary impaired fecundability increases from ~15% between the ages of 15 and 29 to 18% between the ages of 30 and 35, and 40% between the ages of 35 and 44. It is estimated that 12% of women in the United States have received medical assistance for infertility, although this represents

<50% of women with current fertility problems. Both infertility and the use of medical services increase with age and both are affected by race and ethnicity. There is increased infertility in non-Hispanic black women and lower use of fertility services among Hispanic and non-Hispanic black women, suggesting disparities in access to care.

GLOBAL CONSIDERATIONS

The World Health Organization (WHO) considers infertility as a disability (an impairment of function) and thus access to health care for this indication falls under the Convention on the Rights of Persons with Disability. Thirty-four million women, predominantly from developing countries, have infertility resulting from maternal sepsis and unsafe abortion. In populations <60 years old, infertility is ranked the fifth highest serious global disability.

CAUSES OF INFERTILITY

The spectrum of infertility ranges from reduced conception rates or the need for medical intervention to irreversible causes of infertility. Infertility can be attributed primarily to male factors in 20% of couples and female factors in 38% of couples and is unexplained in about 15% of couples **(Fig. 389-1)**. Both male and female factors contribute to infertility in 25% of couples. Decreases in the ability to conceive as a function of age in women has led to recommendations that not only should women ≥34 years old seek attention sooner, but that they also receive an expedited workup and approach to treatment.

APPROACH TO THE PATIENT

Infertility

INITIAL EVALUATION

In all couples presenting with infertility, the initial evaluation includes discussion of the appropriate timing of intercourse and discussion of modifiable risk factors such as smoking, alcohol, caffeine, and obesity. The range of required investigations should be reviewed as well as a brief description of infertility treatment options, including adoption. Initial investigations are focused on determining whether the primary cause of the infertility is male, female, or both. These investigations include a semen analysis in the male, confirmation of ovulation in the female, and, in the majority of situations, documentation of tubal patency in the female. In some cases, after an extensive workup excluding all male and female factors, a specific cause cannot be identified, and infertility may ultimately be classified as unexplained.

PSYCHOLOGICAL ASPECTS OF INFERTILITY

Infertility is invariably associated with psychological stress related not only to the diagnostic and therapeutic procedures themselves but also to repeated cycles of hope and loss associated with each new procedure or cycle of treatment that does not result in the birth of a child. These feelings are often combined with a sense of isolation from friends and family. Counseling and stress-management techniques should be introduced early in the evaluation of infertility. Importantly, infertility and its treatment do not appear to be associated with long-term psychological sequelae.

FEMALE CAUSES

Abnormalities in menstrual function constitute the most common cause of female infertility. These disorders, which include ovulatory dysfunction and abnormalities of the uterus or outflow tract, may present as amenorrhea or as irregular or short menstrual cycles. A careful history and physical examination and a limited number of laboratory tests will help to determine whether the abnormality is (1) hypothalamic or pituitary (low follicle-stimulating hormone [FSH], luteinizing hormone [LH], and estradiol with or without an increase in prolactin), (2) polycystic ovary syndrome (PCOS; irregular cycles, hyperandrogenism and/or polycystic ovarian pathology assessed by ultrasound in the absence of other causes of androgen excess), (3) ovarian (low estradiol with increased FSH), or (4) a uterine or outflow tract abnormality. The frequency of these diagnoses depends on whether the amenorrhea is primary or occurs after normal puberty and menarche **(see Fig. 386-2)**.

The approach to further evaluation of these disorders is described in detail in Chap. 386.

Ovulatory Dysfunction In women with a history of regular menstrual cycles, *evidence of ovulation* should be sought **(Chap. 385)**. Even in the presence of ovulatory cycles, evaluation of *ovarian*

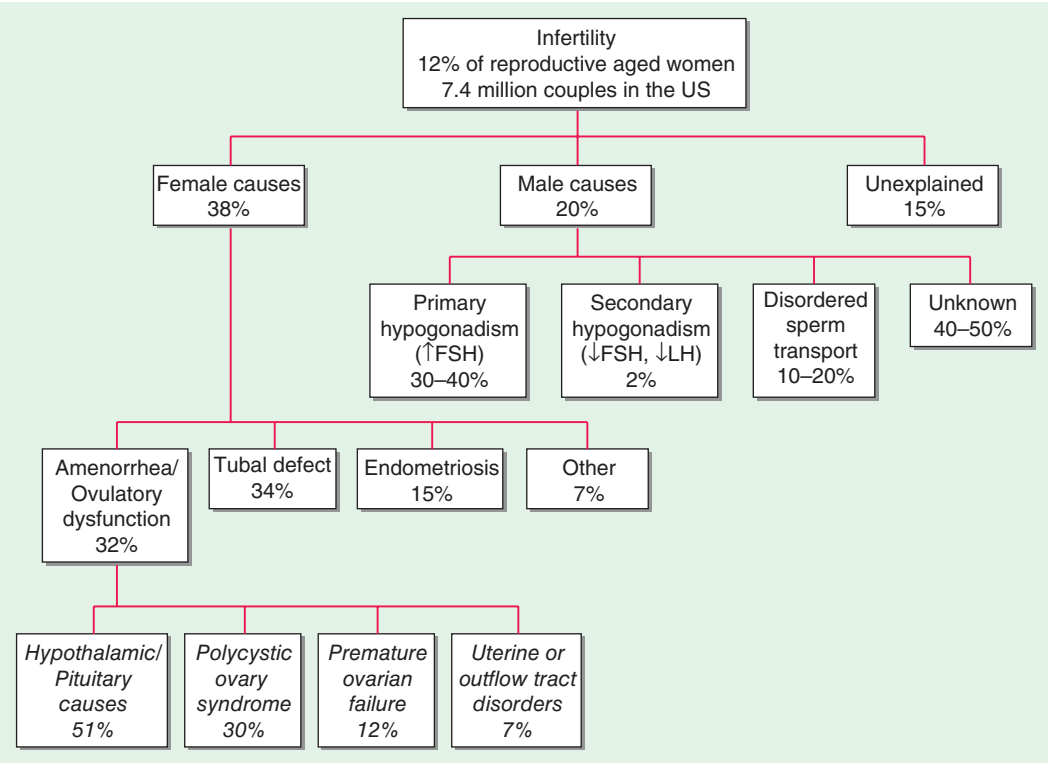


FIGURE 389-1 Causes of infertility. FSH, follicle-stimulating hormone; LH, luteinizing hormone.

reserve is recommended for women age >35 years if they are interested in fertility. Measurement of FSH on day 3 of the cycle (an FSH level <10 IU/mL on cycle day 3 predicts adequate ovarian oocyte reserve) is the most cost-effective test. Other tests antral follicle count on ultrasound, and anti-müllerian hormone (AMH; <0.5 ng/mL predicts reduced ovarian reserve although there is variability between labs). Importantly, tests of ovarian reserve help to predict the response to exogenous gonadotropins but do not predict ability to conceive.

Tubal Disease Tubal dysfunction may result from pelvic inflammatory disease (PID), appendicitis, endometriosis, pelvic adhesions, tubal surgery, previous use of an intrauterine device (IUD), and a previous ectopic pregnancy. However, a cause is not identified in up to 50% of patients with documented tubal factor infertility. Because of the high prevalence of tubal disease, evaluation of tubal patency by hysterosalpingogram (HSG) or laparoscopy should occur early in the majority of couples with infertility. Subclinical infections with *Chlamydia trachomatis* may be an underdiagnosed cause of tubal infertility and require the treatment of both partners.

Endometriosis *Endometriosis* is defined as the presence of endometrial glands or stroma outside the endometrial cavity and uterine musculature. Its presence is suggested by a history of dyspareunia (painful intercourse), worsening dysmenorrhea that often begins before menses, or a thickened rectovaginal septum or deviation of the cervix on pelvic examination. Mild endometriosis does not appear to impair fertility; the pathogenesis of the infertility associated with moderate and severe endometriosis may be multifactorial with impairments of folliculogenesis, fertilization, and implantation, as well as adhesions. Endometriosis is often clinically silent, however, and can only be excluded definitively by laparoscopy.

MALE CAUSES **(SEE ALSO CHAP. 384)**

Known causes of male infertility include primary testicular disease, genetic disorders (particularly Y chromosome microdeletions), disorders of sperm transport, and hypothalamic-pituitary disease resulting in secondary hypogonadism. However, the etiology is not ascertained in up to one-half of men with suspected male factor infertility. The key initial diagnostic test is a *semen analysis*. Testosterone levels should be measured if the sperm count is low on repeated examination or if there is clinical evidence of hypogonadism. Gonadotropin levels will help to determine a gonadal versus a central cause of hypogonadism.

TREATMENT

Infertility

In addition to addressing the negative impact of smoking on fertility and pregnancy outcome, counseling about nutrition and weight is a fundamental component of infertility and pregnancy management. Both low and increased body mass index (BMI) are associated with infertility in women and with increased morbidity during pregnancy. Obesity has also been associated with reduced fertility in men. The treatment of infertility should be tailored to the problems unique to each couple. In many situations, including unexplained infertility, mild-to-moderate endometriosis, and/or borderline semen parameters, a stepwise approach to infertility is optimal, beginning with low-risk interventions and moving to more invasive, higher risk interventions only if necessary. After determination of all infertility factors and their correction, if possible, this approach might include, in increasing order of complexity: (1) expectant management, (2) clomiphene citrate or an aromatase inhibitor (see below) with or without intrauterine insemination (IUI), (3) gonadotropins with or without IUI, and (4) in vitro fertilization (IVF). The time used for evaluation, correction of problems identified, and expectant management can be longer in women age <30 years, but this process should be advanced rapidly in women aged >35 years. In some situations, expectant management will not be appropriate.

OVULATORY DYSFUNCTION

Treatment of ovulatory dysfunction should first be directed at identification of the etiology of the disorder to allow specific management when possible. Dopamine agonists, for example, may be indicated in patients with hyperprolactinemia **(Chap. 373)**; lifestyle modification may be successful in women with low body weight, a history of intensive exercise or obesity.

Medications used for ovulation induction include agents that increase FSH through alteration of negative feedback, gonadotropins, and pulsatile GnRH. *Clomiphene citrate* is a nonsteroidal estrogen antagonist that increases FSH and LH levels by blocking estrogen negative feedback at the hypothalamus. The efficacy of clomiphene for ovulation induction is highly dependent on patient selection. It induces ovulation in ~60% of women with PCOS and has traditionally been the initial treatment of choice. Combination with agents that modify insulin levels such as metformin does not appear to improve outcomes. Clomiphene citrate is less successful in patients with hypogonadotropic hypogonadism. *Aromatase inhibitors* are also used for treatment of infertility. Studies suggest they may have advantages over clomiphene in some populations. *Estrogen receptor blockade* using Tamoxifen has been used in conjunction with gonadotropins in breast cancer patients undergoing in vitro fertilization (IVF) for embryo banking. *Gonadotropins* are highly effective for ovulation induction in women with hypogonadotropic hypogonadism and PCOS and are used to induce the development of multiple follicles in unexplained infertility and in older reproductive-age women. Disadvantages include a significant risk of multiple gestation and the risk of ovarian hyperstimulation, particularly in women with polycystic ovaries, with or without other features of PCOS. Careful monitoring and a conservative approach to ovarian stimulation reduce these risks. Currently available gonadotropins include urinary preparations of LH and FSH, highly purified FSH, and recombinant FSH. Although FSH is the key component, LH is essential for steroidogenesis in hypogonadotropic patients, and LH or human chorionic gonadotropin (hCG) may improve results through effects on terminal differentiation of the oocyte. These methods are commonly combined with IUI.

None of these methods are effective in women with premature ovarian failure, in whom donor oocyte or adoption are the methods of choice.

TUBAL DISEASE

If hysterosalpingography suggests a tubal or uterine cavity abnormality or if a patient is aged ≥35 at the time of initial evaluation, laparoscopy with tubal lavage is recommended, often with a hysteroscopy. Although tubal reconstruction may be attempted if tubal disease is identified, it is generally being replaced by the use of IVF. These patients are at increased risk of developing an ectopic pregnancy.

ENDOMETRIOSIS

Although 60% of women with minimal or mild endometriosis may conceive within 1 year without treatment, laparoscopic resection or ablation appears to improve conception rates. Medical management of advanced stages of endometriosis is widely used for symptom control but has not been shown to enhance fertility. In moderate and severe endometriosis, conservative surgery is associated with pregnancy rates of 50 and 39%, respectively, compared with rates of 25 and 5% with expectant management alone. In some patients, IVF may be the treatment of choice.

MALE FACTOR INFERTILITY

The treatment options for male factor infertility have expanded in recent years **(Chap. 384)**. Secondary hypogonadism is highly amenable to treatment with gonadotropins or pulsatile gonadotropin-releasing hormone (GnRH) where available. In vitro techniques have provided new opportunities for patients with primary testicular

failure and disorders of sperm transport. Choice of initial treatment options depends on sperm concentration and motility. Expectant management should be attempted initially in men with mild male factor infertility (sperm count of 15 to 20 × 10^6/mL and normal motility). Moderate male factor infertility (10–15 × 10^6/mL and 20–40% motility) should begin with IUI alone or in combination with treatment of the female partner with ovulation induction, but it may require IVF with or without intracytoplasmic sperm injection (ICSI). For men with a severe defect (sperm count of <10 × 10^6/mL, 10% motility), IVF with ICSI or donor sperm should be used. If ICSI is performed because of azoospermia due to congenital bilateral absence of the vas deferens, genetic testing for *CFTR* gene mutations and counseling should be provided because of the risk of cystic fibrosis.

ASSISTED REPRODUCTIVE TECHNOLOGIES

The development of assisted reproductive technologies (ARTs) has dramatically altered the treatment of male and female infertility. IVF is indicated for patients with many causes of infertility that have not been successfully managed with more conservative approaches. IVF or ICSI is often the treatment of choice in couples with a significant male factor or tubal disease, whereas IVF using donor oocytes is used in patients with premature ovarian failure and in women of advanced reproductive age. Success rates are influenced by cause of infertility and age, varying between 48% in women <35 to ≤10% in women >40. Success rates are highest in anovulatory women and lowest in women with decreased ovarian reserve. In the United States, success rates are higher in white than in black, Asian, or Hispanic women. Although often effective, IVF is costly and requires careful monitoring of ovulation induction and the use of invasive techniques, including the aspiration of multiple follicles. IVF is associated with a significant risk of multiple gestation, particularly in women age <35, in whom the rate can be as high as 30%. However, improved techniques and recognition of the risk associated with even twin pregnancies has led to adoption of age-specific guidelines by many clinics and a significant decline in the rate of twins (<25%) and very few higher order multiple births.

CONTRACEPTION

Only 15% of married couples in the United States report having unprotected sexual intercourse in the past 3 months. Although recent statistics indicate a decrease in unintended pregnancy in the United States, 45% of births are still the result of unintended pregnancy; approximately one-third of these result from incorrect use or failure of contraceptives, and >50% result in induced abortion. Unintended pregnancy is higher in Latina than white women with black women having the highest rates. Teenage pregnancies continue to represent a serious public health problem in the United States, with >1 million unintended pregnancies each year—a significantly greater incidence than in other industrialized nations. However, changes in teen behaviors are occurring, with an increase in contraceptive use at both first and most recent sexual encounter.

GLOBAL CONSIDERATIONS

The use of contraception in women aged 15–49 years who were married or in a union doubled worldwide from 36% in 1970 to 64% in 2015. The absolute number of married women who use contraception is projected to increase to nearly 800 million by 2030. However, there remains an unmet need for family planning in at least 10% of the population in most regions of the world.

Of the contraceptive methods available **(Table 389-1)**, a reversible form of contraception is used by >50% of couples with a significant increase in the use of long-acting forms such IUDs in the past decade. Sterilization (male or female) is used as a permanent form of contraception by over one-third of couples. Pregnancy termination is relatively safe when directed by health care professionals but is rarely the option of choice.

No single contraceptive method is ideal, although all are safer than carrying a pregnancy to term. The effectiveness of a given method of contraception does not just depend on the efficacy of the method itself. Discrepancies between theoretical and actual effectiveness emphasize the importance of patient education and adherence when considering various forms of contraception (Table 389-1). Contraceptive use is stratified by race/ethnicity with higher oral contraceptive use in white women and greater use of long-acting reversible contraceptive (LARC) methods in Latina women. For oral contraceptives, discontinuation

TABLE 389-1 Effectiveness of Different Forms of Contraception

METHOD OF CONTRACEPTION	THEORETICAL EFFECTIVENESS (%)	ACTUAL EFFECTIVENESS (%)	CONTINUED USE AT 1 YEAR (%)	USE OF METHOD BY U.S. WOMEN AT RISK OF UNINTENDED PREGNANCY (%)
No Method	15	15		10
Fertility Awareness	96	76	47	1.2
Withdrawal	96	78	46	4.4
Barrier Methods				
Condoms	98	82	43	13.7
Diaphragm	94	82	57	2
Spermicides	82	72	43	1
Sterilization				
Female	99.5	99.5	100	22.6
Male	99.5	99.9	100	7.4
Intrauterine Device				9.3
Copper T	99.4	99.8	85	
Progestin-containing	99.8	99.8	88	
Hormonal Contraceptives				
Combined and Progestin only	99.7	91	67	23.3
Transdermal Patch	99.7	91	67	0.5
Vaginal Ring	99.7	91	67	1.8
Implant				1.2
Depot Provera	99.8	94	56	
Nexplanon	99.5	99.5	84	
Emergency Contraception				0.3

Adapted from J Trussell et al: Contraceptive Efficacy, in *Contraceptive Technology*, 20th revised ed, RA Hatcher et al (eds). New York, Ardent Media, 2011; CDC. NCHS National Survey of Family Growth, 2011-2013; J Jones, WD Mosher, K Daniels: Current contraceptive use in the United States, 2006-2010, and changes in patterns of use since 1995, National Health Statistics Reports, 2012, No. 60, *http://www.cdc.gov/nchs/data/nhsr/nhsr060.pdf*; and NE Birgisson et al: Preventing unintended pregnancy: The contraceptive CHOICE project in review. J Womens Health (Larchmt) 24:349, 2015.

rates are highest among Black women whereas there is no racial/ethnic differences for LARC methods. Knowledge of the advantages and disadvantages of each contraceptive is essential for counseling an individual about the methods that are safest and most consistent with his or her lifestyle. The WHO has extensive family planning resources for the clinician and patient that can be accessed online. Similar resources for determining medical eligibility are available through the Centers for Disease Control and Prevention (CDC). Considerations for contraceptive use in obese patients and after bariatric surgery are discussed below.

BARRIER METHODS

Barrier contraceptives (such as condoms, diaphragms, and cervical caps) and spermicides are easily available, reversible, and have fewer side effects than hormonal methods. However, their effectiveness is highly dependent on adherence and proper use (Table 389-1). A major advantage of barrier contraceptives is the protection provided against sexually transmitted infections (STIs) **(Chap. 131)**. Consistent use is associated with a decreased risk of HIV, gonorrhea, nongonococcal urethritis, and genital herpes, probably due in part to the concomitant use of spermicides. Natural membrane condoms may be less effective than latex condoms for prevention of sexually transmitted diseases, and petroleum-based lubricants can degrade condoms and decrease their efficacy for preventing HIV infection. Barrier methods used by women include the diaphragm, cervical cap, and contraceptive sponge. There is some evidence that the diaphragm is more effective when used in conjunction with a spermicide. The cervical cap and sponge are less effective than the diaphragm, and there have been rare reports of toxic shock syndrome with the diaphragm and contraceptive sponge.

STERILIZATION

Sterilization procedures are highly effective for both men and women (Table 389-1) and are commonly chosen by fertile men and multiparous women >30 years old. Sterilization refers to a procedure that prevents fertilization by surgical interruption of the fallopian tubes in women or the vas deferens in men. Although tubal ligation and vasectomy are potentially reversible, these procedures should be considered permanent and should not be undertaken without patient counseling.

Tubal ligation methods are highly effective with a 10-year cumulative pregnancy rate of 1.85 per 100 women with both postpartum or interval procedures. However, when pregnancy does occur, the risk of ectopic pregnancy may be as high as 30%. The success rate of tubal reanastomosis depends on the method of ligation used, but even after successful reversal, the risk of ectopic pregnancy remains high. The use of salpingectomy has increased to 33% with emerging evidence that ovarian cancer originates from dysplastic cells in the fallopian tube. Hysteroscopic sterilization has been used, particularly in women with pelvic adhesions or other co-morbidities. Essure is the most commonly used commercially available product and involves insertion of a nickel-titanium double coil which results in tubal fibrosis. Data indicate very low unintended pregnancy rates (1.5 per 1000 women) with ultrasound and/or HSG confirmation of correct placement. Although still available, the U.S. Food and Drug Administration (FDA) issued a black box warning in 2015 due to post-marketing reports of long-term pain, abnormal bleeding, and allergic reactions. Intrauterine quinacrine is also effective and has been used for many years in resource-poor settings. *Vasectomy* is a highly effective and low risk outpatient surgical procedure. The no-scalpel technique, which is used in the United States, results in fewer complications, but has not been accepted worldwide. The development of azoospermia may be delayed for 2–6 months, and other forms of contraception must be used until two sperm-free ejaculations provide proof of sterility. Current data indicate that reanastomosis may restore fertility in 50–70% of men, but the success rate declines with time after vasectomy and may be influenced by factors such as the development of antisperm antibodies which occurs in 60–80% of men.

INTRAUTERINE DEVICES

IUDs inhibit pregnancy through a spermicidal effect (copper IUDs) or by inhibiting ovulation (progestin containing devices). IUDs provide a high level of efficacy in the absence of systemic metabolic effects, and ongoing motivation is not required to ensure efficacy once the device has been placed. IUD use is greatest in Europe and Canada (33%) but is increasing in other parts of the world, including the United States. An IUD should not be used in women at high risk for development of STI or in women at high risk for bacterial endocarditis. Progestin-containing IUDs are contraindicated in women with breast cancer. The IUD may not be effective in women with uterine leiomyomas because they alter the size or shape of the uterine cavity. IUD use is associated with increased menstrual blood flow, although this is less pronounced with the progestin-releasing IUD.

HORMONAL METHODS

Oral Contraceptive Pills Because of their ease of use and efficacy, oral contraceptive pills are the most widely used form of hormonal contraception. They act by suppressing ovulation, changing cervical mucus, and altering the endometrium. The current formulations are made from synthetic estrogens and progestins. The estrogen component of the pill consists of ethinyl estradiol or mestranol, which is metabolized to ethinyl estradiol. Multiple synthetic progestins are used. Norethindrone and its derivatives are used in many formulations. Low-dose norgestimate and the more recently developed (third-generation) progestins (desogestrel, gestodene, drospirenone) have a less androgenic profile; levonorgestrel appears to be the most androgenic of the progestins and should be avoided in patients with hyperandrogenism. The three major formulations of oral contraceptives are (1) fixed-dose estrogen-progestin combination, (2) phasic estrogen-progestin combination, and (3) progestin only. Each of these formulations is administered daily for 3 weeks followed by a week of no medication during which menstrual bleeding generally occurs. Two extended oral contraceptives are approved for use in the United States; Seasonale is a 3-month preparation with 84 days of active drug and 7 days of placebo, whereas Lybrel is a continuous preparation. Current doses of ethinyl estradiol range from 10 to 50 μg. However, indications for the 50-μg dose are rare, and the majority of formulations contain 20–35 μg of ethinyl estradiol. The reduced estrogen and progestin content in the second- and third-generation pills has decreased both side effects and risks associated with oral contraceptive use **(Table 389-2)**. At the currently used doses, patients must be cautioned not to miss pills due to the potential for ovulation and this may be particularly important in obese women. Side effects, including breakthrough bleeding, amenorrhea, breast tenderness, and weight gain, often respond to a change in formulation. Oral contraceptive use is associated with a decreased risk of endometrial, ovarian and colon cancer. However, even the lower dose oral contraceptives have been associated with an increased risk of breast cancer, cardiovascular disease (myocardial infarction, stroke, venous thromboembolism [VTE]), but the absolute excess risk is extremely low. VTE risk is higher with the third-generation than the second-generation progestins, and the risk of stroke and VTE is also higher with drospirenone (although not cyproterone). Again, the absolute excess risk is small. In addition to their use as highly effective contraceptives, estrogen-progestin combinations are used for treatment of amenorrhea and oligoamenorrhea and continuous formulations are commonly used for treatment of premenstrual syndrome and premenstrual dysphoric disorder, menstrual migraine, leiomyomas, and endometriosis.

The microdose progestin-only minipill is less effective as a contraceptive, having a pregnancy rate of 2–7 per 100 women-years. However, it may be appropriate for women at increased risk for cardiovascular disease or for women who cannot tolerate synthetic estrogens.

Alternative Methods A *weekly contraceptive patch* (Ortho Evra) is available and has similar efficacy to oral contraceptives. Approximately 2% of patches fail to adhere, and a similar percentage of women have skin reactions. Efficacy is lower in women weighing >90 kg. The amount of estrogen delivered may be comparable to that of a 40-μg ethinyl estradiol oral contraceptive, raising the possibility of increased risk of VTE, which must be balanced against potential benefits for women not able to successfully use other methods. A *monthly*

TABLE 389-2 Oral Contraceptives: Contraindications and Disease Risk

Contraindications
Absolute
Women age >35 years who smoke ≥15 cigarettes per day
Known ischemic heart disease or multiple risk factors for cardiovascular disease (older age, smoking, diabetes, and hypertension)
Previous thromboembolic event, stroke or known thrombogenic mutations
Complicated valvular heart disease
Complicated solid organ transplantation
Hypertension (systolic ≥160 mmHg or diastolic ≥100 mmHg)
Systemic lupus erythematous (positive or unknown antiphospholipid antibodies
Cirrhosis, hepatic adenoma or hepatoma
Pregnancy and early postpartum (<21 days)
Undiagnosed abnormal uterine bleeding
Relative
Hypertension (adequately controlled or systolic 140–159 or diastolic 90–99)
Women receiving anticonvulsant drug therapy
Women following bariatric surgery (malabsorptive procedure)
Disease Risks
Increased
Coronary heart disease—increased in smokers >35; no relation to progestin type
Hypertension—relative risk 1.8 (current users) and 1.2 (previous users)
Venous thrombosis—relative risk ~4; may be higher with third-generation progestin, drospirenone, and patch; compounded by obesity (tenfold increased risk compared with nonobese, no OCP); markedly increased with factor V Leiden or prothrombin gene mutations
Stroke—slight increase; unclear relation to migraine headache
Cerebral vein thrombosis—relative risk ~13–15; synergistic with prothrombin gene mutation
Cervical cancer—relative risk 2–4
Breast cancer—may increase risk, particularly in carriers of *BRCA1* and possibly *BRCA2*
Decreased
Ovarian cancer—50% reduction in risk
Endometrial cancer—40% reduction in risk

Abbreviation: OCP, oral contraceptive pill.

contraceptive estrogen/progestin injection (Lunelle) is highly effective, with a first-year failure rate of <0.2%, but it may be less effective in obese women. Its use is associated with bleeding irregularities that diminish over time. Fertility returns rapidly after discontinuation. A *monthly vaginal ring* (NuvaRing) that is intended to be left in place during intercourse is also available for contraceptive use. It is highly effective, with a 12-month failure rate of 0.7%. Ovulation returns within the first recovery cycle after discontinuation.

Long-Term Hormonal Contraceptives There are two forms of long-acting hormonal contraceptives in addition to the progestin-containing IUD, both of which act by inhibiting ovulation and causing changes in the endometrium and cervical mucus that result in decreased implantation and sperm transport. Depot medroxyprogesterone acetate (Depo-Provera, DMPA) is effective for 3 months when administered SC or IM, but return of fertility after discontinuation may be delayed for up to 12–18 months. Irregular bleeding is common initially with amenorrhea in almost 70% of users at 2 years. Two percent of women discontinue use due to weight gain. Nexplanon is a hormonal implant that slowly releases the progestin, etonogestrel, and is approved by the FDA for 3 years of use. It is associated with unscheduled bleeding but has very favorable continuation rates, and is associated with rapid return of ovulation after removal. Both forms of contraception may be particularly good for women in whom an estrogen-containing contraceptive is contraindicated (e.g., migraine exacerbation, sickle cell anemia, fibroids). DMPA can induce bone loss and may be contraindicated in areas of high HIV prevalence because of potential negative effects on both immunologic regulation and genital track permeability. Neither of these side effects occur with use of the progestin implant. DMPA and Nexplanon are contraindicated in women with current breast cancer because of theoretical concerns about the adverse effects of progestins on breast cancer.

POSTCOITAL CONTRACEPTION

The probability of pregnancy without relation to time of the month is 8%, but the probability varies significantly in relation to proximity to ovulation and may be as high has 30%. In order of efficacy, methods of postcoital contraception include the following:

1. Copper IUD insertion within a maximum of 5 days has a reported efficacy of 99–100% and prevents pregnancy by its spermicidal effect; insertion is frequently available through family planning clinics, but may be associated with a higher risk of abdominal pain compared with other methods.
2. Oral antiprogestins (ulipristal acetate, 30 mg single dose, available worldwide, or mifepristone, 600 mg single dose, not available for this indication in the United States) prevent pregnancy by delaying or preventing ovulation; when administered, ideally within 72 h but up to 120 h after intercourse, they have an efficacy of 98–99%; require a prescription.
3. Levonorgestrel (1.5 mg as a single dose) delays or prevents ovulation and is not effective after ovulation; should be taken within 72 h of unprotected intercourse, and has an efficacy that varies between 60 and 94%; it is available over the counter.

Combined estrogen and progestin regimens have lower efficacy and are no longer recommended. A pregnancy test is not necessary before the use of oral methods, but pregnancy should be excluded before IUD insertion. Risk factors for failure of oral regimens include close proximity to ovulation and unprotected intercourse after use. In addition, there is an increased risk of pregnancy in obese and overweight women using levonorgestrel for postcoital contraception and an increased risk in obese women using an antiprogestin.

IMPACT OF OBESITY ON CONTRACEPTIVE CHOICE

Approximately one-third of adults in the United States are obese. Although obesity is associated with some reduction in fertility, the vast majority of obese women can conceive. The risk of pregnancy-associated complications is higher in obese women. Intrauterine contraception may be more effective than oral or transdermal methods for obese women. The WHO guidelines provide no restrictions (class 1) for the use of intrauterine contraception, DMPA, and progestin-only pills for obese women (BMI ≥30) in the absence of coexistent medical problems, whereas methods that include estrogen (pill, patch, ring) are considered class 2 (advantages generally outweigh theoretical or proven risks) due to the increased risk of thromboembolic disease. There are no restrictions to the use of any contraceptive methods following restrictive bariatric surgery procedures, but both combined and progestin-only pills are relatively less effective following procedures associated with malabsorption.

FURTHER READING

Birgisson NE et al: Preventing unintended pregnancy: The contraceptive CHOICE project in review. J Womens Health (Larchmt) 24:349, 2015.

Broughton DE, Moley KH: Obesity and female infertility: Potential mediators of obesity's impact. Fertil Steril 107:840, 2017.

Curtis KM, Peipert JF: Long-acting reversible contraception. N Engl J Med 376:461, 2017.

Donnez J, Dolmans M-M: Fertility preservation in women. N Engl J Med 377:1657, 2017.

Jackson AV et al: Racial and ethnic differences in contraception use and obstetric outcomes: A review. Semin Perinatol 41:273, 2017.

Lee JK, Schwartz EB: The safety of available and emerging options for emergency contraception. Expert Opinion on Drug Safety 16:1163, 2017.

Morch LS et al: Contemporary hormonal contraception and the risk of breast cancer. N Engl J Med 377:2228, 2017.
Simmons KB, Elelman AB: Hormonal contraception and obesity. Fertil Steril 106:1282, 2016.
Steiner AZ et al: Association between biomarkers of ovarian reserve and infertility among older women of reproductive age. JAMA 318:1367, 2017.
Tepper NK et al: Update to CDC's U.S. medical eligibility Criteria for contraceptive use, 2016: Revised recommendations for the use of hormonal contraception among women at high risk for HIV infection. MMWR 66:990, 2017.

390 Sexual Dysfunction

Kevin T. McVary

Male sexual dysfunction affects 10–25% of middle-aged and elderly men, and female sexual dysfunction occurs with a similar frequency. Demographic changes, the popularity of newer treatments, and greater awareness of sexual dysfunction by patients and society have led to increased diagnosis and associated health care expenditures for the management of this common disorder. Sexual health and satisfaction with sex life are important aspects of quality of life for many, including those in poor health. Because many patients are reluctant to initiate discussion of their sex lives, physicians should address this topic directly to elicit a history of sexual dysfunction. Specifically addressing sexual health should be a routine part of the clinical encounter.

MALE SEXUAL DYSFUNCTION

PHYSIOLOGY OF MALE SEXUAL RESPONSE

Normal male sexual function requires (1) an intact libido, (2) the ability to achieve and maintain penile erection, (3) ejaculation, and (4) detumescence. *Libido* refers to sexual desire and is influenced by a variety of visual, olfactory, tactile, auditory, imaginative, and hormonal stimuli. Sex steroids, particularly testosterone, act to increase libido. Libido can be diminished by hormonal or psychiatric disorders and by medications.

Penile tumescence leading to erection depends on an increased flow of blood into the lacunar network accompanied by complete relaxation of the arteries and corporal smooth muscle. The microarchitecture of the corpora is composed of a mass of smooth muscle (trabecula) that contains a network of endothelial-lined vessels (lacunar spaces). Subsequent compression of the trabecular smooth muscle against the fibroelastic tunica albuginea causes a passive closure of the emissary veins and accumulation of blood in the corpora. In the presence of a full erection and a competent valve mechanism, the corpora become noncompressible cylinders from which blood does not escape.

The central nervous system (CNS) exerts an important influence by either stimulating or antagonizing spinal pathways that mediate erectile function and ejaculation. The erectile response is mediated by a combination of central (psychogenic) innervation and peripheral (reflexogenic) innervation. Sensory nerves that originate from receptors in the penile skin and glans converge to form the dorsal nerve of the penis, which travels to the S2-S4 dorsal root ganglia via the pudendal nerve. Parasympathetic nerve fibers to the penis arise from neurons in the intermediolateral columns of the S2-S4 sacral spinal segments. Sympathetic innervation originates from the T-11 to the L-2 spinal segments and descends through the hypogastric plexus.

Neural input to smooth-muscle tone is crucial to the initiation and maintenance of an erection. There is also an intricate interaction between the corporal smooth-muscle cell and its overlying endothelial cell lining **(Fig. 390-1*A*)**. Nitric oxide, which induces vascular relaxation, promotes erection and is opposed by endothelin 1 (ET-1) and Rho kinase, which mediate vascular contraction. Nitric oxide is synthesized from L-arginine by nitric oxide synthase and is released from the nonadrenergic, noncholinergic (NANC) autonomic nerve supply to act postjunctionally on smooth-muscle cells. Nitric oxide increases the production of cyclic 3′,5′-guanosine monophosphate (cyclic GMP), which induces relaxation of smooth muscle **(Fig. 390-1*B*)**. Cyclic GMP is gradually broken down by phosphodiesterase type 5 (PDE-5). Inhibitors of PDE-5 such as the oral medications sildenafil, vardenafil, and tadalafil maintain erections by reducing the breakdown of cyclic GMP **(Fig. 390-2)**. However, if nitric oxide is not produced at some level, PDE-5 inhibitors are ineffective, as these drugs facilitate, but do not initiate, the initial enzyme cascade. In addition to nitric oxide, vasoactive prostaglandins (PGE_1, $PGF_{2\alpha}$) are synthesized within the cavernosal tissue and increase cyclic AMP levels, also leading to relaxation of cavernosal smooth-muscle cells.

Ejaculation is stimulated by the sympathetic nervous system; this results in contraction of the epididymis, vas deferens, seminal vesicles, and prostate, causing seminal fluid to enter the urethra. Seminal fluid emission is followed by rhythmic contractions of the bulbocavernosus and ischiocavernosus muscles, leading to ejaculation. This is followed by expulsion, characterized by stereotypic rhythmic contractions of the striated perineal muscles, leading to forceful expulsion of semen with the bladder neck closed. This emission and expulsion are controlled by the autonomic (parasympathetic and sympathetic) and somatic spinal centers, respectively. The synchronization between autonomic and somatic spinal centers is orchestrated by interneurons that form a spinal ejaculation generator which is present in mammals including man.

Premature ejaculation usually is related to anxiety or a learned behavior and is amenable to behavioral therapy or treatment with medications such as selective serotonin reuptake inhibitors (SSRIs). *Retrograde ejaculation* results when the internal urethral sphincter does not close; it may occur in men with diabetes or after surgery involving the bladder neck.

Detumescence is mediated by norepinephrine from the sympathetic nerves, endothelin from the vascular surface, and smooth-muscle contraction induced by postsynaptic α-adrenergic receptors and activation of Rho kinase. These events increase venous outflow and restore the flaccid state. Venous leak can cause premature detumescence and is caused by insufficient relaxation of the corporal smooth muscle rather than a specific anatomic defect. *Priapism* refers to a persistent and painful erection and may be associated with sickle cell anemia, hypercoagulable states, spinal cord injury, or injection of vasodilator agents into the penis.

ERECTILE DYSFUNCTION

Epidemiology Erectile dysfunction (ED) is not considered a normal part of the aging process. Nonetheless, it is associated with certain physiologic and psychological changes related to age. In the Massachusetts Male Aging Study (MMAS), a community-based survey of men aged 40–70, 52% of responders reported some degree of ED. Complete ED occurred in 10% of respondents, moderate ED in 25%, and minimal ED in 17%. The incidence of moderate or severe ED more than doubled between the ages of 40 and 70. In the National Health and Social Life Survey (NHSLS), which included a sample of men and women aged 18–59, 10% of men reported being unable to maintain an erection (corresponding to the proportion of men in the MMAS reporting severe ED). Incidence was highest among men in the age group 50–59 (21%) and men who were poor (14%), divorced (14%), and less educated (13%).

The incidence of ED is also higher among men with certain medical disorders, such as diabetes mellitus, obesity, lower urinary tract symptoms secondary to benign prostatic hyperplasia (BPH), heart disease, hypertension, decreased high-density lipoprotein (HDL) levels and diseases associated with general systemic inflammation (e.g., rheumatoid arthritis). Cardiovascular disease and ED share etiologies as well as pathophysiology (e.g., endothelial dysfunction), and the degree of ED appears to correlate with the severity of cardiovascular disease. Consequently, ED represents a "sentinel symptom" in patients with occult cardiovascular and peripheral vascular disease.

Smoking is also a significant risk factor in the development of ED. Medications used in treating diabetes or cardiovascular disease are

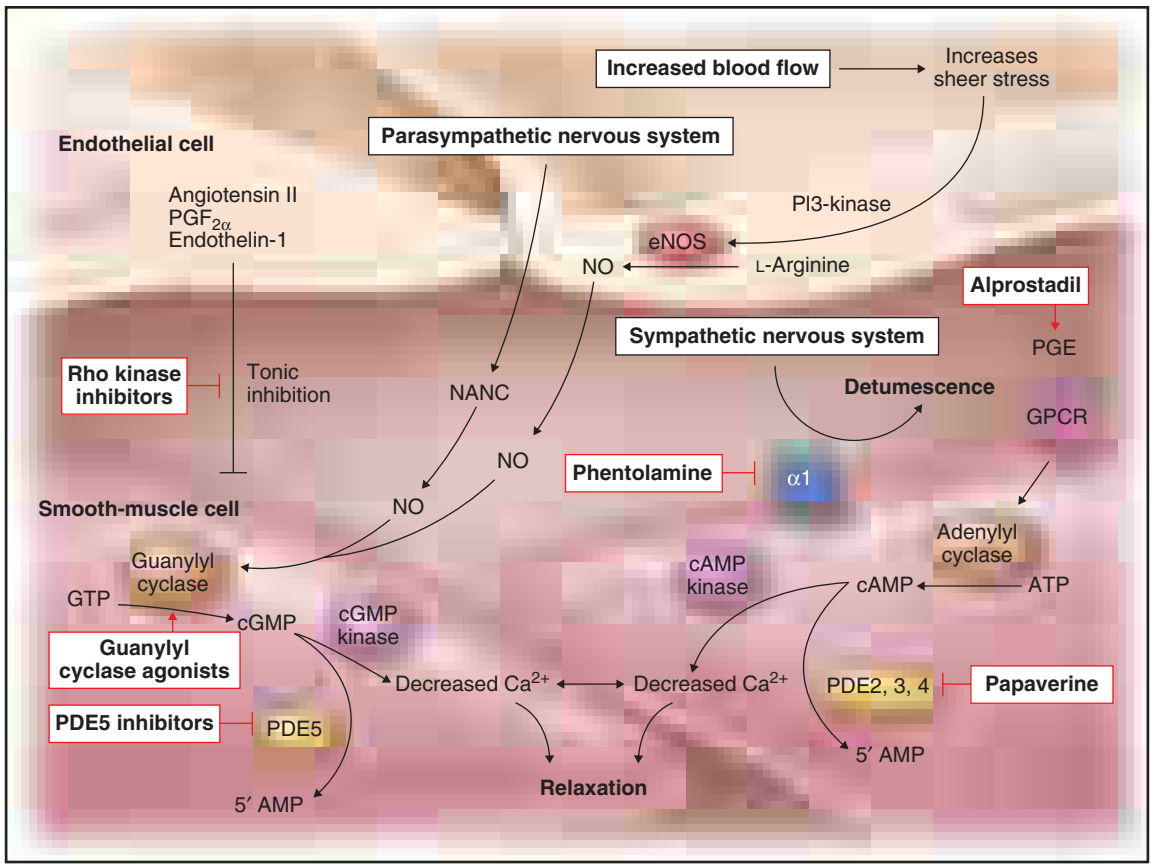


FIGURE 390-1 Pathways that control erection and detumescence. ***A.*** Outflow from the parasympathetic nervous system leads to relaxation of the cavernous sinusoids in two ways, both of which increase the concentration of nitric oxide (NO) in smooth-muscle cells. First, NO is the neurotransmitter in nonadrenergic, noncholinergic (NANC) fibers; second, stimulation of endothelial nitric oxide synthase (eNOS) through cholinergic output causes increased production of NO. The NO produced in the endothelium then diffuses into the smooth-muscle cells and decreases its intracellular calcium concentration through a pathway mediated by cyclic guanosine monophosphate (cGMP), leading to relaxation. A separate mechanism that decreases the intracellular calcium level is mediated by cyclic adenosine monophosphate (cAMP). With increased cavernosal blood flow, as well as increased levels of vascular endothelial growth factor (VEGF), the endothelial release of NO is further sustained through the phosphatidylinositol 3 (PI3) kinase pathway. Active treatments (red boxes) include drugs that affect the cGMP pathway (phosphodiesterase type 5 [PDE-5] inhibitors and guanylyl cyclase agonists), the cAMP pathway (alprostadil), or both pathways (papaverine), along with neural-tone mediators (phentolamine and Rho kinase inhibitors). Agents that are being developed include guanylyl cyclase agonists (to bypass the need for endogenous NO) and Rho kinase inhibitors (to inhibit tonic contraction of smooth-muscle cells mediated through endothelin). ***B.*** Biochemical pathways of NO synthesis and action. Sildenafil, vardenafil, tadalafil, and avanafil enhance erectile function by inhibiting PDE-5, thereby maintaining high levels of cyclic 3′,5′-guanosine monophosphate (cyclic GMP). α1, α-adrenergic receptor; GPCR, G-protein–coupled receptor; GTP, guanosine triphosphate; iCa^{2+}, intracellular calcium; NOS, nitric oxide synthase; PGE, prostaglandin E; PGF, prostaglandin F. *(Part A from K McVary: N Engl J Med 357:2472, 2007; with permission.)*

additional risk factors (see below). There is a higher incidence of ED among men who have undergone radiation or surgery for prostate cancer and in those with a lower spinal cord injury. Psychological causes of ED include depression, anger, stress from unemployment, anxiety (especially younger men), and other stress-related causes.

Pathophysiology ED may result from three basic mechanisms: (1) failure to initiate (psychogenic, endocrinologic, or neurogenic), (2) failure to fill (arteriogenic), and (3) failure to store adequate blood volume within the lacunar network (venoocclusive dysfunction). These categories are not mutually exclusive, and multiple factors contribute to ED in many patients. For example, diminished filling pressure can lead secondarily to venous leak. Psychogenic factors frequently coexist with other etiologic factors and should be considered in all cases. Diabetic, atherosclerotic, and drug-related causes account for >80% of cases of ED in older men.

Vasculogenic The most common organic cause of ED is a disturbance of blood flow to and from the penis. Atherosclerotic or traumatic arterial disease can decrease flow to the lacunar spaces, resulting in decreased rigidity and an increased time to full erection. Excessive outflow through the veins despite adequate inflow also may contribute to ED. Structural alterations to the fibroelastic components of the corpora may cause a loss of compliance and inability to compress the tunical veins. This condition may result from aging, increased cross-linking of collagen fibers induced by nonenzymatic glycosylation, hypoxemia, or altered synthesis of collagen associated with hypercholesterolemia.

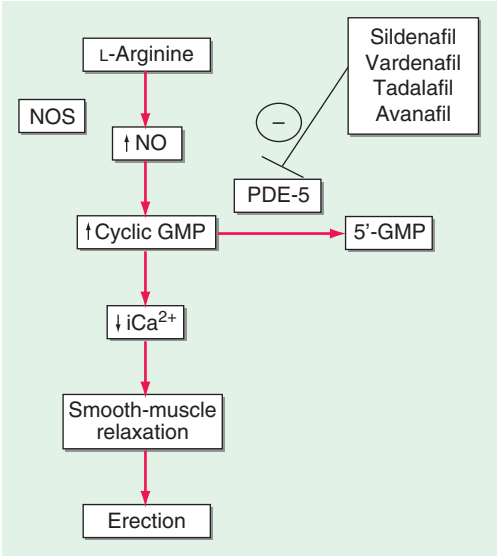


FIGURE 390-2 Biochemical pathways modified by phosphodiesterase type 5 (PDE-5) inhibitors. Sildenafil, vardenafil, tadalafil, and avanafil enhance erectile function by inhibiting PDE-5, thereby maintaining high levels of cyclic 3′,5′-guanosine monophosphate (cyclic GMP). iCa^{2+}, intracellular calcium; NO, nitric oxide; NOS, nitric oxide synthase.

Neurogenic Disorders that affect the sacral spinal cord or the autonomic fibers to the penis preclude nervous system relaxation of penile smooth muscle, thus leading to ED. In patients with spinal cord injury, the degree of ED depends on the completeness and level of the lesion. Patients with incomplete lesions or injuries to the upper part of the spinal cord are more likely to retain erectile capabilities than are those with complete lesions or injuries to the lower part. Although 75% of patients with spinal cord injuries have some erectile capability, only 25% have erections sufficient for penetration. Other neurologic disorders commonly associated with ED include multiple sclerosis and peripheral neuropathy. The latter is often due to either diabetes or alcoholism. Pelvic surgery may cause ED through disruption of the autonomic nerve supply.

Endocrinologic Androgens increase libido, but their exact role in erectile function is unclear. Individuals with castrate levels of testosterone can achieve erections from visual or sexual stimuli. Nonetheless, normal levels of testosterone appear to be important for erectile function, particularly in older males. Androgen replacement therapy can improve depressed erectile function when it is secondary to hypogonadism; however, it is not useful for ED when endogenous testosterone levels are normal. Increased prolactin may decrease libido by suppressing gonadotropin-releasing hormone (GnRH), and it also leads to decreased testosterone levels. Treatment of hyperprolactinemia with dopamine agonists can restore libido and testosterone.

Diabetic ED occurs in 35–75% of men with diabetes mellitus. Pathologic mechanisms are related primarily to diabetes-associated vascular and neurologic complications. Diabetic macrovascular complications are related mainly to age, whereas microvascular complications correlate with the duration of diabetes and the degree of glycemic control **(Chap. 396)**. Individuals with diabetes also have reduced amounts of nitric oxide synthase in both endothelial and neural tissues.

Psychogenic Two mechanisms contribute to the inhibition of erections in psychogenic ED. First, psychogenic stimuli to the sacral cord may inhibit reflexogenic responses, thereby blocking activation of vasodilator outflow to the penis. Second, excess sympathetic stimulation in an anxious man may increase penile smooth-muscle tone. The most common causes of psychogenic ED are performance anxiety, depression, relationship conflict, loss of attraction, sexual inhibition, conflicts over sexual preference, sexual abuse in childhood, and fear of pregnancy or sexually transmitted disease. Almost all patients with ED, even when it has a clear-cut organic basis, develop a psychogenic component as a reaction to ED.

TABLE 390-1 Drugs Associated with Erectile Dysfunction

CLASSIFICATION	DRUGS
Diuretics	Thiazides
	Spironolactone
Antihypertensives	Calcium channel blockers
	Methyldopa
	Clonidine
	Reserpine
	Beta blockers
	Guanethidine
Cardiac/antihyperlipidemics	Digoxin
	Gemfibrozil
	Clofibrate
Antidepressants	Selective serotonin reuptake inhibitors
	Tricyclic antidepressants
	Lithium
	Monoamine oxidase inhibitors
Tranquilizers	Butyrophenones
	Phenothiazines
H_2 antagonists	Ranitidine
	Cimetidine
Hormones	Progesterone
	Estrogens
	Corticosteroids
	GnRH agonists
	5α-Reductase inhibitors
	Cyproterone acetate
Cytotoxic agents	Cyclophosphamide
	Methotrexate
	Roferon-A
Anticholinergics	Disopyramide
	Anticonvulsants
Recreational	Ethanol
	Cocaine
	Marijuana

Abbreviation: GnRH, gonadotropin-releasing hormone.

Medication-Related Medication-induced ED **(Table 390-1)** is estimated to occur in 25% of men seen in general medical outpatient clinics. The adverse effects related to drug therapy are additive, especially in older men. In addition to the drug itself, the underlying disease being treated is likely to contribute to sexual dysfunction. Among the antihypertensive agents, the thiazide diuretics and beta blockers have been implicated most frequently. Calcium channel blockers and angiotensin converting-enzyme inhibitors are cited less frequently. These drugs may act directly at the corporal level (e.g., calcium channel blockers) or indirectly by reducing pelvic blood pressure, which is important in the development of penile rigidity. α-Adrenergic blockers are less likely to cause ED. Estrogens, GnRH agonists, H_2 antagonists, and spironolactone cause ED by suppressing gonadotropin production or by blocking androgen action. Antidepressant and antipsychotic agents—particularly neuroleptics, tricyclics, and SSRIs—are associated with erectile, ejaculatory, orgasmic, and sexual desire difficulties. Among the SSRIs, paroxetine and escitalopram have been associated with the highest risk of sexual dysfunction. Bupropion, nefazodone, and mirtazapine appear less likely to cause sexual dysfunction. A number of molecular pathways have been implicated in antidepressant-induced sexual adverse events. Serotonin has been hypothesized to inhibit normal sexual response by decreasing dopamine-enhanced libido, arousal and erection, and by increasing prolactin release. SSRIs have also been shown to be potent inhibitors of nitric oxide synthesis.

If there is a strong association between the institution of a drug and the onset of ED, alternative medications should be considered. Otherwise, it is often practical to treat the ED without attempting multiple changes in medications, as it may be difficult to establish a causal role for a drug.

APPROACH TO THE PATIENT

Erectile Dysfunction

A good physician-patient relationship helps unravel the possible causes of ED, many of which require discussion of personal and sometimes embarrassing topics. For this reason, a primary care provider is often ideally suited to initiate the evaluation. However, a significant percentage of men experience ED and remain undiagnosed unless specifically questioned about this issue. By far the two most common reasons for underreporting of ED are patient embarrassment and perceptions of physicians' inattention to the disorder. Once the topic is initiated by the physician, patients are more willing to discuss their potency issues. A complete medical and sexual history should be taken in an effort to assess whether the cause of ED is organic, psychogenic, or multifactorial **(Fig. 390-3)**.

Both the patient and his sexual partner should be interviewed regarding sexual history. ED should be distinguished from other sexual problems, such as premature ejaculation. Lifestyle factors such as sexual orientation, the patient's distress from ED, performance anxiety, and details of sexual techniques should be addressed. Standardized questionnaires are available to assess ED, including the International Index of Erectile Function (IIEF) and the more easily administered Sexual Health Inventory for Men (SHIM), a validated abridged version of the IIEF.The initial evaluation of ED begins with a review of the patient's medical, surgical, sexual, and psychosocial histories. The history should note whether the patient has experienced pelvic trauma, surgery, or radiation. In light of the increasing recognition of the relationship between lower urinary tract symptoms and ED, it is advisable to evaluate for the presence of symptoms of bladder outlet obstruction. Questions should focus on the onset of symptoms, the presence and duration of partial erections, and the progression of ED. A history of nocturnal or early morning erections is useful for distinguishing physiologic ED from psychogenic ED. Nocturnal erections occur during rapid eye movement (REM) sleep and require intact neurologic and circulatory systems. Organic causes of ED generally are characterized by a gradual and persistent change in rigidity or the inability to sustain nocturnal, coital, or self-stimulated erections. The patient should be questioned about the presence of penile curvature or pain with coitus. It is also important to address libido, as decreased sexual drive and ED are sometimes the earliest signs of endocrine abnormalities (e.g., increased prolactin, decreased testosterone levels). It is useful to ask whether the problem is confined to coitus with one partner or also involves other partners; ED not uncommonly arises in association with new or extramarital sexual relationships. Situational ED, as opposed to consistent ED, suggests psychogenic causes. Ejaculation is much less commonly affected than erection, but questions should be asked about whether ejaculation is normal, premature, delayed, or absent. Relevant risk factors should be identified, such as diabetes mellitus, coronary artery disease (CAD), and neurologic disorders. The patient's surgical history should be explored with an emphasis on bowel, bladder, prostate, and vascular procedures. A complete drug history is also important. Social changes that may precipitate ED are also crucial to the evaluation, including health worries, spousal death, divorce, relationship difficulties, and financial concerns.

Because ED commonly involves a host of endothelial cell risk factors, men with ED report higher rates of overt and silent myocardial infarction. Therefore, ED in an otherwise asymptomatic male warrants consideration of other vascular disorders, including CAD.

Men who suffer from ED are at high risk for concomitant lower urinary tract symptoms (LUTS) from BPH and vice versa. Given that some treatments of one disorder will impact the other, the clinician should consider an assessment of LUTS in any man with ED.

The physical examination is an essential element in the assessment of ED. Signs of hypertension as well as evidence of thyroid, hepatic, hematologic, cardiovascular, or renal diseases should be sought. An assessment should be made of the endocrine and vascular systems, the external genitalia, and the prostate gland. The penis should be palpated carefully along the corpora to detect fibrotic plaques. Reduced testicular size and loss of secondary sexual characteristics are suggestive of hypogonadism. Neurologic examination should include assessment of anal sphincter tone, investigation of the bulbocavernosus reflex, and testing for peripheral neuropathy.

Although hyperprolactinemia is uncommon, a serum prolactin level should be measured, as decreased libido and/or ED may be the presenting symptoms of a prolactinoma or another mass lesion of the sella **(Chap. 373)**. The serum testosterone level should be measured, and if it is low, gonadotropins should be measured to determine whether hypogonadism is primary (testicular) or secondary (hypothalamic-pituitary) in origin **(Chap. 384)**. If not performed recently, serum chemistries, complete blood count (CBC), and lipid profiles may be of value, as they can yield evidence of anemia, diabetes, hyperlipidemia, or other systemic diseases associated with ED. Determination of serum prostate-specific antigen (PSA) should be conducted according to recommended clinical guidelines **(Chap. 83)**.

Additional diagnostic testing is rarely necessary in the evaluation of ED. However, in selected patients, specialized testing may provide insight into pathologic mechanisms of ED and aid in the selection of treatment options. Optional specialized testing includes (1) studies of nocturnal penile tumescence and rigidity, (2) vascular testing (in-office injection of vasoactive substances, penile Doppler ultrasound, penile angiography, dynamic infusion cavernosography/cavernosometry), (3) neurologic testing (biothesiometry-graded vibratory perception, somatosensory-evoked potentials), and (4) psychological diagnostic tests. The information potentially gained from these procedures must be balanced against their invasiveness and cost.

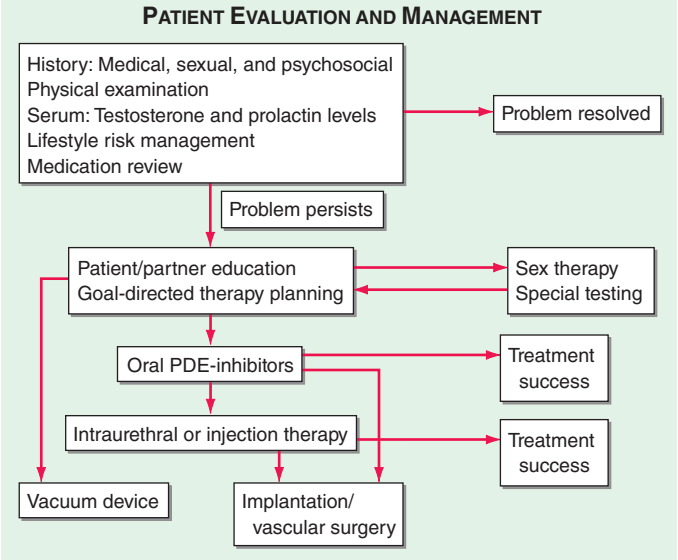


FIGURE 390-3 Algorithm for the evaluation and management of patients with erectile dysfunction. PDE, phosphodiesterase.

TREATMENT

Male Sexual Dysfunction

PATIENT EDUCATION

Patient and partner education is essential in the treatment of ED. In goal-directed therapy, education facilitates understanding of

the disease, the results of the tests, and the selection of treatment. Discussion of treatment options helps clarify how treatment is best offered and stratify first- and second-line therapies. Patients with high-risk lifestyle issues such as obesity, smoking, alcohol abuse, and recreational drug use should be counseled on the role those factors play in the development of ED.

Therapies currently employed for the treatment of ED include oral phosphodiesterase type 5 inhibitor therapy (most commonly used), injection therapies, testosterone therapy, penile devices, and psychological therapy. In addition, limited data suggest that treatments for underlying risk factors and comorbidities—for example, weight loss, exercise, stress reduction, and smoking cessation—may improve erectile function. Decisions regarding therapy should take into account the preferences and expectations of patients and their partners.

ORAL AGENTS

Sildenafil, tadalafil, vardenafil, and avanafil are the only approved and effective oral agents for the treatment of ED. These four medications have markedly improved the management of ED because they are effective for the treatment of a broad range of causes, including psychogenic, diabetic, vasculogenic, post-radical prostatectomy (nerve-sparing procedures), and spinal cord injury. They belong to a class of medications that are selective and potent inhibitors of PDE-5, the predominant phosphodiesterase isoform found in the penis. They are administered in graduated doses and enhance erections after sexual stimulation (Fig. 390-2). The onset of action is ~30–120 min, depending on the medication used and other factors, such as recent food intake. Reduced initial doses should be considered for patients who are elderly, are taking concomitant alpha blockers, have renal insufficiency, or are taking medications that inhibit the CYP3A4 metabolic pathway in the liver (e.g., erythromycin, cimetidine, ketoconazole, and possibly itraconazole and mibefradil), as they may increase the serum concentration of the PDE-5 inhibitors (PDE-5i) or promote hypotension.

Initially there were concerns about the cardiovascular safety of these drugs. It is known that these agents can act as mild vasodilators and warnings exist about orthostatic hypotension with concomitant use of alpha blockers. The use of PDE5i is not contraindicated in men who are also receiving alpha blockers, but they must be stabilized on this blood pressure medication prior to initiating therapy. Earlier concerns that the use of PDE5i would increase cardiovascular events have been mitigated by the results of several controlled trials showing no increase in myocardial ischemic events or overall mortality compared to the general population.

TABLE 390-3 Issues to Consider if Patients Report Failure of PDE-5i to Improve Erectile Dysfunction

1. A trial of medication on at least 6 different days at the maximal dose should be performed before declaring patient nonresponsive to PDE-5i use.
2. Confirm that the patient did not partake in a high-fat meal prior to taking medication.
3. Failure to include physical and psychic stimulation at the time of foreplay to induce endogenous NO.
4. Unrecognized hypogonadism

Abbreviations: NO, nitric oxide; PDE-5i, phosphodiesterase.

Several randomized trials have demonstrated the efficacy of this class of medications. There are no compelling data to support the superiority of one PDE-5i over another. Subtle differences between agents have variable clinical relevance **(see Table 390-2)**.

Patients may fail to respond to a PDE-5i for several reasons **(Table 390-3)**. Some patients may not tolerate PDE-5i secondary to adverse events from vasodilation in nonpenile tissues expressing PDE-5 or from the inhibition of homologous nonpenile isozymes (i.e., PDE-6 found in the retina). Abnormal vision attributed to the effects of PDE-5i on retinal PDE-6 is of short duration, reported only with sildenafil and not thought to be clinically significant. A more serious concern is the possibility that PDE-5i may cause nonarteritic anterior ischemic optic neuropathy (NAION); although data to support that association are limited, it is prudent to avoid the use of these agents in men with a prior history of nonarteritic anterior ischemic optic neuropathy.

Testosterone supplementation combined with a PDE-5i may be beneficial in improving erectile function in hypogonadal men with ED who are unresponsive to PDE-5i alone. These drugs do not affect ejaculation, orgasm, or sexual drive. Side effects associated with PDE-5i include headaches (19%), facial flushing (9%), dyspepsia (6%), and nasal congestion (4%). Approximately 7% of men using sildenafil may experience transient altered color vision (blue halo effect), and 6% of men taking tadalafil may experience loin pain. PDE-5i is contraindicated in men receiving nitrate therapy for cardiovascular disease, including agents delivered by the oral, sublingual, transnasal, and topical routes. These agents can potentiate its hypotensive effect and may result in profound shock. Likewise, amyl/butyl nitrate "poppers" may have a fatal synergistic effect on blood pressure. PDE-5i also should be avoided in patients with congestive heart failure and cardiomyopathy because of the risk of vascular collapse. Because sexual activity leads to an increase in

TABLE 390-2 PDE5 Inhibitor

DRUG	ONSET OF ACTION	$T_{1/2}$	DOSE	ADVERSE EFFECTS	CONTRAINDICATIONS
Sildenafil	T_{max} 30–120 min Duration 4 h High-fat meal decreases absorption. Alcohol use may affect efficacy.	2–5 h	25–100 mg Starting dose 50 mg	Headache, flushing, dyspepsia, nasal congestion, altered vision	Nitrates Hypotension Cardiovascular risk factors Retinitis pigmentosa Change dose with some antiretrovirals Should be on stable dose of alpha blockers
Vardenafil	T_{max} 30–120 min Duration 4–5 h High-fat meal decreases absorption. ETOH may affect efficacy.	4.5 h	5–10 mg	Headache, flushing, rhinitis, dyspepsia	Same as sildenafil May have minor prolongation of QT interval Concomitant use of Class I anti-arrhythmic
Tadalafil	T_{max} 30–60 min Duration 12–36 h Plasma concentration not affected by food or ETOH	17.5 h	10 mg, 20 mg; 2.5 or 5 mg for daily dose	Headache, dyspepsia, backpain, nasal congestion, myalgia	Same as sildenafil
Avanafil	T_{max} 30 min Duration 2 h Plasma concentration not affected by food	3–5 h	50 mg 100 mg and 200 mg dose	Headache, flushing, nasal congestion nasopharyngitis back pain	Same as sildenafil

physiologic expenditure [5–6 metabolic equivalent tasks (METs)], physicians have been advised to exercise caution in prescribing any drug for sexual activity to those with active coronary disease, heart failure, borderline hypotension, or hypovolemia and to those on complex antihypertensive regimens.

Although the various forms of PDE-5i have a common mechanism of action, there are a few differences among the four agents (Table 390-2). Tadalafil is unique in its longer half-life and avanafil appears to have the fastest onset of action. All four drugs are effective for patients with ED of all ages, severities, and etiologies. Although there are pharmacokinetic and pharmacodynamic differences among these agents, clinically relevant differences are not clear.

ANDROGEN THERAPY

Testosterone replacement is used to treat both primary and secondary causes of hypogonadism **(Chap. 384)**. Androgen supplementation in the setting of normal testosterone is rarely efficacious in the treatment of ED and is discouraged. Methods of androgen replacement include transdermal patches and gels, including cutaneous nasal and axillary gels. Parenteral administration of long-acting testosterone esters (enanthate and cypionate), long-acting subcutaneous pellets, and oral preparations (17 α-alkylated derivatives) are also available **(Chap. 384)**. Oral androgen preparations have the potential for hepatotoxicity and should be avoided.

The increased scrutiny of testosterone caused the U.S. Food and Drug Administration (FDA) to issue a warning that there is a "weak signal" that testosterone replacement therapy increases the risk of thromboembolic events and may have addictive properties. Though testosterone therapy has known risks, such as water retention in heart failure patients, and worsening sleep apnea, increasing evidence suggests that, when monitored appropriately, this therapy decreases the risk for metabolic syndrome, changes body composition by increasing lean muscle mass, and improves insulin sensitivity and average hemoglobin A1c. This evidence, combined with the fact that hypogonadism is a known risk factor for metabolic syndrome and cardiovascular disease, has led to the conclusion that testosterone therapy for age-related hypogonadism in fact improves overall health and decreases risk of cardiovascular events. It is important to note that men with secondary hypogonadism who desire fertility should not be treated directly with testosterone, but with an alternative such as the selective estrogen-receptor modulator (SERM) clomiphene citrate, which increases gonadotropin levels, stimulating testicular T production.

Testosterone circulates in the body in two forms: free and unbound or bound to proteins such as albumin or sex hormone-binding globulin (SHBG). SHBG has a very high affinity for testosterone and, thus testosterone bound to SHBG does not bind to the androgen receptor and is not bioavailable. Bioavailable testosterone is any testosterone that is not bound to SHBG. Unfortunately, reliable assays to directly measure bioavailable testosterone or free testosterone are expensive and difficult to perform, and are thus not offered by most laboratories. However, direct measurement of SHBG is inexpensive and reliable, allowing free and bioavailable testosterone to be calculated.

Men who receive testosterone should be reevaluated after 3–6 months and at least annually thereafter for testosterone levels, erectile function, and adverse effects, which may include gynecomastia, sleep apnea, development or exacerbation of LUTS or BPH, prostate cancer, lowering of HDL, erythrocytosis, elevations of liver function tests, and reduced fertility. Periodic reevaluation should include measurement of CBC and PSA and digital rectal examination. Therapy should be discontinued in patients who do not respond within 3–6 months without an alternate explanation (e.g., elevated estradiol).

VACUUM CONSTRICTION DEVICES

Vacuum constriction devices (VCDs) are a well-established noninvasive therapy. They are a reasonable treatment alternative for select patients who cannot take sildenafil or do not desire other interventions. VCDs draw venous blood into the penis and use a constriction ring to restrict venous return and maintain tumescence. Adverse events with VCD include pain, numbness, bruising, and altered ejaculation. Additionally, many patients complain that the devices are cumbersome and that the induced erections have a nonphysiologic appearance and feel.

INTRAURETHRAL ALPROSTADIL

If a patient fails to respond to oral agents, a reasonable next choice is intraurethral or self-injection of vasoactive substances. Intraurethral prostaglandin E_1 (alprostadil), in the form of a semisolid pellet (doses of 125–1000 μg), is delivered with an applicator. Approximately 65% of men receiving intraurethral alprostadil respond with an erection when tested in the office, but only 50% achieve successful coitus at home. Intraurethral insertion is associated with a markedly reduced incidence of priapism in comparison to intracavernosal injection.

INTRACAVERNOSAL SELF-INJECTION

Injection of synthetic formulations of alprostadil is effective in 70–80% of patients with ED, but discontinuation rates are high because of the invasive nature of administration. Doses range between 1 and 40 μg. Injection therapy is contraindicated in men with a history of hypersensitivity to the drug and men at risk for priapism (hypercoagulable states, sickle cell disease). Side effects include local adverse events, prolonged erections, pain, and fibrosis with chronic use. Various combinations of alprostadil, phentolamine, and/or papaverine sometimes are used.

SURGERY

A less frequently used form of therapy for ED involves the surgical implantation of a semirigid or inflatable penile prosthesis. The choice of prosthesis is dependent on patient preference and should take into account body habitus and manual dexterity, which may affect the ability of the patient to manipulate the device. Because of the permanence of prosthetic devices, patients should be advised to first consider less invasive options for treatment. These surgical treatments are associated with a low rate of potential complications but generally are reserved for treatment of refractory ED or in men who cannot tolerate less invasive treatments. Despite their cost and the requirement for surgery, penile prostheses are associated with very high rates of patient and partner satisfaction.

SEX THERAPY

A course of sex therapy may be useful for addressing specific interpersonal factors that may affect sexual functioning. Sex therapy generally consists of in-session discussion and at-home exercises specific to the person and the relationship. Psychosexual therapy involves techniques such as sensate focus (nongenital massage), sensory awareness exercises, correction of misconceptions about sexuality, and interpersonal difficulties therapy (e.g., open communication about sexual issues, physical intimacy scheduling, and behavioral interventions). These approaches may be useful in patients who have psychogenic or social components to their ED, although data from randomized trials are scanty and inconsistent. It is preferable to include both partners in therapy if the patient is involved in an ongoing relationship.

FEMALE SEXUAL DYSFUNCTION

Female sexual dysfunction (FSD) has traditionally included disorders of desire, arousal, pain, and muted orgasm. The associated risk factors for FSD are similar to those in males: cardiovascular disease, endocrine disorders, hypertension, neurologic disorders, and smoking **(Table 390-4)**. Women with hypertension report significantly lower sexual satisfaction (especially younger women).

EPIDEMIOLOGY

Epidemiologic data are limited, but the available estimates suggest that as many as 43% of women complain of at least one sexual problem. Despite the recent interest in organic causes of FSD, desire and arousal phase disorders (including lubrication complaints) remain the most

TABLE 390-4 Risk Factors for Female Sexual Dysfunction

Neurologic disease: stroke, spinal cord injury, parkinsonism
Trauma, genital surgery, radiation
Endocrinopathies: diabetes, hyperprolactinemia
Liver and/or renal failure
Cardiovascular disease, especially hypertension
Psychological factors and interpersonal relationship disorders: sexual abuse, life stressors
Medications
Antiandrogens: cimetidine, spironolactone
Antidepressants, alcohol, hypnotics, sedatives
Antiestrogens or GnRH antagonists
Antihistamines, sympathomimetic amines
Antihypertensives: diuretics, calcium channel blockers
Alkylating agents
Anticholinergics

Abbreviation: GnRH, gonadotropin-releasing hormone.

common presenting problems when surveyed in a community-based population.

PHYSIOLOGY OF THE FEMALE SEXUAL RESPONSE

The female sexual response requires the presence of estrogens. A role for androgens is also likely but less well established. In the CNS, estrogens and androgens work synergistically to enhance sexual arousal and response. A number of studies report enhanced libido in women during preovulatory phases of the menstrual cycle, suggesting that hormones involved in the ovulatory surge (e.g., estrogens) increase desire.

Sexual motivation is heavily influenced by context, including the environment and partner factors. Once sufficient sexual desire is reached, sexual arousal is mediated by the central and autonomic nervous systems. Cerebral sympathetic outflow is thought to increase desire, and peripheral parasympathetic activity results in clitoral vasocongestion and vaginal secretion (lubrication).

The neurotransmitters for clitoral corporal engorgement are similar to those in the male, with a prominent role for neural, smooth-muscle, and endothelial released nitric oxide (NO). A fine network of vaginal nerves and arterioles promotes a vaginal transudate. The major transmitters of this complex vaginal response are not certain, but roles for NO and vasointestinal polypeptide (VIP) are suspected. Investigators studying the normal female sexual response have challenged the long-held construct of a linear and unmitigated relationship between initial desire, arousal, vasocongestion, lubrication, and eventual orgasm. Caregivers should consider a paradigm of a positive emotional and physical outcome with one, many, or no orgasmic peak and release.

Although there are anatomic differences as well as variation in the density of vascular and neural beds in males and females, the primary effectors of sexual response are strikingly similar. Intact sensation is important for arousal. Thus, reduced levels of sexual functioning are more common in women with peripheral neuropathies (e.g., diabetes). Vaginal lubrication is a transudate of serum that results from the increased pelvic blood flow associated with arousal. Vascular insufficiency from a variety of causes may compromise adequate lubrication and result in dyspareunia. Cavernosal and arteriole smooth-muscle relaxation occurs via increased nitric oxide synthase (NOS) activity and produces engorgement in the clitoris and the surrounding vestibule. Orgasm requires an intact sympathetic outflow tract; hence, orgasmic disorders are common in female patients with spinal cord injuries.

APPROACH TO THE PATIENT

Female Sexual Dysfunction

Many women do not volunteer information about their sexual response. Open-ended questions in a supportive atmosphere are helpful in initiating a discussion of sexual fitness in women who are reluctant to discuss such issues. Once a complaint has been voiced, a comprehensive evaluation should be performed, including a medical history, a psychosocial history, a physical examination, and limited laboratory testing.

The history should include the usual medical, surgical, obstetric, psychological, gynecologic, sexual, and social information. Past experiences, intimacy, knowledge, and partner availability should also be ascertained. Medical disorders that may affect sexual health should be delineated. They include diabetes, cardiovascular disease, gynecologic conditions, obstetric history, depression, anxiety disorders, and neurologic disease. Medications should be reviewed as they may affect arousal, libido, and orgasm. The need for counseling and recognizing life stresses should be identified. The physical examination should assess the genitalia, including the clitoris. Pelvic floor examination may identify prolapse or other disorders. Laboratory studies are needed, especially if menopausal status is uncertain. Estradiol, follicle-stimulating hormone (FSH), and luteinizing hormone (LH) are usually obtained, and dehydroepiandrosterone (DHEA) should be considered as it reflects adrenal androgen secretion. A CBC, liver function assessment, and lipid studies may be useful, if not otherwise obtained. Complicated diagnostic evaluation such as clitoral Doppler ultrasonography and biothesiometry require expensive equipment and are of uncertain utility. It is important for the patient to identify which symptoms are most distressing.

The evaluation of FSD previously occurred mainly in a psychosocial context. However, inconsistencies between diagnostic categories based only on psychosocial considerations and the emerging recognition of organic etiologies have led to a new classification of FSD. This diagnostic scheme is based on four components that are not mutually exclusive: (1) *Hypoactive sexual desire*—the persistent or recurrent lack of sexual thoughts and/or receptivity to sexual activity, which causes personal distress. Hypoactive sexual desire may result from endocrine failure or may be associated with psychological or emotional disorders, (2) *Sexual interest arousal disorder*—the persistent or recurrent inability to attain or maintain sexual excitement, which causes personal distress, (3) *Orgasmic disorder*—the persistent or recurrent loss of orgasmic potential after sufficient sexual stimulation and arousal, which causes personal distress, and (4) *Sexual pain disorder*—persistent or recurrent genital pain associated with noncoital sexual stimulation, which causes personal distress. This newer classification emphasizes "personal distress" as a requirement for dysfunction and provides clinicians with an organized framework for evaluation before or in conjunction with more traditional counseling methods.

TREATMENT

Female Sexual Dysfunction

GENERAL

An open discussion with the patient is important as couples may need to be educated about normal anatomy and physiologic responses, including the role of orgasm, in sexual encounters. Physiologic changes associated with aging and/or disease should be explained. Couples may need to be reminded that clitoral stimulation rather than coital intromission may be more beneficial.

Behavioral modification and nonpharmacologic therapies should be a first step. Patient and partner counseling may improve communication and relationship strains. Lifestyle changes involving known risk factors can be an important part of the treatment process. Emphasis on maximizing physical health and avoiding lifestyles (e.g., smoking, alcohol abuse) and medications likely to produce FSD is important (Table 390-3). The use of topical lubricants may address complaints of dyspareunia and dryness. Contributing medications such as antidepressants may need to be altered, including the use of medications with less impact on sexual function, dose reduction, medication switching, or drug holidays.

HORMONAL THERAPY

In postmenopausal women, estrogen replacement therapy may be helpful in treating vaginal atrophy, decreasing coital pain, and improving clitoral sensitivity **(Chap. 388)**. Menopause and its transition represent significant risk factors for the development of vulvovaginal atrophy-related sexual dysfunction. Available vaginal estrogen preparations include conjugated equine estrogens, estradiol vaginal cream, a sustained-release intravaginal estradiol ring, or a low-dose estradiol tablet. Vaginal estrogen preparations with the lowest systemic absorption rate may be preferred in women with history of breast cancer and severe vaginal atrophy. Vaginal lubricants and moisturizers applied on a regular basis have an efficacy comparable to that of local estrogen therapy and should be offered to women wishing to avoid the use of vaginal estrogens. If a hormonal supplement is chosen, then estrogen replacement in the form of local cream is the preferred method, as it avoids systemic side effects. Androgen levels in women decline substantially before menopause. However, low levels of testosterone or DHEA are not effective predictors of a positive therapeutic outcome with androgen therapy. The widespread use of exogenous androgens is not supported by the literature except in select circumstances (premature ovarian failure or menopausal states) and in secondary arousal disorders.

Atrophic vaginitis is very common in postmenopausal women and is most commonly treated with estrogen-based treatments. However, many women are hesitant to use estrogen-based treatments due to health concerns or are unable to use them due to a history of breast cancer or endometrial cancer. Hyaluronic acid vaginal gel has been found to be efficacious in treating atrophic vaginitis.

ORAL AGENTS

Flibanserin, originally developed as an antidepressant, has been approved by the FDA as a treatment for low sexual desire in premenopausal women. Flibanserin, a postsynaptic agonist of serotonin receptor 1A and antagonist of serotonin receptor 2A, increases sexual desire and reduces resultant stress in women with HypoSexual Desire Disorder (HSDD) with few adverse effects. Flibanserin may boost sex drive in women who experience low sexual desire and who find the experience distressing. The drug should be discontinued if there is no improvement in sex drive after 8 weeks. Potentially serious side effects include low blood pressure, dizziness, and fainting, particularly if it is mixed with alcohol. Other common adverse events include dizziness, nausea, fatigue, sleepiness, and insomnia. Health care professionals and pharmacies dealing with flibanserin have to undergo a certification process, and patients need to submit a written agreement to abstain from alcohol.

The efficacy of PDE-5i in FDS has been a marked disappointment in light of the proposed role of NO-dependent physiology in the normal female sexual response. The use of PDE-5i for FSD should be discouraged pending proof that it is effective.

CLITORAL VACUUM DEVICE

In patients with arousal and orgasmic difficulties, the option of using a clitoral vacuum device may be explored. This handheld battery-operated device has a small soft plastic cup that applies a vacuum over the stimulated clitoris. This causes increased cavernosal blood flow, engorgement, and vaginal lubrication.

FURTHER READING

AUA Clinical Guideline: The management of erectile dysfunction. Available from *https://www.auanet.org/common/pdf/education/clinical-guidance/Erectile-Dysfunction.pdf*. Accessed January 12, 2017.

Bhasin S: A perspective on the evolving landscape in male reproductive medicine. J Clin Endocrinol Metab 101:827, 2016.

Bhasin S et al: Sexual dysfunction in men and women with endocrine disorders. Lancet 369:597, 2007.

Cappelleri JC, Rosen RC: The Sexual Health Inventory for Men (SHIM): A 5-year review of research and clinical experience. Int J Impot Res 17:307, 2005.

McVary KT: Clinical practice. Erectile dysfunction. N Engl J Med 357:2472, 2007.

Parish SJ, Hahn SR: Hypoactive sexual desire disorder: A review of epidemiology, biopsychology, diagnosis, and treatment. Sex Med Rev 4:103, 2016.

391 Women's Health

Andrea Dunaif

The clinical discipline of women's health is well established. Indeed, its emphasis on greater attention to patient education and medical decision-making is a paradigm for what has become known as patient-centered health care. Moreover, the recognition of sex differences in disease processes and health outcomes is an important example of precision medicine. Sex difference refers to the biologic differences conferred by sex chromosomes and hormones. In contrast, gender differences are related to psychosocial roles and cultural expectations. The study of sex differences continues to grow as a scientific discipline. In 2016, the National Institutes of Health recognized its importance by implementing the expectation that sex should be considered as a biologic variable in study designs, analyses, and reporting in not only human but also vertebrate animal research. Strong scientific justification must be provided to limit research to only one sex.

DISEASE RISK: REALITY AND PERCEPTION

The leading causes of death are the same in women and men: (1) heart disease and (2) cancer **(Fig. 391-1)**. The leading cause of cancer death, lung cancer, is the same in both sexes. Breast cancer is the second leading cause of cancer death in women, but it causes about 60% fewer deaths than does lung cancer. Men are more likely than women to die from suicide and accidents.

Women's risk for many diseases increases at menopause. The median age of menopause in Caucasian women from industrialized countries is between 50 and 52 years, where women spend one-third of their lives in the postmenopausal period. Menopause occurs at earlier ages in Hispanic and African-American women as well as in women of lower socioeconomic status. Estrogen levels fall abruptly at menopause, inducing a variety of physiologic and metabolic responses. Rates of cardiovascular disease (CVD) increase and bone density begins to decrease rapidly after menopause.

In the United States, women live on average 4.8 years longer than men, with a life expectancy at birth in 2014 of 81.2 years in women compared with 76.4 years in men of all races. Life expectancy was lower in African Americans of both sexes and higher in Hispanics of both sexes than their Caucasian counterparts. The difference in life expectancy between men and women has decreased since its peak of 7.8 years in 1979 but has remained unchanged at 4.8 years since 2010. Accordingly, elderly women outnumber elderly men, so that age-related conditions, such as hypertension, have a female preponderance.

Public awareness campaigns have resulted in a marked increase in the percentage of U.S. women knowing that CVD is the leading cause of death in women. In 1997, the majority of U.S. women surveyed thought that cancer (35%) rather than heart disease (30%) was the leading cause of death in women **(Fig. 391-2)**. In 2012, these perceptions were reversed, with 56% of U.S. women surveyed recognizing that heart disease rather than cancer (24%) was the leading cause of death in women (Fig. 391-2). Although awareness of heart disease has improved substantially among black and Hispanic women over this time period,

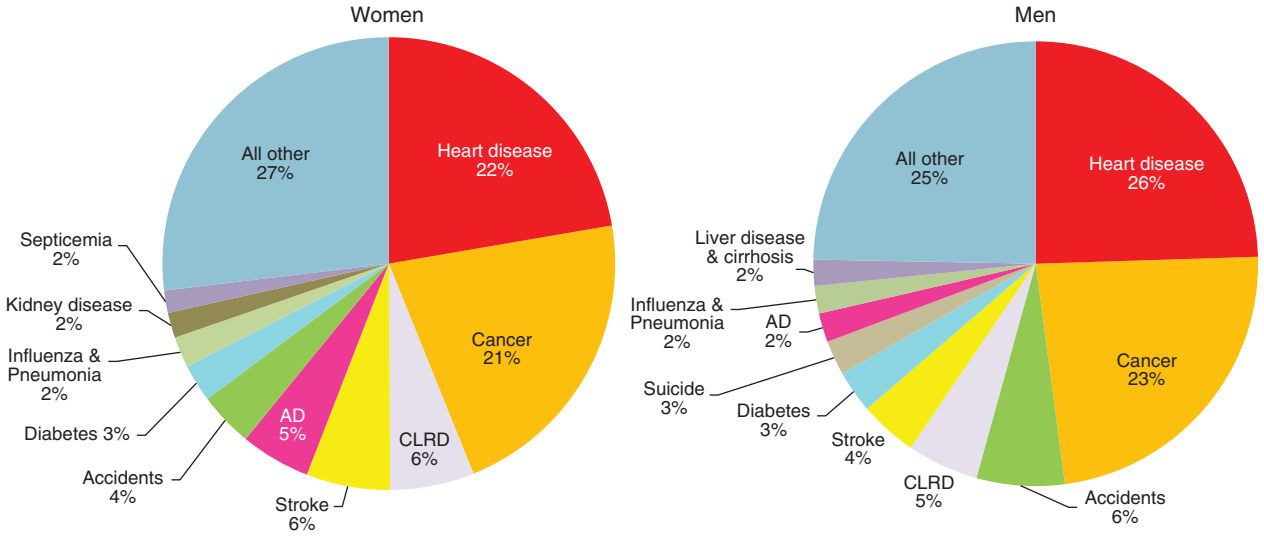


FIGURE 391-1 Percent distribution of 10 leading causes of death in women compared to men in the United States in 2014. In both women and men, the first and second leading causes of death are the same, heart disease and cancer, respectively. Causes of death then diverge by sex. For example, accidents are the third leading cause of death in men but the sixth leading cause of death in women. Stroke, chronic lower respiratory disease (CLRD) and Alzheimer's disease (AD) cause a larger percentage of deaths in women than in men. Suicide is among the 10 leading causes of death in men but not in women. *(Data from https://www.cdc.gov/nchs/data/nvsr/nvsr65/nvsr65_04.pdf.)*

these groups were 66% less likely than white women to recognize that heart disease is the leading cause of death in women.

Nevertheless, women aged <65 years still consider breast cancer to be their leading health risk, despite the fact that death rates from breast cancer have been falling since the 1990s. In 2011, 1 in 30.8 deaths in women was due to breast cancer, whereas 1 in 7.5 deaths was due to CVD. Although a woman's lifetime risk of developing breast cancer if she lives past 85 years is about 1 in 9, it is much more likely that she will die from CVD than from breast cancer. In other words, many elderly women have breast cancer but die from other causes. Similarly, a minority of women are aware that lung cancer is the leading cause of cancer death in women. Physicians are also less likely to recognize women's risk for CVD. Even in 2012, only 21% of U.S. women surveyed reported that their physicians had counseled them about their risk for heart disease. These misconceptions are unfortunate as they perpetuate inadequate attention to modifiable risk factors such as dyslipidemia, hypertension, and cigarette smoking.

SEX DIFFERENCES IN HEALTH AND DISEASE

ALZHEIMER'S DISEASE

(See also Chap. 423) Alzheimer's disease (AD) affects approximately twice as many women as men. Because the risk for AD increases with age, part of this sex difference is accounted for by the fact that women live longer than men. However, additional factors probably contribute to the increased risk for AD in women, including sex differences in brain size, structure, and functional organization. There is emerging evidence for sex-specific differences in gene expression, not only for genes on the X and Y chromosomes but also for some autosomal genes. Estrogens have pleiotropic genomic and nongenomic effects on the central nervous system, including neurotrophic actions in key areas involved in cognition and memory. There are sex differences in the severity of AD with women experiencing greater deficits in cognition.

Women with AD have lower endogenous estrogen levels than do women without AD. These observations have led to the hypothesis that estrogen is neuroprotective. Some studies have suggested that estrogen administration improves cognitive function in nondemented postmenopausal women as well as in women with AD, and several observational studies have suggested that postmenopausal hormone therapy (HT) may decrease the risk of AD. The Women's Health Initiative Memory Study (WHIMS), an ancillary study in the Women's Health Initiative (WHI) in women aged ≥65 years, found significantly increased risk for both dementia and mild cognitive impairment in women receiving estrogen alone (combined continuous equine estrogen [CEE], 0.625 mg daily) or estrogen with progestin (CEE, 0.625 mg daily, and medroxyprogesterone acetate [MPA], 2.5 mg daily) compared to placebo. However, the Kronos Early Estrogen Prevention Study (KEEPS), a randomized clinical trial of early initiation of HT after menopause that compared CEE 0.45 mg daily, 50 μg of weekly transdermal estradiol (both estrogen arms included cyclic oral micronized progesterone 200 mg daily for 12 days each month), or placebo, found no adverse effects of HT on cognitive function. In summary, there is

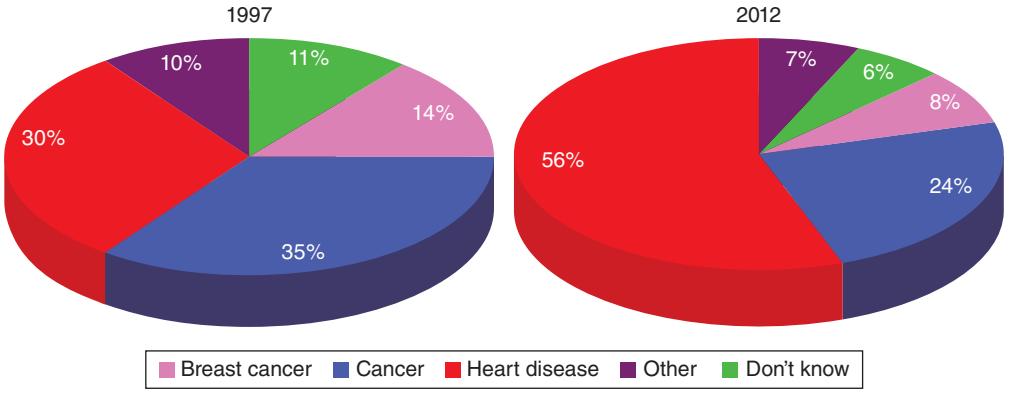


FIGURE 391-2 Changes in perceived leading causes of death among women surveyed in 1997 compared with those surveyed in 2012. In 1997, cancer was cited as the leading cause of death in women, not heart disease. In 2012, this trend had reversed. The rate of awareness that heart disease is the leading cause of death in women was significantly higher in 2012 (56% vs 30%, $p < .001$) than in 1997. *(Data adapted from L Mosca et al: Circulation 127:1254, 2013.)*

no evidence from placebo-controlled trials that HT improves cognitive function.

CVD AND STROKE

(See also Chap. 267) There are major sex differences in CVD, the leading cause of death in men and women in developed countries. However, there are also major gender differences because of perceptions by both women and their health care providers that women are at lower risk for CVD. As a result of these misconceptions, women are less likely to seek medical help when they experience symptoms of CVD. Health care providers are less likely to suspect CVD, so women receive fewer interventions for modifiable risk factors as well as fewer acute interventions than do men. Women and their health care providers are also less aware that prodromal symptoms of cardiac disease differ in women compared to men. Women are less likely than men to present with chest pain and more likely to present with fatigue, shortness of breath, indigestion/nausea, and anxiety.

Sex steroids have major effects on the cardiovascular system and lipid metabolism. Estrogen increases high-density lipoprotein (HDL) and lowers low-density lipoprotein (LDL), whereas androgens have the opposite effect. Estrogen has direct vasodilatory effects on the vascular endothelium, enhances insulin sensitivity, and has antioxidant and anti-inflammatory properties. There is a striking increase in CVD after both natural and surgical menopause, suggesting that endogenous estrogens are cardioprotective. Women also have longer QT intervals on electrocardiograms, and this increases their susceptibility to certain arrhythmias.

CVD presents differently in women, who are usually 10–15 years older than their male counterparts and are more likely to have comorbidities such as hypertension, congestive heart failure, and diabetes mellitus (DM). In the Framingham study, angina was the most common initial symptom of CVD in women, whereas myocardial infarction (MI) was the most common initial presentation in men. Women more often have atypical symptoms such as fatigue, anxiety, nausea, indigestion, and upper back pain. Although awareness that heart disease is the leading cause of death in women has nearly doubled over the last 15 years, women remain less aware that its symptoms are often atypical, and they are less likely to contact 9-1-1 when they experience such symptoms. The recent availability of a high-specificity troponin assay with sex-specific cutoffs has increased diagnostic accuracy for MI in women but not in men.

Deaths from CVD have decreased markedly in men since 1980, whereas CVD deaths only began to decrease substantially in women beginning in 2000. Women with MI are more likely to present with cardiac arrest or cardiogenic shock, whereas men are more likely to present with ventricular tachycardia. Further, younger women with MI are more likely to die than are men of similar age. However, this mortality gap has decreased in recent years because younger women have experienced greater improvements in survival after MI than men (Fig. 391-3). The improvement in survival is due largely to a reduction in comorbidities, suggesting a greater attention to modifiable risk factors in women. Nevertheless, 1 year after MI, 26% of women aged >45 years will die compared to 19% of men. Within 5 years of the first MI, 47% of women compared to 36% of men will die.

Physicians are less likely to suspect heart disease in women with chest pain and less likely to perform diagnostic and therapeutic cardiac procedures in women. Women are less likely to receive therapies such as angioplasty, thrombolytic therapy, coronary artery bypass grafts (CABGs), beta blockers, and aspirin. There are also sex differences in outcomes when women with CVD do receive therapeutic interventions. Women undergoing CABG surgery have more advanced disease, a higher perioperative mortality rate, less relief of angina, and less graft patency; however, 5- and 10-year survival rates are similar. Women undergoing percutaneous transluminal coronary angioplasty have lower rates of initial angiographic and clinical success than men, but they also have a lower rate of restenosis and a better long-term outcome. Women may benefit less and have more frequent serious bleeding complications from thrombolytic therapy compared with men. Factors such as older age, more comorbid conditions, smaller body size, and more severe CVD in women at the time of events or procedures account in part for the observed sex differences.

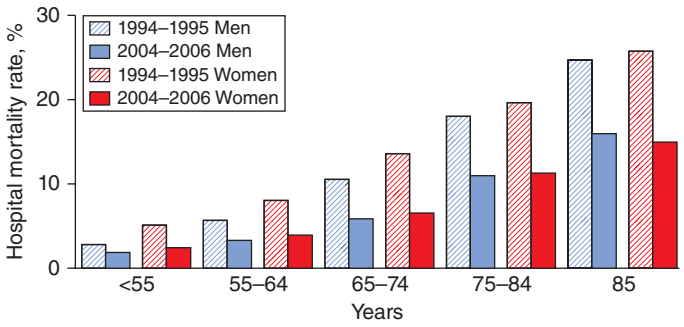


FIGURE 391-3 Hospital mortality rates in men and women for acute myocardial infarction (MI) in 1994–1995 compared with 2004–2006. Women aged <65 years had substantially greater mortality than men of similar age in 1994–1995. Mortality rates declined markedly for both sexes across all age groups in 2004–2006 compared with 1994–1995. However, there was a more striking decrease in mortality in women aged <75 years compared with men of similar age. The mortality rate reduction was largest in women <55 years (52.9%) and lowest in men of similar age (33.3%). *(Data adapted from V Vaccarino et al: Arch Intern Med 169:1767, 2009.)*

Elevated cholesterol levels, hypertension, smoking, obesity, low HDL cholesterol levels, DM, and lack of physical activity are important risk factors for CVD in both men and women. Total triglyceride levels are an independent risk factor for CVD in women but not in men. Low HDL cholesterol and DM are more important risk factors for CVD in women than in men. Smoking is an important risk factor for CVD in women—it accelerates atherosclerosis, exerts direct negative effects on cardiac function, and is associated with an earlier age of menopause. Several disorders affect women exclusively, including pregnancy-associated hypertension, preeclampsia, gestational DM, polycystic ovary syndrome, or predominantly, including rheumatoid arthritis (RA) and systemic lupus erythematosus (SLE). Cholesterol-lowering drugs are equally effective in men and women for primary and secondary prevention of CVD. In contrast to men, randomized trials showed that aspirin was not effective in the primary prevention of CVD in women; it did significantly reduce the risk of ischemic stroke.

The sex differences in CVD prevalence, beneficial biologic effects of estrogen on the cardiovascular system, and reduced risk for CVD in observational studies led to the hypothesis that HT was cardioprotective. However, the WHI, which studied >16,000 women on CEE plus MPA or placebo and >10,000 women with hysterectomy on CEE alone or placebo, did not demonstrate a benefit of HT for the primary or secondary prevention of CVD. In addition, CEE plus MPA was associated with an increased risk for CVD, particularly in the first year of therapy, whereas CEE alone neither increased nor decreased CVD risk. Both CEE plus MPA and CEE alone were associated with an increased risk for ischemic stroke.

In the WHI, there was a suggestion of a reduction in CVD risk in women who initiated HT closer to menopause. This finding suggests that the time at which HT is initiated is critical for cardioprotection. According to this "timing" hypothesis, HT has differential effects, depending on the stage of atherosclerosis; adverse effects are seen with advanced, unstable lesions. This hypothesis was supported by data from the Danish Osteoporosis Prevention Study (DOPS), an open-label randomized trial of triphasic oral estradiol compared with no treatment in recently menopausal or perimenopausal women (a cyclic oral synthetic progestin, norethisterone acetate, was added in women who had a uterus), that found significantly reduced mortality and CVD after 10 years of HT. However, DOPS was designed to investigate HT for the primary prevention of osteoporotic bone fractures, and CVD outcomes were not prespecified endpoints.

KEEPS was designed to directly test the "timing" hypothesis. Seven hundred twenty-seven recently menopausal women aged 42–58 years (mean 52.7 years) were randomized to oral CEE (lower dose than WHI), transdermal estradiol, or placebo for 4 years; both estrogen arms included oral cyclical micronized progesterone (see above section

on AD for dosing details). There were no significant beneficial or deleterious effects on the progression of atherosclerosis by computed tomography assessment of coronary artery calcification in either HT arm. Adverse events including stroke, MI, venous thromboembolism, and breast cancer were not increased in the HT arms compared with the placebo arm. There were improvements in hot flashes, night sweats, mood, sexual function, and bone density in the HT arms. This relatively small study does not suggest that early HT administration reduces atherosclerosis. However, the study suggests that short-term HT may be safely administered for symptom relief in recently menopausal women. **HT is discussed further in Chap. 388.**

DIABETES MELLITUS

(See also Chap. 396) Women are more sensitive to insulin than men are. Despite this, the prevalence of type 2 DM is similar in men and women. There is a sex difference in the relationship between endogenous androgen levels and DM risk. Higher bioavailable testosterone levels are associated with increased risk in women, whereas lower bioavailable testosterone levels are associated with increased risk in men. Polycystic ovary syndrome, preeclampsia, pregnancy-associated hypertension, and gestational DM—common conditions in premenopausal women—are associated with a significantly increased risk for type 2 DM. Among individuals with DM, women have a greater risk for MI than do men. Women with DM have a sixfold greater risk of dying of CVD compared to women without DM.

Premenopausal women with DM lose the cardioprotective effect of female sex and have rates of CVD identical to those in males. These women have impaired endothelial function and reduced coronary vasodilatory responses, which may predispose to cardiovascular complications. Women with DM are more likely to have left ventricular hypertrophy. Women with DM receive less aggressive treatment for modifiable CVD risk factors than men with DM. In the WHI, CEE plus MPA significantly reduced the incidence of DM, whereas with CEE alone, there was only a trend toward decreased DM incidence.

HYPERTENSION

(See also Chap. 271) After age 60, hypertension is more common in U.S. women than in men, largely because of the high prevalence of hypertension in older age groups and the longer survival of women. Isolated systolic hypertension is present in 30% of women >60 years old. Sex hormones affect blood pressure. Both normotensive and hypertensive women have higher blood pressure levels during the follicular phase than during the luteal phase. In the Nurses' Health Study, the relative risk of hypertension was 1.8 in current users of oral contraceptives, but this risk is lower with the newer low-dose contraceptive preparations. HT is not associated with hypertension. Among secondary causes of hypertension, there is a female preponderance of renal artery fibromuscular dysplasia.

The benefits of treatment for hypertension have been dramatic in both women and men. A meta-analysis of the effects of hypertension treatment, the Individual Data Analysis of Antihypertensive Intervention Trial, found a reduction of risk for stroke and for major cardiovascular events in women. The effectiveness of various antihypertensive drugs appears to be comparable in women and men; however, women may experience more side effects. For example, women are more likely to develop cough with angiotensin-converting enzyme inhibitors.

AUTOIMMUNE DISORDERS

(See also Chap. 348) Most autoimmune disorders occur more commonly in women than in men; they include autoimmune thyroid and liver diseases, SLE, RA, scleroderma, multiple sclerosis (MS), and idiopathic thrombocytopenic purpura. However, there is no sex difference in the incidence of type 1 DM, and ankylosing spondylitis occurs more commonly in men. Women may be more resistant to bacterial infections than men. Sex differences in both immune responses and adverse reactions to vaccines have been reported. For example, there is a female preponderance of postvaccination arthritis.

Adaptive immune responses are more robust in women than in men; this may be explained by the stimulatory actions of estrogens and the inhibitory actions of androgens on the cellular mediators of immunity. Consistent with an important role for sex hormones, there is variation in immune responses during the menstrual cycle, and the activity of certain autoimmune disorders is altered by castration or pregnancy (e.g., RA and MS may remit during pregnancy). Nevertheless, the majority of studies show that exogenous estrogens and progestins in the form of HT or oral contraceptives do not alter autoimmune disease incidence or activity. Exposure to fetal antigens, including circulating fetal cells that persist in certain tissues, has been speculated to increase the risk of autoimmune responses. There is clearly an important genetic component to autoimmunity, as indicated by the familial clustering and HLA association of many such disorders. X chromosome genes also contribute to sex differences in immunity. Indeed, nonrandom X chromosome inactivation may be a risk factor for autoimmune diseases.

HIV INFECTION

(See also Chap. 197) Women accounted for almost 19% of the ~40,000 new HIV diagnoses in the United States in 2015. Of these newly diagnosed women, 61% were African American, 19% were Caucasian, and 15% were Hispanic. Annual HIV diagnoses declined by 20% among women from 2010 to 2014. Nevertheless, AIDS remains an important cause of death in younger women, particularly African-American women aged 25–44 years. Heterosexual contact with an at-risk partner is the fastest-growing transmission category, and women are more susceptible to HIV infection during vaginal sex than men. This increased susceptibility is accounted for in part by an increased prevalence of sexually transmitted diseases, i.e., gonorrhea and syphilis, in women.

Some studies have suggested that hormonal contraceptives may increase the risk of HIV transmission. Progesterone has been shown to increase susceptibility to infection in nonhuman primate models of HIV. Women are also more likely to be infected by multiple variants of the virus than men. Women with HIV have more rapid decreases in their CD4 cell counts than do men. Compared with men, HIV-infected women more frequently develop candidiasis, but Kaposi's sarcoma is less common than it is in men. Women have more adverse reactions, such as lipodystrophy, dyslipidemia, and rash, with antiretroviral therapy than do men. This observation is explained in part by sex differences in the pharmacokinetics of certain antiretroviral drugs, resulting in higher plasma concentrations in women.

OBESITY

(See also Chap. 395) The prevalence of both obesity (body mass index [BMI] ≥30 kg/m^2) and abdominal obesity (waist circumference ≥88 cm in women) is higher in U.S. women than in men. Between 2005 and 2014, the prevalence of obesity (BMI ≥30 kg/m^2) and class 3 obesity (BMI ≥40 kg/m^2) increased significantly in women but not in men. In 2014, the prevalence of obesity was 40.4% and class 3 obesity 9.9% for women, and 35.0% and class 3 obesity 5.5% for men. The prevalence of abdominal obesity increased over this time period in both sexes. More than 80% of patients who undergo bariatric surgery are women. Pregnancy and menopause are risk factors for obesity.

There are major sex differences in body fat distribution. Women characteristically have gluteal and femoral or gynoid pattern of fat distribution, whereas men typically have a central or android pattern. Women have more subcutaneous fat than men. In women, endogenous androgen levels are positively associated with abdominal obesity, and androgen administration increases visceral fat. In contrast, there is an inverse relationship between endogenous androgen levels and abdominal obesity in men. Further, androgen administration decreases visceral fat in these obese men. The reasons for these sex differences in the relationship between visceral fat and androgens are unknown. Studies in humans also suggest that sex steroids play a role in modulating food intake and energy expenditure.

In men and women, abdominal obesity characterized by increased visceral fat is associated with an increased risk for CVD and DM. Obesity increases a woman's risk for certain cancers, in particular postmenopausal breast and endometrial cancer, in part because adipose tissue provides an extragonadal source of estrogen through aromatization

of circulating adrenal and ovarian androgens, especially the conversion of androstenedione to estrone. Obesity increases the risk of infertility, miscarriage, and complications of pregnancy.

OSTEOPOROSIS

(See also Chap. 404) Osteoporosis is about five times more common in postmenopausal women than in age-matched men, and osteoporotic hip fractures are a major cause of morbidity in elderly women. Men accumulate more bone mass and lose bone more slowly than do women. Sex differences in bone mass are found as early as infancy. Calcium intake, vitamin D, and estrogen all play important roles in bone formation and bone loss. Particularly during adolescence, calcium intake is an important determinant of peak bone mass. Vitamin D deficiency is surprisingly common in elderly women, occurring in >40% of women living in northern latitudes. Receptors for estrogens and androgens have been identified in bone. Estrogen deficiency is associated with increased osteoclast activity and a decreased number of bone-forming units, leading to net bone loss. The aromatase enzyme, which converts androgens to estrogens, is also present in bone. Estrogen is an important determinant of bone mass in men (derived from the aromatization of androgens) as well as in women.

PHARMACOLOGY

On average, women have lower body weights, smaller organs, a higher percentage of body fat, and lower total-body water than men. There are also important sex differences in drug action and metabolism that are not accounted for by these differences in body size and composition. Sex steroids alter the binding and metabolism of a number of drugs. Further, menstrual cycle phase and pregnancy can alter drug action. Women also take more medications than men, including over-the-counter formulations and supplements. The greater use of medications combined with these biologic differences may account for the reported higher frequency of adverse drug reactions in women than in men.

Two-thirds of cases of drug-induced torsades des pointes, a rare, life-threatening ventricular arrhythmia, occur in women because they have a longer, more vulnerable QT interval. These drugs, which include certain antihistamines, antibiotics, antiarrhythmics, and antipsychotics, can prolong cardiac repolarization by blocking cardiac voltage-gated potassium channels. Women require lower doses of neuroleptics to control schizophrenia. Women awaken from anesthesia faster than do men given the same doses of anesthetics. In 2013, the Food and Drug Administration recommended that doses of the drug zolpidem be lowered for women because of slower drug clearance than in men.

PSYCHOLOGICAL DISORDERS

(See also Chap. 444) Depression, anxiety, and affective and eating disorders (bulimia and anorexia nervosa) are more common in women than in men. Epidemiologic studies from both developed and developing nations consistently find major depression to be twice as common in women as in men, with the sex difference becoming evident in early adolescence. Depression occurs in 10% of women during pregnancy and in 10–15% of women during the postpartum period. There is a high likelihood of recurrence of postpartum depression with subsequent pregnancies. The incidence of major depression diminishes after the age of 45 years and does not increase with the onset of menopause. Depression in women appears to have a worse prognosis than does depression in men; episodes last longer, and there is a lower rate of spontaneous remission. Schizophrenia and bipolar disorders occur at equal rates in men and women, although there may be sex differences in symptoms.

Both biologic and social factors account for the greater prevalence of depressive disorders in women. Men have higher levels of the neurotransmitter serotonin. Sex steroids also affect mood, and fluctuations during the menstrual cycle have been linked to symptoms of premenstrual syndrome. Sex hormones differentially affect the hypothalamic-pituitary-adrenal responses to stress. Testosterone appears to blunt cortisol responses to corticotropin-releasing hormone. Both low and high levels of estrogen can activate the hypothalamic-pituitary-adrenal axis.

SLEEP DISORDERS

(See also Chap. 27) There are striking sex differences in sleep and its disorders. During sleep, women have an increased amount of slow-wave activity, differences in timing of delta activity, and an increase in the number of sleep spindles. Testosterone modulates neural control of breathing and upper airway mechanics. Men have a higher prevalence of sleep apnea. Testosterone administration to hypogonadal men as well as to women increases apneic episodes during sleep. Women with the hyperandrogenic disorder polycystic ovary syndrome have an increased prevalence of obstructive sleep apnea, and apneic episodes are positively correlated with their circulating testosterone levels. In contrast, progesterone accelerates breathing, and in the past, progestins were used for treatment of sleep apnea.

SUBSTANCE ABUSE AND TOBACCO

(See also Chaps. 445 and 448) Substance abuse is more common in men than in women. However, one-third of Americans who suffer from alcoholism are women. Women alcoholics are less likely to be diagnosed than men. A greater proportion of men than women seek help for alcohol and drug abuse. Men are more likely to go to an alcohol or drug treatment facility, whereas women tend to approach a primary care physician or mental health professional for help under the guise of a psychosocial problem. Late-life alcoholism is more common in women than in men. On average, alcoholic women drink less than alcoholic men but exhibit the same degree of impairment. Blood alcohol levels are higher in women than in men after drinking equivalent amounts of alcohol, adjusted for body weight. This greater bioavailability of alcohol in women is due to both the smaller volume of distribution and the slower gastric metabolism of alcohol secondary to lower activity of gastric alcohol dehydrogenase than is the case in men. In addition, alcoholic women are more likely to abuse tranquilizers, sedatives, and amphetamines. Women alcoholics have a higher mortality rate than do nonalcoholic women and alcoholic men. Women also appear to develop alcoholic liver disease and other alcohol-related diseases with shorter drinking histories and lower levels of alcohol consumption. Alcohol abuse also poses special risks to a woman, adversely affecting fertility and the health of the baby (fetal alcohol syndrome). Even moderate alcohol use increases the risk of breast cancer, hypertension, and stroke in women.

More men than women smoke tobacco, but this sex difference continues to decrease. Women have a much larger burden of smoking-related disease. Smoking markedly increases the risk of CVD in premenopausal women and is also associated with a decrease in the age of menopause. Women who smoke are more likely to develop chronic obstructive pulmonary disease and lung cancer than men and at lower levels of tobacco exposure. Postmenopausal women who smoke have lower bone density than women who never smoked. Smoking during pregnancy increases the risk of preterm deliveries and low birth weight infants.

VIOLENCE AGAINST WOMEN

More than one in three women in the United States have experienced rape, physical violence, and/or stalking by an intimate partner. Adult women are much more likely to be raped by a spouse, ex-spouse, or acquaintance than by a stranger. Domestic or intimate partner violence is a leading cause of death among young women. Domestic violence may be an unrecognized feature of certain clinical presentations, such as chronic abdominal pain, headaches, and eating disorders, in addition to more obvious manifestations such as trauma. Intimate partner violence is an important risk factor for depression, substance abuse, and suicide in women. Screening instruments can accurately identify women experiencing intimate partner violence. Such screening by health care providers is acceptable to women in settings ensuring adequate privacy and safety.

SUMMARY

Women's health is now a mature discipline, and the importance of sex differences in biologic processes is well recognized. Nevertheless, ongoing misperceptions about disease risk, not only among women but

also among their health care providers, result in inadequate attention to modifiable risk factors. Research into the fundamental mechanisms of sex differences will provide important biologic insights. Further, those insights will have an impact on both women's and men's health.

FURTHER READING

Carter CL et al: Sex and gender differences in Alzheimer's disease: Recommendations for future research. J Womens Health 21:1018, 2012.

Clayton JA, Collins FS: Policy: NIH to balance sex in cell and animal studies. Nature 509:282, 2014.

Flegal KM et al: Trends in obesity among adults in the United States, 2005 to 2014. JAMA 315:2284, 2016.

How sex/gender influence health & disease (A-Z) *http://orwh.od.nih.gov/resources/sexgenderhealth/*.

Mcsweeney JC et al: Preventing and experiencing ischemic heart disease as a woman: State of the science: A scientific statement from the American Heart Association. Circulation 133:1302, 2016.

Ritz SA et al: First steps for integrating sex and gender considerations into basic experimental biomedical research. FASEB J 28:4, 2014.

Siegel RL et al: Cancer statistics, 2017. CA Cancer J Clin 67:7, 2017.

392 Men's Health

Shalender Bhasin, Shehzad Basaria

The emergence of men's health as a distinct discipline within internal medicine is founded on the wide consensus that men and women differ across their lifespan in their susceptibility to disease, in the clinical manifestations of the disease, and in their response to treatment. Furthermore, men and women weigh the health consequences of illness differently and have different motivation for seeking care. Men and women experience different types of disparities in access to healthcare services, and in the manner in which health care is delivered to them because of a complex array of socioeconomic and cultural factors. Attitudinal and institutional barriers to accessing care, fear, and embarrassment due to the perception that it is not manly to seek medical help, and reticence on the part of patients and physicians in discussing issues related to sexuality, drug use, and aging have heightened the need for programs tailored to address the specific health needs of men.

The sex differences in disease prevalence, susceptibility, and clinical manifestations of the disease were discussed in **Chap. 391** (Women's Health) and will not be discussed here. It is notable that the two leading causes of death in both men and women—heart disease and cancer—are the same. However, men have higher prevalence of neurodevelopmental and degenerative disorders, substance abuse disorders, including the use of performance enhancing drugs and alcohol dependence, diabetes, and cardiovascular disease, and women have higher prevalence of autoimmune disorders, depression, rheumatologic disorders, and osteoporosis, The men are substantially more likely to die from accidents, suicides, and homicides than women. Among men, 15–34 years of age, unintentional injuries, homicides, and suicides account for over three-fourths of all deaths. Among men, 35–64 years of age, heart disease, cancer, and unintentional injuries are the leading causes of death. Among men ≥65 years of age, heart disease, cancer, lower respiratory tract infections, and stroke are the major causes of death.

The biologic bases of sex differences in disease susceptibility, progression, and manifestation remain incompletely understood, and are likely multifactorial. Undoubtedly, sex-specific differences in the genetic architecture and circulating sex hormones influence disease phenotype; additionally, epigenetic effects of sex hormones during fetal life, early childhood, and during pubertal development may epigenetically imprint sexual and nonsexual behaviors, body composition, and disease susceptibility. The circulating and tissue concentrations of sex hormones differ substantially in men and women, and these hormonal differences may affect gene expression in cells of males and females in all parts of the body. The presence of only one X chromosome in men renders them more susceptible to X-linked disorders than women. Due to the X inactivation of one randomly chosen X chromosome, women's bodies contain two epigenetically different cell populations. The genes that do not undergo X inactivation exhibit dosage differences between male and female cells. Expression of the Y chromosome genes in men may affect the function of somatic cells containing the Y chromosome. The differences in the imprinting of maternally and paternally derived genes may also contribute to sex differences in the expression of disease. Reproductive load and physiologic changes during pregnancy, including profound hormonal and metabolic shifts, and microchimerism (transfer of cells from the mother to the fetus and from the fetus to the mother) may affect disease susceptibility and disease severity in women. Sociocultural norms of child-rearing practices, societal expectations of gender roles, and the long-term economic impact of these practices and gender roles influence health behaviors and disease risk. Furthermore, the trajectories of age-related changes in sex hormones during the reproductive and postreproductive years vary substantially between men and women, and influence the sex-specific patterns of the temporal evolution of age-related conditions such as osteoporosis, breast cancer, and autoimmune disease.

In a reflection of the growing attention on issues related to men's health, men's health clinics have mushroomed all over the country. Although the major threats to men's health have not changed—heart disease, cancer, and unintentional injury continue to dominate the list of major medical causes of morbidity and mortality in men—the men who attend men's health clinics do so largely for sexual, reproductive, and urologic health concerns involving common conditions, such as androgen deficiency syndromes, age-related decline in testosterone levels, sexual dysfunction, muscle dysmorphia and anabolic-androgenic steroid (AAS) use, lower urinary tract symptoms (LUTS), and medical complications of prostate cancer therapy, which are the subjects of this chapter. Additionally, we are witnessing the emergence of new categories of body image disorders in men that had not been recognized until the 1980s, such as the body dysmorphia syndrome and the use of performance enhancing drugs to increase muscularity and lean appearance. Although menopause has been the subject of intense investigation for more than five decades, these issues that are specific to men's health are just beginning to gain attention that they deserve because of their high prevalence and impact on overall health, well-being, and quality of life.

AGING-RELATED CHANGES IN MALE REPRODUCTIVE FUNCTION

A number of cross-sectional and longitudinal studies (e.g., the Baltimore Longitudinal Study of Aging, the Framingham Heart Study [FHS], the Massachusetts Male Aging Study, and the European Male Aging Study [EMAS]) have established that testosterone concentrations decrease with advancing age. This age-related decline starts in the third decade of life and progresses slowly **(Fig.392-1)**; the rate of decline in testosterone concentrations is greater in obese men, in men with chronic illness and in those taking medications than in healthy older men. Because sex-hormone binding globulin (SHBG) concentrations are higher in older men than in younger men, free or bioavailable testosterone concentrations decline with aging to a greater extent than total testosterone concentrations. The age-related decline in testosterone is due to defects at all levels of the hypothalamic-pituitary-testicular (HPT) axis: pulsatile gonadotropin-releasing hormone (GnRH) secretion is attenuated, luteinizing hormone (LH) response to GnRH is reduced, and testicular response to LH is impaired. However, the gradual rise of LH with aging suggests that testis dysfunction is the main cause of declining androgen levels. The term *andropause* has been used to denote age-related decline in testosterone concentrations; this term is a misnomer because there is no discrete time when testosterone concentrations decline abruptly.

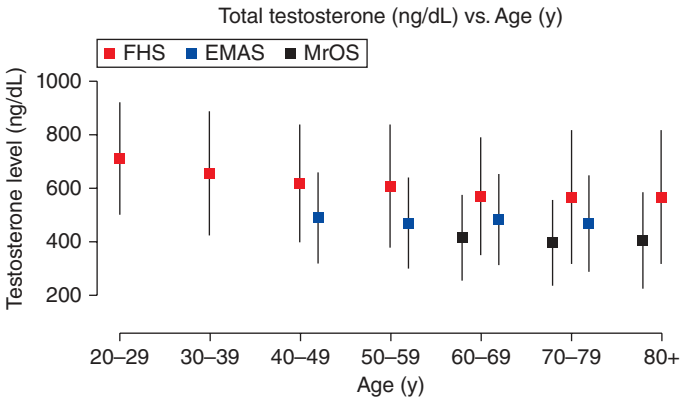


FIGURE 392-1 Age-related decline in total testosterone levels. Total testosterone levels measured using liquid chromatography tandem mass spectrometry in men of the Framingham Heart Study (FHS), the European Male Aging Study (EMAS), and the Osteoporotic Fractures in Men Study (MrOS). *(Reproduced with permission from S Bhasin et al: J Clin Endocrinol Metab 96:2430, 2011.)*

In epidemiologic surveys, low total and bioavailable testosterone concentrations have been associated with decreased appendicular skeletal muscle mass and strength, decreased self-reported physical function, higher visceral fat mass, insulin resistance, and increased risk of coronary artery disease and mortality **(Table 392-1)**. An analysis of signs and symptoms in older men in the EMAS revealed a syndromic association of sexual symptoms with total testosterone levels <320 ng/dL and free testosterone levels <64 pg/mL in community-dwelling older men.

In systematic reviews of randomized controlled trials, testosterone therapy of healthy older men with low or low-normal testosterone levels was associated with greater increments in lean body mass, grip strength, and self-reported physical function than that associated with placebo **(Fig. 392-2)**. Testosterone therapy also induced greater improvement in vertebral but not femoral bone mineral density (BMD). Testosterone therapy of older men with sexual dysfunction and unequivocally low testosterone levels improves libido, but testosterone effects on erectile function and response to selective phosphodiesterase inhibitors have been inconsistent. Testosterone therapy has not been shown to improve depression scores, fracture risk, cognitive function, response to phosphodiesterase inhibitors, or clinical outcomes in older men. Furthermore, neither the long-term risks nor clinical benefits of testosterone therapy in older men have been demonstrated in adequately powered trials. While there is no evidence that testosterone causes prostate cancer, there is concern that testosterone therapy might cause subclinical prostate cancers to grow. Testosterone therapy is associated with increased risk of detection of prostate events **(Fig. 392-3)**.

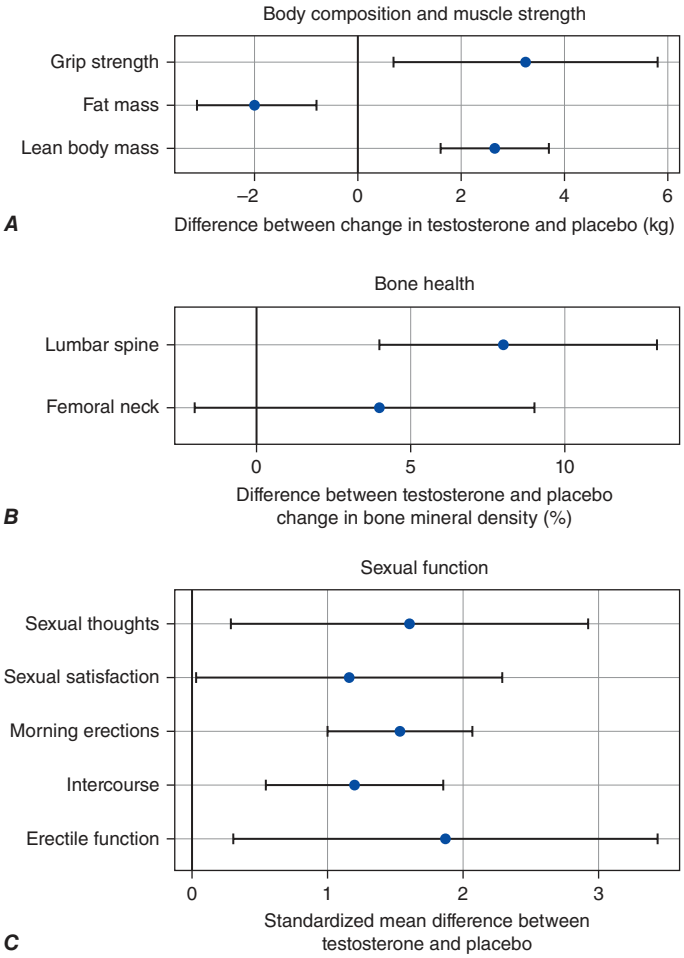


FIGURE 392-2 The effects of testosterone therapy on body composition, muscle strength, bone mineral density (BMD), and sexual function in intervention trials. The point estimates and the associated 95% confidence intervals are shown. ***A.*** The effects of testosterone therapy on lean body mass, grip strength, and fat mass in a meta-analysis of randomized trials. *(Data derived from S Bhasin et al: Nat Clin Pract Endocrinol Metab 2:146, 2006.)* ***B.*** The effects of testosterone therapy on lumbar and femoral BMD in a meta-analysis of randomized trials. *(Data derived from a meta-analysis by MJ Tracz et al: J Clin Endocrinol Metab 91:2011, 2006.)* ***C.*** The effects of testosterone therapy on measures of sexual function in men with baseline testosterone <10 nmol/L (290 ng/dL). *(Data derived from a meta-analysis by AM Isidori et al: Clin Endocrinol (Oxf) 63:381, 2005.) (Reproduced with permission from M Spitzer et al: Nat Rev Endocrinol 9:414, 2013.)*

TABLE 392-1 Association of Testosterone Levels with Outcomes in Older Men

1. Positively associated with:
   - Muscle mass and muscle strength
   - Physical function
   - Sexual desire
   - Bone mineral density, bone geometry, and volumetric bone mineral density
2. Negatively associated with:
   - Coronary artery disease
   - Visceral fat
   - Diabetes mellitus
   - Metabolic syndrome
   - Mortality
   - Falls and fracture risk
   - Frailty
3. Not associated with:
   - Lower urinary tract symptoms
   - Erectile dysfunction
   - Dementia
   - Major depression

One randomized testosterone trial in older men with mobility limitation and high burden of chronic conditions such as diabetes, heart disease, hypertension, and hyperlipidemia, reported a greater number of cardiovascular events in men randomized to the testosterone arm of the study than in those randomized assigned to the placebo arm. Since then, two large retrospective analyses of patient databases have reported higher frequency of cardiovascular events, including myocardial infarction in older men with preexisting heart disease. A meta-analysis of randomized testosterone trials in older men found increased risk of cardiovascular events in men assigned to testosterone arms of the trials (Fig. 392-3).

Population screening of all older men for low testosterone levels is not recommended; testing should be restricted to men who have symptoms or physical features attributable to androgen deficiency. Testosterone therapy is not recommended for all older men with low testosterone levels. In older men with significant symptoms of androgen deficiency who have consistently low testosterone levels, testosterone therapy may be considered on an individualized basis and should be instituted after careful discussion of the risks and benefits.

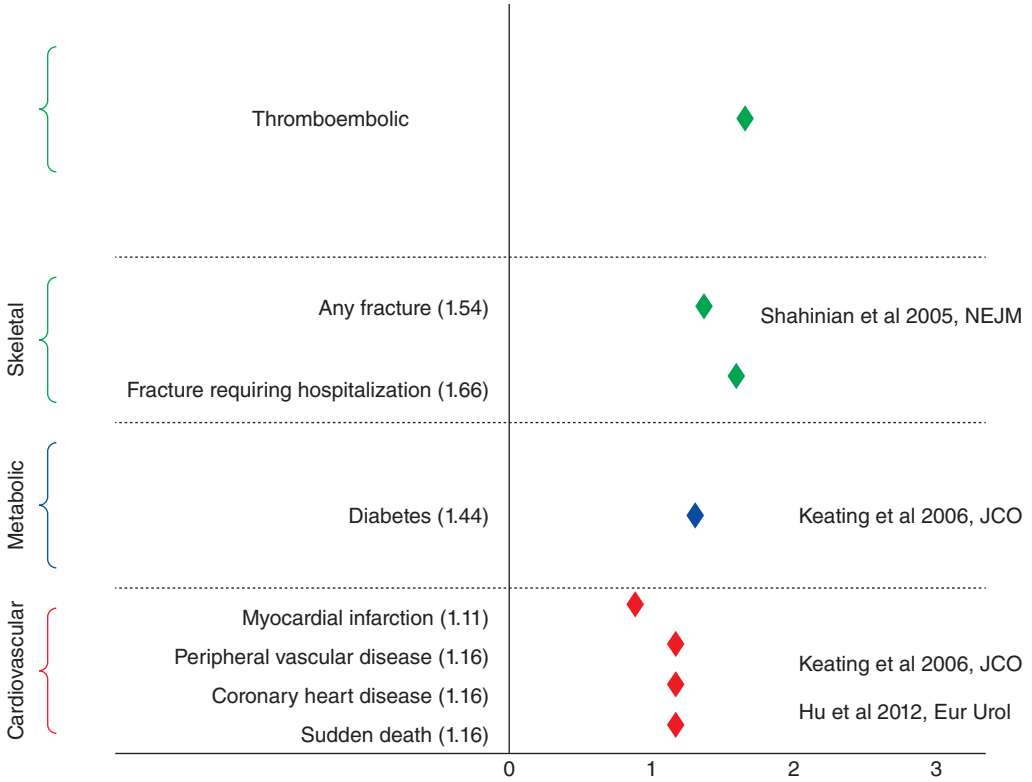


FIGURE 392-3 Adverse cardiometabolic and skeletal effects of androgen deprivation therapy (ADT) in men receiving ADT for prostate cancer. Administration of ADT has been associated with increased risk of thromboembolic events, fractures, and diabetes. Some, but not all, studies have reported increased risk of cardiovascular events in men receiving ADT. *(Data on relative risk were derived from VB Shahinian et al: N Engl J Med 352:154, 2005; NL Keating et al: J Clin Oncol 24:4448, 2006; and JC Hu et al: Eur Urol 61:1119, 2012.)*

Testicular morphology, semen production, and fertility are maintained up to a very old age in men. Although concern has been expressed about age-related increases in germ cell mutations and impairment of DNA repair mechanisms, there is no clear evidence that the frequency of chromosomal aneuploidy is increased in the sperm of older men. However, the incidence of autosomal dominant diseases, such as achondroplasia, polyposis coli, Marfan syndrome, and Apert's syndrome, increases in the offspring of men who are advanced in age, consistent with transmission of sporadic missense mutations. Advanced paternal age may be associated with increased rates of de novo mutations, which may contribute to an increased risk of neurodevelopmental diseases such as schizophrenia and autism. The somatic mutations in male germ cells that enhance the proliferation of germ cells could lead to within-testis expansion of mutant clonal lines, thus favoring the propagation of germ cells carrying these pathogenic mutations, and increasing the risk of mutations in the offspring of older fathers (the "selfish spermatogonial selection" hypothesis).

Sexual Dysfunction Various forms of sexual dysfunction are a major motivating factor for men seeking care at men's health clinics. The landmark descriptions of the human sexual response cycle by Masters and Johnson demonstrating that men and women display predictable physiologic responses after sexual stimulation provided the basis for rational classification of human sexual disorders. Accordingly, sexual disorders have been classified into four categories depending on phase of sexual response cycle in which the abnormality exists:

1. Hypoactive sexual desire disorder
2. Erectile dysfunction
3. Ejaculatory and orgasmic disorders
4. Disorders of pain

Classification of the patient's disorder into these categories is important as the etiologic factors, diagnostic tests, and the therapeutic strategies vary for each class of sexual disorder. Historically, the classification and nomenclature for sexual disorders were based on Diagnostic and Statistical Manual (DSM), based on the erroneous belief that sexual disorders in men are largely psychogenic in their origin. However, the recognition of erectile dysfunction as a manifestation of systemic disease and the availability of easy-to-use oral selective phosphodiesterase-5 (PDE5) inhibitors have placed sexual disorders in men within the purview of the primary care provider. These disorders have been discussed in Chap. 390 (Sexual Dysfunction).

MUSCLE DYSMORPHIA SYNDROME IN MEN—A FORM OF BODY IMAGE DISORDER

Muscle dysmorphia is a form of body image disorder characterized by a pathological preoccupation with muscularity and leanness. The men with muscle dysmorphia express a strong desire to be more muscular and lean. These men describe shame and embarrassment about their body size and shape and often report aversive symptoms such as dissatisfaction with appearance, preoccupation with bodybuilding and muscularity, and functional impairment. Patients with muscle dysmorphia also report higher rates of mood and anxiety disorders, and obsessive and compulsive behaviors than individuals with no history of muscle dysmorphia. These men often experience impairment of social and occupational functioning.

The patients with muscle dysmorphia syndrome—nearly all men—are almost always engaged in weightlifting and body building and are more likely to use performance enhancing drugs, especially AASs than men in the general population or even weightlifters without body dysmorphia. The muscle dysmorphia disorder renders men to an increased risk of disease due to the combined interactive effects of the intensity of physical exercise, the use of performance enhancing drugs, and other lifestyle factors associated with weightlifting and the use of performance enhancing drugs. These patients are also at increased risk of functioning poorly in their occupation and social life than men without this disorder. No randomized trials of any treatment modalities have been conducted; anecdotally, behavioral and cognitive therapies have been tried with varying degrees of success.

AAS Abuse by Athletes and Recreational Bodybuilders The illicit use of AASs to enhance athletic performance first surfaced

in the 1950s among powerlifters and spread rapidly to other sports, professional as well as high school athletes, and recreational bodybuilders. In the early 1980s, the use of AAS spreads beyond the athletic community into the general population. As many as 3 million Americans—most of them men—have likely used these compounds. Most AAS users are not athletes, but rather recreational weightlifters, who use these drugs to look lean and more muscular.

The most commonly used AASs include testosterone esters, nandrolone, stanozolol, methandienone, and methenolol. AAS users generally use increasing doses of multiple steroids in a practice known as stacking.

The adverse effects of long-term AAS abuse remain poorly understood. Most of the information about the adverse effects of AAS has emerged from case reports, uncontrolled studies, or from clinical trials that used replacement doses of testosterone **(Table 392-2)**. The adverse event data from clinical trials using physiologic replacement doses of testosterone have been extrapolated unjustifiably to AAS users who may administer 10–100 times the replacement doses of testosterone over many years and to support the claim that AAS use is safe. A substantial fraction of AAS users also use other drugs that are perceived to be muscle-building or performance-enhancing, such as growth hormone; erythropoiesis stimulating agents; insulin; stimulants such as amphetamine, clenbuterol, cocaine, ephedrine, and thyroxine; and drugs perceived to reduce adverse effects such as human chorionic gonadotropin (hCG), aromatase inhibitors, or estrogen antagonists. The men who abuse AAS are more likely to engage in other high-risk behaviors than nonusers. The adverse events associated with AAS use may be due to AAS themselves, concomitant use of other drugs, high-risk behaviors, and host characteristics that may render these individuals more susceptible to AAS use or to other high-risk behaviors.

TABLE 392-2 Potential Adverse Effects Associated with the Use of Anabolic-Androgenic Steroids (AASs)

ORGAN SYSTEM	EFFECT
Cardiovascular	Dyslipidemia
	Atherosclerotic disease
	Sudden death
	Myocardial fibrosis, cardiomyopathy
	Cardiac conduction abnormalities
	Hypertension
Neuroendocrine	HPT axis suppression
	Hypogonadism after AAS withdrawal
	Gynecomastia
Females	Virilizing effects
Neuropsychiatric	Major mood disorders (mania, hypomania, depression)
	Aggression, violence
	AAS dependence
	Neuronal apoptosis
	Cognitive deficits
Hematologic	Polycythemia
	Hypercoagulability and thrombosis
Hepatic	Inflammatory and cholestatic effects
	Peliosis hepatis (rare)
	Neoplasms (rare)
Musculoskeletal	Premature epiphyseal closure (in adolescents)
	Tendon rupture
Kidney	Renal failure secondary to rhabdomyolysis
	Focal segmental glomerulosclerosis
Dermatologic	Acne
	Striae

Abbreviation: HPT axis, hypothalamic-pituitary-testicular axis.

Source: Modified with permission from HG Pope Jr et al: Adverse health consequences of performance-enhancing drugs: an endocrine society scientific statement. *Endocr Rev* 35:341, 2014.

The high rates of mortality and morbidities observed in AAS users are alarming. One Finnish study reported 4.6 times the risk of death among elite power lifters than in age-matched men from the general population. The causes of death among power lifters included suicides, myocardial infarction, hepatic coma, and non-Hodgkin's lymphoma. A retrospective review of patient records in Sweden also reported higher standardized mortality ratios for AAS users than for nonusers. Studies indicate that 32% of deaths among AAS users were suicidal, 26% homicidal, and 35% accidental. The median age of death among AAS users—24 years—is even lower than that for heroin or amphetamine users.

Numerous reports of cardiac death among young AAS users raise concerns about the adverse cardiovascular effects of AAS. High doses of AAS may induce proatherogenic dyslipidemia, increase thrombosis risk via effects on clotting factors and platelets, induce vasospasm through their effects on vascular nitric oxide, and induce myocardial hypertrophy and fibrosis.

Replacement doses of testosterone, when administered parenterally, are associated with only a small decrease in high-density lipoprotein (HDL) cholesterol and little or no effect on total cholesterol, low-density lipoprotein (LDL) cholesterol, and triglyceride levels. In contrast, supraphysiologic doses of testosterone and orally administered, 17-α-alkylated, nonaromatizable AAS are associated with marked reductions in HDL cholesterol and increases in LDL cholesterol.

Recent studies of AAS users using tissue Doppler and strain imaging and magnetic resonance imaging have reported diastolic and systolic dysfunction, including significantly lower early and late diastolic tissue velocities, reduced E/A ratio, and reduced peak systolic strain in AAS users than in nonusers. Power athletes using AAS often have short QT intervals but increased QT dispersion, which may predispose them to ventricular arrhythmias. Long-term AAS use may be associated with myocardial hypertrophy and fibrosis. Myocardial tissue of power lifters using AAS has been shown to be infiltrated with fibrous tissue and fat droplets. The finding of androgen receptors on myocardial cells suggests that AAS might be directly toxic to myocardial cells.

Long-term AAS use suppresses LH and follicle-stimulating hormone (FSH) secretion and inhibits endogenous testosterone production and spermatogenesis. Men who have used AAS for more than a few months experience marked suppression of the HPT axis after stopping AAS that may be associated with sexual dysfunction, fatigue, infertility, and depressive symptoms. In some AAS users, HPT suppression may last more than a year, and in a few individuals, complete recovery may not occur. The symptoms of androgen deficiency during AAS withdrawal may cause some men to revert back to using AAS, leading to continued use and AAS dependence. As many as 30% of AAS users develop a syndrome of AAS dependence, characterized by long-term AAS use despite adverse medical and psychiatric effects.

Supraphysiologic doses of testosterone may also impair insulin sensitivity. Orally administered androgens have been associated with insulin resistance and diabetes.

Unsafe injection practices, high-risk behaviors, and increase rates of incarceration render AAS users at increased risk of HIV, and hepatitis B and C. In one survey, nearly one in 10 gay men had injected AAS or other substances, and AAS users were more likely to report high-risk unprotected anal sex than other men.

Some AAS users develop hypomanic and manic symptoms during AAS exposure (irritability, aggressiveness, reckless behavior, and occasional psychotic symptoms, sometimes associated with violence) and major depression (sometimes associated with suicidality) during AAS withdrawal. Users may also develop other forms of illicit drug use, which may be potentiated or exacerbated by AAS.

Elevated liver enzymes, cholestatic jaundice, hepatic neoplasms, and peliosis hepatis have been reported with oral, 17-α-alkylated AAS. AAS use may cause muscle hypertrophy without compensatory adaptations in tendons, ligaments, and joints, thus increasing the risk of tendon and joint injuries. AAS use is associated with acne, baldness, as well as increased body hair.

TABLE 392-3 Detection of the Use of Anabolic-Androgenic Steroids

Clinical indicators that should raise suspicion of anabolic-androgenic steroid (AAS) use
Very muscular phenotype
Reduced testicular volume (<15 mL)
Laboratory indicators
Suppressed LH and FSH levels
Increased hematocrit
Detection of AASs
LC-MS/MS analysis of urine
Detection of exogenous testosterone use
Urinary testosterone-to-epitestosterone ratio
Isotope ratio mass spectrometry analysis to detect differences in the ^{13}C:^{12}C ratio in exogenous and endogenous testosterone

Abbreviations: FSH, follicle-stimulating hormone; LC-MS/MS, liquid chromatography and tandem mass spectrometry; LH, luteinizing hormone.

APPROACH TO THE PATIENT

Detection of AAS Use

AAS users generally mistrust physicians and seek medical help infrequently; when they do seek medical help, it is often for the treatment of AAS withdrawal syndrome, infertility, gynecomastia, or other medical or psychiatric complications of AAS use. The suspicion of AAS use should be raised by the increased hemoglobin and hematocrit levels, suppressed luteinizing hormone and follicle-stimulating hormone and testosterone levels, low high-density lipoproteins cholesterol, and low testicular volume and sperm density in a person who looks highly muscular **(Table 392-3)**. A combination of these findings along with self-report of their use by the patient—which usually can be elicited by a tactful interview—are often sufficient to establish a diagnosis in clinical practice.

Accredited laboratories use gas chromatography and mass spectrometry or liquid chromatography and mass spectrometry to detect AAS abuse. In recent years, the availability of high-resolution mass spectrometry and tandem mass spectrometry has further improved the sensitivity of detecting AAS abuse. Illicit testosterone use is detected generally by the application of the measurement of urinary testosterone to epitestosterone ratio and further confirmed by the use of the ^{13}C:^{12}C ratio in testosterone by the use of isotope ratio combustion mass spectrometry. Exogenous testosterone administration increases urinary testosterone glucuronide excretion and consequently the testosterone to epitestosterone ratio. Ratios >4 suggest exogenous testosterone use but can also reflect genetic variation. Genetic variations in the uridine diphospho-glucuronyl transferase 2B17 (*UGT2B17*), the major enzyme for testosterone glucuronidation, affect testosterone to epitestosterone ratio. Synthetic testosterone has a lower ^{13}C:^{12}C ratio than endogenously produced testosterone and these differences in the ^{13}C:^{12}C ratio can be detected by isotope ratio combustion mass spectrometry, which is used to confirm exogenous testosterone use in individuals with a high testosterone to epitestosterone ratio.

TREATMENT

Integrated Management of Patients with AAS Use

The nonathlete weightlifters who abuse AAS frequently do not seek medical treatment and often mistrust physicians. They also do not view these drugs and the associated lifestyle as deleterious to their health. In turn, many internists erroneously view AAS abuse as largely a problem of cheating in competitive sports, while, in fact, most AAS users are not athletes at all. Also, physicians often have a poor understanding of the factors motivating the use of these performance-enhancing drugs, the long-term health effects of AAS, and the associated psychopathologies which may affect treatment choices.

In addition to treating the underlying body dysmorphia disorder which motivates the use of these drugs, the treatment should be directed at the symptoms or the condition for which the patient seeks therapy, such as infertility, sexual dysfunction, gynecomastia, or depressive symptoms. Accordingly, therapy may include some combination of the cognitive and behavioral therapy for the muscle dysmorphia syndrome, antidepressant therapy for depression, selective PDE5 inhibitors for erectile dysfunction, or the use of selective estrogen receptor modulators or aromatase inhibitors to reactivate HPT axis or hCG to restore testosterone levels.

As discussed above, AASs suppress the male hypothalamic-pituitary-gonadal axis and men with long-term AAS use may experience symptoms of profound androgen deficiency such as sexual dysfunction, fatigue, and depressive symptoms during AAS withdrawal. Some of these patients may resume the use of AAS or start using other drugs to combat the distressing withdrawal symptoms. There are no randomized trials of any therapies for AAS withdrawal. Case reports and clinical experience suggest that administration of selective estrogen receptor modulators, CYP19 aromatase inhibitors, or hCG may restore circulating testosterone levels. Clomiphene citrate, a partial estrogen receptor agonist, administered in a dose of 25–50 mg on alternate days can increase LH and FSH levels and restore testosterone levels in a vast majority of men with AAS withdrawal syndrome. However, the recovery of sexual function during clomiphene administration is variable in spite of improvements in testosterone levels. Anecdotally, other aromatase inhibitors such as anastrozole have also been used. hCG, administered by intramuscular injections of 750–1500 international units three times each week, can raise testosterone levels into the normal range. Some patients may not respond to either clomiphene or hCG therapy raising the possibility of irreversible long-term toxic effects of AAS on Leydig cell function.

Adjunctive cognitive and behavioral therapy or antidepressants to treat depression inadequately responsive to endocrine therapies alone may be needed. Emerging human and animal evidence suggests AAS and opioids likely promote dependence via common mechanisms. The opioid antagonist naltrexone blocks AAS dependence in animals. Therefore, treatments for human opioid dependence might also benefit AAS-dependence. Many patients who abuse AAS suffer from body-image disorder such as "muscle dysmorphia" and require psychiatric treatment for this underlying disorder.

LUTS IN MEN

LUTS in men include storage symptoms (urgency, daytime as well as nighttime frequency, and urgency incontinence), voiding disturbances (slow or intermittent stream, difficulty in initiating micturition, straining to void, pain or discomfort during the passage of urine, and terminal dribbling), or postmicturition symptoms (a sense of incomplete voiding after passing urine, and postmicturition dribble). The overactive bladder syndrome refers to urgency with or without urgency incontinence, usually with urinary frequency and nocturia, and is often due to detrusor muscle overactivity. A presumptive diagnosis of benign prostatic hyperplasia should be made only in men with LUTS, who have demonstrable evidence of prostate enlargement and obstruction based on the size of the prostate. LUTS have historically been attributed to benign prostatic hyperplasia although it has become apparent that the pathophysiologic mechanisms of LUTS are complex and multifactorial and may include structural or functional abnormalities of the bladder, bladder neck, prostate, distal sphincter mechanism, and urethra, as well as abnormalities in the neural control of the lower urinary tract. Diuretics, antihistamines, antidepressants, and other medications that have anticholinergic properties can cause or exacerbate LUTS in older men. The intensity of LUTS tends to fluctuate over time.

LUTS is highly prevalent in older men, affecting nearly 50% of >65 and 70% of men >80. The LUTS adversely affects quality of life because of its impact on sleep, ability to perform activities of daily living, and depressive symptoms. LUTS is often associated with erectile dysfunction.

APPROACH TO THE PATIENT

Lower Urinary Tract Symptoms

Medical evaluation should include assessment of potential causes of symptoms; medications including herbal and over-the-counter products that might contribute symptoms; the symptom severity and bother using an International Prostate Symptom Score, and in some patients a frequency-volume chart. The impact of LUTS on sleep, activities of daily living, and quality of life should be evaluated. Evaluation should also include digital prostate examination, neurological examination focused on perineum and lower extremities, urinalysis, fasting blood glucose, electrolytes, creatinine, and prostate-specific antigen (PSA). Urodynamic studies are not required in most patients, but are recommended when invasive surgical therapies are being considered. A urological referral may be appropriate if the patient has hydronephrosis, renal insufficiency, recurrent urinary tract infections, hematuria, or history of acute urinary retention.

TREATMENT

Patients with LUTS

Considerations of the severity of symptoms; the impact of symptoms on sleep, activities of daily living, and quality of life; the natural history of the disease; and potential adverse effects of the intervention should guide the decision of whether to intervene or not. In men with mild to moderately severe LUTS, the symptoms typically progress slowly over many years, and may remain stable or even improve in some men. The men who have mild symptoms can usually be reassured and followed. Several simple steps such as reducing caffeine and alcohol intake, especially late in the day, taking the diuretic medication early in the day, avoiding excessive water intake close to bedtime, double voiding to ensure complete emptying of the bladder may be helpful in reducing the severity of symptoms. Men with mild to moderate bothersome LUTS can be treated effectively using α-adrenergic antagonists, steroid 5α-reductase inhibitors, PDE5 inhibitors, or anticholinergic agents alone or in combination. Selective α-adrenergic antagonists are typically the first line of therapy; their side effects may include hypotension, dizziness, nasal congestion, headache, and floppy iris syndrome. In men with probable benign prostate obstruction with gland enlargement and LUTS, therapy with steroid 5α-reductase inhibitors, finasteride, or dutasteride, for one or more years improves urinary symptoms and flow rate and reduces prostatic volume. Long-term treatment with 5α-reductase inhibitors can reduce the risk of acute urinary retention and need for prostate surgery. Combined administration of steroid 5α-reductase inhibitor and α(1)-adrenergic blocker can rapidly improve urinary symptoms and reduce the relative risk of acute urinary retention and surgery. PDE5 inhibitors when administered chronically alone or in combination with α-adrenergic blockers are effective in improving LUTS and erectile dysfunction (ED) through their effects on nitric oxide—cyclic guanosine monophosphate (cGMP) in the bladder, urethra, and prostate. PDE5 inhibitors do not improve uroflow parameters, and their hypotensive effect may be potentiated by α(1)-adrenergic blockers. Anticholinergic drugs are used for the treatment of overactive bladder in men with prominent urgency symptoms and no evidence of elevated postvoid residual urine. Containment products, such as pads, can help improve social life in men who have severe storage symptoms, including incontinence. Surgery is indicated when medical therapy fails, symptoms progress in spite of medical therapy, or the patient develops acute urinary retention, hydronephrosis, renal insufficiency, or recurrent urinary tract infections, or if the patient has postvoid residual urine volume >25% of the urinary bladder volume.

MEDICAL COMPLICATIONS OF PROSTATE CANCER THERAPY

Prostate cancer is the most common malignancy in American men, accounting for 19% of all diagnosed cancers and ~8% of all cancer deaths; its incidence is on the rise, partly due to increased screening with PSA. In 2017, ~161,360 new cases of prostate cancer were diagnosed in the United States, and there were 26,730 deaths related to prostate cancer. The majority of these men have low-grade, organ-confined prostate cancer and excellent prospects of long-term survival. Substantial improvement in survival in men with prostate cancer has focused attention on the high prevalence of sexual dysfunction, physical dysfunction, and low vitality in the men, which are important contributors to poor quality of life among the patients treated for prostate cancer. The pathophysiology of these symptoms after radical prostatectomy is multifactorial, but denervation and androgen deficiency are important contributors to these symptoms.

Androgen deficiency is common in men with prostate cancer. Testosterone levels decline with age and men with prostate cancer are at risk of having low testosterone levels simply by virtue of their age. However, total and free testosterone levels are even lower in men with prostate cancer, who have undergone prostatectomy, when compared to noncancer age-matched controls. This age-related androgen deficiency in men with prostate cancer is associated with fatigue, sexual dysfunction, mobility limitation, and decreased physical function. Even with bilateral nerve-sparing procedure, >50% of men develop sexual dysfunction after surgery. Although there is some recovery of sexual function with passage of time, 40–50% of men undergoing radical prostatectomy find their sexual performance to be a moderate-to-large problem 18 months after surgery. Sexual problems are a source of psychosocial distress in men with localized prostate cancer. The men with locally advanced or metastatic prostate cancer who undergo androgen deprivation therapy (ADT) encounter even more distressing symptoms because of the profound androgen deficiency. In addition to fatigue and sexual dysfunction, and hot flushes, these men are at increased risk for diabetes, metabolic syndrome, coronary heart disease, and frailty.

Testosterone Therapy in Men with History of Prostate Cancer A history of prostate cancer has historically been considered a contraindication for testosterone therapy. This guidance is based on observations that testosterone promotes the growth of metastatic prostate cancer. Metastatic prostate cancer generally regresses after orchidectomy and ADT. Androgen receptor signaling plays a central role in maintaining growth of normal prostate and prostate cancer. PSA levels are lower in hypogonadal men and increase after testosterone therapy. Prostate volume is lower in hypogonadal men and increases after testosterone therapy to levels seen in age-matched controls.

However, the role of testosterone in prostate cancer is complex. Epidemiological studies and their meta-analyses have not revealed a consistent relationship between serum testosterone and prostate cancer. Others have reported that low testosterone levels are associated with high-grade cancers. In a landmark randomized trial, testosterone therapy of older men with low testosterone did not affect intraprostatic androgen levels or the expression of androgen-dependent prostatic genes. The suppression of circulating testosterone levels by a GnRH antagonist also does not affect intraprostatic androgen concentrations. Open label trials and retrospective analyses of testosterone therapy in men with prostate cancer, who have undergone radical prostatectomy and have undetectable PSA levels after radical prostatectomy, have found very low rates of PSA recurrence. Even in men with high-grade prostatic intraepithelial neoplasia (HGPIN)—a group at high risk of developing prostate cancer—testosterone therapy for 1 year did not increase PSA or rates of prostate cancer.

A majority of men diagnosed with prostate cancer today has localized disease that can be potentially cured by radical prostatectomy. The men with organ-confined prostate cancer (pT2,N0,M0), Gleason score <6, are at a very low risk of disease recurrence after radical prostatectomy with 0.5% biochemical recurrence rate and 0.2% local recurrence rate >10–15 years. Similarly, preoperative PSA <10 ng/mL is associated with lower risk of disease recurrence than PSA >10 ng/mL. After

radical prostatectomy, in the absence of residual cancer, PSA becomes undetectable within a month. An undetectable PSA after radical prostatectomy is a good indicator of biochemical recurrence-free survival at 5 years. Therefore, men with organ-confined prostate cancer (pT2), Gleason score <6, and a preoperative PSA of <10 ng/mL, who have had undetectable PSA levels (<0.1 ng/mL) for >2 years after radical prostatectomy, have very low risk of disease recurrence (<0.5% at 10 years) and may be considered for testosterone therapy on an individualized basis. If testosterone therapy is instituted, it should be associated with careful monitoring of PSA levels and in consultation with a urologist.

MEDICAL COMPLICATIONS OF ADT

In patients with prostate cancer and distant metastases, ADT improves survival. In patients with locally advanced disease, ADT in combination with external beam radiation or as an adjuvant therapy (postprostatectomy and pelvic lymphadenectomy) also has been shown to improve survival. However, ADT is being increasingly used as primary therapy in men with localized disease and in men encountering biochemical recurrence without clear evidence of survival advantage. The overall use of ADT in men with prostate cancer has increased in the past two decades and its use in men with localized disease and biochemical recurrence accounts for a substantial fraction of this increase. Since most men with prostate cancer die of conditions other than their primary malignancy, recognition and management of these adverse effects is paramount.

Profound hypogonadism resulting from ADT is associated with sexual dysfunction, vasomotor symptoms, gynecomastia, decreased muscle mass and strength, frailty, increased fat mass, anemia, fatigue, bone loss, loss of body hair, depressive symptoms, and reduced quality of life. Diabetes and cardiovascular disease have recently been added to the list of these complications (Fig. 392-3). Treatment with GnRH agonists in men with prostate cancer is associated with rapid induction of insulin resistance, hyperinsulinemia, and a significant increase in the risk of incident diabetes. Metabolic syndrome is prevalent in >50% of men undergoing long-term ADT when compared to age-matched men with prostate cancer not undergoing ADT (22%) and their age-matched eugonadal counterparts (20%). Some but not all studies have reported an increased risk of cardiovascular events, death due to cardiovascular events, and peripheral vascular disease in men undergoing ADT. Some reports suggest that men receiving ADT are at an increased risk of thromboembolic events and cognitive dysfunction. The rates of acute kidney injury are higher in men currently receiving ADT than in men not receiving ADT; the increased risk appears to be particularly associated with the use of combined regiments of a GnRH agonist plus and an antiandrogen. ADT also is associated with substantially increased risk of osteoporosis and bone fractures.

APPROACH TO THE PATIENT

Men Receiving ADT

The benefits of ADT in treating nonmetastatic prostate cancer should be carefully weighed against the risks of ADT-induced adverse events (Table 392-4). If ADT is medically indicated, consider whether intermittent ADT is a feasible option. Men being considered for ADT should undergo assessment of cardiovascular, diabetes, and fracture risk; this assessment may include measurement of blood glucose, plasma lipids, and BMD by dual energy x-ray absorptiometry. Institute measures to prevent bone loss, including physical activity, adequate calcium and vitamin D intake, and pharmacological therapy in men with a previous minimal trauma fracture and those with 10-year risk of a major osteoporotic fracture >20%, unless contraindicated. Bisphosphonates and denosumab have been shown to reduce fracture risk in men undergoing ADT. Men with prostate cancer who are receiving ADT should be monitored for weight gain and diabetes. Encourage lifestyle intervention, including physical activity and exercise, and attention to weight, blood pressure, lipid profile, blood glucose, and smoking cessation, to reduce the risk of cardiometabolic complications. In randomized trials, medroxyprogesterone, cyproterone acetate, and a serotonin uptake inhibitor, venlafaxine, have been shown to be more efficacious than placebo in alleviating hot flushes. The side effects of these medications—increased appetite and weight gain with medroxyprogesterone, gynecomastia with estrogenic compounds, and dry mouth with venlafaxine—should be weighed against their relative efficacy. Acupuncture, soy products, vitamin E, herbal medicines, and transdermal estradiol have been used empirically for the treatment of vasomotor symptoms without clear evidence of efficacy. Gynecomastia can be prevented by local radiation therapy or the use of an antiestrogen or an aromatase inhibitor; these therapies are effective in alleviating pain and tenderness, but are less effective in reducing established gynecomastia. For long-standing gynecomastia that persists after cessation of ADT and is bothersome, mammoplasty is an effective treatment option.

TABLE 392-4 Checklist for Men Undergoing Androgen Deprivation Therapy (ADT)

1. Weigh the risks and benefits of ADT and whether intermittent ADT is a feasible and safe option.
2. Perform a baseline assessment including fasting glucose, plasma lipids, blood pressure, bone mineral density, and FRAX® score.
3. Optimize calcium and vitamin D intake, encourage structured physical activity and exercise, and consider pharmacologic therapy in men with a previous minimal trauma fracture and those with a 10-year risk of a major osteoporotic fracture >20%, unless contraindicated.
4. Monitor body weight, fasting glucose, plasma lipids, blood pressure, and bone mineral density, and encourage smoking cessation and physical activity.
5. In men who are receiving ADT and who experience bothersome hot flushes, as indicated by sleep disturbance or interference with work or activities of daily living, consider initial therapy with venlafaxine. If in effective, add medroxyprogesterone acetate.
6. In men who experience painful breast enlargement, consider therapy with an estrogen receptor antagonist, such as tamoxifen.

PREVENTION OF SEXUALLY TRANSMITTED DISEASES

Adolescent boys and young men 15–24 years; men who have sex with men, or have multiple sex partners, or have unprotected sex without condom, or have sex with sex workers; men who use illicit drugs; men who have history of previous sexually transmitted infection (STI); and transgender men are at increased risk for STIs. STIs increase the risk of oropharyngeal and anogenital cancers, liver disease, pelvic pain, infertility, inadvertent transmission of infection to others, and emergency department visits, and are a preventable cause of excess morbidity and mortality. HIV, hepatitis B and C infections, and syphilis can have additional disease-specific complications. The prevention and treatment of STIs are discussed in Chap. 131. Additionally, the Centers for Disease Control (CDC) and U.S. Preventive Health Services Task Force (USPHSTF) have published guidelines on the prevention, treatment, and pre- and postexposure prophylaxis of STIs. The approach to the prevention of STIs includes a structured risk assessment; counseling about safe sex practices including condom use; immunization of individuals at risk; diagnosis and treatment of infected individuals whether or not they are symptomatic; detection and treatment of sexual partners; and targeted sex education of adolescents and young men who are at high risk for STIs. The USPHSTF recommends screening for HIV in all men, 15–65 years, regardless of risk, and for hepatitis B virus and syphilis in men at increased risk. Because more than half of STIs occur in persons, 15–24 years, the USPHSTF also recommends behavioral counseling for all sexually active adolescents and adult men at increased risk of STIs to encourage condom use and other protective behaviors, including consideration of abstinence, reducing the number of sex partners, and avoidance of unsafe sex practices. Consistent and correct condom use is the most important method of preventing STIs. Effective immunizations are available against hepatitis B, human papillomavirus (HPV), and *Neisseria meningitides*. The CDC's Advisory Committee on Immunization Practices (ACIP) recommends universal hepatitis B immunization for all unvaccinated adults presenting to an

STI clinic, all HIV-infected adults, and health workers. Although ACIP recommends HPV vaccination in males 9–21 years, and in men 9–26 if they have sex with men or have an immunecompromising condition, recent data suggest that the prevalence of HPV and its complications continues to increase until middle age and some experts recommend extending the age limit for HPV vaccination. Meningococcal vaccination is indicated for men who have sex with men from an area of outbreak and for all HIV-infected men.

Because men seeking care in men's health clinics often do so for sexual and urogenital problems, these visits offer opportunities for counseling, screening, and treatment of STIs, and institution of immunization and other preventive measures for STIs.

FURTHER READING

Abrams P et al: Evaluation and treatment of lower urinary tract symptoms in older men. J Urol 189:S93, 2013.

Basaria S: Cardiovascular disease associated with androgen deprivation therapy: Time to give it due respect. J Clin Oncol 33:1232, 2015.

Bhasin S et al: Testosterone therapy in men with androgen deficiency syndromes: An Endocrine Society clinical practice guideline. J Clin Endocrinol Metab 95:2536–59, 2010.

Centers for Disease Control and Prevention. Sexually Transmitted Diseases Treatment Guidelines. MMWR 64:51, 2015.

Choi PY et al: Muscle dysmorphia: A new syndrome in weightlifters. Br J Sports Med 36:375, 2002.

Jones C et al: Management of lower urinary tract symptoms in men: summary of NICE guidance. BMJ 340:c2354, 2010.

López AM et al: Fracture risk in patients with prostate cancer on androgen deprivation therapy. Osteoporos Int 16:707, 2005.

Mortality Tables at *http://www.cdc.gov/nchs/deaths.htm* (HHS, CDC, NCHS).

Pope HG Jr et al: Adverse health consequences of performance-enhancing drugs: An endocrine society scientific statement. Endocr Rev 35:341, 2014.

Shahani S et al: Androgen deprivation therapy in prostate cancer and metabolic risk for atherosclerosis. J Clin Endocrinol Metab 93:2042, 2008.

Siegel RL et al: Cancer statistics. CA Cancer J Clin 67:7, 2017.

Smith MR et al: Denosumab in men receiving androgen deprivation therapy for prostate cancer. N Engl J Med 361:745, 2009.

Spitzer M et al: Risks and benefits of testosterone therapy in older men. Nat Rev Endocrinol 9:414, 2013.

United States Preventive Health Services Task Force. Final recommendation statement sexually transmitted infections: behavioral counseling. *https://www.uspreventiveservicestaskforce.org/Page/Document/RecommendationStatementFinal/sexually-transmitted-infections-behavioral-counseling1*. Accessed June 21, 2017.

393 Lesbian, Gay, Bisexual, and Transgender (LGBT) Health

Baligh R. Yehia, Harvey J. Makadon

UNDERSTANDING LGBT HEALTH DISPARITIES

Numerous studies highlight health disparities involving the care of lesbian, gay, bisexual, and transgender (LGBT) people. Lesbian and bisexual women are less likely to receive recommended preventive screenings such as breast, cervical, and colorectal cancer screenings. Among men who have sex with men, rates of human papillomavirus–associated anal cancers are 17 times higher than those of heterosexual men. In addition, gay and bisexual men accounted for 67% of all new HIV diagnoses in the United States in 2014, and they disproportionately contract sexually transmitted infections. In 2014, men who have sex with men accounted for 83% of primary and secondary syphilis infections in the United States where the sex of the sexual partner was known. Transgender individuals have a higher prevalence of HIV infection and suicide compared with other groups; 41% of transgender adults report ever attempting suicide compared with 1.6% of the general population.

Research has found that LGBT individuals are more likely to experience depression, anxiety, and alcohol and drug use than their counterparts. Most concerning are the rates of suicide attempts and ideation among the LGBT community, particularly youth. LGB youth are more than twice as likely to attempt suicide than their heterosexual peers, and approximately 30% of LGB students report having attempted suicide over a 12-month period. In addition, U.S. studies indicate that substance abuse is twice as common in LGBT youth compared with their counterparts. These findings are mirrored among LGBT adults: the prevalence of substance abuse disorders is 20–30% compared with approximately 9% in the general population.

These disparities are compounded by structural barriers to healthcare, including decreased access to medical care, lack of awareness to the unique health needs of LGBT individuals, and stigma and discrimination toward the LGBT community. Many LGB individuals perceive the healthcare setting and providers as threatening, which may lead to avoiding needed medical care or withholding important medical information. A large U.S. survey identified that 8% of LGB and 27% of transgender individuals were refused needed healthcare, and almost 11% of LGB and 21% of transgender people reported being subjected to harsh or abusive language by healthcare professionals. Apart from healthcare settings, more than two-thirds of LGB people report discrimination in their personal lives, and 90% of transgender individuals report harassment, mistreatment, or discrimination at work. Chronic exposure to high levels of stress from real or anticipated discrimination, referred to as "minority stress," may be an important factor contributing to the poor health outcomes experienced by LGBT populations.

While some research on LGBT health has been conducted, the Institute of Medicine has called for more study to better understand the needs and experiences of LGBT individuals. Moreover, many LGBT individuals experience health disparities across their life cycle (e.g., LGBT youth are at greater risk of suicide and homelessness, while elderly LGBT individuals face barriers to health because of isolation and fewer family supports), necessitating a longitudinal approach to examining LGBT health issues. We have limited data on the health of LGBT individuals outside the United States and Europe. However, studies demonstrate that problems are greatest where people cannot be open about their sexual orientation and gender identity. Encouraging greater LGBT acceptance and access to healthcare will be critical to improving outcomes and experiences for LGBT communities.

CREATING POSITIVE HEALTH EXPERIENCES FOR LGBT PATIENTS

Understanding Gender Identify and Sexual Orientation

Addressing health disparities and creating positive healthcare experiences requires an understanding of the diversity of cultural expression and lives of LGBT persons. Foremost, providers must be able to distinguish *gender identity* from *sexual orientation*. Gender identity is a person's internal sense of their gender. It should not be confused with sex assigned at birth, which is based on anatomy and biology. Gender identity expands beyond the binary male and female, and includes persons who think of their gender as containing elements of both or neither. Many individuals who do not identify with the gender that correlates with their sex assigned at birth often use the terms *transgender* or *transmale* or *trans-female* to identify themselves. Sexual orientation refers to how one thinks of their physical or emotional attraction to others. Sexual orientation has three dimensions: attraction, behavior, and identity. *Attraction* refers to one's desire to be with someone, regardless of one's behavior or stated identity. For example, a woman may be attracted to another woman, but this attraction may never be acted upon and may not form part of her sexual identity. *Behavior* refers to a person's sexual and romantic partners. Although sexual identity often aligns with behavior, some individuals who identify as heterosexual

TABLE 393-1 Common LGBT Terminology and Definitions

TERM	DEFINITION
Agender	Identifying as having no gender.
Asexual	Experiencing little or no sexual attraction to others.
Assigned sex at birth	The sex (male or female) assigned to a child at birth, most often based on the child's external anatomy. Also referred to as birth sex, natal sex, biological sex, or sex.
Bisexual	A sexual orientation that describes a person who is emotionally and sexually attracted to people of their own gender and people of other genders.
Cisgender	A person whose gender identity and assigned sex at birth correspond (i.e., a person who is not transgender).
Gay	A sexual orientation that describes a person who is emotionally and sexually attracted to people of their own gender. It can be used regardless of gender identity, but it is more commonly used to describe men.
Gender dysphoria	Distress experienced by some individuals whose gender identity does not correspond with their assigned sex at birth. Manifests as clinically significant distress or impairment in social, occupational, or other important areas of functioning.
Gender expression	The way a person acts, dresses, speaks, and behaves (i.e., feminine, masculine, androgynous). Gender expression does not necessarily correspond to assigned sex at birth or gender identity.
Gender identity	A person's internal sense of being a man/male, woman/female, both, neither, or another gender.
Gender nonconforming	Expressing a gender that differs from a given society's norms for males and females.
Heterosexual	A sexual orientation that describes women who are emotionally and sexually attracted to men, and men who are emotionally and sexually attracted to women.
Intersex (disorders of sexual development)	A group of rare conditions where the reproductive organs and genitals do not develop as expected.
Lesbian	A sexual orientation that describes a woman who is emotionally and sexually attracted to other women.
Men who have sex with men (MSM)/Women who have sex with women (WSW)	Categories used in research and public health to describe those who engage in same-sex sexual behavior, regardless of their sexual orientation. Individuals rarely use the terms MSM or WSW to describe themselves.
Pangender	Describes a person whose gender identity comprises many genders.
Pansexual	A sexual orientation that describes a person who is emotionally and sexually attracted to people regardless of gender.
Queer	An umbrella term used by some to describe people who think of their sexual orientation or gender identity as outside of societal norms. Some people view the term as more fluid and inclusive than traditional categories for sexual orientation and gender identity. Due to its history as a derogatory term, it is not embraced or used by all members of the LGBT community.
Questioning	Describes an individual who is unsure about or is exploring their own sexual orientation and/or gender identity.
Same-sex attraction	Describes the experience of a person who is emotionally and/or sexually attracted to people of the same gender. Use of this term is not indicative of a person's sexual behavior.
Sexual orientation	Describes how a person characterizes their physical and emotional attraction to others. Sexual orientation is distinct from sex, gender identity, and gender expression.
Trans man/transgender man/female-to-male (FTM)	A transgender person whose gender identity is male may use these terms to describe themselves. Some will just use the term *man*.
Trans woman/transgender woman/male-to-female (MTF)	A transgender person whose gender identity is female may use these terms to describe themselves. Some will just use the term *woman*.
Transgender	Describes a person whose gender identity and assigned sex at birth do not correspond. Also used as an umbrella term to include gender identities outside of male and female.
Transition/Affirmation	For transgender persons, the process of coming to recognize, accept, and express one's gender identity. Most often, this refers to the period when a person makes social, legal, and/or medical changes, such as changing their clothing, name, and sex designation, as well as using medical interventions.

Note: It is important to note that definitions vary across communities, that they change over time, and that not all LGBT people agree with all these definitions.

may have same-gender partners and some individuals who identify as lesbian or gay may have different-gender partners. Lastly, *identity* refers to how a person defines their own sexuality. Common terms for sexual identity include *gay*, *lesbian*, *bisexual*, *straight*, *heterosexual*, *homosexual*, and *asexual* **(Table 393-1)**. As individuals go through the process of understanding their sexuality and self-identity over time, they may change how they define their sexual identity.

The creation of a welcoming environment requires not making any assumptions about an individual's gender identity or sexual orientation. Both front-line staff and clinicians should be cognizant of patient communication. For example:

- *Instead of saying* "How may I help you, sir?"
- *Say* "How may I help you?"
- *Instead of saying* "She is here for her appointment."
- *Say* "The patient is here in the waiting room."
- *Instead of saying* "Do you have a wife?"
- *Say* "Are you in a relationship?
- *Instead of saying* "What are your mother's and fathers' names?"
- *Say* "What are your parents' names?"

Developing Comfort and Competency in Sexual Health

Developing comfort discussing sexual health and intimacy is critical to providing appropriate care. A good starting place is to ask if a patient is sexually active—and if so, with whom, how often, and what they do with their partner(s). These discussions can allow providers to focus subsequent conversations on issues most relevant to a patient's health. For example, a gay man with multiple sexual partners who engages in receptive anal sex without condoms is at high risk for HIV and sexually transmitted infections. It will be important to recommend more frequent screenings and discuss use of pre-exposure prophylaxis (PrEP) and condoms to prevent HIV and sexually transmitted infections. If you are seeing a transgender man, it will be important to know if he still has natal female genitalia to ensure appropriate cancer screening. Notably, many if not most transgender people have not had genital affirmation surgery and retain their natal sex organs.

Creating a Welcoming and Safe Healthcare Environment

Hospitals and clinics can take a number of steps to create a welcoming and safe space for LGBT patients. This starts by establishing and communicating a nondiscrimination policy that clearly includes gender identity, gender expression, and sexual orientation protections.

Additionally, hospitals and clinics can develop and implement an equal visitation policy to ensure equal visitation for LGBT patients from same-sex partners, parents, and other family and friends. Staff training in LGBT patient-centered care also is a key component of creating inclusive health environments. This includes covering LGBT cultural competency, caring for LGBT patients, creating an inclusive environment for LGBT patients and staff, and other topics important for LGBT health.

As hospitals and clinics continue to adopt electronic health records, collecting sexual orientation and gender identity information becomes increasingly important to delivering personalized care to LGBT individuals. It allows providers to monitor quality of care and track population-based outcomes. This information can be captured by three questions:

- *Do you think of yourself as:* straight or heterosexual; lesbian, gay, or homosexual; bisexual; something else; don't know; choose not to disclose.
- *What is your current gender identity?* Male; female; transgender male/trans man/female-to-male (FTM); transgender female/trans woman/male-to-female (MTF); genderqueer, neither exclusively male nor female; additional category, please specify; choose not to disclose.
- *What sex were you assigned at birth on your original birth certificate?* Male; female; choose not to disclose.

The physical environment of a hospital or clinic is important, but the majority of clinical spaces do not signal that they are safe spaces for LGBT patients. Most healthcare posters, pamphlets, and materials feature heterosexual individuals or couples; adding LGBT-friendly images and text can help signal that the hospital or clinic is a safe space for sexual and gender minorities. In addition, easily identifying LGBT-competent providers by using websites, buttons, and pins can help patients select providers and feel at ease when attending appointments. Lastly, designating all-gender bathrooms is important to creating welcoming spaces, particularly for transgender and gender-nonconforming individuals.

FUTURE DIRECTION IN LGBT HEALTH

While social and cultural acceptance of the LGBT community has improved in certain parts of the world, many LGBT individuals continue to experience discrimination, stigmatization, and violence. Inequitable healthcare policies and practices, lack of awareness to LGBT health issues, and limited understanding of the unique health needs of LGBT individuals contribute to decreased access to care and disparities in health outcomes for LGBT individuals. Addressing these barriers will require improved data collection on the LGBT population; understanding of the intersectionality of gender identity, sexual orientation, race/ethnicity, and other determinants of health; and outcomes-focused research across the life course. In striving to deliver high-quality care experiences for all patients, hospitals, clinics, and providers will have to focus on meeting the needs of the LGBT community.

FURTHER READING

Centers for Disease Control and Prevention: Lesbian, Gay, Bisexual, and Transgender Health. 2016. Available at *www.cdc.gov/lgbthealth/*. Accessed September 5, 2017.

Fenway Health: Glossary of LGBT Terms for Health Care Teams. 2016. Available at *www.lgbthealtheducation.org/wp-content/uploads/LGBT-Glossary_March2016.pdf*. Accessed September 5, 2017.

Institute of Medicine: The Health of Lesbian, Gay, Bisexual, and Transgender (LGBT) People: Building a Foundation for Better Understanding. 2011. Available at *www.nap.edu/catalog.php?record_id=13128*. Accessed September 5, 2017.

Institute of Medicine: Collecting Sexual Orientation and Gender Identity Data in Electronic Health Records. 2012. Available at *www.nationalacademies.org/hmd/Reports/2012/Collecting-Sexual-Orientation-and-Gender-Identity-Data-in-Electronic-Health-Records.aspx*. Accessed September 5, 2017.

Joint Commission: Advancing Effective Communication, Cultural Competence, Patient- and Family-Centered Care for the Lesbian, Gay, Bisexual, and Transgender Community: A Field Guide; 2011. Available at *www.jointcommission.org/lgbt*. Accessed September 5, 2017.

Substance Abuse and Mental Health Services Administration: Top Health Issues for LGBT Populations Information & Resource Kit. 2012. Available at *www.store.samhsa.gov/shin/content/SMA12-4684/SMA12-4684.pdf*. Accessed September 5, 2017.

Section 3 Obesity, Diabetes Mellitus, and Metabolic Syndrome

394 Pathobiology of Obesity

Jeffrey S. Flier, Eleftheria Maratos-Flier

In a world where food supplies are intermittent, the ability to store energy in excess of what is required for immediate use is essential for survival. Fat cells, residing within widely distributed adipose tissue depots, are adapted to store excess energy efficiently as triglyceride and, when needed, to release stored energy as free fatty acids for use at other sites. This physiologic system, orchestrated through endocrine and neural pathways, permits humans to survive starvation for as long as several months. However, in the presence of nutritional abundance and a sedentary lifestyle, and influenced importantly by genetic endowment, this system increases adipose energy stores and produces adverse health consequences.

DEFINITION AND MEASUREMENT

Obesity is a state of excess adipose tissue mass. Although often viewed as equivalent to increased body weight, this need not be the case—lean but very muscular individuals may be overweight by numerical standards without having increased adiposity. Body weights are distributed continuously in populations, so that choice of a medically meaningful distinction between lean and obese is somewhat arbitrary. Obesity is therefore defined by assessing its linkage to morbidity or mortality.

Although not a direct measure of adiposity, the most widely used method to gauge obesity is the *body mass index* (BMI), which is equal to weight/height^2 (in kg/m^2) **(Fig. 394-1)**. Other approaches to quantifying obesity include anthropometry (skinfold thickness), densitometry (underwater weighing), computed tomography (CT) or magnetic resonance imaging (MRI), and electrical impedance. Using data from the Metropolitan Life Tables, BMIs for the midpoint of all heights and frames among both men and women range from 19 to 26 kg/m^2; at a similar BMI, women have more body fat than men. Based on data of substantial morbidity, a BMI of 30 is most commonly used as a threshold for obesity in both men and women. Most but not all large-scale epidemiologic studies suggest that all-cause, metabolic, cancer, and cardiovascular morbidity begin to rise (albeit at a slow rate) when BMIs are ≥25. Most authorities use the term *overweight* (rather than obese) to describe individuals with BMIs between 25 and 30. A BMI between 25 and 30 should be viewed as medically significant and worthy of therapeutic intervention in the presence of risk factors that are influenced by adiposity, such as hypertension and glucose intolerance.

The distribution of adipose tissue in different anatomic depots also has substantial implications for morbidity. Specifically, intraabdominal and abdominal subcutaneous fat have more significance than subcutaneous fat present in the buttocks and lower extremities. This distinction is most easily made clinically by determining the waist-to-hip ratio, with a ratio >0.9 in women and >1.0 in men being abnormal. Many of the most important complications of obesity, such as insulin resistance, diabetes, hypertension, hyperlipidemia, and hyperandrogenism in women, are linked more strongly to intraabdominal and/or

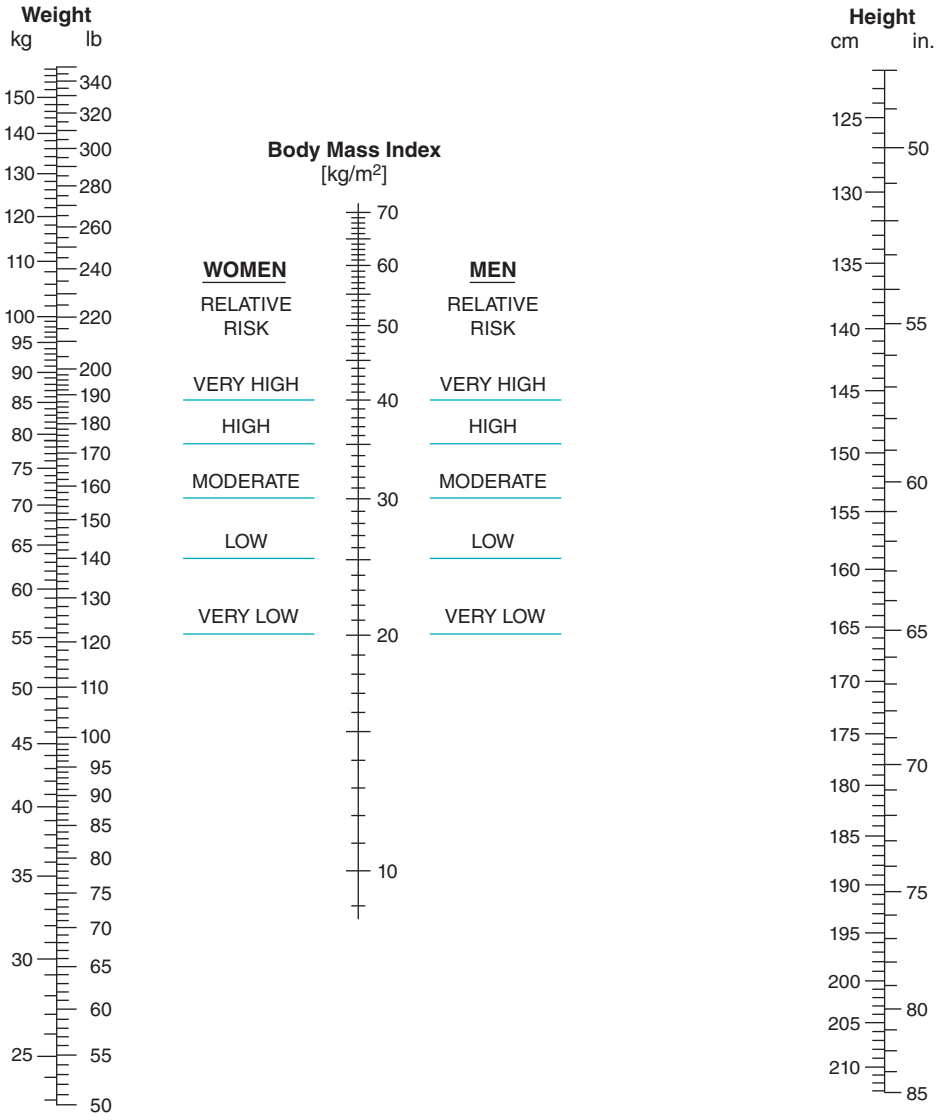


FIGURE 394-1 Nomogram for determining body mass index. To use this nomogram, place a ruler or other straight edge between the body weight (without clothes) in kilograms or pounds located on the left-hand line and the height (without shoes) in centimeters or inches located on the right-hand line. The body mass index is read from the middle of the scale and is in metric units. *(Copyright 1979, George A. Bray, MD; used with permission.)*

upper body fat than to overall adiposity **(Chap. 401)**. The mechanism underlying this association is unknown but may relate to the fact that intraabdominal adipocytes are more lipolytically active than those from other depots. Release of free fatty acids into the portal circulation has adverse metabolic actions, especially on the liver. Adipokines and cytokines that are differentially secreted by adipocyte depots may play a role in the systemic complications of obesity.

PREVALENCE

Data from the National Health and Nutrition Examination Surveys (NHANES) show that the percentage of the American adult population with obesity (BMI >30) has increased from 14.5% (between 1976 and 1980) to 36.5% (between 2011 and 2014). As many as 70% of U.S. adults aged ≥20 years were either overweight (defined as BMI >25) or obese (BMI >30) between the years of 2013 and 2014. Extreme obesity (BMI ≥40) has also increased and affects 5.7% of the population. The increasing prevalence of medically significant obesity raises great concern. Overall, the prevalence of obesity is higher in women (38%) than in men (34%). In women, poverty is associated with increased prevalence. Obesity is more common among blacks and Hispanics, and less common in Asians. Among some Asian subgroups, health and mortality risks may begin at lower BMIs, associated with greater intraabdomnal obesity. The prevalence in children and adolescents has been rising at a worrisome rate, reaching 17.0% in 2011–2014, but may be leveling off.

GLOBAL CONSIDERATIONS

According to the World Health Organization, global obesity almost doubled between 1980 and 2008. There were >200 million obese men and almost 300 million obese women, 11% of adults worldwide, in 2008. A majority (65%) of the world's population lives in countries where being overweight or obese kills more people than being underweight. In 2014, the U.S. led the Organization for Economic Co-operation and Development (OECD) countries in obesity prevalence (34%), with New Zealand, Australia, the UK, Mexico, and Canada close behind in the high 20% range. Worldwide, the highest prevalence of obesity is in several Pacific Island states including the Cook Islands (50%), Nauru (45.6%), and Tonga (43.3%) and Gulf States including Qatar (42%) and Kuwait (38%). In developing countries, such as India, obesity prevalence is rising (5%) with a greater tendency to harmful intraabdominal obesity at lower BMIs in this population, and the consequences for metabolic and cardiovascular health are disproportionate to obesity prevalence. The causes of increased obesity globally are complicated, and may vary from country to country, but contributory factors include rising incomes, changing food supplies, and reduced activity.

PHYSIOLOGIC REGULATION OF ENERGY BALANCE

Substantial evidence suggests that body weight is regulated by both endocrine and neural components that ultimately influence the effector arms of energy intake and expenditure. This complex regulatory system is necessary because even small imbalances between energy intake and expenditure will ultimately have large effects on body weight. For example, a 0.3% positive imbalance over 30 years would result in a 9-kg (20-lb) weight gain. This exquisite regulation of energy balance cannot be monitored easily by calorie-counting in relation to physical activity. Rather, body weight regulation or dysregulation depends on a complex interplay of hormonal and neural signals. Alterations in stable weight by forced overfeeding or food deprivation induce physiologic changes that resist these perturbations: with weight loss, appetite increases and energy expenditure falls; with overfeeding, appetite falls and energy expenditure increases. This latter compensatory mechanism frequently fails, however, permitting obesity to develop when food is abundant and physical activity is limited. A major regulator of these adaptive responses is the adipocyte-derived hormone leptin, which acts through brain circuits (predominantly in the hypothalamus) to influence appetite, energy expenditure, and neuroendocrine function (see below).

Appetite is influenced by many factors that are integrated by the brain, most importantly within the hypothalamus **(Fig. 394-2)**. Signals that impinge on the hypothalamic center include neural afferents, hormones, and metabolites. Vagal inputs are particularly important, bringing information from viscera, such as gut distention. Hormonal signals include leptin, insulin, cortisol, and gut peptides. Among the latter is ghrelin, which is made in the stomach and stimulates feeding, and peptide YY (PYY) and cholecystokinin, which is made in the small intestine and signals to the brain through direct action on hypothalamic control centers and/or via the vagus nerve. Metabolites, including glucose, can influence appetite, as seen by the effect of hypoglycemia

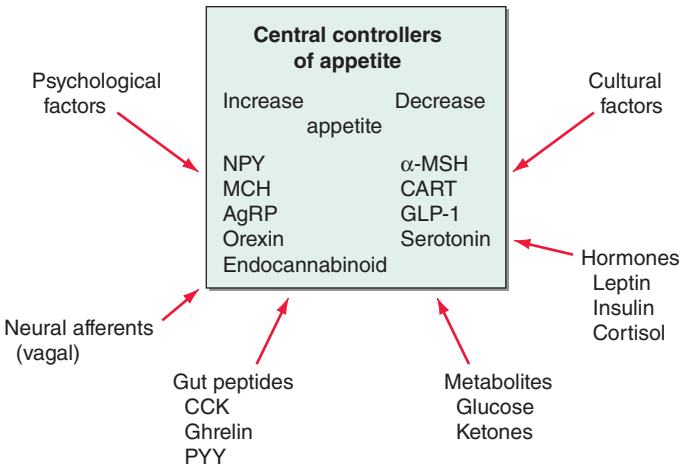


FIGURE 394-2 The factors that regulate appetite through effects on central neural circuits. Some factors that increase or decrease appetite are listed. AgRP, Agouti-related peptide; CART, cocaine- and amphetamine-related transcript; CCK, cholecystokinin; GLP-1, glucagon–related peptide-1; MCH, melanin-concentrating hormone; α-MSH, α-melanocyte-stimulating hormone; NPY, neuropeptide Y.

to induce hunger; however, glucose is not normally a major regulator of appetite. These diverse hormonal, metabolic, and neural signals act by influencing the expression and release of various hypothalamic peptides (e.g., neuropeptide Y [NPY], Agouti-related peptide [AgRP], α-melanocyte-stimulating hormone [α-MSH], and melanin-concentrating hormone [MCH]) that are integrated with serotonergic, catecholaminergic, endocannabinoid, and opioid-signaling pathways (see below). Psychological and cultural factors also play a role in the final expression of appetite. Apart from rare genetic syndromes involving leptin, its receptor, and the melanocortin system, specific defects in this complex appetite control network that influence common cases of obesity are not well defined.

Energy expenditure includes the following components: (1) resting or basal metabolic rate; (2) the energy cost of metabolizing and storing food; (3) the thermic effect of exercise; and (4) adaptive thermogenesis, which varies in response to long-term caloric intake (rising with increased intake). Basal metabolic rate accounts for ~70% of daily energy expenditure, whereas active physical activity contributes 5–10%. Thus, a significant component of daily energy consumption is fixed.

Genetic models in mice indicate that mutations in certain genes (e.g., targeted deletion of the insulin receptor in adipose tissue) protect against obesity, apparently by increasing energy expenditure. Adaptive thermogenesis occurs in *brown adipose tissue* (BAT), which plays an important role in energy metabolism in many mammals. In contrast to white adipose tissue, which is used to store energy in the form of lipids, BAT expends stored energy as heat. A mitochondrial *uncoupling protein* (UCP-1) in BAT dissipates the hydrogen ion gradient in the oxidative respiration chain and releases energy as heat. The metabolic activity of BAT is increased by a central action of leptin, acting through the sympathetic nervous system that heavily innervates this tissue. In rodents, BAT deficiency causes obesity and diabetes; stimulation of BAT with a specific adrenergic agonist (β_3 agonist) protects against diabetes and obesity. BAT exists in humans (especially neonates), and although its physiologic role is not yet established, identification of functional BAT in many adults using positron emission tomography (PET) imaging has increased interest in the implications of the tissue for pathogenesis and therapy of obesity. Beige fat cells, recently described, resemble BAT cells in expressing UCP-1. They are scattered through white adipose tissue, and their thermogenic potential is uncertain.

THE ADIPOCYTE AND ADIPOSE TISSUE

Adipose tissue is composed of the lipid-storing adipose cell and a stromal/vascular compartment in which cells including preadipocytes and macrophages reside. Adipose mass increases by enlargement of adipose cells through lipid deposition, as well as by an increase in the number of adipocytes. Obese adipose tissue is also characterized by increased numbers of infiltrating macrophages. The process by which adipose cells are derived from a mesenchymal preadipocyte involves an orchestrated series of differentiation steps mediated by a cascade of specific transcription factors. One of the key transcription factors is *peroxisome proliferator-activated receptor γ* (PPARγ), a nuclear receptor that binds the thiazolidinedione class of insulin-sensitizing drugs used in the treatment of type 2 diabetes **(Chap. 397)**.

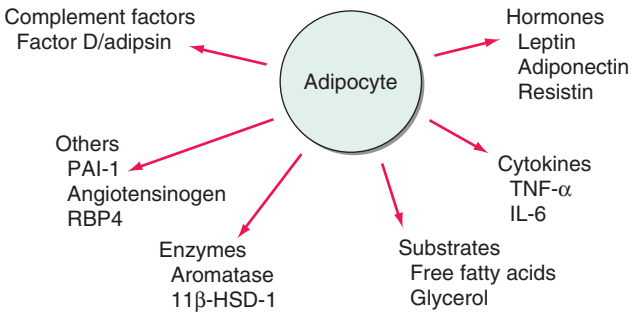


FIGURE 394-3 Factors released by the adipocyte that can affect peripheral tissues. IL-6, interleukin 6; PAI, plasminogen activator inhibitor; RBP4, retinal binding protein 4; TNF, tumor necrosis factor.

Although the adipocyte has generally been regarded as a storage depot for fat, it is also an endocrine cell that releases numerous molecules in a regulated fashion **(Fig. 394-3)**. These include the energy balance–regulating hormone leptin, cytokines such as tumor necrosis factor (TNF)-α and interleukin (IL)-6, complement factors such as factor D (also known as *adipsin*), prothrombotic agents such as plasminogen activator inhibitor I, and a component of the blood pressure–regulating system, angiotensinogen. Adiponectin, an abundant adipose-derived protein whose levels are reduced in obesity, enhances insulin sensitivity and lipid oxidation and has vascular-protective effects, whereas RBP4, whose levels are increased in obesity, may induce insulin resistance. Obesity is accompanied by increased fat storage in tissues such as muscle and liver, and this ectopic lipid has been linked to metabolic disturbances. These factors, and others not yet identified, play a role in the physiology of lipid homeostasis, insulin sensitivity, blood pressure control, coagulation, and vascular health, and are likely to contribute to obesity-related pathologies.

ETIOLOGY OF OBESITY

Although the molecular pathways regulating energy balance are beginning to be illuminated, the causes of obesity remain elusive. In part, this reflects the fact that obesity is a heterogeneous group of disorders. At one level, the pathophysiology of obesity seems simple: a chronic excess of nutrient intake relative to the level of energy expenditure. However, due to the complexity of the neuroendocrine and metabolic systems that regulate energy intake, storage, and expenditure, it has been difficult to quantitate all the relevant parameters (e.g., food intake and energy expenditure) over time in human subjects.

Role of Genes Versus Environment Obesity is commonly seen in families, and the heritability of body weight is similar to that for height. Inheritance is usually not Mendelian, however, and it is difficult to distinguish the role of genes and environmental factors. Adoptees more closely resemble their biologic than adoptive parents with respect to obesity, providing strong support for genetic influences. Likewise, identical twins have very similar BMIs whether reared together or apart, and their BMIs are much more strongly correlated than those of dizygotic twins. These genetic effects appear to relate to both energy intake and expenditure. Currently, identified genetic variants, both common and rare, account for <5% of the variance of body weight.

Whatever the role of genes, it is clear that the environment plays a key role in obesity, as evidenced by the fact that famine limits obesity in even the most obesity-prone individual. In addition, the recent increase in the prevalence of obesity in the United States is far too rapid to be due to changes in the gene pool. Undoubtedly, genes influence the susceptibility to obesity in response to specific diets and availability

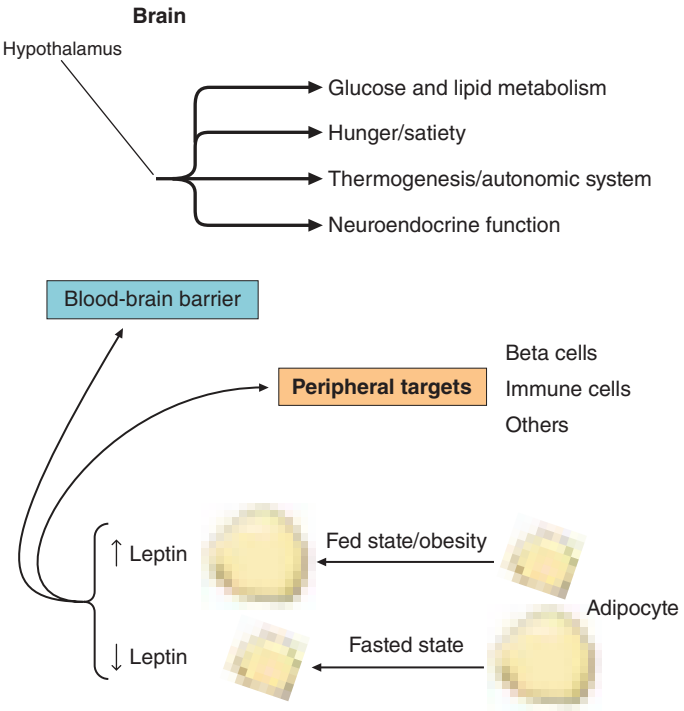


FIGURE 394-4 The physiologic system regulated by leptin. Rising or falling leptin levels act through the hypothalamus to influence appetite, energy expenditure, and neuroendocrine function and through peripheral sites to influence systems such as the immune system.

of nutrition. Cultural factors are also important—these relate to both availability and composition of the diet and to changes in the level of physical activity. In industrial societies, obesity is more common among poor women, whereas in underdeveloped countries, wealthier women are more often obese. In children, obesity correlates to some degree with time spent watching television. Although the role of diet composition in obesity continues to generate controversy, it appears that high-fat diets may, when combined with simple, rapidly absorbed carbohydrates, promote obesity. Specific genes are likely to influence the response to specific diets, but these genes are largely unidentified.

Additional environmental factors may contribute to the increasing obesity prevalence. Both epidemiologic correlations and experimental data suggest that sleep deprivation leads to increased obesity. Changes in gut microbiome with capacity to alter energy balance are receiving experimental support from animal studies, and a possible role for obesigenic viral infections continues to receive sporadic attention.

Specific Genetic Syndromes For many years, obesity in rodents has been known to be caused by a number of distinct mutations distributed through the genome. Most of these single-gene mutations cause both hyperphagia and diminished energy expenditure, suggesting a physiologic link between these two parameters of energy homeostasis. Identification of the *ob* gene mutation in genetically obese (ob/ob) mice represented a major breakthrough in the field. The ob/ob mouse develops severe obesity, insulin resistance, and hyperphagia, as well as efficient metabolism (e.g., it gets fat even when ingesting the same number of calories as lean litter mates). The product of the *ob* gene is the peptide leptin, a name derived from the Greek root *leptos*, meaning thin. Leptin is secreted by adipose cells and acts primarily through the hypothalamus. Its level of production provides an index of adipose energy stores **(Fig. 394-4)**. Raising leptin levels can decrease food intake and increase energy expenditure. Another mouse mutant, db/db, which is resistant to leptin, has a mutation in the leptin receptor and develops a similar syndrome. The *ob* gene is present in humans where it is also expressed in fat. Several families with morbid, early-onset obesity caused by inactivating mutations in either leptin or the leptin receptor have been described, thus demonstrating the biologic relevance of the leptin pathway in humans. Obesity in these individuals begins shortly after birth, is severe, and is accompanied by neuroendocrine abnormalities. The most prominent of these is hypogonadotropic hypogonadism, which is reversed by leptin replacement in the leptin-deficient subset. Central hypothyroidism and growth retardation are seen in the mouse model, but their occurrence in leptin-deficient humans is less clear. Mutations in the leptin or leptin receptor genes do not play a prominent role in common forms of obesity.

Mutations in several other genes cause severe obesity in humans **(Table 394-1)**; each of these syndromes is rare. Mutations in the gene encoding proopiomelanocortin (POMC) cause severe obesity through failure to synthesize α-MSH, and potentially β-MSH, key neuropeptides that inhibit appetite in the hypothalamus. The absence of POMC also causes secondary adrenal insufficiency due to absence of adrenocorticotropic hormone (ACTH), as well as pale skin and red hair due to absence of α-MSH. Proenzyme convertase 1 (PC-1) mutations are thought to cause obesity by preventing synthesis of α-MSH from its precursor peptide, POMC. α-MSH binds to the type 4 melanocortin receptor (MC4R), a key hypothalamic receptor that inhibits eating. Heterozygous loss-of-function mutations of this receptor account for as much as 5% of severe obesity. Loss of function of MRAP2, a protein required for normal MC4R signaling, has been found in rare cases of severe obesity. These six genetic defects define a pathway through which leptin (by stimulating POMC and increasing α-MSH) restricts food intake and limits weight **(Fig. 394-5)**. The results of genomewide association studies to identify genetic loci responsible for obesity in the general population have so far been disappointing. More than 100 replicated loci linked to obesity have been identified, but together they account for <3% of interindividual variation in BMI. The most replicated of these involves a region which includes several genes including *FTO*, which is of unknown function, but like many of the other recently described candidates, is expressed in the brain. Because the heritability of obesity is estimated to be 40–70%, it is likely that many more loci remain to be identified. It is possible that epistatic interactions between

TABLE 394-1 Selected Obesity Genes in Humans and Mice

GENE	GENE PRODUCT	MECHANISM OF OBESITY	IN HUMAN	IN RODENT
Lep (ob)	Leptin, a fat-derived hormone	Mutation prevents leptin from delivering satiety signal; brain perceives starvation	Yes	Yes
LepR (db)	Leptin receptor	Same as above	Yes	Yes
POMC	Proopiomelanocortin, a precursor of several hormones and neuropeptides	Mutation prevents synthesis of melanocyte-stimulating hormone (MSH), a satiety signal	Yes	Yes
MC4R	Type 4 receptor for MSH	Mutation prevents reception of satiety signal from MSH	Yes	Yes
AgRP	Agouti-related peptide, a neuropeptide expressed in the hypothalamus	Overexpression inhibits signal through *MC4R*	No	Yes
PC-1	Prohormone convertase 1, a processing enzyme	Mutation prevents synthesis of neuropeptide, probably MSH	Yes	No
Fat	Carboxypeptidase E, a processing enzyme	Same as above	No	Yes
Tub	Tub, a hypothalamic protein of unknown function	Hypothalamic dysfunction	No	Yes
TrkB	TrkB, a neurotrophin receptor	Hyperphagia due to uncharacterized hypothalamic defect	Yes	Yes

causative loci or unknown gene-environment interactions explain the missing heritability of obesity.

In addition to these human obesity genes, studies in rodents reveal several other molecular candidates for hypothalamic mediators of human obesity or leanness. The *tub* gene encodes a hypothalamic peptide of unknown function; mutation of this gene causes late-onset obesity. The *fat* gene encodes carboxypeptidase E, a peptide-processing enzyme; mutation of this gene is thought to cause obesity by disrupting production of one or more neuropeptides. Mutations in TUB and Carboxypeptidase-E (CPE) have recently been identified in humans.

AgRP is coexpressed with NPY in arcuate nucleus neurons. AgRP antagonizes α-MSH action at MC4 receptors, and its overexpression induces obesity. In contrast, a mouse deficient in the peptide MCH, whose administration causes feeding, is lean.

A number of complex human syndromes with defined inheritance are associated with obesity **(Table 394-2)**. Although specific genes have limited definition at present, their identification may enhance our understanding of more common forms of human obesity. In the Prader-Willi syndrome, a multigenic neurodevelopmental disorder, obesity coexists with short stature, mental retardation, hypogonadotropic hypogonadism, hypotonia, small hands and feet, fish-shaped mouth, and hyperphagia. Most patients have reduced expression of imprinted paternally inherited genes encoded in the 15q11-13 chromosomal region. Reduced expression of Snord116, a small nucleolar RNA highly expressed in hypothalamus, may be an important cause of defective hypothalamic function in this disorder. Bardet-Biedl syndrome (BBS) is a genetically heterogeneous disorder characterized by obesity, mental retardation, retinitis pigmentosa, diabetes, renal and cardiac malformations, polydactyly, and hypogonadotropic hypogonadism. At least 16 genetic loci have been identified, and most of the encoded proteins form two multiprotein complexes that are involved in ciliary function and microtubule-based intracellular transport. Some evidence suggests that mutations might disrupt leptin receptor trafficking in key hypothalamic neurons, causing leptin resistance.

Other Specific Syndromes Associated with Obesity

CUSHING'S SYNDROME Although obese patients commonly have central obesity, hypertension, and glucose intolerance, they lack other specific stigmata of Cushing's syndrome **(Chap. 379)**. Nonetheless, a potential diagnosis of Cushing's syndrome is often entertained. Cortisol production and urinary metabolites (17OH steroids) may be increased in simple obesity. Unlike in Cushing's syndrome, however, cortisol levels in blood and urine in the basal state and in response to corticotropin-releasing hormone (CRH) or ACTH are normal; the overnight 1-mg dexamethasone suppression test is normal in 90%, with the remainder being normal on a standard 2-day low-dose dexamethasone suppression test. Obesity may be associated with excessive local reactivation of cortisol in fat by 11β-hydroxysteroid dehydrogenase 1, an enzyme that converts inactive cortisone to cortisol.

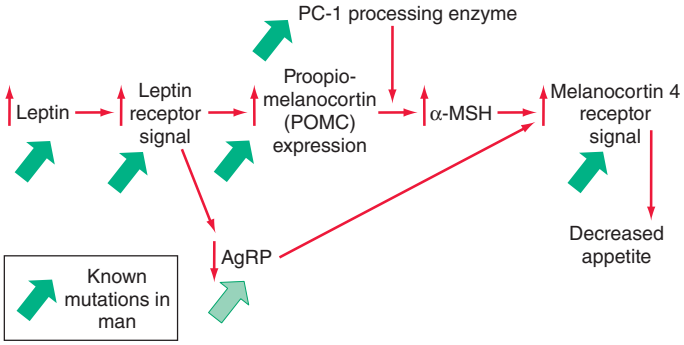


FIGURE 394-5 A central pathway through which leptin acts to regulate appetite and body weight. Leptin signals through proopiomelanocortin (POMC) neurons in the hypothalamus to induce increased production of α-melanocyte-stimulating hormone (α-MSH), requiring the processing enzyme PC-1 (proenzyme convertase 1). α-MSH acts as an agonist on melanocortin-4 receptors to inhibit appetite, and the neuropeptide AgRp (Agouti-related peptide) acts as an antagonist of this receptor. Mutations that cause obesity in humans are indicated by the *solid green arrows*.

HYPOTHYROIDISM The possibility of hypothyroidism should be considered, but it is an uncommon cause of obesity; hypothyroidism is easily ruled out by measuring thyroid-stimulating hormone (TSH). Much of the weight gain that occurs in hypothyroidism is due to myxedema **(Chap. 375)**.

INSULINOMA Patients with insulinoma often gain weight as a result of overeating to avoid hypoglycemic symptoms **(Chap. 399)**. The increased substrate plus high insulin levels promote energy storage in fat. This can be marked in some individuals but is modest in most.

CRANIOPHARYNGIOMA AND OTHER DISORDERS INVOLVING THE HYPOTHALAMUS Whether through tumors, trauma, or inflammation, hypothalamic dysfunction of systems controlling satiety, hunger, and energy expenditure can cause varying degrees of obesity **(Chap. 372)**. It is uncommon to identify a discrete anatomic basis for these disorders. Subtle hypothalamic dysfunction is probably a more common cause of obesity than can be documented using currently available imaging techniques. Growth hormone (GH), which exerts lipolytic activity, is diminished in obesity and is increased with weight loss. Despite low GH levels, insulin-like growth factor (IGF) I (somatomedin) production is normal, suggesting that GH suppression may be a compensatory response to increased nutritional supply.

Pathogenesis of Common Obesity

Obesity can result from increased energy intake, decreased energy expenditure, or a combination of the two. Thus, identifying the etiology of obesity should involve measurements of both parameters. However, it is difficult to perform direct and accurate measurements of energy intake in free-living individuals; and the obese, in particular, often underreport intake. Measurements of chronic energy expenditure are possible using doubly labeled water or metabolic chamber/rooms. In subjects at stable weight and body composition, energy intake equals expenditure. Consequently, these techniques allow assessment of energy intake in free-living individuals. The level of energy expenditure differs in established obesity, during periods of weight gain or loss, and in the pre- or postobese state. Studies that fail to take note of this phenomenon are not easily interpreted.

There is continued interest in the concept of a body weight "set point." This idea is supported by physiologic mechanisms centered around a sensing system in adipose tissue that reflects fat stores and a receptor, or "adipostat," that is in the hypothalamic centers. When fat stores are depleted, the adipostat signal is low, and the hypothalamus responds by stimulating hunger and decreasing energy expenditure to conserve energy. Conversely, when fat stores are abundant, the signal is increased, and the hypothalamus responds by decreasing hunger and increasing energy expenditure. The recent discovery of the *ob* gene, and its product leptin, and the *db* gene, whose product is the leptin receptor, provides important elements of a molecular basis for this physiologic concept (see above).

What Is the Status of Food Intake in Obesity? (Do the Obese Eat More Than the Lean?)

This question has stimulated much debate, due in part to the methodologic difficulties inherent in determining food intake. Many obese individuals believe that they eat small quantities of food, and this claim has often been supported by the results of food intake questionnaires. However, it is now established that average energy expenditure increases as individuals get more obese, due primarily to the fact that metabolically active lean tissue mass increases with obesity. Given the laws of thermodynamics, the obese person must therefore eat more than the average lean person to maintain their increased weight. It may be the case, however, that a subset of individuals who are predisposed to obesity have the capacity to become obese initially without an absolute increase in caloric consumption.

What Is the State of Energy Expenditure in Obesity?

The average total daily energy expenditure is higher in obese than lean individuals when measured at stable weight. However, energy expenditure falls as weight is lost, due in part to loss of lean body mass and to decreased sympathetic nerve activity. When reduced to near-normal weight and maintained there for a while, (some) obese individuals

TABLE 394-2 A Comparison of Syndromes of Obesity—Hypogonadism and Mental Retardation

	SYNDROME				
FEATURE	PRADER-WILLI	LAURENCE-MOON-BIEDL	AHLSTROM'S	COHEN'S	CARPENTER'S
Inheritance	Sporadic; two-thirds have defect	Autosomal recessive	Autosomal recessive	Probably autosomal recessive	Autosomal recessive
Stature	Short	Normal; infrequently short	Normal; infrequently short	Short or tall	Normal
Obesity	Generalized Moderate to severe Onset 1–3 years	Generalized Early onset, 1–2 years	Truncal Early onset, 2–5 years	Truncal Mid-childhood, age 5	Truncal, gluteal
Craniofacies	Narrow bifrontal diameter Almond-shaped eyes Strabismus V-shaped mouth High-arched palate	Not distinctive	Not distinctive	High nasal bridge Arched palate Open mouth Short philtrum	Acrocephaly Flat nasal bridge High-arched palate
Limbs	Small hands and feet Hypotonia	Polydactyly	No abnormalities	Hypotonia Narrow hands and feet	Polydactyly Syndactyly Genu valgum
Reproductive status	1° Hypogonadism	1° Hypogonadism	Hypogonadism in males but not in females	Normal gonadal function or hypogonadotropic hypogonadism	2° Hypogonadism
Other features	Enamel hypoplasia Hyperphagia Temper tantrums Nasal speech			Dysplastic ears Delayed puberty	
Mental retardation	Mild to moderate		Normal intelligence	Mild	Slight

have lower energy expenditure than (some) lean individuals. There is also a tendency for those who will develop obesity as infants or children to have lower resting energy expenditure rates than those who remain lean. The physiologic basis for variable rates of energy expenditure (at a given body weight and level of energy intake) is essentially unknown.

Another component of thermogenesis, called *nonexercise activity thermogenesis* (NEAT), has been linked to obesity. It is the thermogenesis that accompanies physical activities other than volitional exercise such as the activities of daily living, fidgeting, spontaneous muscle contraction, and maintaining posture. NEAT accounts for about two-thirds of the increased daily energy expenditure induced by overfeeding. The wide variation in fat storage seen in overfed individuals is predicted by the degree to which NEAT is induced. The molecular basis for NEAT and its regulation is unknown.

Leptin in Typical Obesity The vast majority of obese persons have increased leptin levels but do not have mutations of either leptin or its receptor. They appear, therefore, to have a form of functional "leptin resistance." Data suggesting that some individuals produce less leptin per unit fat mass than others or have a form of relative leptin deficiency that predisposes to obesity are at present contradictory and unsettled. The mechanism for leptin resistance, and whether it can be overcome by raising leptin levels or combining leptin with other treatments in a subset of obese individuals, is not yet established. Some data suggest that leptin may not effectively cross the blood-brain barrier as levels rise. It is also apparent from animal studies that leptin-signaling inhibitors, such as SOCS3 and PTP1b, are involved in the leptin-resistant state.

PATHOLOGIC CONSEQUENCES OF OBESITY (SEE ALSO CHAP. 395)

Obesity has major adverse effects on health. Obesity is associated with an increase in mortality, with a 50–100% increased risk of death from all causes compared to normal-weight individuals, mostly due to cardiovascular causes. Obesity and overweight together are the second leading cause of preventable death in the United States, accounting for 300,000 deaths per year. Mortality rates rise as obesity increases, particularly when obesity is associated with increased intraabdominal fat (see above). Life expectancy of a moderately obese individual could be shortened by 2–5 years, and a 20- to 30-year-old male with a BMI >45 may lose 13 years of life. It is likely that the degree to which obesity affects particular organ systems is influenced by susceptibility genes that vary in the population.

Insulin Resistance and Type 2 Diabetes Mellitus Hyperinsulinemia and insulin resistance are pervasive features of obesity, increasing with weight gain and diminishing with weight loss (Chap. 401). Insulin resistance is more strongly linked to intraabdominal fat than to fat in other depots. Molecular links between obesity and insulin resistance in fat, muscle, and liver have been sought for many years. Major factors include: (1) insulin itself, by inducing receptor downregulation; (2) free fatty acids that are increased and capable of impairing insulin action; (3) intracellular lipid accumulation; and (4) several circulating peptides produced by adipocytes, including the cytokines TNF-α and IL-6, RBP4, and the "adipokine" adiponectin, which have altered expression in obese adipocytes and can modify insulin action. Additional mechanisms are obesity-linked inflammation, including infiltration of macrophages into tissues including fat, and induction of the endoplasmic reticulum stress response, which can bring about resistance to insulin action in cells. Despite the prevalence of insulin resistance, most obese individuals do not develop diabetes, suggesting that diabetes requires an interaction between obesity-induced insulin resistance and other factors such as impaired insulin secretion (Chap. 396). Obesity, however, is a major risk factor for diabetes, and as many as 80% of patients with type 2 diabetes mellitus are obese. Weight loss and exercise, even of modest degree, increase insulin sensitivity and often improve glucose control in diabetes.

Reproductive Disorders Disorders that affect the reproductive axis are associated with obesity in both men and women. Male hypogonadism is associated with increased adipose tissue, often distributed in a pattern more typical of females. In men whose weight is >160% ideal body weight (IBW), plasma testosterone and sex hormone–binding globulin (SHBG) are often reduced, and estrogen levels (derived from conversion of adrenal androgens in adipose tissue) are increased (Chap. 384). Gynecomastia may be seen. However, masculinization, libido, potency, and spermatogenesis are preserved in most of these individuals. Free testosterone may be decreased in morbidly obese men whose weight is >200% IBW.

Obesity has long been associated with menstrual abnormalities in women, particularly in women with upper body obesity (Chap. 385). Common findings are increased androgen production, decreased SHBG, and increased peripheral conversion of androgen to estrogen. Most obese women with oligomenorrhea have polycystic ovarian syndrome (PCOS), with its associated anovulation and ovarian hyperandrogenism; 40% of women with PCOS are obese. Most nonobese women with PCOS are also insulin-resistant, suggesting that insulin resistance, hyperinsulinemia, or the combination of the two are causative or contribute to the ovarian pathophysiology in PCOS in both obese and lean individuals. In obese women with PCOS, weight loss often restores normal menses. The increased conversion of androstenedione to estrogen, which occurs to a greater degree in women with lower body obesity, may contribute to the increased incidence of uterine cancer in postmenopausal women with obesity.

Cardiovascular Disease The Framingham Study revealed that obesity was an independent risk factor for the 26-year incidence of cardiovascular disease in men and women (including coronary disease, stroke, and congestive heart failure). The waist-to-hip ratio may be the best predictor of these risks. When the additional effects of hypertension and glucose intolerance associated with obesity are included, the adverse impact of obesity is even more evident. The effect of obesity on cardiovascular mortality in women may be seen at BMIs as low as 25. Obesity, especially abdominal obesity, is associated with an atherogenic lipid profile; with increased low-density lipoprotein cholesterol, very-low-density lipoprotein, and triglyceride; and with decreased high-density lipoprotein cholesterol and decreased levels of the vascular protective adipokine adiponectin (Chap. 400). Obesity is also associated with hypertension. Measurement of blood pressure in the obese requires use of a larger cuff size to avoid artifactual increases. Obesity-induced hypertension is associated with increased peripheral resistance and cardiac output, increased sympathetic nervous system tone, increased salt sensitivity, and insulin-mediated salt retention; it is often responsive to modest weight loss.

Pulmonary Disease Obesity may be associated with a number of pulmonary abnormalities. These include reduced chest wall compliance, increased work of breathing, increased minute ventilation due to increased metabolic rate, and decreased functional residual capacity and expiratory reserve volume. Severe obesity may be associated with obstructive sleep apnea and the "obesity hypoventilation syndrome" with attenuated hypoxic and hypercapnic ventilatory responses. Sleep apnea can be obstructive (most common), central, or mixed and is associated with hypertension. Weight loss (10–20 kg) can bring substantial improvement, as can major weight loss following gastric bypass or restrictive surgery. Continuous positive airway pressure has been used with some success.

Hepatobiliary Disease Obesity is frequently associated with nonalcoholic fatty liver disease (NAFLD), and this association represents one of the most common causes of liver disease in industrialized countries. The hepatic fatty infiltration of NAFLD progresses in a subset to inflammatory nonalcoholic steatohepatitis (NASH) and more rarely to cirrhosis and hepatocellular carcinoma. Steatosis typically improves following weight loss, secondary to diet or bariatric surgery. The mechanism for the association remains unclear. Obesity is associated with enhanced biliary secretion of cholesterol, supersaturation of bile, and a higher incidence of gallstones, particularly cholesterol gallstones (Chap. 339). A person 50% above IBW has about a sixfold increased incidence of symptomatic gallstones. Paradoxically, fasting increases supersaturation of bile by decreasing the phospholipid component. Fasting-induced cholecystitis is a complication of extreme diets.

Cancer Obesity is associated with increased risk of several cancer types, and in addition can lead to poorer treatment outcomes and increased cancer mortality. Obesity in males is associated with higher mortality from cancer of the esophagus, colon, rectum, pancreas, liver, and prostate; obesity in females is associated with higher mortality from cancer of the gallbladder, bile ducts, breasts, endometrium, cervix, and ovaries. Some of the latter may be due to increased rates of conversion of androstenedione to estrone in adipose tissue of obese individuals. Other possible mechanistic links may involve hormones, growth factors, and cytokines whose levels are linked to nutritional state, including insulin, leptin, adiponectin, and IGF-I, as well as activation of signaling pathways linked to both obesity and cancer. It has been estimated that obesity accounts for 14% of cancer deaths in men and 20% in women in the United States.

Bone, Joint, and Cutaneous Disease Obesity is associated with an increased risk of osteoarthritis, no doubt partly due to the trauma of added weight bearing, but potentially linked as well to activation of inflammatory pathways that could promote synovial pathology. The prevalence of gout may also be increased (Chap. 365). One of the skin problems associated with obesity is acanthosis nigricans, manifested by darkening and thickening of the skinfolds on the neck, elbows, and dorsal interphalangeal spaces. Acanthosis reflects the severity of underlying insulin resistance and diminishes with weight loss. Friability of skin may be increased, especially in skinfolds, enhancing the risk of fungal and yeast infections. Finally, venous stasis is increased in the obese.

FURTHER READING

Allison MB, Myers MG: 20 years of leptin: Connecting leptin signaling to biological function. J Endocrinol 223:T25, 2014.

Farooqi IS, O'Rahilly S: 20 years of leptin: Human disorders of leptin action. J Endocrinol 223:T63, 2014.

Heymsfield SB, Wadden TA: Mechanisms, pathophysiology and management of obesity. N Engl J Med 376:254, 2017.

Ogden CL et al: Prevalence of obesity in the United States. JAMA 312:189, 2014.

395 Evaluation and Management of Obesity

Robert F. Kushner

More than 66% of U.S. adults are categorized as overweight or obese, and the prevalence of obesity is increasing rapidly in most of the industrialized world. Children and adolescents also are becoming more obese, indicating that the current trends will accelerate over time. Obesity is associated with an increased risk of multiple health problems, including hypertension, type 2 diabetes, dyslipidemia, obstructive sleep apnea, nonalcoholic fatty liver disease, degenerative joint disease, and some malignancies. Thus, it is important for physicians to identify, evaluate, and treat patients for obesity and associated comorbid conditions.

EVALUATION

Physicians should screen all adult patients for obesity and offer intensive counseling and behavioral interventions to promote sustained weight loss. The five main steps in the evaluation of obesity, as described below, are (1) a focused obesity-related history, (2) a physical examination to determine the degree and type of obesity, (3) assessment of comorbid conditions, (4) determination of fitness level, and (5) assessment of the patient's readiness to adopt lifestyle changes.

The Obesity-Focused History Information from the history should address the following seven questions:

- What factors contribute to the patient's obesity?
- How is the obesity affecting the patient's health?
- What is the patient's level of risk from obesity?
- What does the patient find difficult about managing weight?

- What are the patient's goals and expectations?
- Is the patient motivated to begin a weight management program?
- What kind of help does the patient need?

Although the vast majority of cases of obesity can be attributed to behavioral factors that affect diet and physical activity patterns, the history may suggest secondary causes that merit further evaluation. Disorders to consider include polycystic ovarian syndrome, hypothyroidism, Cushing's syndrome, and hypothalamic disease. Drug-induced weight gain also should be considered. Common causes include medications for diabetes (insulin, sulfonylureas, thiazolidinediones); steroid hormones; antipsychotic agents (clozapine, olanzapine, risperidone); mood stabilizers (lithium); antidepressants (tricyclics, monoamine oxidase inhibitors, paroxetine, mirtazapine); and antiepileptic drugs (valproate, gabapentin, carbamazepine). Other medications, such as nonsteroidal anti-inflammatory drugs and calcium channel blockers, may cause peripheral edema but do not increase body fat.

The patient's current diet and physical activity patterns may reveal factors that contribute to the development of obesity and may identify behaviors to target for treatment. This type of historic information is best obtained by the combination of a questionnaire and an interview.

Body Mass Index (BMI) and Waist Circumference Three key anthropometric measurements are important in evaluating the degree of obesity: weight, height, and waist circumference. The BMI, calculated as weight (kg)/height $(m)^2$ or as weight (lb)/height $(in)^2 \times 703$, is used to classify weight status and risk of disease (Table 395-1). BMI provides an estimate of body fat and is related to disease risk. Lower BMI thresholds for overweight and obesity have been proposed for the Asia-Pacific region since this population appears to be at risk for glucose and lipid abnormalities at lower body weights.

Excess abdominal fat, assessed by measurement of waist circumference or waist-to-hip ratio, is independently associated with a higher risk for diabetes mellitus and cardiovascular disease. Measurement of the waist circumference is a surrogate for visceral adipose tissue and should be performed in the horizontal plane above the iliac crest (Table 395-2).

Physical Fitness Several prospective studies have demonstrated that physical fitness, reported by questionnaire or measured by a maximal treadmill exercise test, is an important predictor of all-cause mortality rate independent of BMI and body composition. These observations highlight the importance of taking a physical activity and exercise history during examination as well as emphasizing physical activity as a treatment approach.

Obesity-Associated Comorbid Conditions The evaluation of comorbid conditions should be based on presentation of symptoms, risk factors, and index of suspicion. For all patients, a fasting lipid panel should be performed (total, low-density lipoprotein, and high-density lipoprotein cholesterol and triglyceride levels) and a fasting blood glucose level and blood pressure determined. Symptoms and diseases that are directly or indirectly related to obesity are listed in Table 395-3. Although individuals vary, the number and severity of organ-specific comorbid conditions usually rise with increasing levels of obesity. Patients at very high absolute risk include those with the following: established coronary heart disease; presence of other atherosclerotic diseases, such as peripheral arterial disease, abdominal aortic aneurysm, and symptomatic carotid artery disease; type 2 diabetes; and sleep apnea.

Identifying the High-Risk Patient Efforts are under way to develop more practical and useful assessments to identify patients

TABLE 395-1 Classification of Weight Status and Disease Risk

CLASSIFICATION	BODY MASS INDEX (kg/m^2)	OBESITY CLASS	DISEASE RISK
Underweight	<18.5	—	—
Healthy weight	18.5–24.9	—	—
Overweight	25.0–29.9	—	Increased
Obesity	30.0–34.9	I	High
Obesity	35.0–39.9	II	Very high
Extreme obesity	≥40	III	Extremely high

Source: Adapted from the National Institutes of Health, National Heart, Lung, and Blood Institute: *Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults.* U.S. Department of Health and Human Services, U.S. Public Health Service, 1998.

TABLE 395-2 Ethnic-Specific Cutpoint Values for Waist Circumference

ETHNIC GROUP	WAIST CIRCUMFERENCE
Europeans	
Men	>94 cm (>37 in)
Women	>80 cm (>31.5 in)
South Asians and Chinese	
Men	>90 cm (>35 in)
Women	>80 cm (>31.5 in)
Japanese	
Men	>85 cm (>33.5 in)
Women	>90 cm (>35 in)
Ethnic South and Central Americans	Use South Asian recommendations until more specific data are available.
Sub-Saharan Africans	Use European data until more specific data are available.
Eastern Mediterranean and Middle Eastern (Arab) populations	Use European data until more specific data are available.

Source: From KGMM Alberti et al for the IDF Epidemiology Task Force Consensus Group: Lancet 366:1059, 2005.

TABLE 395-3 Obesity-Related Organ Systems Review

Cardiovascular	**Respiratory**
Hypertension	Dyspnea
Congestive heart failure	Obstructive sleep apnea
Cor pulmonale	Hypoventilation syndrome
Varicose veins	Pickwickian syndrome
Pulmonary embolism	Asthma
Coronary artery disease	**Gastrointestinal**
Endocrine	Gastroesophageal reflux disease
Metabolic syndrome	Nonalcoholic fatty-liver disease
Type 2 diabetes	Cholelithiasis
Dyslipidemia	Hernias
Polycystic ovarian syndrome	Colon cancer
Musculoskeletal	**Genitourinary**
Hyperuricemia and gout	Urinary stress incontinence
Immobility	Obesity-related glomerulopathy
Osteoarthritis (knees and hips)	Hypogonadism (male)
Low back pain	Breast and uterine cancer
Carpal tunnel syndrome	Pregnancy complications
Psychological	**Neurologic**
Depression/low self-esteem	Stroke
Body image disturbance	Idiopathic intracranial hypertension
Social stigmatization	Meralgia paresthetica
Integument	Dementia
Striae distensae	
Stasis pigmentation of legs	
Lymphedema	
Cellulitis	
Intertrigo, carbuncles	
Acanthosis nigricans	
Acrochordons (skin tags)	
Hidradenitis suppurativa	

who are at high risk in addition to using BMI alone. Analogous to other staging systems commonly used for congestive heart failure or chronic kidney disease, the American Society of Clinical Endocrinologists (AACE) and the American College of Endocrinology (ACE) have proposed an obesity disease staging system that is based on ethnic-specific BMI cutoffs in conjunction with assessment for adiposity-related complications. Stage 0 is assigned to individuals who are overweight or obese by BMI classification but have no complications, whereas stages 1 and 2 are defined as individuals who are overweight or obese by BMI classification and have one or more mild-moderate complications (stage 1) or at least one severe complication (stage 2). A different functional staging system for obesity, called the Edmonton Obesity Staging System (EOSS), classifies individuals with obesity into five graded categories (0–4), based on their morbidity and health-risk profile along three domains—medical, functional, and mental. In this system, staging occurs independent of BMI.

Assessing the Patient's Readiness to Change An attempt to initiate lifestyle changes when the patient is not ready usually leads to frustration and may hamper future weight-loss efforts. Assessment includes patient motivation and support, stressful life events, psychiatric status, time availability and constraints, and appropriateness of goals and expectations. Readiness can be viewed as the balance of two opposing forces: (1) motivation, or the patient's desire to change; and (2) resistance, or the patient's resistance to change.

A helpful method to begin a readiness assessment is to use the motivational interviewing technique of "anchoring" the patient's interest and confidence to change on a numerical scale. With this technique, the patient is asked to rate—on a scale from 0 to 10, with 0 being not so important (or confident) and 10 being very important (or confident)—his or her level of interest in and confidence about losing weight at this time. This exercise helps establish readiness to change and also serves as a basis for further dialogue.

TREATMENT

Obesity

THE GOAL OF THERAPY

The primary goals of treatment are to improve obesity-related comorbid conditions and reduce the risk of developing future comorbidities. Information obtained from the history, physical examination, and diagnostic tests is used to determine risk and develop a treatment plan **(Fig. 395-1)**. The decision of how aggressively to treat the patient and which modalities to use is determined by the patient's risk status, expectations, and available resources. Not all patients who are deemed obese by BMI alone need to be treated, as exemplified by the concepts of obesity paradox or the metabolically healthy obese. However, patients who present with obesity-related comorbidities and who would benefit from weight-loss intervention should be managed proactively. Therapy for obesity always begins with lifestyle management and may include pharmacotherapy or surgery, depending on BMI risk category **(Table 395-4)**. Setting an initial weight-loss goal of 8–10% over 6 months is a realistic target.

LIFESTYLE MANAGEMENT

Obesity care involves attention to three essential elements of lifestyle: dietary habits, physical activity, and behavior modification. Because obesity is fundamentally a disease of energy imbalance, all patients must learn how and when energy is consumed (diet), how and when energy is expended (physical activity), and how to incorporate this information into their daily lives (behavioral therapy). Lifestyle management has been shown to result in a modest (typically 3–5 kg) weight loss when compared with no treatment or usual care.

Diet Therapy The primary focus of diet therapy is to reduce overall calorie consumption. Guidelines from the American Heart Association/American College of Cardiology/The Obesity Society (AHA/ACC/TOS) recommend initiating treatment with a calorie deficit of 500–750 kcal/d compared with the patient's habitual diet. Alternatively, a diet of 1200–1500 kcal/d for women and 1500–1800 kcal/d for men (adjusted for the individual's body weight) can be prescribed. This reduction is consistent with a goal of losing ~1–2 lb/week. The calorie deficit can be instituted through dietary substitutions or alternatives. Examples include choosing smaller portion sizes, eating more fruits and vegetables, consuming more whole-grain cereals, selecting leaner cuts of meat and skimmed dairy products, reducing consumption of fried foods and other foods with added fats and oils, and drinking water instead of sugar-sweetened beverages. It is important that dietary counseling remains patient centered and that the selected goals are SMART (specific, measurable, agreed upon, realistic, timely).

The macronutrient composition of the diet will vary with the patient's preference and medical condition. The 2015 U.S. Department of Agriculture Dietary Guidelines for Americans **(Chap. 325)**, which focus on health promotion and risk reduction, can be applied to treatment of patients who are overweight or obese. The recommendations include maintaining a diet rich in whole grains, fruits, vegetables, and dietary fiber; decreasing sodium intake to <2300 mg/d; consuming fat-free or low-fat dairy products; and keeping added sugars and saturated fat intake to <10% of daily calories. Application of these guidelines to specific calorie goals can be found on the website *www.choosemyplate.gov.* Since portion control is one of the most difficult strategies for patients to manage, the use of preprepared products such as meal replacements is a simple and convenient suggestion. Examples include frozen entrees, canned beverages, and bars. Use of meal replacements in the diet has been shown to result in a 7–8% weight loss.

Numerous randomized trials comparing diets of different macronutrient composition (e.g., low-carbohydrate, low-fat, Mediterranean) have shown that weight loss depends primarily on reduction of total caloric intake and adherence to the prescribed diet, not the specific proportions of carbohydrate, fat, and protein in the diet. The macronutrient composition will ultimately be determined by the patient's taste preferences, cooking style, and culture. However, the patient's underlying medical problems are also important in guiding the recommended dietary composition. The dietary prescription will vary according to the patient's metabolic profile and risk factors. A consultation with a registered dietitian for medical nutrition therapy is particularly useful in considering patient preference and treatment of comorbid diseases.

Another dietary approach to consider is based on the concept of *energy density*, which refers to the number of calories (i.e., amount of energy) a food contains per unit of weight. People tend to ingest a constant volume of food regardless of caloric or macronutrient content. Adding water or fiber to a food decreases its energy density by increasing weight without affecting caloric content. Examples of foods with low-energy density include soups, fruits, vegetables, oatmeal, and lean meats. Dry foods and high-fat foods such as pretzels, cheese, egg yolks, potato chips, and red meat have a high-energy density. Diets containing low-energy-dense foods have been shown to control hunger and thus to result in decreased caloric intake and weight loss.

Occasionally, very low-calorie diets (VLCDs) are prescribed as a form of aggressive dietary therapy. The primary purpose of a VLCD is to promote a rapid and significant (13- to 23-kg) short-term weight loss over a 3- to 6-month period. The proprietary formulas designed for this purpose typically supply ≤800 kcal, 50–80 g of protein, and 100% of the recommended daily intake for vitamins and minerals. According to a review by the National Task Force on the Prevention and Treatment of Obesity, indications for initiating a VLCD include the involvement of well-motivated individuals who are moderately to severely obese (BMI, >30 kg/m^2), have failed at more conservative approaches to weight loss, and have a medical condition that would be immediately improved with rapid weight loss. These conditions include poorly controlled type 2 diabetes, hypertriglyceridemia, obstructive sleep apnea, and symptomatic peripheral edema. The risk for gallstone formation increases

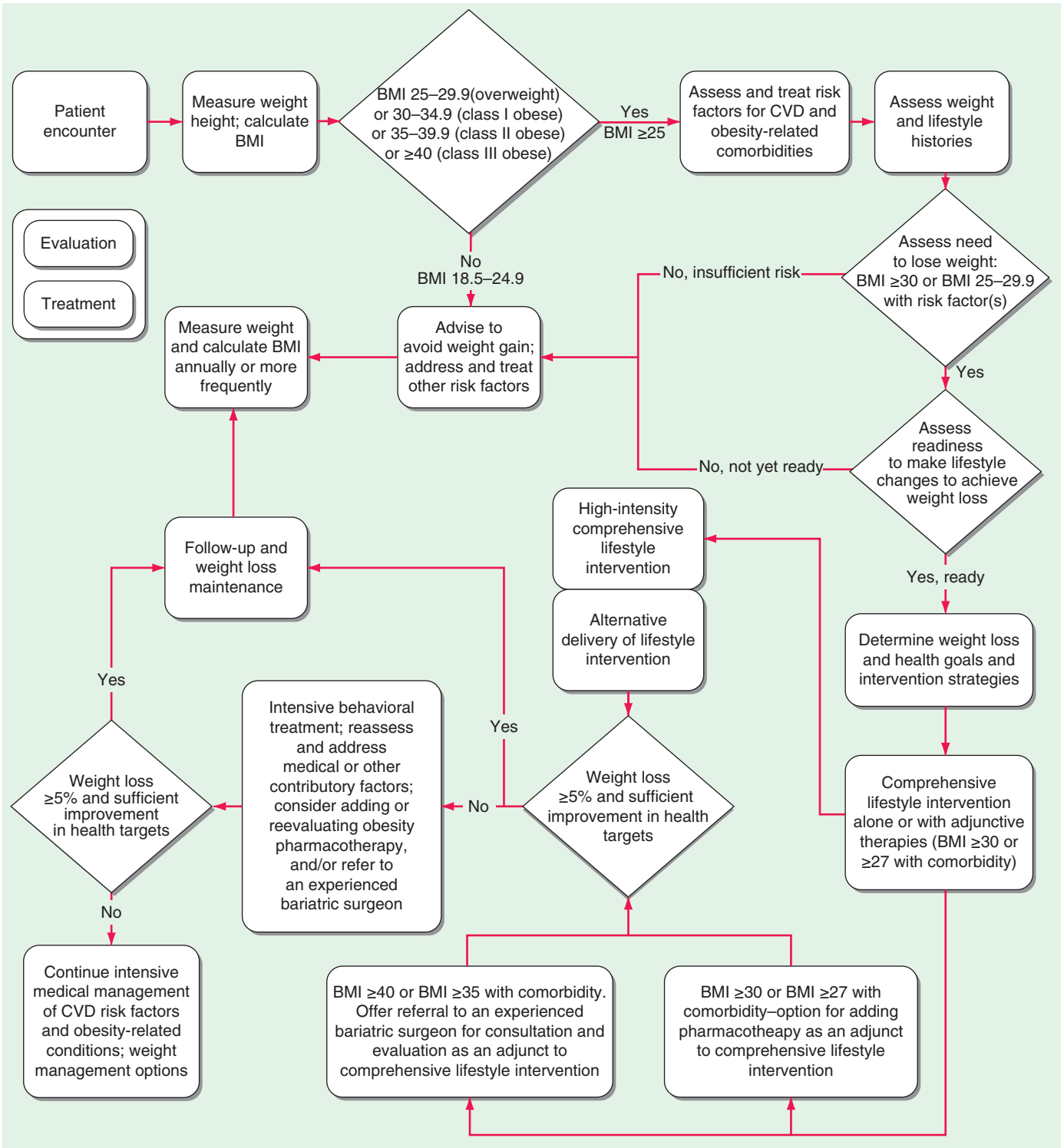


FIGURE 395-1 Treatment algorithm—chronic disease management model for primary care of patients with overweight and obesity. This algorithm applies to the assessment of overweight and obesity and subsequent decisions based on that assessment. BMI indicates body mass index; CVD, cardiovascular disease; FDA, U.S. Food and Drug Administration. *(From Jensen MD et al: 2013 AHA/ACC/TOS guideline for the management of overweight and obesity in adults: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and The Obesity Society. Circulation. 129[suppl 2]:S102, 2014.)*

TABLE 395-4 A Guide to Opting for Treatment for Obesity

	BMI CATEGORY (kg/m²)				
TREATMENT	**25–26.9**	**27–29.9**	**30–34.9**	**35–39.9**	**≥40**
Diet, exercise, behavioral therapy	With comorbidities	With comorbidities	+	+	+
Pharmacotherapy	—	With comorbidities	+	+	+
Surgery	—	—	—	With comorbidities	+

Source: From the National Heart, Lung, and Blood Institute, North American Association for the Study of Obesity (2000).

exponentially at rates of weight loss >1.5 kg/week (3.3 lb/week). Prophylaxis against gallstone formation with ursodeoxycholic acid (600 mg/d) is effective in reducing this risk. VLCDs should be used only in limited circumstances and only when provided by trained practitioners in a medical care setting where medical monitoring and high intensity lifestyle intervention can be provided. Medical supervision is required because of the rapid rate of weight loss and potential for health complications.

Physical Activity Therapy Although exercise alone is only moderately effective for weight loss, the combination of dietary modification and exercise is the most effective behavioral approach for the treatment of obesity. The most important role of exercise appears to be in the maintenance of the weight loss. The 2008 Physical Activity Guidelines for Americans (*www.health.gov/paguidelines*) recommend that adults should engage in 150 min of moderate-intensity or 75 min a week of vigorous-intensity aerobic physical activity per week, performed in episodes of at least 10 min and preferably spread throughout the week. Focusing on simple ways to add physical activity into the normal daily routine through leisure activities, travel, and domestic work should be suggested. Examples include brisk walking, using the stairs, doing housework and yard work, and engaging in sports. Asking the patient to wear a pedometer or accelerometer to monitor total accumulation of steps or kcal expended as part of the activities of daily living is a useful strategy. Step counts are highly correlated with activity level. Studies have demonstrated that lifestyle activities are as effective as structured exercise programs for improving cardiorespiratory fitness and weight loss. A high level of physical activity (>300 min of moderate-intensity activity per week) is often needed to lose weight and sustain weight loss. These exercise recommendations are daunting to most patients and need to be implemented gradually. Consultation with an exercise physiologist or personal trainer may be helpful.

Behavioral Therapy Cognitive behavioral therapy is used to help change and reinforce new dietary and physical activity behaviors. Strategies include self-monitoring techniques (e.g., journaling, weighing, and measuring food and activity); stress management; stimulus control (e.g., using smaller plates, not eating in front of the television or in the car); social support; problem solving; and cognitive restructuring to help patients develop more positive and realistic thoughts about themselves. When recommending any behavioral lifestyle change, the patient should be asked to identify what, when, where, and how the behavioral change will be performed. The patient should keep a record of the anticipated behavioral change so that progress can be reviewed at the next office visit. Because these techniques are time consuming to implement, their supervision is often undertaken by ancillary office staff, such as a nurse-clinician or registered dietitian.

PHARMACOTHERAPY

Adjuvant pharmacologic treatments should be considered for patients with a BMI ≥30 kg/m^2 or for patients with a BMI ≥27 kg/m^2 who have concomitant obesity-related diseases and for whom dietary and physical activity therapy has not been successful. When an antiobesity medication is prescribed, patients should be actively engaged in a lifestyle program that provides the strategies and skills needed to use the drug effectively, since such support increases total weight loss.

Medications for obesity have traditionally fallen into two major categories: appetite suppressants (*anorexiants*) and gastrointestinal fat blockers.. Four new antiobesity medications have been approved by the U.S. Food and Drug Administration (FDA) since 2012: lorcaserin, phentermine/topiramate (PHEN/TPM) extended release, naltrexone sustained release (SR)/bupropion SR, and liraglutide. Gastrointestinal fat blockers reduce the absorption of selective macronutrients, such as fat, from the gastrointestinal tract.

Centrally Acting Anorexiant Medications Anorexiants affect *satiety* (the absence of hunger after eating) and hunger (the biologic sensation that prompts eating). By increasing satiety and decreasing hunger, these agents help patients reduce caloric intake without a sense of deprivation. The target site for the actions of anorexiants is the ventromedial and lateral hypothalamic regions in the central nervous system **(Chap. 394)**. The biologic effect of these agents on appetite regulation is produced by augmentation of the neurotransmission of three monoamines: norepinephrine; serotonin (5-hydroxytryptamine [5-HT]); and, to a lesser degree, dopamine. The classic sympathomimetic adrenergic agents (benzphetamine, phendimetrazine, diethylpropion, mazindol, and phentermine) function by stimulating norepinephrine release or by blocking its reuptake. Among the anorexiants, phentermine is the most commonly prescribed; there are limited long-term data on its effectiveness. A 2002 review of six randomized, placebo-controlled trials of phentermine for weight control found that patients lost 0.6–6.0 additional kg of weight over 2–24 weeks of treatment. The most common side effects of the amphetamine-derived anorexiants are restlessness, insomnia, dry mouth, constipation, and increased blood pressure and heart rate.

PHEN/TPM is a combination drug that contains a catecholamine releaser (phentermine) and an anticonvulsant (topiramate). Topiramate is approved by the FDA as an anticonvulsant for the treatment of epilepsy and for the prophylaxis of migraine headaches. Weight loss was identified as an unintended side effect of topiramate during clinical trials for epilepsy. The mechanism responsible for weight loss is uncertain but is thought to be mediated through the drug's modulation of γ-aminobutyric acid receptors, inhibition of carbonic anhydrase, and antagonism of glutamate. PHEN/TPM has undergone two 1-year pivotal randomized, placebo-controlled, double-blind trials of efficacy and safety: EQUIP and CONQUER. In a third study, SEQUEL, 78% of CONQUER participants continued to receive their blinded treatment for an additional year. All participants received diet and exercise counseling. Participant numbers, eligibility, characteristics, and weight-loss outcomes are displayed in **Table 395-5**. Intention-to-treat 1-year placebo-subtracted weight loss for PHEN/TPM was 9.3% (15-mg/92-mg dose) and 6.6% (7.5-mg/46-mg dose), respectively, in the EQUIP and CONQUER trials. Clinical and statistical dose-dependent improvements were seen in selected cardiovascular and metabolic outcome measurements that were related to the weight loss. The most common adverse events experienced by the drug-randomized group were paresthesias, dry mouth, constipation, dysgeusia, and insomnia. Because of an increased risk of congenital fetal oral-cleft formation from topiramate, women of childbearing age should have a negative pregnancy test before treatment and monthly thereafter, and use effective contraception consistently during medication therapy.

Lorcaserin is a selective 5-HT2$_C$ receptor agonist with a functional selectivity ~15 times that of 5-HT2$_A$ receptors and 100 times that of 5-HT2$_B$ receptors. This selectivity is important, since the drug-induced valvulopathy documented with two other serotonergic agents that were removed from the market—fenfluramine and dexfenfluramine—was due to activation of the 5-HT2$_B$ receptors expressed on cardiac valvular interstitial cells. By activating the 5-HT2$_C$ receptor, lorcaserin is thought to decrease food intake through the pro-opiomelanocortin (POMC) system of neurons.

Lorcaserin has undergone two randomized, placebo-controlled, double-blind trials for efficacy and safety. Participants were randomized to receive lorcaserin (10 mg bid) or placebo in the BLOOM study and lorcaserin (10 mg bid or qd) or placebo in the BLOSSOM study. All participants received diet and exercise counseling. Participant numbers, eligibility, characteristics, and weight-loss outcomes are displayed in Table 395-5. Patients who were overweight or obese had at least one coexisting condition (hypertension, dyslipidemia, cardiovascular disease, impaired glucose tolerance, or sleep apnea)—medical conditions that are commonly seen in the office setting. Intention-to-treat 1-year placebo-subtracted weight loss was 3.6% and 3.0%, respectively, in the BLOOM and BLOSSOM trials. Echocardiography was performed at the screening visit and

TABLE 395-5 Clinical Trials for Antiobesity Medications

	PHEN/TPM		LORCASERIN		NALTREXONE SR/BUPROPION SR			LIRAGLUTIDE	
	EQUIP	CONQUER	BLOOM	BLOSSOM	COR-I	COR-II	COR-BMOD	SCALE	SCALE MAINTENANCE
No. of participants (ITT-LOCF)	1230	2487	3182	4008	1742	1496	793	3731	422
BMI (kg/m^2)	≥35	27–45	27–45	30–45	30–45	30–45	30–45	≥27	≥27
Age (y)	18–70	18–70	18–65	18–65	18–65	18–65	18–65	≥18	≥18
Comorbid conditions (cardiovascular and metabolic)	≥1	≥2	≥1	≥1	≥1	≥1	≥1	≥1	≥1
Mean weight loss (%) with treatment vs placebo	10.9 vs 1.6	7.8 vs 1.2	5.8 vs 2.2	4.8 vs 2.8	6.1 vs 1.3	6.5 vs 1.9	9.3 vs 5.1	8.0 vs 2.6	6.2 vs 0.2
Placebo-subtracted weight loss (%)	9.3	6.6	3.6	3.0	4.8	4.6	4.2	5.4	6.0
Categorical change in 5% weight loss with treatment vs placebo	66.7 vs 17.3	62 vs 21	47.5 vs 20.3	47.2 vs 25	48 vs 16	50.5 vs 17.1	66.4 vs 42.5	63.2 vs 27.1	81.4 vs 48.9
Study completion rate, treatment vs placebo (%)	66.4 vs 52.9	69 vs 57	55.4 vs 45.1	57.2 vs 52	50	54	57.9 vs 58.4	71.9 vs 64.4	75 vs 69.5

Note: EQUIP (PHEN/TPM = 15/92 mg dose; CONQUER (PHEN/TPM = 7.5/46 mg dose).

Abbreviations: BMI, body mass index; ITT-LOCF, intention to treat, last observation carried forward; PHEN-TPM, phentermine-topiramate extended release.

at scheduled time points over the course of the studies. There was no difference in the development of FDA-defined valvulopathy between drug-treated and placebo-treated participants at 1 or 2 years. Modest statistical improvements consistent with the weight loss were seen in selected cardiovascular and metabolic outcome measurements. The most common adverse events experienced by the drug group were headache, dizziness, and nausea.

Naltrexone SR/bupropion SR (NB) is a combination of an opioid antagonist and a mild reuptake inhibitor of dopamine and norepinephrine, respectively. Individually, naltrexone is approved by the FDA for the treatment of alcohol dependence and for the blockade of the effects of exogenously administered opioids, whereas bupropion is approved as an antidepressant and smoking cessation aid. As a combination drug, each component works in consort: bupropion stimulates secretion of α-melanocyte stimulating hormone (MSH) from POMC whereas naltrexone blocks the feedback inhibitory effects of opioid receptors activated by the β-endorphin released in the hypothalamus, thus allowing the inhibitory effects of MSH to reduce food intake.

The medication has undergone three randomized, placebo-controlled, double-blind trials for efficacy and safety. Participants were randomized to receive NB (8 mg/90 mg two tablets bid) or placebo in the three COR studies. Whereas participants received standardized nutritional and exercise counseling in COR-I and COR-II, a more intensive behavior modification program was provided in COR-BMOD (Table 395-5). Intention-to-treat 1-year placebo-subtracted weight loss was 4.8%, 5.1%, and 4.2%, respectively, in the COR-I, COR-II, and COR-BMOD trials. Clinical and statistical dose-dependent improvements were seen in selected cardiovascular and metabolic outcome measurements that were related to the weight loss. However, the medication led to slight increased or smaller decreases in blood pressure and pulse than placebo. The most common adverse events experienced by the drug-randomized groups were nausea, constipation, headache, vomiting, dizziness, diarrhea, insomnia, and dry mouth.

Liraglutide, the fourth new medication, is a glucagon-like peptide-1 (GLP-1) analogue with 97% homology to human GLP-1 that was previously approved for the treatment of type 2 diabetes at doses up to 1.8 mg once daily. In addition to its effect as an incretin hormone (glucose-induced insulin secretion), liraglutide inhibits both gastric emptying and glucagon secretion and stimulates GLP-1 receptors in the arcuate nucleus of the hypothalamus to reduce feeding.

Liraglutide has undergone three randomized, placebo-controlled, double-blind trials for efficacy and safety. Participants were randomized to receive liraglutide (3.0 mg sc daily) or placebo for initial weight loss—SCALE (patients without diabetes) and SCALE Diabetes (patients with diabetes), or for weight maintenance after initial weight loss (SCALE Maintenance) (Table 395-5). All participants received diet and exercise counseling. For SCALE and SCALE Maintenance, patients were overweight or obese and had treated or untreated hypertension or dyslipidemia. Intention-to-treat 1-year placebo-subtracted weight loss was 5.4%, and 6.1%, respectively, in the SCALE and SCALE Maintenance trials. Clinical and statistical dose-dependent improvements were seen in selected cardiovascular and metabolic outcome measurements; however, there is a small increase in heart rate. The most common adverse effects include nausea, diarrhea, constipation, and vomiting. GLP-1 agonists should not be prescribed in patients with a family or personal history of medullary thyroid cancer or multiple endocrine neoplasia.

In approving the four new antiobesity medications, the FDA introduced a new provision with important clinical relevance: a prescription trial period to assess effectiveness. Response to these medications should be assessed after 12 weeks of treatment for PHEN/TPM and lorcaserin (or 16 weeks for naltrexone SR/bupropion SR and liraglutide since these medications are uptitrated during the first month). Determining responsiveness at 3 or 4 months is based on the post hoc observed trial data that patients who did not lose a prespecified amount of weight early in treatment were less successful at 1 year. For PHEN/TPM, if the patient has not lost at least 3% of body weight at 3 months, the clinician can either escalate the dose and reassess progress at 6 months or discontinue treatment entirely. For lorcaserin and naltrexone SR/bupropion SR, the medication should be discontinued if the patient has not lost at least 5% of body weight. The corresponding responsive target for liraglutide is a 4% weight loss.

Peripherally Acting Medications Orlistat (Xenical™) is a synthetic hydrogenated derivative of a naturally occurring lipase inhibitor, lipostatin, that is produced by the mold *Streptomyces toxytricini*. This drug is a potent, slowly reversible inhibitor of pancreatic, gastric, and carboxylester lipases and phospholipase A_2, which are required for the hydrolysis of dietary fat into fatty acids and

monoacylglycerols. Orlistat acts in the lumen of the stomach and small intestine by forming a covalent bond with the active site of these lipases. Taken at a therapeutic dose of 120 mg tid, orlistat blocks the digestion and absorption of ~30% of dietary fat. After discontinuation of the drug, fecal fat content usually returns to normal within 48–72 h.

Multiple randomized, double-blind, placebo-controlled studies have shown that, after 1 year, orlistat produces a weight loss of ~9–10%, whereas placebo recipients have a 4–6% weight loss. Because orlistat is minimally (<1%) absorbed from the gastrointestinal tract, it has no systemic side effects. The drug's tolerability is related to the malabsorption of dietary fat and the subsequent passage of fat in the feces. Adverse gastrointestinal effects, including flatus with discharge, fecal urgency, fatty/oily stool, and increased defecation, are reported in at least 10% of orlistat-treated patients. These side effects generally are experienced early, diminish as patients control their dietary fat intake, and only infrequently cause patients to withdraw from clinical trials. When taken concomitantly, psyllium mucilloid is helpful in controlling orlistat-induced gastrointestinal side effects. Because serum concentrations of the fat-soluble vitamins D and E and β-carotene may be reduced by orlistat treatment, vitamin supplements are recommended to prevent potential deficiencies. Orlistat was approved for over-the-counter use in 2007.

SURGERY

Bariatric surgery **(Fig. 395-2)** can be considered for patients with severe obesity (BMI, ≥40 kg/m^2) or for those with moderate obesity (BMI, ≥35 kg/m^2) associated with a serious medical condition. Weight-loss surgeries have traditionally been classified into three categories on the basis of anatomic changes: restrictive, restrictive malabsorptive, and malabsorptive. More recently, however, the clinical benefits of bariatric surgery in achieving weight loss and alleviating metabolic comorbidities have been attributed largely to changes in the physiologic responses of gut hormones, bile acid metabolism, the microbiota, and in adipose tissue metabolism. Metabolic effects resulting from bypassing the foregut include altered responses of ghrelin, glucagon-like peptide 1, peptide YY3-36, and oxyntomodulin. Additional effects on food intake and body weight control may be attributed to changes in vagal signaling. The loss of fat mass, particularly visceral fat, is associated with multiple metabolic, adipokine, and inflammatory changes that include improved insulin sensitivity and glucose disposal; reduced free fatty acid flux; increased adiponectin levels; and decreased interleukin 6, tumor necrosis factor α, and high-sensitivity C-reactive protein levels.

Restrictive surgeries limit the amount of food the stomach can hold and slow the rate of gastric emptying. *Laparoscopic adjustable gastric banding* is the prototype of this category. The first banding device, the LAP-BAND, was approved for use in the United States in 2001 and the second, the REALIZE band, in 2007. In contrast to previous devices, these bands have diameters that are adjustable by way of their connection to a reservoir that is implanted under the skin. Injection of saline into the reservoir and removal of saline from the reservoir tighten and loosen the band's internal diameter, respectively, thus changing the size of the gastric opening. Although the mean percentage of total body weight lost at 5 years is estimated at 20–25%, longer-term follow-up has been more disappointing leading to near abandonment of the procedure. In the *laparoscopic sleeve gastrectomy*, the stomach is restricted by stapling and dividing it vertically, removing ~80% of the greater curvature and leaving a slim banana-shaped remnant stomach along the lesser curvature. Weight loss after this procedure is superior to that after laparoscopic adjustable gastric banding.

The three restrictive-malabsorptive bypass procedures combine the elements of gastric restriction and selective malabsorption: Roux-en-Y gastric bypass, biliopancreatic diversion, and biliopancreatic diversion with duodenal switch (Fig. 395-2). Roux-en-Y is the

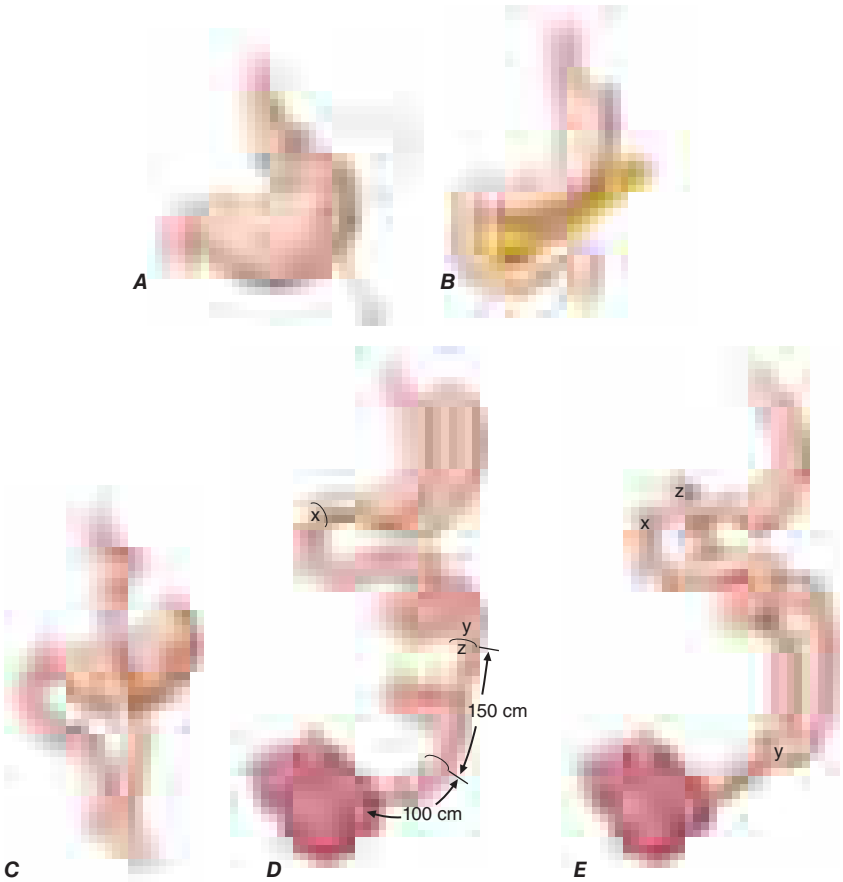


FIGURE 395-2 Bariatric surgical procedures. Examples of operative interventions used for surgical manipulation of the gastrointestinal tract. ***A.*** Laparoscopic adjustable gastric banding. ***B.*** Laparoscopic sleeve gastrectomy. ***C.*** The Roux-en-Y gastric bypass. ***D.*** Biliopancreatic diversion with duodenal switch. ***E.*** Biliopancreatic diversion. *(From ML Kendrick, GF Dakin: Mayo Clin Proc 815:518, 2006; with permission.)*

most commonly undertaken and most accepted bypass procedure. They are routinely performed by laparoscopy.

These procedures generally produce a 30–35% average total body weight loss that is maintained in ~60% of patients at 5 years. Significant improvement in multiple obesity-related comorbid conditions, including type 2 diabetes, hypertension, dyslipidemia, obstructive sleep apnea, quality of life, and long-term cardiovascular events, has been reported. A meta-analysis of controlled clinical trials comparing bariatric surgery versus no surgery showed that surgery was associated with a reduced odds ratio (OR) risk of global mortality (OR = 0.55), cardiovascular death (OR = 0.58), and all-cause mortality (OR = 0.70).

Among the observed improvements in comorbidities, the prevention and treatment of type 2 diabetes resulting from bariatric surgery has garnered the most attention. Fifteen-year data from the Swedish Obese Subjects study demonstrated a marked reduction (i.e., by 78%) in the incidence of type 2 diabetes development among obese patients who underwent bariatric surgery. Several randomized controlled studies have shown greater weight loss and more improved glycemic control at 1 and 3 years among surgical patients than among patients receiving conventional medical therapy. A retrospective cohort study of >4000 adults with diabetes found that overall 68.2% of patients experienced an initial complete type 2 diabetes remission within 5 years after surgery. However, among these patients, one-third redeveloped type 2 diabetes within 5 years. The rapid improvement seen in diabetes after restrictive-malabsorptive procedures is thought to be due to caloric restriction, reduced insulin resistance, and surgery-specific effects on glucose homeostasis brought about by alteration of gut hormones.

The mortality rate from bariatric surgery is generally <1% but varies with the procedure, the patient's age and comorbid conditions, and the experience of the surgical team. The most common surgical complications include stomal stenosis or marginal ulcers (occurring in 5–15% of patients) that present as prolonged nausea and vomiting after eating or inability to advance the diet to solid foods. These complications typically are treated by endoscopic balloon dilation and acid suppression therapy, respectively. For patients who undergo laparoscopic adjustable gastric banding, there are no intestinal absorptive abnormalities other than mechanical reduction in gastric size and outflow. Therefore, selective deficiencies are uncommon unless eating habits become unbalanced. In contrast, the restrictive-malabsorptive procedures carry an increased risk for micronutrient deficiencies of vitamin B_{12}, iron, folate, calcium, and vitamin D. Patients with restrictive-malabsorptive procedures require lifelong supplementation with these micronutrients.

Intraluminal Gastric Balloons Recently, the FDA approved two gastric balloon devices for weight loss that are placed in the stomach endoscopically. The RESHAPE device consists of two silicone balloons attached to a central silicone shaft, whereas the ORBERA is a single-balloon device. Mean weight loss of 7.2 kg and 8.8 kg, respectively, was seen for these devices in short-term pivotal trials. Both systems are approved only for up to 6 months of use in adults with a BMI of 30–40 kg/m^2. Adverse effects include nausea, vomiting, and abdominal pain.

FURTHER READING

Apovian CM et al: Pharmacological management of obesity: An Endocrine Society clinical practice guideline. J Clin Endocrinol Metab 100:342, 2015.

Garvey WT et al: American Association of Clinical Endocrinologists and American College of Endocrinology Comprehensive Clinical Practice Guidelines for Medical Care of Patients with Obesity. Endocrine Practice 22 (suppl 3):1, 2016.

Jensen MD et al: 2013 AHA/ACC/TOS guideline for the management of overweight and obesity in adults: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and The Obesity Society. Circulation 129 (suppl 2):S102, 2014.

396 Diabetes Mellitus: Diagnosis, Classification, and Pathophysiology

Alvin C. Powers, Kevin D. Niswender, Carmella Evans-Molina

Diabetes mellitus (DM) refers to a group of common metabolic disorders that share the phenotype of hyperglycemia. Several distinct types of DM are caused by a complex interaction of genetics and environmental factors. Depending on the etiology of the DM, factors contributing to hyperglycemia include reduced insulin secretion, decreased glucose utilization, and increased glucose production. The metabolic dysregulation associated with DM causes secondary pathophysiologic changes in multiple organ systems that impose a tremendous burden on the individual with diabetes and on the health care system. In the United States, DM is the leading cause of end-stage renal disease (ESRD), nontraumatic lower extremity amputations, and adult blindness. It also predisposes to cardiovascular diseases. With an increasing incidence worldwide, DM is likely to continue to be a leading cause of morbidity and mortality in the future.

CLASSIFICATION

DM is classified on the basis of the pathogenic process leading to hyperglycemia, as opposed to earlier criteria such as age of onset or type of therapy **(Fig. 396-1)**. There are two broad categories of DM, designated as either type 1 or type 2 DM **(Table 396-1)**. However, there is increasing recognition of other forms of diabetes in which the molecular pathogenesis is better understood and may be associated with a single gene defect. These alternative forms may share features of type 1 and/or type 2 DM. Type 1 DM develops as a result of autoimmunity against the insulin-producing beta cells, resulting in complete

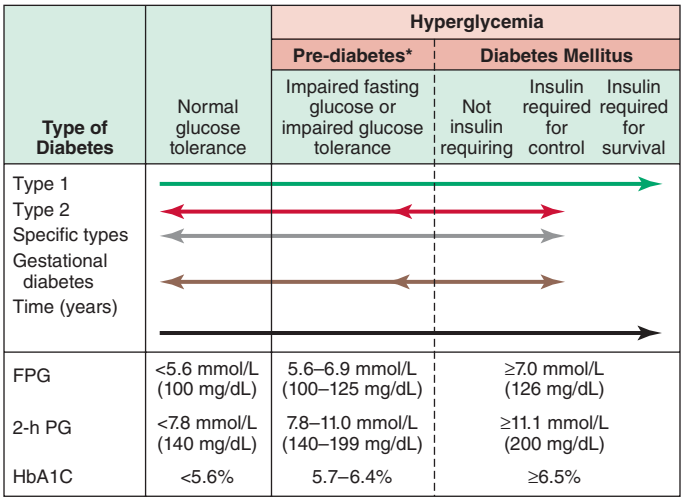


FIGURE 396-1 Spectrum of glucose homeostasis and diabetes mellitus (DM). The spectrum from normal glucose tolerance to diabetes in type 1 DM, type 2 DM, specific types of diabetes, and gestational DM is shown from left to right. In most types of DM, the individual traverses from normal glucose tolerance to impaired glucose tolerance to overt diabetes (these should be viewed not as abrupt categories but as a spectrum). *Arrows* indicate that changes in glucose tolerance may be bidirectional in some types of diabetes. For example, individuals with type 2 DM may return to the impaired glucose tolerance category with weight loss; in gestational DM, diabetes may revert to impaired glucose tolerance or even normal glucose tolerance after delivery. The fasting plasma glucose (FPG), the 2-h plasma glucose (PG) after a glucose challenge, and the hemoglobin A_{1c} (HbA_{1c}) for the different categories of glucose tolerance are shown at the lower part of the figure. These values do not apply to the diagnosis of gestational DM. *Some use the term *increased risk for diabetes* or *intermediate hyperglycemia* (World Health Organization) rather than *prediabetes*. *(Adapted from the American Diabetes Association, 2017.)*

TABLE 396-1 Etiologic Classification of Diabetes Mellitus

I. Type 1 diabetes (immune-mediated beta cell destruction, usually leading to absolute insulin deficiency)

II. Type 2 diabetes (may range from predominantly insulin resistance with relative insulin deficiency to a predominantly insulin secretory defect with insulin resistance)

III. Specific types of diabetes

- A. Genetic defects of beta cell development or function characterized by mutations in:
  1. Hepatocyte nuclear transcription factor (HNF) 4α (MODY 1)
  2. Glucokinase (MODY 2)
  3. HNF-1α (MODY 3)
  4. Insulin promoter factor-1, HNF-1β, NeuroD1, and others leading to other forms of MODY
  5. Insulin, subunits of ATP-sensitive potassium channel leading to permanent neonatal diabetes
  6. Mitochondrial DNA
  7. Other pancreatic islet regulators/proteins such as *KLF11, PAX4, BLK, GATA4, GATA6, SLC2A2* (GLUT2), *RFX6, GLIS3*
- B. Transient neonatal diabetes
- C. Diseases of the exocrine pancreas—pancreatitis, pancreatectomy, neoplasia, cystic fibrosis, hemochromatosis, fibrocalculous pancreatopathy, mutations in carboxyl ester lipase
- D. Genetic defects in insulin action, including type A insulin resistance, Leprechaunism, Rabson-Mendenhall syndrome, Lipodystrophy syndromes
- E. Endocrinopathies—acromegaly, Cushing's syndrome, glucagonoma, pheochromocytoma, hyperthyroidism, somatostatinoma, aldosteronoma
- F. Drug- or chemical-induced—glucocorticoids, vacor (a rodenticide), pentamidine, nicotinic acid, diazoxide, β-adrenergic agonists, thiazides, calcineurin and mTOR inhibitors, hydantoins, asparaginase, α-interferon, protease inhibitors, antipsychotics (atypicals and others), epinephrine
- G. Infections—congenital rubella, cytomegalovirus, coxsackievirus
- H. Uncommon forms of immune-mediated diabetes—"stiff-person" syndrome, anti-insulin receptor antibodies
- I. Other genetic syndromes sometimes associated with diabetes—Wolfram's syndrome, Down's syndrome, Klinefelter's syndrome, Turner's syndrome, Friedreich's ataxia, Huntington's chorea, Laurence-Moon-Biedl syndrome, myotonic dystrophy, porphyria, Prader-Willi syndrome

IV. Gestational diabetes mellitus (GDM)

Abbreviation: MODY, maturity-onset diabetes of the young.

Source: Adapted from American Diabetes Association: Diabetes Care 37(Suppl 1): S14, 2014 and updated from American Diabetes Association: Diabetes Care 40(Suppl 1):S11–24, 2017.

or near-total insulin deficiency. Type 2 DM is a heterogeneous group of disorders characterized by variable degrees of insulin resistance, impaired insulin secretion, and increased hepatic glucose production. Distinct genetic and metabolic defects in insulin action and/or secretion give rise to the common phenotype of hyperglycemia in type 2 DM and have important therapeutic implications now that pharmacologic agents are available to target specific metabolic derangements. Both type 1 and type 2 diabetes are preceded by a period of progressive worsening of glucose homeostasis, followed by the development of hyperglycemia that exceeds the threshold for clinical diagnosis. In terms of type 2 diabetes, this phase is referred to as prediabetes and are more specifically classified as impaired fasting glucose (IFG) or impaired glucose tolerance (IGT) (Fig. 396-1). Recently, three distinct stages of type 1 DM have been defined based on the development of autoantibodies against pancreatic beta cell antigens or the development of worsening dysglycemia (discussed below).

Two features of the current classification of DM merit emphasis from previous classifications. First, the terms *insulin-dependent diabetes mellitus* (IDDM) or *non-insulin-dependent diabetes mellitus* (NIDDM) are no longer used. Because many individuals with type 2 DM eventually require insulin treatment for control of glycemia, the use of the term NIDDM generated considerable confusion. A second difference is that age or treatment modality is not a criterion. Although type 1 DM most commonly develops before the age of 30, autoimmunity against beta cells can develop at any age. It is estimated that between 5 and 10% of individuals who develop DM after age 30 have type 1 DM. Although type 2 DM more typically develops with increasing age, it is now being diagnosed more frequently in children and young adults, particularly in obese adolescents.

OTHER TYPES OF DM

Other etiologies of DM include specific genetic defects in insulin secretion or action, metabolic abnormalities that impair insulin secretion, mitochondrial abnormalities, and a host of conditions that impair glucose tolerance (Table 396-1). *Maturity-onset diabetes of the young* (MODY) and *monogenic diabetes* are subtypes of DM characterized by autosomal dominant inheritance, early onset of hyperglycemia (usually <25 years; sometimes in neonatal period), and impaired insulin secretion (discussed below). Mutations in the insulin receptor cause a group of rare disorders characterized by severe insulin resistance.

DM may also develop as a result of cystic fibrosis or chronic pancreatitis, in which the islets become damaged from a primary pathologic process originating in the pancreatic exocrine tissue. Hormones that antagonize insulin action can lead to DM. Thus, DM is often a feature of endocrinopathies such as acromegaly and Cushing's disease. Viral infections have been implicated in pancreatic islet destruction but are an extremely rare cause of DM. A form of acute onset of type 1 diabetes, termed *fulminant diabetes*, has been noted in Japan and may be related to viral infection of the islets.

GESTATIONAL DM

Glucose intolerance developing during the second or third trimester of pregnancy is classified as gestational diabetes mellitus (GDM). Insulin resistance is related to the metabolic changes of pregnancy, during which the increased insulin demands may lead to IGT or diabetes. The American Diabetes Association (ADA) recommends that diabetes diagnosed within the first trimester be classified as preexisting pregestational diabetes rather than GDM. In 2015, the International Diabetes Federation (IDF) estimated that one in seven pregnancies worldwide was affected by either GDM or preexisting DM. Most women with GDM revert to normal glucose tolerance postpartum but have a substantial risk (35–60%) of developing DM in the next 10–20 years. In addition, children born to a mother with GDM also have an increased risk of developing metabolic syndrome and type 2 DM later in life. Currently, the ADA recommends that women with a history of GDM undergo lifelong screening for the development of diabetes or prediabetes at least every 3 years.

EPIDEMIOLOGY AND GLOBAL CONSIDERATIONS

The worldwide prevalence of DM has risen dramatically over the past two decades, from an estimated 30 million cases in 1985 to 415 million in 2017 **(Fig. 396-2)**. Based on current trends, the IDF projects that 642 million individuals will have diabetes by the year 2040 (see *http://www.idf.org/*). Although the prevalence of both type 1 and type 2 DM is increasing worldwide, the prevalence of type 2 DM is rising much more rapidly, presumably because of increasing obesity, reduced activity levels as countries become more industrialized, and the aging of the population. The incidence of type 1 diabetes has been increasing at a rate of 3–5% per year worldwide. The cause for this increase is not well understood, but type 1 DM is increasingly being diagnosed at younger ages. In 2015, the prevalence of diabetes in individuals aged 20–79 ranged from 7.2–11.4%. The countries with the greatest number of individuals with diabetes in 2015 are China (109.6 million), India (73 million), the United States (30.3 million), Brazil (14 million), and the Russian Federation (9 million). In the most recent estimate for the United States (2017), the Centers for Disease Control and Prevention (CDC) estimated that 9.4% of the population had diabetes, and as many as 34% of U.S. adults had prediabetes. Approximately 25% of the individuals with diabetes in the United States were undiagnosed; globally, it is estimated that as many of 50% of individuals with diabetes may be undiagnosed. The prevalence of DM increases with age. In 2015, the prevalence of DM in

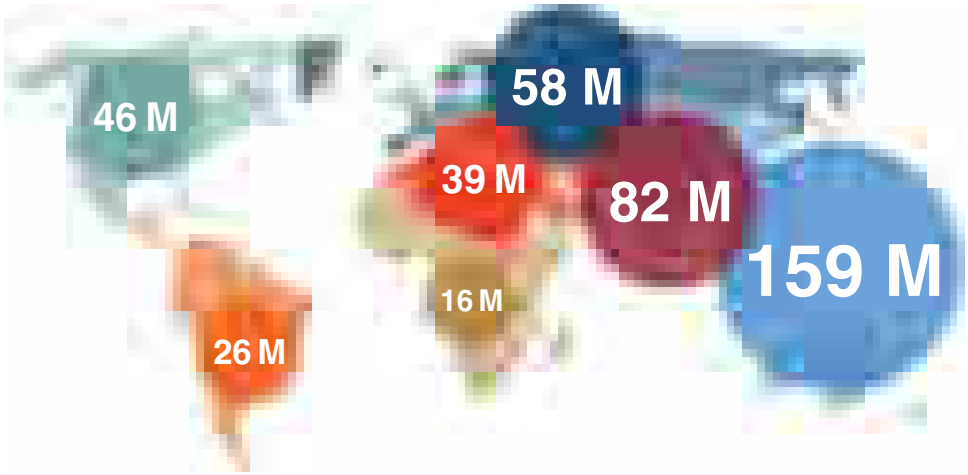


FIGURE 396-2 Worldwide prevalence of diabetes mellitus. Global estimate is 415 million individuals with diabetes in 2017. Regional estimates of the number of individuals with diabetes (20–79 years of age) are shown (2017). *(Adapted from the IDF Diabetes Atlas, the International Diabetes Federation, 2017.)*

the United Sates was estimated to be 0.25% in individuals age <20 years, 4.1% in persons aged 20–44 years, and 16.2% in persons 45–64 years old. In individuals aged >65 years, the prevalence of DM was 25.9%. Similar age-related trends have been observed worldwide. The prevalence of diabetes is similar among men and women, but diabetes-related mortality rates are higher in men compared to women.

There is considerable geographic variation in the incidence of both type 1 and type 2 DM. Currently, Scandinavia has the highest incidence of type 1 DM; the lowest incidence is in the Pacific Rim where it is twenty- to thirtyfold lower. Northern Europe and the United States have an intermediate rate. Much of the increased risk of type 1 DM is believed to reflect the frequency of high-risk human leukocyte antigen (HLA) alleles among ethnic groups in different geographic locations.

However, now populations less enriched with these classic high-risk HLA alleles are experiencing more rapid increases in type 1 DM incidence, suggesting an influence of environmental factors.

The prevalence of type 2 DM and its harbinger, IGT, is highest in certain Pacific islands and the Middle East and intermediate in countries such as India and the United States. This variability is likely due to genetic, behavioral, and environmental factors. DM prevalence also varies among different ethnic populations within a given country, with indigenous populations usually having a greater incidence of diabetes than the general population of the country. For example, the CDC estimated that the age-adjusted prevalence of DM in the United States (age >20 years; 2010–2012) was 8% in non-Hispanic whites, 9% in Asian Americans, 13% in Hispanics, 13% in non-Hispanic blacks, and 16% in American-Indian and Alaskan native populations. The onset of type 2 DM occurs, on average, at an earlier age in ethnic groups other than non-Hispanic whites. In Asia, the prevalence of diabetes is increasing rapidly, and the diabetes phenotype appears to be somewhat different from that in the United States and Europe, with an onset at a lower body mass index (BMI) and younger age, greater visceral adiposity, and reduced insulin secretory capacity.

Diabetes is a major cause of mortality. In recent years, diabetes has been listed as the seventh leading cause of death in the United States, but several studies indicate that diabetes-related deaths are likely underreported. Data from the IDF suggests that diabetes was responsible for almost 5 million deaths worldwide, accounting for 14.5% of global all-cause mortality in adults aged 20–79 years of age. Diabetes also has important economic implications. In 2015, it was estimated that $673 billion or 12% of health care expenditures worldwide were spent on diabetes (range 5–20%). Up to 75% of individuals with diabetes live in low or middle-income countries.

DIAGNOSIS

Glucose tolerance is classified into three broad categories: normal glucose homeostasis, impaired glucose homeostasis, or DM. Glucose tolerance can be assessed using the fasting plasma glucose (FPG), the response to oral glucose challenge, or the hemoglobin A_{1c} (HbA_{1c}). An FPG <5.6 mmol/L (100 mg/dL), a plasma glucose <7.9 mmol/L (140 mg/dL) following an oral glucose challenge, and an HbA_{1c} <5.7% are considered to define normal glucose tolerance. The International Expert Committee with members appointed by the ADA, the European Association for the Study of Diabetes, and the IDF have issued diagnostic criteria for DM **(Table 396-2)** based on the following premises: (1) the FPG, the response to an oral glucose challenge (oral glucose tolerance test [OGTT]), and HbA_{1c} differ among individuals, and (2) DM is defined as the level of glycemia at which diabetes-specific complications occur rather than deviation from a population-based mean. For example, the prevalence of retinopathy in Native Americans (Pima Indian population) begins to increase at an FPG >6.4 mmol/L (116 mg/dL) **(Fig. 396-3)**.

Abnormal glucose homeostasis (Fig. 396-1) is defined as (1) FPG = 5.6–6.9 mmol/L (100–125 mg/dL), which is defined as *impaired fasting glucose* (IFG); the World Health Organization uses 6.1–6.9 mmol/L (110–125 mg/dL) for IFG; (2) plasma glucose levels between 7.8 and 11 mmol/L (140 and 199 mg/dL) following an oral glucose challenge, which is termed *impaired glucose tolerance* (IGT); or (3) HbA_{1c} of 5.7–6.4%. An HbA_{1c} of 5.7–6.4%, IFG, and IGT do not identify the same individuals, but individuals in all three groups are at greater risk of progressing to type 2 DM, have an increased risk of cardiovascular disease, and should be counseled about ways to decrease these risks (see below). Some use the terms *prediabetes*, *increased risk of diabetes*, or *intermediate hyperglycemia* (World Health Organization) and slightly different metrics for this category.

These values for the FPG, the glucose following an oral glucose challenge, and HbA_{1c} are continuous rather than discrete variables. A FPG ≥7.0 mmol/L (126 mg/dL), a glucose ≥11.1 mmol/L (200 mg/dL) 2 h after an oral glucose challenge, or an HbA_{1c} ≥6.5% meets the criteria for the diagnosis of DM (Table 396-2). A random plasma glucose concentration ≥11.1 mmol/L (200 mg/dL) accompanied by classic symptoms of DM (polyuria, polydipsia, weight loss) is also sufficient for the diagnosis of DM (Table 396-2). The current criteria for the diagnosis of DM emphasize the HbA_{1c} or the FPG as the most reliable and convenient tests for identifying DM in asymptomatic individuals. However, some individuals may meet criteria for one test but not the other. Also, it is important to note that race and ethnicity may impact the reliability of HbA_{1c} levels. For example, African Americans have a higher HbA_{1c}

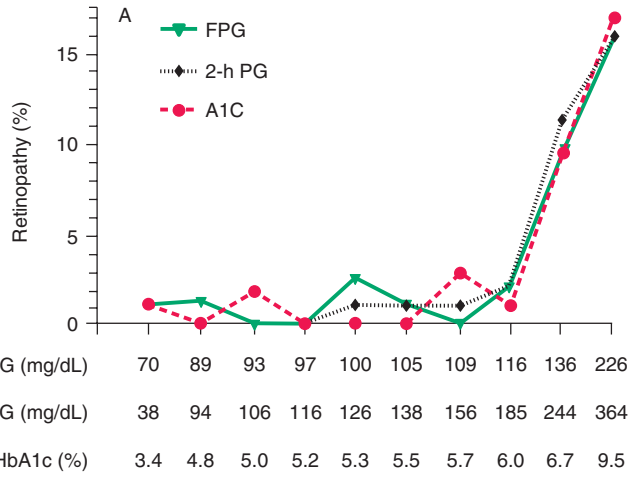
TABLE 396-2 Criteria for the Diagnosis of Diabetes Mellitus

- Symptoms of diabetes plus random blood glucose concentration ≥11.1 mmol/L (200 mg/dL)[a] *or*
- Fasting plasma glucose ≥7.0 mmol/L (126 mg/dL)[b] *or*
- Hemoglobin A_{1c} ≥ 6.5%[c] *or*
- 2-h plasma glucose ≥11.1 mmol/L (200 mg/dL) during an oral glucose tolerance test[d]

[a]Random is defined as without regard to time since the last meal. [b]Fasting is defined as no caloric intake for at least 8 h. [c]Hemoglobin A_{1c} test should be performed in a laboratory using a method approved by the National Glycohemoglobin Standardization Program and correlated to the reference assay of the Diabetes Control and Complications Trial. Point-of-care hemoglobin A_{1c} should not be used for diagnostic purposes. [d]The test should be performed using a glucose load containing the equivalent of 75 g anhydrous glucose dissolved in water, not recommended for routine clinical use.

Note: In the absence of unequivocal hyperglycemia and acute metabolic decompensation, these criteria should be confirmed by repeat testing on a different day.

Source: Adapted from American Diabetes Association: Diabetes Care 40(Suppl 1): S13-27, 2018.



FPG (mg/dL)	70	89	93	97	100	105	109	116	136	226
2-h PG (mg/dL)	38	94	106	116	126	138	156	185	244	364
HbA1c (%)	3.4	4.8	5.0	5.2	5.3	5.5	5.7	6.0	6.7	9.5

FIGURE 396-3 Relationship of diabetes-specific complication and glucose tolerance. This figure shows the incidence of retinopathy in Pima Indians as a function of the fasting plasma glucose (FPG), the 2-h plasma glucose after a 75-g oral glucose challenge (2-h PG), or the hemoglobin A_{1c} (HbA_{1c}). Note that the incidence of retinopathy greatly increases at a fasting plasma glucose >116 mg/dL, a 2-h plasma glucose of 185 mg/dL, or an HbA_{1c} >6.5%. (Blood glucose values are shown in mg/dL; to convert to mmol/L, divide value by 18.) *(Copyright 2002, American Diabetes Association. From Diabetes Care 25[Suppl 1]: S5–S20, 2002.)*

value compared to non-Hispanic whites with a similar level of glycemia. An OGTT, although a valid means for diagnosing DM, is not often used in routine clinical care with the exception of pregnancy care and screening for gestational diabetes.

The diagnosis of DM has profound implications for an individual from both a medical and a financial standpoint. Thus, abnormalities on screening tests for diabetes should be repeated before making a definitive diagnosis of DM, unless acute metabolic derangements or a markedly elevated plasma glucose are present (Table 396-2). These criteria also allow for the diagnosis of DM to be withdrawn in situations when the glucose intolerance reverts to normal.

SCREENING

Widespread use of the FPG or the HbA_{1c} as a screening test for type 2 DM is recommended because (1) a large number of individuals who meet the current criteria for DM are asymptomatic and unaware that they have the disorder, (2) epidemiologic studies suggest that type 2 DM may be present for up to a decade before diagnosis, (3) some individuals with type 2 DM have one or more diabetes-specific complications at the time of their diagnosis, (4) treatment of type 2 DM may favorably alter the natural history of DM, (5) diagnosis of prediabetes should spur efforts for diabetes prevention. The ADA recommends screening all individuals aged >45 years every 3 years and screening individuals at an earlier age if they are overweight (BMI >25 kg/m^2 or ethnically relevant definition for overweight) and have one additional risk factor for diabetes **(Table 396-3)**. Although a number of immunologic markers for type 1 DM are becoming available (discussed below), their routine use outside a clinical trial is discouraged, pending the identification of clinically beneficial interventions for individuals at high risk for developing type 1 DM.

REGULATION OF GLUCOSE HOMEOSTASIS

OVERALL REGULATION OF GLUCOSE HOMEOSTASIS

Glucose homeostasis reflects a balance between energy intake from ingested food, hepatic glucose production (gluconeogenesis), and peripheral tissue glucose uptake and utilization. Insulin is the most important regulator of this metabolic equilibrium, but neural input, metabolic signals, and other hormones (e.g., glucagon) result in integrated control of glucose supply and utilization **(Fig. 396-4)**. The organs that regulate glucose and lipids communicate by neural and humoral mechanisms with fat and muscle producing adipokines, myokines, and metabolites that influence liver function. In the fasting state, low insulin levels, together with modest increases in glucagon, increase glucose

TABLE 396-3 Risk Factors for Type 2 Diabetes Mellitus

Family history of diabetes (i.e., parent or sibling with type 2 diabetes)
Overweight or obese (BMI ≥25 kg/m^2, ≥23 kg/m^2 in Asian Americans, or other ethnically relevant definition for overweight)
Physical inactivity
Race/ethnicity (e.g., African American, Latino, Native American, Asian American, Pacific Islander)
Previously identified with IFG, IGT, or an hemoglobin A_{1c} of 5.7–6.4%
History of GDM
Hypertension (blood pressure ≥140/90 mmHg)
HDL cholesterol level <35 mg/dL (0.90 mmol/L) and/or a triglyceride level >250 mg/dL (2.82 mmol/L)
Polycystic ovary syndrome or acanthosis nigricans
History of cardiovascular disease

Abbreviations: BMI, body mass index; GDM, gestational diabetes mellitus; HDL, high-density lipoprotein; IFG, impaired fasting glucose; IGT, impaired glucose tolerance.

Source: Adapted from American Diabetes Association: Diabetes Care 40(Suppl 1): S13, 2018.

production by promoting hepatic gluconeogenesis and glycogen breakdown (glycogenolysis) and reducing glucose uptake in insulin-sensitive tissues (skeletal muscle and fat), thereby promoting mobilization of stored precursors such as amino acids and free fatty acids (lipolysis). Glucagon, secreted by pancreatic alpha cells normally only when blood glucose or insulin levels are low or during exercise, is increased in DM and stimulates glycogenolysis and gluconeogenesis by the liver and to a small degree by the renal medulla **(Chap. 399)**. Postprandially, the glucose load elicits a rise in insulin and fall in glucagon, leading to a reversal of these processes. Insulin, an anabolic hormone, promotes the storage of carbohydrate and fat and protein synthesis. The major portion of postprandial glucose is used by skeletal muscle, an effect of insulin-stimulated glucose uptake. Other tissues, most notably the brain, use glucose in an insulin-independent fashion. Factors secreted by skeletal myocytes, adipocytes (leptin, resistin, adiponectin, etc.), and bone also influence glucose homeostasis.

INSULIN BIOSYNTHESIS

Insulin, produced by the beta cells of the pancreatic islets, is initially synthesized as a single-chain 86-amino-acid precursor polypeptide, preproinsulin. Subsequent proteolytic processing removes the amino-terminal signal peptide, giving rise to proinsulin. Proinsulin is structurally related to insulin-like growth factors I and II, which bind weakly to the insulin receptor. Cleavage of an internal 31-residue fragment from proinsulin generates C-peptide with the A (21 amino acids) and B (30 amino acids) chains of insulin being connected by disulfide bonds. The mature insulin molecule and C-peptide are stored together

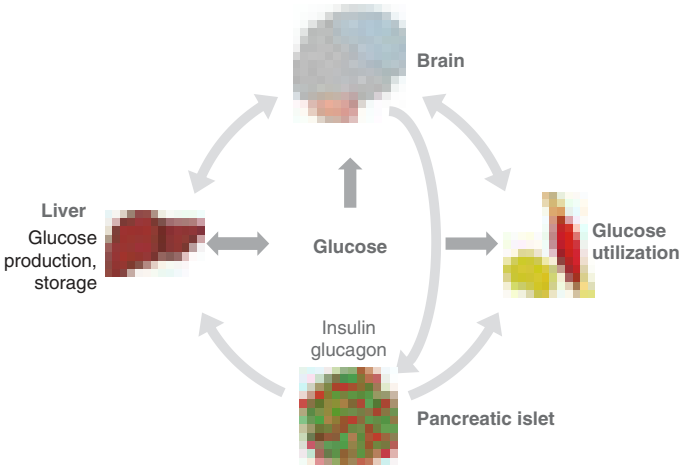


FIGURE 396-4 Regulation of glucose homeostasis. The organs shown contribute to glucose utilization, production, or storage. See text for a description of the communications (*arrows*), which can be neural or humoral. Although not shown, the GI tract and bone produce factors that influence glucose homeostasis.

and co-secreted from secretory granules in the beta cells. Because C-peptide is cleared more slowly than insulin, it is a useful marker of insulin secretion and allows discrimination of endogenous and exogenous sources of insulin in the evaluation of hypoglycemia (Chaps. 399 and 80). Elevated levels of serum proinsulin have been observed in both type 1 and 2 DM and are thought to be indicative of beta cell dysfunction. Pancreatic beta cells co-secrete islet amyloid polypeptide (IAPP) or amylin, a 37-amino-acid peptide, along with insulin. The role of IAPP in normal physiology is incompletely defined, but it is the major component of the amyloid fibrils found in the islets of patients with type 2 diabetes, and an analogue is sometimes used in treating type 1 and type 2 DM. Human insulin is produced by recombinant DNA technology; structural modifications at one or more amino acid residues modify insulin's physical and pharmacologic characteristics (Chap. 397).

INSULIN SECRETION

Glucose is the key regulator of insulin secretion by the pancreatic beta cell, although amino acids, ketones, various nutrients, gastrointestinal peptides, and neurotransmitters also influence insulin secretion. Glucose levels >3.9 mmol/L (70 mg/dL) stimulate insulin synthesis, primarily by enhancing protein translation and processing. Glucose stimulation of insulin secretion begins with its transport into the beta cell by a facilitative glucose transporter (Fig. 396-5). Glucose phosphorylation by glucokinase is the rate-limiting step that controls glucose-regulated insulin secretion. Further metabolism of glucose-6-phosphate via glycolysis generates ATP, which inhibits the activity of an ATP-sensitive K^+ channel. This channel consists of two separate proteins: one is the binding site for certain oral hypoglycemics (e.g., sulfonylureas, meglitinides); the other is an inwardly rectifying K^+ channel protein (Kir6.2). Inhibition of this K^+ channel induces beta cell membrane depolarization, which opens voltage-dependent calcium channels (leading to an influx of calcium) and stimulates insulin secretion. Insulin secretory profiles reveal a pulsatile pattern of hormone release, with small secretory bursts occurring about every 10 min, superimposed upon greater amplitude oscillations of about 80–150 min. A number of metabolic pathways internal to the beta cell as well as external hormonal cues amplify glucose-stimulated insulin secretion. Glucagon-like peptide-1 (GLP-1) and glucose-dependent insulinotropic peptide (GIP) are incretin hormones that bind specific receptors on the beta cell to stimulate insulin secretion through cyclic AMP production, but have this effect only when the blood glucose is above the fasting level. Incretin hormones also suppress glucagon production and secretion. Incretin analogues or pharmacologic agents that prolong the activity of endogenous GLP-1 are used therapeutically in type 2 DM. Classically, GLP-1 release was thought to occur solely from neuroendocrine L-cells of the gastrointestinal tract following food ingestion. However, recent pre-clinical studies suggest that intraislet production of GLP-1 from alpha cells may play a role in the regulation of insulin secretion.

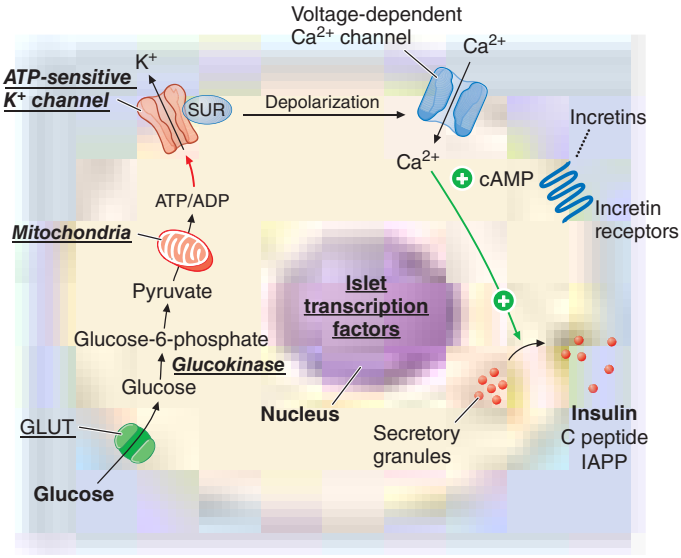


FIGURE 396-5 Mechanisms of glucose-stimulated insulin secretion and abnormalities in diabetes. Glucose and other nutrients regulate insulin secretion by the pancreatic beta cell. Glucose is transported by a glucose transporter (GLUT1 and/or GLUT2 in humans, GLUT2 in rodents); subsequent glucose metabolism by the beta cell alters ion channel activity, leading to insulin secretion. The SUR receptor is the binding site for some drugs that act as insulin secretagogues. Mutations in the events or proteins underlined are a cause of monogenic forms of diabetes. ADP, adenosine diphosphate; ATP, adenosine triphosphate; cAMP, cyclic adenosine monophosphate; IAPP, islet amyloid polypeptide or amylin; SUR, sulfonylurea receptor.

INSULIN ACTION

Once insulin is secreted into the portal venous system, ~50% is removed and degraded by the liver. Unextracted insulin enters the systemic circulation where it binds to receptors in target sites. Insulin binding to its receptor stimulates intrinsic tyrosine kinase activity, leading to receptor autophosphorylation and the recruitment of intracellular signaling molecules, such as insulin receptor substrates (IRS). IRS and other adaptor proteins initiate a complex cascade of phosphorylation and dephosphorylation reactions, resulting in the widespread metabolic and mitogenic effects of insulin. As an example, activation of the phosphatidylinositol-3′-kinase (PI-3-kinase) pathway stimulates translocation of a facilitative glucose transporter (e.g., GLUT4) to the cell surface, an event that is crucial for glucose uptake by skeletal muscle and fat. Activation of other insulin receptor signaling pathways induces glycogen synthesis, protein synthesis, lipogenesis, and regulation of various genes in insulin-responsive cells.

PATHOGENESIS

TYPE 1 DM

Type 1 DM is the result of interactions of genetic, environmental, and immunologic factors that ultimately lead to immune-mediated destruction of the pancreatic beta cells and insulin deficiency. Type 1 DM can develop at any age, but most commonly develops before 20 years of age. Most, but not all, individuals with type 1 DM have evidence of islet-directed autoimmunity. However, some individuals who have the clinical phenotype of type 1 DM lack immunologic markers indicative of an autoimmune process involving the beta cells and the genetic markers of type 1 DM. These individuals are thought to develop insulin deficiency by unknown, nonimmune mechanisms and may be ketosis prone; many are African American or Asian in heritage. The temporal decline of beta cell function and mass preceding the development of type 1 DM is shown schematically in Fig. 396-6. In susceptible individuals, the autoimmune process is thought to be triggered by an infectious or environmental stimulus. In the majority of patients, autoantibodies against beta cell antigens appear after this triggering event, followed by progressive loss of insulin secretion. The rate of decline in beta cell function varies widely among individuals, with some patients progressing rapidly to clinical diabetes and others evolving to diabetes more slowly and over a period of several years. Features of diabetes do not become evident until a threshold loss of insulin secretion and beta cell mass occurs. Autopsy studies suggest the degree of loss of beta cell mass is variable at the time of disease presentation but may be as high as 70–80%. At this point, residual, functional beta cells exist but are insufficient in number and quality to maintain glucose tolerance. The events that trigger the transition from glucose intolerance to frank diabetes are often associated with increased insulin requirements, as might occur during infections or at puberty. After the initial clinical presentation of type 1 DM, a "honeymoon" phase may ensue during which time glycemic control is achieved with modest doses of insulin or, rarely, insulin is not needed. However, this fleeting phase of endogenous insulin production from residual beta cells disappears and the individual becomes insulin deficient. Many individuals with long-standing type 1 DM produce a small amount of insulin (as reflected by C-peptide production), and some individuals with 50 years of type 1 DM have insulin-positive cells in the pancreas at autopsy.

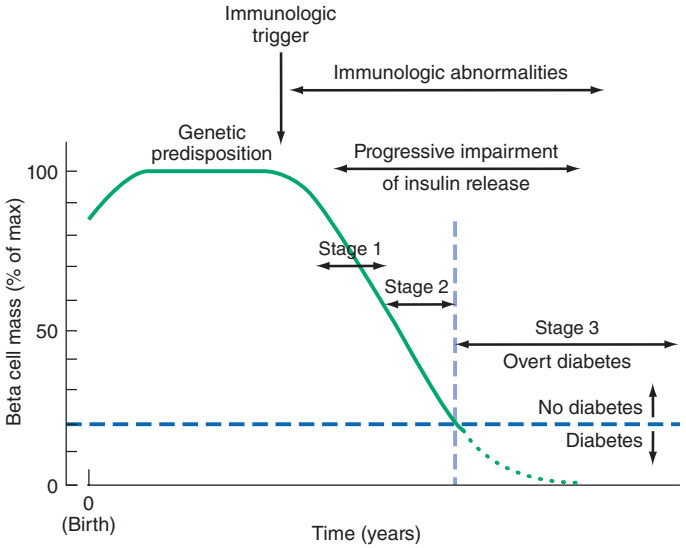


FIGURE 396-6 Temporal model for development of type 1 diabetes. Individuals with a genetic predisposition are exposed to a trigger that initiates an autoimmune process, resulting in the development of islet autoantibodies and a gradual decline in beta cell function and mass. Stage 1 disease is characterized by the development of two or more islet cell autoantibodies but the maintenance of normoglycemia. Stage 2 disease is defined by continued autoimmunity and the development of dysglycemia. Stage 3 is defined by the development of hyperglycemia that exceeds the diagnostic criteria for the diagnosis of diabetes. The downward slope of the beta cell function varies among individuals and may not be continuous. A "honeymoon" phase may be seen in the first 1 or 2 years after the onset of diabetes and is associated with reduced insulin requirements. *(Adapted and modified from ER Kaufman: Medical Management of Type 1 Diabetes, 6th ed. American Diabetes Association, Alexandria, VA, 2012.)*

GENETIC CONSIDERATIONS

Susceptibility to type 1 DM involves multiple genes. The concordance of type 1 DM in identical twins ranges between 40 and 60%, indicating that additional modifying factors are likely involved in determining whether diabetes develops. The major susceptibility gene for type 1 DM is located in the HLA region on chromosome 6. Polymorphisms in the HLA complex account for 40–50% of the genetic risk of developing type 1 DM. This region contains genes that encode the class II major histocompatibility complex (MHC) molecules, which present antigen to helper T cells and thus are involved in initiating the immune response **(Chap. 343)**. The ability of class II MHC molecules to present antigen is dependent on the amino acid composition of their antigen-binding sites. Amino acid substitutions may influence the specificity of the immune response by altering the binding affinity of different antigens for class II molecules.

Most individuals with type 1 DM have the HLA DR3 and/or DR4 haplotype. Refinements in genotyping of HLA loci have shown that the haplotypes DQA1*0301, DQB1*0302, and DQB1*0201 are most strongly associated with type 1 DM. These haplotypes are present in 40% of children with type 1 DM as compared to 2% of the normal U.S. population. However, most individuals with predisposing haplotypes do not develop diabetes.

In addition to MHC class II associations, genome association studies have identified at least 20 additional genetic loci that contribute susceptibility to type 1 DM (i.e., polymorphisms in the promoter region of the insulin gene, the CTLA-4 gene, interleukin 2 receptor, and PTPN22, etc.). Among recent cohorts of individuals with new-onset type 1 diabetes, there is a decreased representation of the highest risk HLA alleles and increasing penetrance of disease in those genotypes classically associated with lower risk. Genes that confer protection against the development of the disease also exist. The haplotype DQA1*0102, DQB1*0602 is extremely rare in individuals with type 1 DM (<1%) and appears to provide protection from type 1 DM.

Although the risk of developing type 1 DM is increased tenfold in relatives of individuals with the disease, the risk is relatively low: 3–4% if the parent has type 1 DM and 5–15% in a sibling (depending on which HLA haplotypes are shared). Hence, most individuals with type 1 DM (75%) do not have a first-degree relative with this disorder.

Pathophysiology Although other islet cell types (alpha cells [glucagon-producing], delta cells [somatostatin-producing], or PP cells [pancreatic polypeptide-producing]) are functionally and embryologically similar to beta cells, they are spared from the autoimmune destruction. However, altered patterns of hormone secretion from these other cell types in type 1 DM likely contributes to metabolic instability. Alpha cell dysfunction as reflected by fasting hyperglucagonemia, hyperglucagonemia in the post-prandial state, and an impaired glucagon response to hypoglycemia. Pathologically, the pancreatic islets have modest infiltration of lymphocytes (a process termed *insulitis*). After beta cells are destroyed, it is thought that the inflammatory process abates and the islets become atrophic. Studies of the autoimmune process in humans and in animal models of type 1 DM (NOD mouse and BB rat) have identified the following abnormalities in the humoral and cellular arms of the immune system: (1) islet cell autoantibodies; (2) activated lymphocytes in the islets, peripancreatic lymph nodes, and systemic circulation; (3) T lymphocytes that proliferate when stimulated with islet proteins; and (4) release of cytokines within the insulitis. Beta cells seem to be particularly susceptible to the toxic effect of some cytokines (tumor necrosis factor α [TNF-α], interferon γ, and interleukin 1 [IL-1]). The precise mechanisms of beta cell death are not known but may involve formation of nitric oxide metabolites, apoptosis, and direct CD8+ T cell cytotoxicity. The islet destruction is mediated by T lymphocytes rather than islet autoantibodies, as these antibodies do not generally react with the cell surface of islet cells and are not capable of transferring DM to animals. Efforts to suppress the autoimmune process at the time of diagnosis of diabetes have largely been ineffective or only temporarily effective in slowing beta cell destruction. Thus, increased emphasis has now been placed on interventions earlier in the disease course (i.e., during Stage 1 and 2 disease; Fig. 396-6).

Pancreatic islet molecules targeted by the autoimmune process include proinsulin, insulin, glutamic acid decarboxylase (GAD; the biosynthetic enzyme for the neurotransmitter GABA), ICA-512/IA-2 (homology with tyrosine phosphatases), and a beta cell–specific zinc transporter (ZnT-8). Most of the autoantigens are not beta cell–specific, which raises the question of how the beta cells are selectively destroyed. Current theories favor initiation of an autoimmune process directed at one beta cell molecule, which then spreads to other islet molecules as the immune process destroys beta cells and creates a series of secondary autoantigens. Stress pathways and processes arising within the beta cell may exacerbate autoimmunity through the development of modified proteins or "neoantigens" that serve as additional immune targets.

Immunologic Markers Islet cell autoantibodies (ICAs) are a composite of several different antibodies directed at pancreatic islet molecules such as GAD, insulin, IA-2/ICA-512, and ZnT-8, and serve as a marker of the autoimmune process of type 1 DM. Assays for autoantibodies to GAD-65 and insulin are commercially available. Testing for ICAs can be useful in classifying the type of DM as type 1 and in identifying nondiabetic individuals at risk for developing type 1 DM. ICAs are present in the majority of individuals (>85%) diagnosed with new-onset type 1 DM, in a significant minority of individuals with newly diagnosed type 2 DM (5–10%), and occasionally in individuals with GDM (<5%). ICAs are present in 3–4% of first-degree relatives of individuals with type 1 DM. In combination with impaired insulin secretion after IV glucose tolerance testing, they predict a >50% risk of developing type 1 DM within 5 years. Increasing numbers of autoantibodies are associated with an increased risk of diabetes development. In children with multiple autoantibodies, ~70% developed type 1 DM after 10 years of follow-up, with 80% developing diabetes after 15-year of follow-up. At present, the measurement of ICAs in nondiabetic individuals remains a research tool because no treatments have been demonstrated to prevent the occurrence or progression to type 1 DM.

Environmental Factors Numerous environmental events have been proposed to trigger the autoimmune process in genetically

susceptible individuals; however, none have been conclusively linked to diabetes. Identification of an environmental trigger has been difficult because the event may precede the onset of DM by several years (Fig. 396-6). Putative environmental triggers include viruses (coxsackie, rubella, enteroviruses most prominently), bovine milk proteins, nitrosourea compounds, vitamin D deficiency, and environmental toxins. There is increasing interest in the microbiome and type 1 diabetes (Chap. 459).

Prevention of Type 1 DM A number of interventions have prevented diabetes in animal models, but relatively few interventions have been tested in humans in Stages 1 and 2 of type 1 DM. The Diabetes Prevention Trial–Type 1 concluded that administering insulin (IV or PO) to individuals at high risk for developing type 1 DM did not prevent type 1 DM. However, this is an area of active clinical investigation with several trials evaluating interventions that target different aspects of the immune response in early stage type 1 DM.

TYPE 2 DM

Insulin resistance and abnormal insulin secretion are central to the development of type 2 DM. Although the primary defect is controversial, most studies support the view that insulin resistance precedes an insulin secretory defect but that diabetes develops only when insulin secretion becomes inadequate. Type 2 DM likely encompasses a range of disorders with the common phenotype of hyperglycemia. Most of our current understanding (and the discussion below) of the pathophysiology and genetics is based on studies of individuals of European descent. It is becoming increasing apparent that DM in other ethnic groups (Asian, African, and Latin American) has a somewhat different, but yet undefined, pathophysiology. In general, Latinos have greater insulin resistance and East Asians and South Asians have more beta cell dysfunction, but both defects are present in both populations. East and South Asians appear to develop type 2 DM at a younger age and a lower BMI. In some groups, DM that is ketosis prone (often in obese individuals) or ketosis-resistant (often lean) is sometimes seen.

GENETIC CONSIDERATIONS

Type 2 DM has a strong genetic component. The concordance of type 2 DM in identical twins is between 70 and 90%. Individuals with a parent with type 2 DM have an increased risk of diabetes; if both parents have type 2 DM, the risk approaches 40%. Insulin resistance, as demonstrated by reduced glucose utilization in skeletal muscle, is present in many nondiabetic, first-degree relatives of individuals with type 2 DM. The disease is polygenic and multifactorial, because in addition to genetic susceptibility, environmental factors (such as obesity, poor nutrition, and physical inactivity) modulate the phenotype. The in utero environment also contributes, and either increased or reduced birth weight increases the risk of type 2 DM in adult life. Children of pregnancies complicated by gestational hyperglycemia also exhibit an increased risk of type 2 DM. The genes that predispose to type 2 DM are incompletely identified, but genome-wide association studies have identified a large number of genes that convey a relatively small risk for type 2 DM (>70 genes, each with a relative risk of 1.06–1.5). Most prominent is a variant of the transcription factor 7–like 2 gene that has been associated with type 2 DM in several populations and with IGT. Genetic polymorphisms associated with type 2 DM have also been found in the genes encoding the peroxisome proliferator–activated receptor γ, inward rectifying potassium channel, zinc transporter, IRS, and calpain 10. The mechanisms by which these genetic loci increase the susceptibility to type 2 DM are not clear, but most are predicted to alter islet function or development or insulin secretion. Although the genetic susceptibility to type 2 DM is under active investigation (it is estimated that <10% of genetic risk is determined by loci identified thus far), it is currently not possible to use a combination of known genetic loci to predict type 2 DM.

Pathophysiology Type 2 DM is characterized by impaired insulin secretion, insulin resistance, excessive hepatic glucose production, abnormal fat metabolism, and systemic low-grade inflammation. Obesity, particularly visceral or central (as evidenced by the hip-waist ratio), is very common in type 2 DM (≥80% of patients are obese). In the early stages of the disorder, glucose tolerance remains near-normal, despite insulin resistance, because the pancreatic beta cells compensate by increasing insulin output (Fig. 396-7). As insulin resistance and compensatory hyperinsulinemia progress, the pancreatic islets in certain individuals are unable to sustain the hyperinsulinemic state. IGT, characterized by elevations in postprandial glucose, then develops. A further decline in insulin secretion and an increase in hepatic glucose production lead to overt diabetes with fasting hyperglycemia. Ultimately, beta cell failure ensues. Possibly because of inadequate insulin suppression, glucagon is relatively overproduced and secreted, further augmenting hepatic glucose production. Although both insulin resistance and impaired insulin secretion contribute to the pathogenesis of type 2 DM, the relative contribution of each varies from individual to individual.

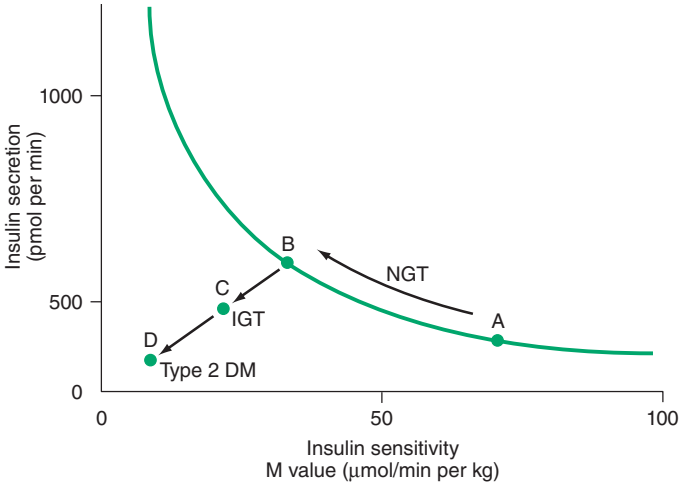


FIGURE 396-7 Metabolic changes during the development of type 2 diabetes mellitus (DM). Insulin secretion and insulin sensitivity are related, and as an individual becomes more insulin resistant (by moving from point A to point B), insulin secretion increases. A failure to compensate by increasing the insulin secretion results initially in impaired glucose tolerance (IGT; point C) and ultimately in type 2 DM (point D). NGT, normal glucose tolerance. *(Adapted from SE Kahn: J Clin Endocrinol Metab 86:4047, 2001; RN Bergman, M Ader: Trends Endocrinol Metab 11:351, 2000.)*

Metabolic Abnormalities • **ABNORMAL MUSCLE AND FAT METABOLISM** Insulin resistance, the decreased ability of insulin to act effectively on target tissues (especially muscle, liver, and fat), is a prominent feature of type 2 DM and results from a combination of genetic susceptibility and obesity. Insulin resistance is relative, however, because supranormal levels of circulating insulin will normalize the plasma glucose. Insulin dose-response curves exhibit a rightward shift, indicating reduced sensitivity, and a reduced maximal response, indicating an overall decrease in maximum glucose utilization (30–60% lower than in normal individuals). Insulin resistance impairs glucose utilization by insulin-sensitive tissues and increases hepatic glucose output; both effects contribute to the hyperglycemia. Increased hepatic glucose output predominantly accounts for increased FPG levels, whereas decreased peripheral glucose utilization results in postprandial hyperglycemia. In skeletal muscle, there is a greater impairment in nonoxidative glucose usage (glycogen formation) than in oxidative glucose metabolism through glycolysis. Glucose metabolism in insulin-independent tissues is not altered in type 2 DM.

The precise molecular mechanism leading to insulin resistance in type 2 DM has not been elucidated. Insulin receptor levels and tyrosine kinase activity in skeletal muscle are reduced, but these alterations are most likely secondary to hyperinsulinemia and are not a primary defect. Therefore, "postreceptor" defects in insulin-regulated phosphorylation/dephosphorylation appear to play the predominant role in insulin resistance. Abnormalities include the accumulation of lipid intermediates within skeletal myocytes, which may impair mitochondrial oxidative phosphorylation and reduce insulin-stimulated mitochondrial ATP production. Impaired fatty acid oxidation and lipid accumulation

within skeletal myocytes also may generate reactive oxygen species such as lipid peroxides. Of note, not all insulin signal transduction pathways are resistant to the effects of insulin (e.g., those controlling cell growth and differentiation using the mitogenic-activated protein kinase pathway). Consequently, hyperinsulinemia may increase the insulin action through these pathways, potentially accelerating diabetes-related conditions such as atherosclerosis.

The obesity accompanying type 2 DM, particularly in a central or visceral location, is thought to be part of the pathogenic process **(Chap. 394)**. In addition to these white fat depots, humans now are recognized to have brown fat, which has much greater thermogenic capacity. Efforts are under way to increase the activity or quantity of brown fat. The increased adipocyte mass leads to increased levels of circulating free fatty acids and other fat cell products. For example, adipocytes secrete a number of biologic products (nonesterified free fatty acids, retinol-binding protein 4, leptin, TNF-α, resistin, IL-6, and adiponectin). In addition to regulating body weight, appetite, and energy expenditure, adipokines also modulate insulin sensitivity. The increased production of free fatty acids and some adipokines may cause insulin resistance in skeletal muscle and liver. The venous drainage of the visceral adipose beds is the portal circulation and this likely contributes to hepatic dysfunction. Free fatty acids also impair glucose utilization in skeletal muscle, promote glucose production by the liver, and impair beta cell function. In contrast, the production by adipocytes of adiponectin, an insulin-sensitizing peptide, is reduced in obesity, and this may contribute to hepatic insulin resistance. Adipocyte products and adipokines also produce an inflammatory state and may explain why markers of inflammation such as IL-6 and C-reactive protein are often elevated in type 2 DM. In addition, inflammatory cells have been found infiltrating adipose tissue.

IMPAIRED INSULIN SECRETION Insulin secretion and sensitivity are interrelated (Fig. 396-7). In type 2 DM, insulin secretion initially increases in response to insulin resistance to maintain normal glucose tolerance. Initially, the insulin secretory defect is mild and selectively involves glucose-stimulated insulin secretion, including a greatly reduced first secretory phase. The response to other nonglucose secretagogues, such as arginine, is preserved, but overall beta cell function is reduced by as much as 50% at the onset of type 2 DM. Abnormalities in proinsulin processing are reflected by increased secretion of proinsulin in type 2 DM. Eventually, the insulin secretory defect is progressive.

The reason(s) for the decline in insulin secretory capacity in type 2 DM is unclear. The assumption is that a second genetic defect—superimposed upon insulin resistance—leads to defects in beta cell function, mass, and potentially cellular identity and differentiation status. Beta cell mass is decreased by ~50% in individuals with long-standing type 2 DM. Islet amyloid polypeptide or amylin, co-secreted by the beta cell, forms amyloid fibrillar deposits found in the islets of individuals with long-standing type 2 DM. Whether such islet amyloid deposits are a primary or secondary event is not known. The metabolic environment of diabetes also negatively impacts islet function. For example, chronic hyperglycemia paradoxically impairs islet function ("glucose toxicity") and leads to a worsening of hyperglycemia. Improvement in glycemic control is often associated with improved islet function. In addition, elevated levels of free fatty acids ("lipotoxicity"), and systemic and local elevations in pro-inflammatory cytokines from increased numbers of islet-associated macrophages, may also worsen islet function. Reduced GLP-1 action may contribute to the reduced insulin secretion.

INCREASED HEPATIC GLUCOSE AND LIPID PRODUCTION In type 2 DM, insulin resistance in the liver reflects the failure of hyperinsulinemia to suppress gluconeogenesis, which results in fasting hyperglycemia and decreased glycogen storage by the liver in the postprandial state. Increased hepatic glucose production occurs early in the course of diabetes, although likely after the onset of insulin secretory abnormalities and insulin resistance in skeletal muscle. As a result of insulin resistance in adipose tissue, lipolysis and free fatty acid flux from adipocytes are increased and efficiently cleared by liver leading to increased very-low-density lipoprotein [VLDL]-triglyceride) synthesis in hepatocytes and secretion from liver. This is also responsible for the dyslipidemia found in type 2 DM (elevated triglycerides, reduced high-density lipoprotein [HDL], and increased small dense low-density lipoprotein [LDL] particles). If this lipid is retained, steatosis in the liver may lead to nonalcoholic fatty liver disease and abnormal liver function tests.

Insulin Resistance Syndromes The insulin resistance condition comprises a spectrum of disorders, with hyperglycemia representing one of the most readily diagnosed features. The *metabolic syndrome*, the *insulin resistance syndrome*, and *syndrome X* are terms used to describe a constellation of metabolic derangements that includes insulin resistance, hypertension, dyslipidemia (decreased HDL and elevated triglycerides), central or visceral obesity, type 2 DM or IGT/IFG, and accelerated cardiovascular disease. This syndrome is discussed in **Chap. 401**.

A number of relatively rare forms of severe insulin resistance include features of type 2 DM or IGT (Table 396-1). Mutations in the insulin receptor that interfere with binding or signal transduction are a rare cause of insulin resistance. Acanthosis nigricans and signs of hyperandrogenism (hirsutism, acne, and oligomenorrhea in women) are also common physical features. Two distinct syndromes of severe insulin resistance have been described in adults: (1) type A, which affects young women and is characterized by severe hyperinsulinemia, obesity, and features of hyperandrogenism; and (2) type B, which affects middle-aged women and is characterized by severe hyperinsulinemia, features of hyperandrogenism, and autoimmune disorders. Individuals with the type A insulin resistance syndrome have an undefined defect in the insulin-signaling pathway; individuals with the type B insulin resistance syndrome have autoantibodies directed at the insulin receptor. These receptor autoantibodies may block insulin binding or may stimulate the insulin receptor, leading to intermittent hypoglycemia.

Polycystic ovary syndrome (PCOS) is a common disorder that affects premenopausal women and is characterized by chronic anovulation and hyperandrogenism **(Chap. 385)**. Insulin resistance is seen in a significant subset of women with PCOS, and the disorder substantially increases the risk for type 2 DM, independent of the effects of obesity.

Lipodystrophies are group of heterogeneous disorders characterized by selective loss of adipose tissue, leading to severe insulin resistance and hypertriglyceridemia. Lipodystrophies can be inherited or acquired and associated with variable degrees of adipose tissue loss.

Prevention Type 2 DM is preceded by a period of IGT or IFG, and a number of lifestyle modifications and pharmacologic agents prevent or delay the onset of DM. Individuals with prediabetes or increased risk of diabetes should be referred to a structured program to reduce body weight and increase physical activity as well as being screened for cardiovascular disease. The Diabetes Prevention Program (DPP) demonstrated that intensive changes in lifestyle (diet and exercise for 30 min/d five times/week) in individuals with IGT prevented or delayed the development of type 2 DM by 58% compared to placebo. This effect was seen in individuals regardless of age, sex, or ethnic group. In the same study, metformin prevented or delayed diabetes by 31% compared to placebo. The lifestyle intervention group lost 5–7% of their body weight during the 3 years of the study; the effects of the intervention persisted for at least 15 years. Studies in Finnish and Chinese populations noted similar efficacy of diet and exercise in preventing or delaying type 2 DM. A number of agents, including α-glucosidase inhibitors, metformin, thiazolidinediones, GLP-1 receptor pathway modifiers, and orlistat, prevent or delay type 2 DM but are not approved by the Food and Drug Administration for this purpose. Individuals with a strong family history of type 2 DM and individuals with IFG or IGT should be strongly encouraged to maintain a normal BMI and engage in regular physical activity. Pharmacologic therapy for individuals with prediabetes is currently controversial because its cost-effectiveness and safety profile are not known. The ADA suggests that metformin be considered in individuals with both IFG and IGT who are at very high risk for progression to diabetes (age <60 years, BMI ≥35 kg/m^2, and women with a history of GDM). Individuals with IFG, IGT, or an HbA_{1c} of 5.7–6.4% should be monitored annually to determine if diagnostic criteria for diabetes are present.

GENETICALLY DEFINED, MONOGENIC FORMS OF DM RELATED TO REDUCED INSULIN SECRETION

Several monogenic forms of DM have been identified. More than 10 different variants of MODY, caused by mutations in genes encoding islet-enriched transcription factors or glucokinase (Fig. 396-5; Table 396-1), are transmitted as autosomal dominant disorders. MODY 1, MODY 3, and MODY 5 are caused by mutations in hepatocyte nuclear transcription factor (HNF) 4α, HNF-1α, and HNF-1β, respectively. As their names imply, these transcription factors are expressed in the liver but also in other tissues, including the pancreatic islets and kidney. These factors most likely affect islet development or the expression of genes important in glucose-stimulated insulin secretion or the maintenance of beta cell mass. For example, individuals with an HNF-1α mutation (MODY 3) have a progressive decline in glycemic control but may respond to sulfonylureas. In fact, some of these patients were initially thought to have type 1 DM but were later shown to respond to a sulfonylurea, and insulin was discontinued. Individuals with a HNF-1β mutation have progressive impairment of insulin secretion and hepatic insulin resistance, and require insulin treatment with minimal response to sulfonylureas. These individuals often have other abnormalities such as renal cysts, mild pancreatic exocrine insufficiency, and abnormal liver function tests. Individuals with MODY 2, the result of mutations in the glucokinase gene, have mild-to-moderate, but stable hyperglycemia that does not respond to oral hypoglycemic agents. Glucokinase catalyzes the formation of glucose-6-phosphate from glucose, a reaction that is important for glucose sensing by the beta cells (Fig. 396-5) and for glucose utilization by the liver. As a result of glucokinase mutations, higher glucose levels are required to elicit insulin secretory responses, thus altering the set point for insulin secretion. MODY 4 is a rare variant caused by mutations in pancreatic and duodenal homeobox 1, a transcription factor that regulates pancreatic development and insulin gene transcription. Homozygous inactivating mutations cause pancreatic agenesis, whereas heterozygous mutations may result in DM. Studies of populations with type 2 DM suggest that mutations in MODY-associated genes are an uncommon (<5%) cause of type 2 DM.

Transient or permanent neonatal diabetes (onset <6 months of age) occurs. Permanent neonatal diabetes is a heterogeneous group of disorders caused by genetic mutations that impact beta cell function and/or pancreatic development (Fig. 396-5). Affected individuals typically require treatment with insulin and exhibit phenotypic overlap with type 1 DM. Activating mutations in the ATP-sensitive potassium channel subunits (Kir6.2 and ABCC8) impair glucose-stimulated insulin secretion. However, these individuals may respond to sulfonylureas and can be treated with these agents. Mutations in the transcription factor *GATA6* are the most common cause of pancreatic agenesis. Homozygous glucokinase mutations cause a severe form of neonatal diabetes, while mutations in mitochondrial DNA are associated with diabetes and deafness. A number of mutations identified in the coding sequence of the insulin gene have been found to interfere with proinsulin folding, processing, and bioactivity and are designated as Mutant *Ins*-gene-induced Diabetes of Youth (MIDYs). Some of the neonatal diabetes syndromes are associated with a spectrum of neurologic dysfunction and a variety of extrapancreatic manifestations. Any individual who developed diabetes at 6 months of age or who has atypical features of type 1 or type 2 diabetes should be screened for forms of monogenic diabetes.

APPROACH TO THE PATIENT

Diabetes Mellitus

Once the diagnosis of DM is made, attention should be directed to symptoms related to diabetes (acute and chronic) and classifying the type of diabetes. DM and its complications produce a wide range of symptoms and signs; those secondary to acute hyperglycemia may occur at any stage of the disease, whereas those related to chronic hyperglycemia typically begin to appear during the second decade of hyperglycemia **(Chap. 398)**. Because of long delays in clinical recognition, individuals with previously undetected type 2 DM may present with chronic complications of DM at the time of diagnosis. The history and physical examination should assess for symptoms or signs of acute hyperglycemia and screen for chronic microvascular and macrovascular complications and conditions associated with DM **(Chap. 398)**.

HISTORY

A complete medical history should be obtained with special emphasis on DM-relevant aspects such as current weight as well as any recent changes in weight, family history of DM and its complications, sleep history, risk factors for cardiovascular disease, exercise, smoking status, history of pancreatic disease, and ethanol use. Symptoms of hyperglycemia include polyuria, polydipsia, weight loss, fatigue, weakness, blurry vision, frequent superficial infections (vaginitis, fungal skin infections), and slow healing of skin lesions after minor trauma. Metabolic derangements relate mostly to hyperglycemia (osmotic diuresis) and to the catabolic state of the patient (urinary loss of glucose and calories, muscle breakdown due to protein degradation and decreased protein synthesis). Blurred vision results from changes in the water content of the lens and resolves as hyperglycemia is controlled.

In a patient with established DM, the initial assessment should include a review of symptoms at the time of the initial diabetes diagnosis. This is an essential part of the history that can help define whether the correct type of DM has been diagnosed. Special emphasis should be placed on prior diabetes care, including types of therapies tried, the nature of any intolerance to previous therapies, prior HbA_{1c} levels, self-monitoring blood glucose results, frequency of hypoglycemia (<3.0 mmol/L, <54 mg/dL), presence of DM-specific complications, and assessment of the patient's knowledge about diabetes, exercise, nutrition, and sleep history. Diabetes-related complications may afflict several organ systems, and an individual patient may exhibit some, all, or none of the symptoms related to the complications of DM **(Chap. 398)**. In addition, the presence of DM-related comorbidities should be established (cardiovascular disease, hypertension, dyslipidemia). Pregnancy plans should be ascertained in women of childbearing age. The American Diabetes Association recommends that all women of childbearing age be counseled about the importance of tight glycemic control (HbA_{1c} <6.5%) prior to conception.

PHYSICAL EXAMINATION

In addition to a complete physical examination, special attention should be given to DM-relevant aspects such as weight and BMI, retinal examination, orthostatic blood pressure, foot examination, peripheral pulses, and insulin injection sites. Blood pressure >130/80 mmHg is considered hypertension in individuals with diabetes. Because periodontal disease is more frequent in DM, the teeth and gums should also be examined.

An annual foot examination should (1) assess blood flow (pedal pulses), sensation (vibratory sensation [128-MHz tuning fork at the base of the great toe], the ability to sense touch with a monofilament [5.07, 10-g monofilament], pinprick sensation, ankle reflexes, and nail care; (2) look for the presence of foot deformities such as hammer or claw toes and Charcot foot; and (3) identify sites of potential ulceration. The ADA recommends annual screening for distal symmetric polyneuropathy beginning with the initial diagnosis of diabetes and annual screening for autonomic neuropathy 5 years after diagnosis of type 1 DM and at the time of diagnosis of type 2 DM. This testing is aimed at detecting loss of protective sensation (LOPS) caused by diabetic neuropathy **(Chap. 398)**.

CLASSIFICATION OF DM IN AN INDIVIDUAL PATIENT

The etiology of diabetes in an individual with new-onset disease can usually be assigned on the basis of clinical criteria. Individuals with type 1 DM tend to have the following characteristics: (1) onset of disease prior to age 30 years; (2) lean body habitus; (3) requirement

of insulin as the initial therapy; (4) propensity to develop ketoacidosis; and (5) an increased risk of other autoimmune disorders such as autoimmune thyroid disease, adrenal insufficiency, pernicious anemia, celiac disease, and vitiligo. In contrast, individuals with type 2 DM often exhibit the following features: (1) diabetes onset after the age of 30 years; (2) are usually obese (80% are obese, but elderly individuals may be lean); (3) may not require insulin therapy initially; and (4) may have associated conditions such as insulin resistance, hypertension, cardiovascular disease, dyslipidemia, or polycystic ovarian syndrome. In type 2 DM, insulin resistance is often associated with abdominal obesity (as opposed to hip and thigh obesity) and hypertriglyceridemia. Although most individuals diagnosed with type 2 DM are older, the age of diagnosis is declining, and there is a marked increase among overweight children and adolescents. Some individuals with phenotypic type 2 DM present with diabetic ketoacidosis but lack autoimmune markers and may be later treated with oral glucose-lowering agents rather than insulin (this clinical picture is sometimes referred to as *ketosis-prone type 2 DM*). On the other hand, some individuals (5–10%) with the phenotypic appearance of type 2 DM do not have absolute insulin deficiency but have autoimmune markers (GAD and other ICA autoantibodies) suggestive of type 1 DM (termed *latent autoimmune diabetes of the adult*). Such individuals are more likely to be <50 years of age, thinner, and have a personal or family history of other autoimmune disease than individuals with type 2 DM. They are much more likely to require insulin treatment within 5 years. Monogenic forms of diabetes (discussed above) should be considered in those with diabetes onset in childhood or early adulthood and especially those diagnosed within the first 6 months of life, an autosomal pattern of diabetes inheritance, diabetes without typical features of type 1 or 2 diabetes, and stable mild fasting hyperglycemia. Genetic testing should be considered in individuals suspected of having a monogenic form of diabetes as this may guide therapy selection. Despite recent advances in the understanding of the pathogenesis of diabetes, it often remains difficult to categorize some patients unequivocally. Individuals who deviate from the clinical profile of type 1 and type 2 DM, or who have other associated defects such as deafness, pancreatic exocrine disease (type 3c DM), and other endocrine disorders, should be classified accordingly (Table 396-1).

LABORATORY ASSESSMENT

The laboratory assessment should first determine whether the patient meets the diagnostic criteria for DM (Table 396-2) and then assess the degree of glycemic control **(Chap. 397)**. In addition to the standard laboratory evaluation, the patient should be screened for DM-associated conditions (e.g., albuminuria, dyslipidemia, thyroid dysfunction).

The classification of the type of DM may be facilitated by laboratory assessments. Serum insulin or C-peptide measurements may be useful, but should always be interpreted with a concurrent blood glucose level. A low C-peptide in the setting of an elevated blood glucose level may confirm a patient's need for insulin. However, C-peptide levels are unable to completely distinguish type 1 from type 2 DM, as many individuals with type 1 DM retain some C-peptide production. Measurement of islet cell antibodies at the time of diabetes onset may be useful if the type of DM is not clear based on the characteristics described above.

FURTHER READING

Atkinson MA et al: Type 1 diabetes. Lancet 383:69, 2014.

Classification and Diagnosis of Diabetes. Diabetes Care 41:S13, 2018.

Dall TM et al: The economic burden of elevated blood glucose levels in 2012: Diagnosed and undiagnosed diabetes, gestational diabetes mellitus, and prediabetes. Diabetes Care 37:3172, 2014.

Fuchsberger C et al: The genetic architecture of type 2 diabetes. Nature 536:41, 2016.

Hattersley AT, Patel KA: Precision diabetes: Learning from monogenic diabetes. Diabetologia 60:769, 2017.

Insel RA et al: Staging presymptomatic type 1 diabetes: A scientific statement of JDRF, the Endocrine Society, and the American Diabetes Association. Diabetes Care 38:1964, 2015.

Selph S et al: Screening for type 2 diabetes mellitus: A systematic review for the U.S. Preventive Services Task Force. Ann Intern Med 162:765, 2015.

Skyler JS et al: Differentiation of diabetes by pathophysiology, natural history, and prognosis. Diabetes 66:241, 2017.

397 Diabetes Mellitus: Management and Therapies

Alvin C. Powers, Kevin D. Niswender, Michael R. Rickels

OVERALL GOALS

The goals of therapy for type 1 or type 2 diabetes mellitus (DM) are to (1) eliminate symptoms related to hyperglycemia, (2) reduce or eliminate the long-term microvascular and macrovascular complications of DM **(Chap. 398)**, and (3) allow the patient to achieve as normal a lifestyle as possible. To reach these goals, the physician should identify a target level of glycemic control for each patient, provide the patient with the educational and pharmacologic resources necessary to reach this level, and monitor/treat DM-related complications. Symptoms of diabetes usually resolve when the plasma glucose is <11.1 mmol/L (200 mg/dL), and thus most DM treatment focuses on achieving the second and third goals. This chapter first reviews the ongoing treatment of diabetes in the outpatient setting and then discusses the treatment of severe hyperglycemia, as well as the treatment of diabetes in hospitalized patients.

The care of an individual with either type 1 or type 2 DM requires a multidisciplinary team. Central to the success of this team are the patient's participation, input, and enthusiasm, all of which are essential for optimal diabetes management. Members of the health care team include the primary care provider and/or the endocrinologist or diabetologist, a certified diabetes educator, a nutritionist, a psychologist, and/or social worker. In addition, when the complications of DM arise, subspecialists (including ophthalmologists, neurologists, podiatrists, nephrologists, transplant surgeons, cardiologists, and cardiovascular surgeons) with experience in DM-related complications are essential.

ONGOING ASPECTS OF COMPREHENSIVE DIABETES CARE

A number of names are sometimes applied to different approaches to diabetes care, such as intensive insulin therapy, intensive glycemic control, and "tight control." The current chapter, and other sources, uses the term *comprehensive diabetes care* to emphasize the fact that optimal diabetes therapy involves more than plasma glucose management and medications. Although glycemic control is central to optimal diabetes therapy, comprehensive diabetes care of both type 1 and type 2 DM should also detect and manage DM-specific complications **(Chap. 398)**, and modify risk factors for DM-associated diseases. The key elements of comprehensive diabetes care are summarized in **Table 397-1**. The morbidity and mortality of DM can be greatly reduced by timely and consistent surveillance, including the detection, prevention, and management of DM-related complications (Table 397-1 and **Chap. 398**). Such screening procedures are indicated for all individuals with DM, but many individuals with diabetes do not receive these or comprehensive diabetes care. In addition to the physical aspects of DM, social, family, financial, cultural, and employment-related issues may impact diabetes care. The treatment goals for patients with diabetes summarized in **Table 397-2** should be individualized. The prevention and treatment of clinically significant hypoglycemia (<3.0 mmol/L or 54 mg/dL) is discussed in **Chap. 399**. The International Diabetes Federation (IDF),

TABLE 397-1 Guidelines for Ongoing, Comprehensive Medical Care for Patients with Diabetes

- Individualized glycemic goal and therapeutic plan
- Self-monitoring of blood glucose (individualized frequency)
- HbA_{1c} testing (2–4 times/year)
- Lifestyle management in the care of diabetes, including:
  - Diabetes-self-management education and support
  - Nutrition therapy
  - Physical activity
  - Psychosocial care, including evaluation for depression, anxiety
- Detection, prevention, or management of diabetes-related complications, including:
  - Diabetes-related eye examination (annual or biannual; **Chap. 398**)
  - Diabetes-related foot examination (1–2 times/year by provider; daily by patient; **Chap. 398**)
  - Diabetes-related neuropathy examination (annual; **Chap. 398**)
  - Diabetes-related kidney disease testing (annual; **Chap. 398**)
- Manage or treat diabetes-relevant conditions, including:
  - Blood pressure (assess quarterly; **Chap. 398**)
  - Lipids (annual; **Chap. 398**)
  - Consider antiplatelet therapy **(Chap. 398)**
  - Influenza/pneumococcal/hepatitis B immunizations **(Chap. 4)**

Abbreviation: HbA_{1c}, hemoglobin A_{1c}.

recognizing that resources available for diabetes care vary widely throughout the world, has issued guidelines for "recommended care" (a well-developed service base and with health care funding systems consuming a significant part of their national wealth), "limited care" (health care settings with very limited resources), and "comprehensive care" (health care settings with considerable resources). This chapter provides guidance for this comprehensive level of diabetes care.

LIFESTYLE MANAGEMENT IN DIABETES CARE

The patient with type 1 or type 2 DM should receive education about nutrition, exercise, psychosocial support, care of diabetes during illness, and medications to lower the plasma glucose. The American Diabetes Association (ADA) uses the term "Lifestyle Management" to refer to aspects of diabetes care, including: (1) diabetes self-management education (DSME) and diabetes self-management support (DSMS); (2) nutrition therapy; and (3) psychosocial care. Along with improved compliance, patient education allows individuals with DM to assume greater responsibility for his/her care. Patient education should be viewed as a continuing process with regular visits for reinforcement; it should not be a process that is completed after one or two visits to a nurse educator or nutritionist. DSME and DSMS are ways to improve the patient's knowledge, skills, and abilities necessary for diabetes self-care and should also emphasize psychosocial issues and emotional well-being. More frequent contact between the patient and the diabetes management team (e.g., electronic, telephone) improves glycemic control.

TABLE 397-2 Treatment Goals for Adults with Diabetes[a]

INDEX	GOAL
Glycemic control[b]	
HbA_{1c}	<7.0%[c]
Preprandial capillary plasma glucose	4.4–7.2 mmol/L (80–130 mg/dL)
Postprandial capillary plasma glucose[d]	<10.0 mmol/L (<180 mg/dL)
Blood pressure	<140/90 mmHg[e]

[a]As recommended by the American Diabetes Association; goals should be individualized for each patient (see text) with a different goals for different patients. [b]HbA_{1c} is primary goal. [c]Diabetes Control and Complications Trial-based assay. [d]1–2 h after beginning of a meal. [e]The ADA also advises individualization of the BP goal with consideration of other co-morbidities and adverse events of therapy. A goal of of <130/80 mmHg may be appropriate for younger individuals or individuals with cardiovascular risk factors.

Abbreviation: HbA_{1c}, hemoglobin A_{1c}.

Source: Adapted from American Diabetes Association: Diabetes Care 41(Suppl 1): S86, 2018.

Diabetes Self-Management Education and Support The diabetes educator is a health care professional (nurse, dietician, or pharmacist) with specialized patient education skills who is certified in diabetes education (e.g., American Association of Diabetes Educators). Education topics important for optimal diabetes self-care include self-monitoring of blood glucose (SMBG); urine ketone monitoring (type 1 DM); insulin administration; guidelines for diabetes management during illnesses; prevention and management of hypoglycemia **(Chap. 399)**; foot and skin care; diabetes management before, during, and after exercise; and risk factor-modifying activities. The focus is providing patient-centered, individualized education.

Nutrition Therapy *Medical nutrition therapy* (MNT) is a term used by the ADA to describe the optimal coordination of caloric intake with other aspects of diabetes therapy (insulin, exercise, and weight loss). Primary measures of MNT are directed at preventing or delaying the onset of type 2 DM in high-risk individuals (obese or with prediabetes) by promoting weight reduction. Medical treatment of obesity is a rapidly evolving area and is discussed in **Chap. 395**. Secondary measures of MNT are directed at improving glycemic control. Tertiary measures of MNT are directed at managing diabetes-related complications (cardiovascular disease [CVD], nephropathy). Although the recommendations for all three types of MNT overlap, this chapter emphasizes secondary measures of MNT. Pharmacologic approaches that facilitate weight loss and metabolic surgery should be considered in selected patients **(Chaps. 394 and 395)**.

In general, the components of optimal MNT are similar for individuals with type 1 or type 2 DM and similar to those for the general population—high quality, nutrient-dense without a specific focus on composition (Mediterranean, dietary approaches to stop hypertension, etc.; **Table 397-3**). Historically, nutrition education imposed restrictive, complicated regimens on the patient. Current practices have greatly changed, although many patients and health care providers still view the diabetic diet as monolithic and static. There is not a specific diet for individuals with diabetes, certainly not a one diet for everyone. For example, MNT now includes foods with some sucrose and seeks to modify other risk factors such as hyperlipidemia and hypertension. Using the *glycemic index*, an estimate of the postprandial rise in the

TABLE 397-3 Nutritional Recommendations for Adults with Diabetes or Prediabetes[a]

General dietary guidelines

- Vegetable, fruits, whole grains, legumes, low-fat dairy products in food higher in fiber and lower in glycemic content

Fat in diet (optimal % of diet is not known; should be individualized)

- Mediterranean-style diet rich in monounsaturated fatty acids
- Minimal *trans* fat consumption

Carbohydrate in diet (optimal % of diet is not known; should be individualized)

- Monitor carbohydrate intake in regard to calories
- Sucrose-containing foods may be consumed with adjustments in insulin dose, but minimize intake
- Estimate grams of carbohydrate in diet (type 1 DM)
- Consider using glycemic index to predict how consumption of a particular food may affect blood glucose
- Fructose preferred over sucrose

Protein in diet (optimal % of diet is not known; should be individualized)

Other components

- Reduced-calorie and nonnutritive sweeteners may be useful
- Routine supplements of vitamins, antioxidants, or trace elements not supported by evidence
- Sodium intake as advised for general population

[a]See text for differences for patients with type 1 or type 2 diabetes.

Source: Adapted from American Diabetes Association: Diabetes Care 37(Suppl 1): S14, 2014 with modifications from American Diabetes Association: Diabetes Care 41(Suppl 1):S36, 2018.

blood glucose when a certain amount of that food is consumed, may reduce postprandial glucose excursions and improve glycemic control.

The goal of MNT in type 1 DM is to coordinate and match the caloric intake, both temporally and quantitatively, with the appropriate amount of insulin. MNT in type 1 DM and SMBG should be integrated to define the optimal insulin regimen. The ADA encourages patients and providers to use carbohydrate counting to estimate the nutrient content of a meal or snack. Based on the patient's estimate of the carbohydrate content of a meal, an insulin-to-carbohydrate ratio determines the bolus insulin dose for a meal or snack. MNT must be flexible enough to allow for exercise, and the insulin regimen must allow for deviations in caloric intake. An important component of MNT in type 1 DM is to minimize the weight gain often associated with intensive insulin therapy.

The goals of MNT in type 2 DM should focus on weight loss and address the greatly increased prevalence of cardiovascular risk factors (hypertension, dyslipidemia, obesity) and disease in this population. The majority of these individuals are obese, and weight loss is strongly encouraged. Hypocaloric diets and modest weight loss (5–7%) often result in rapid and dramatic glucose lowering in individuals with new-onset type 2 DM. MNT for type 2 DM should emphasize modest caloric reduction and increased physical activity. Weight loss and exercise improve insulin sensitivity.

Fasting for religious reasons, such as during Ramadan, presents a challenge for individuals with diabetes, especially those taking medications to lower the plasma glucose. Under IDF-issued guidelines on fasting, individuals are risk-stratified as those who can safely fast with medical evaluation and supervision and those in whom fasting is not advised.

Physical Activity Exercise has multiple positive benefits including cardiovascular risk reduction, reduced blood pressure, maintenance of muscle mass, reduction in body fat, and weight loss. For individuals with type 1 or type 2 DM, exercise is also useful for lowering plasma glucose (during and following exercise) and increasing insulin sensitivity. In patients with diabetes, the ADA recommends 150 min/week (distributed over at least 3 days) of moderate aerobic physical activity with no gaps longer than 2 days. Resistance exercise, flexibility and balance training, and reduced sedentary behavior throughout the day are advised.

Despite its benefits, exercise presents challenges for individuals with DM because they lack the normal glucoregulatory mechanisms (normally, insulin falls and glucagon rises during exercise). Skeletal muscle is a major site for metabolic fuel consumption in the resting state, and the increased muscle activity during vigorous, aerobic exercise greatly increases fuel requirements. Individuals with type 1 DM are prone to either hyperglycemia or hypoglycemia during exercise, depending on the preexercise plasma glucose, the circulating insulin level, and the level of exercise-induced catecholamines. If the insulin level is too low, the rise in catecholamines may increase the plasma glucose excessively, promote ketone body formation, and possibly lead to ketoacidosis. Conversely, if the circulating insulin level is excessive, this relative hyperinsulinemia may reduce hepatic glucose production (decreased glycogenolysis, decreased gluconeogenesis) and increase glucose entry into muscle, leading to hypoglycemia.

To avoid exercise-related hyper- or hypoglycemia, individuals with type 1 DM should (1) monitor blood glucose before, during, and after exercise; (2) delay exercise if blood glucose is >14 mmol/L (250 mg/dL) and ketones are present; (3) if the blood glucose is <5.6 mmol/L (100 mg/dL), ingest carbohydrate before exercising; (4) monitor glucose during exercise and ingest carbohydrate to prevent hypoglycemia; (5) decrease insulin doses (based on previous experience) before and after exercise and inject insulin into a nonexercising area; and (6) learn individual glucose responses to different types of exercise. In individuals with type 2 DM, exercise-related hypoglycemia is less common but can occur in individuals taking either insulin or insulin secretagogues. Untreated proliferative retinopathy is a relative contraindication to vigorous exercise, because this may lead to vitreous hemorrhage or retinal detachment (Chap. 398).

Psychosocial Care Because the individual with DM faces challenges that affect many aspects of daily life, psychosocial assessment and support are a critical part of comprehensive diabetes care. Patients should view themselves as essential members of the diabetes care team and not as someone who is cared for by the diabetes management team. Even with considerable effort, normoglycemia can be an elusive goal, and solutions to worsening glycemic control may not be easily identifiable. Depression, anxiety, or "Diabetes Distress," defined by the ADA as "…negative psychological reactions related to emotional burdens…in having to manage a chronic disease like diabetes, should be recognized and may require the care of a mental health specialist." Emotional stress may provoke a change in behavior so that individuals no longer adhere to a dietary, exercise, or therapeutic regimen. The individual with DM must accept that he or she may develop complications related to DM. Eating disorders, including binge eating disorders, bulimia, and anorexia nervosa, appear to occur more frequently in individuals with type 1 or type 2 DM.

MONITORING THE LEVEL OF GLYCEMIC CONTROL

Optimal monitoring of glycemic control involves plasma glucose measurements by the patient and an assessment of long-term control by the providers on the diabetes management team (measurement of hemoglobin A_{1c} [HbA_{1c}] and review of the patient's SMBG). These measurements are complementary: the patient's measurements provide a picture of short-term glycemic control, whereas the HbA_{1c} reflects average glycemic control over the previous 2–3 months.

Self-Monitoring of Blood Glucose SMBG is the standard of care in diabetes management and allows the patient to monitor his or her blood glucose at any time. In SMBG, a small drop of blood and an easily detectable enzymatic reaction allow measurement of the capillary plasma glucose. Many glucose monitors can rapidly and accurately measure glucose (calibrated to provide plasma glucose value even though blood glucose is measured) in small amounts of blood (3–10 μL) obtained from the fingertip; alternative testing sites (e.g., forearm) are less reliable. By combining glucose measurements with diet and exercise history, and medication changes, the diabetes management team and patient can improve the treatment program.

The frequency of SMBG measurements must be individualized and adapted to address the goals of diabetes care. Individuals with type 1 DM or individuals with type 2 DM taking multiple insulin injections each day should routinely measure their plasma glucose three or more times per day (some measure >10 times/day) to estimate and select mealtime boluses of short-acting insulin and to modify long-acting insulin doses. Most individuals with type 2 DM require less frequent monitoring, although the optimal frequency of SMBG has not been clearly defined. Individuals with type 2 DM who are taking insulin should use SMBG more frequently than those on oral agents. Individuals with type 2 DM who are on oral medications should use SMBG as a means of assessing the efficacy of their medication and the impact of dietary choices and exercise. Because plasma glucose levels fluctuate less in these individuals, one or fewer SMBG measurements per day may be sufficient. Most measurements in individuals with type 1 or type 2 DM should be performed prior to a meal and supplemented with postprandial measurements to assist in reaching postprandial glucose targets (Table 397-2).

Devices for continuous glucose monitoring (CGM) usually do not replace the need for traditional glucose measurements and require calibration by SMBG. These rapidly evolving technologies require substantial expertise on the part of the diabetes management team and the patient. Current CGM systems measure the glucose in interstitial fluid, which is in equilibrium with the plasma glucose. These devices provide useful short-term information about the patterns of glucose changes as well as an enhanced ability to detect hypoglycemic episodes. Alerts and alarms (vibration, sound) can notify the patient if the glucose is rising or falling rapidly, or is predicted to cross a hyper- or hypoglycemic threshold. Clinical experience with these devices in type 1 DM is growing rapidly, especially in individuals who have not achieved glycemic targets, those with hypoglycemia unawareness to

decrease the frequency of serious hypoglycemia (especially nocturnal hypoglycemia), and those desiring more frequent glycemic feedback. The combination of an insulin-infusion device (discussed below) and a CGM are currently open-loop, meaning the patient must adjust the insulin-infusion device, but closed-loop systems (insulin-infusion device automatically adjusted by algorithm) may soon be entering clinical practice; one system that adjusts the basal rate has been recently approved by the U.S. Food and Drug Administration (FDA).

Assessment of Long-Term Glycemic Control Measurement of glycated hemoglobin (HbA_{1c}) is the standard method for assessing long-term glycemic control. When plasma glucose is consistently elevated, there is an increase in nonenzymatic glycation of hemoglobin; this alteration reflects the glycemic history over the previous 2–3 months, because erythrocytes have an average life span of 120 days (glycemic level in the preceding month contributes about 50% to the HbA_{1c} value). Measurement of HbA_{1c} at the "point of care" allows for more rapid feedback and may therefore assist in adjustment of therapy.

HbA_{1c} should be measured in all individuals with DM during their initial evaluation and as part of their comprehensive diabetes care. As the primary predictor of long-term complications of DM, the HbA_{1c} should mirror, to a certain extent, the short-term measurements of SMBG. These two measurements are complementary in that recent intercurrent illnesses may impact the SMBG measurements but not the HbA_{1c}. Likewise, postprandial and nocturnal hyperglycemia may not be detected by the SMBG of fasting and preprandial capillary plasma glucose but will be reflected in the HbA_{1c}. The HbA_{1c} is an "average" and thus does not detect glycemic variability in the way SMBG and CGM can. In standardized assays, the HbA_{1c} approximates the following mean plasma glucose values: an HbA_{1c} of 6% = 7.0 mmol/L (126 mg/dL), 7% = 8.6 mmol/L (154 mg/dL), 8% = 10.2 mmol/L (183 mg/dL), 9% = 11.8 mmol/L (212 mg/dL), 10% = 13.4 mmol/L (240 mg/dL), 11% = 14.9 mmol/L (269 mg/dL), and 12% = 16.5 mmol/L (298 mg/dL). However, there is interindividual variability in the HbA_{1c} to mean glucose relationship, and in African-Americans the HbA_{1c} is on average 0.4% higher than in Caucasians for the same mean glucose. Clinical conditions leading to abnormal RBC parameters such as hemoglobinopathies, anemias, reticulocytosis, transfusions, and uremia may alter the HbA_{1c} result. In patients achieving their glycemic goal, the ADA recommends measurement of the HbA_{1c} at least twice per year. More frequent testing (every 3 months) is warranted when glycemic control is inadequate or when therapy has changed. Laboratory standards for the HbA_{1c} test have been established and should be correlated to the reference assay of the Diabetes Control and Complications Trial (DCCT). The degree of glycation of other proteins, such as albumin, or measurement of 1,5-anhydroglucitol can be used as an alternative indicator of glycemic control when the HbA_{1c} is inaccurate. The fructosamine assay (measuring glycated albumin) reflects the glycemic status over the prior 2 weeks.

PHARMACOLOGIC TREATMENT OF DIABETES

Comprehensive care of type 1 and type 2 DM requires an emphasis on nutrition, exercise, and monitoring of glycemic control but also usually involves glucose-lowering medication(s). This chapter discusses classes of such medications but does not describe every glucose-lowering agent available worldwide. The initial step is to select an individualized, glycemic goal for the patient.

ESTABLISHMENT OF TARGET LEVEL OF GLYCEMIC CONTROL

Because the complications of DM are related to glycemic control, normoglycemia or near-normoglycemia is the desired, but often elusive, goal for most patients. Normalization or near-normalization of the plasma glucose for long periods of time is extremely difficult, as demonstrated by the DCCT and United Kingdom Prospective Diabetes Study (UKPDS). Regardless of the level of hyperglycemia, improvement in glycemic control will lower the risk of diabetes-specific complications, most notably the microvascular complications **(Chap. 398)**.

The target for glycemic control (as reflected by the HbA_{1c}) must be individualized, and the goals of therapy should be developed in consultation with the patient after considering a number of medical, social, and lifestyle issues. The ADA calls this a *patient-centered approach*, and other organizations such as the IDF and American Association of Clinical Endocrinologists (AACE) also suggest an individualized glycemic goal. Important factors to consider include the patient's age and ability to understand and implement a complex treatment regimen, presence and severity of complications of diabetes, known CVD, ability to recognize hypoglycemic symptoms, presence of other medical conditions or treatments that might affect survival or the response to therapy, lifestyle and occupation (e.g., possible consequences of experiencing hypoglycemia on the job), and level of support available from family and friends.

In general, the ADA suggests that the goal is to achieve an HbA_{1c} as close to normal as possible without significant hypoglycemia. In most individuals, the target HbA_{1c} should be <7% (Table 397-2) with a more stringent target ≤6.5% for some patients. With modern implementation of intensive insulin therapy for type 1 DM, the level of HbA_{1c} is no longer inversely related to the frequency and severity of hypoglycemia as seen in the DCCT; nevertheless, it may still be appropriate to set a higher HbA_{1c} target <7.5 or 8% for patients with impaired awareness of hypoglycemia. A higher HbA_{1c} goal may also be appropriate for the very young or old or in individuals with limited life span or comorbid conditions. For example, an appropriate HbA_{1c} goal in elderly individuals with multiple, chronic illnesses and impaired activities of daily living might be 8.0 or 8.5%.

More stringent glycemic control (HbA_{1c} of ≤6%) is not beneficial, and may be detrimental, in patients with type 2 DM and a high risk of CVD. Large clinical trials (UKPDS, Action to Control Cardiovascular Risk in Diabetes [ACCORD], Action in Diabetes and Vascular Disease: Preterax and Diamicron MR Controlled Evaluation [ADVANCE], Veterans Affairs Diabetes Trial [VADT]; **Chap. 398**) have examined glycemic control in type 2 DM in individuals with low risk of CVD, with high risk of CVD, or with established CVD and have found that more intense glycemic control is not beneficial and, in some patient populations, may have a negative impact on some outcomes. These divergent outcomes stress the need for individualized glycemic goals based on the following general guidelines: (1) early in the course of type 2 diabetes when the CVD risk is lower, improved glycemic control likely leads to improved cardiovascular outcome, but this benefit may occur more than a decade after the period of improved glycemic control; (2) intense glycemic control in individuals with established CVD or at high risk for CVD is not advantageous, and may be deleterious, over a follow-up of 3–5 years; (3) hypoglycemia in such high-risk populations (elderly, CVD) should be avoided; and (4) improved glycemic control reduces microvascular complications of diabetes **(Chap. 398)** even if it does not improve macrovascular complications like CVD.

TYPE 1 DIABETES MELLITUS

General Aspects The ADA recommendations for fasting and bedtime glycemic goals and HbA_{1c} targets are summarized in Table 397-2. The goal is to design and implement insulin regimens that mimic physiologic insulin secretion. Because individuals with type 1 DM partially or completely lack endogenous insulin production, administration of basal insulin is essential for regulating glycogen breakdown, gluconeogenesis, lipolysis, and ketogenesis (i.e., largely fine-tuning hepatic and adipose metabolism). Likewise, insulin replacement for meals should be appropriate for the carbohydrate intake and promote normal glucose utilization and storage.

Intensive Management Intensive insulin therapy has the goal of achieving near-normal glycemia. This approach requires multiple resources, including thorough and continuing patient education, comprehensive recording of plasma glucose measurements and nutrition intake by the patient, and a variable insulin regimen that matches carbohydrate intake and insulin dose. Insulin regimens include multiple-component insulin regimens, multiple daily injections (MDIs), or continuous subcutaneous (SC) insulin infusion (CSII) (each discussed below).

The benefits of intensive insulin therapy and improved glycemic control include a reduction in the acute metabolic and chronic

microvascular complications of DM. From a psychological standpoint, the patient experiences greater control over his or her diabetes and often notes an improved sense of well-being, greater flexibility in the timing and content of meals, and the capability to alter insulin dosing with exercise. In addition, intensive insulin therapy prior to and during pregnancy reduces the risk of fetal malformations and morbidity. Intensive insulin therapy is encouraged in newly diagnosed patients with type 1 DM because it may prolong the period of C-peptide production, which may result in better glycemic control and a reduced risk of serious hypoglycemia. Although intensive management confers impressive benefits, it is also accompanied by significant personal and financial costs and is therefore not appropriate for all individuals.

Insulin Preparations Current insulin preparations are generated by recombinant DNA technology and consist of the amino acid sequence of human insulin or variations thereof. In the United States, most insulin is formulated as U-100 (100 units/mL); short-acting insulin formulated as U-200 (200 units/mL; lispro) and long-acting as U-300 (300 units/mL; glargine) are available in order to limit injection volumes for patients with high insulin requirements. Regular insulin formulated as U-500 (500 units/mL) is sometimes used in patients with severe insulin resistance. Human insulin has been formulated with distinctive pharmacokinetics (regular and neutral protamine Hagedorn [NPH] insulin have the native insulin amino acid sequence) or genetically modified to alter insulin absorption and hence insulin action. Insulins can be classified as short-acting or long-acting **(Table 397-4)**. For example, one short-acting insulin formulation, insulin lispro, is an insulin analogue in which the 28th and 29th amino acids (lysine and proline) on the insulin B chain have been reversed by recombinant DNA technology. Insulin aspart and insulin glulisine are genetically modified insulin analogues with properties similar to lispro. A biosimilar version of lispro has been approved. All three of these insulin analogues have full biologic activity but less tendency for self-aggregation, resulting in more rapid absorption and onset of action and a shorter duration of action. These characteristics are particularly advantageous for allowing entrainment of insulin injection and action to rising plasma glucose levels following meals. The shorter duration of action also appears to be associated with a decreased number of hypoglycemic episodes, primarily because the decay of insulin action corresponds to the decline in plasma glucose after a meal. Thus, insulin aspart, lispro, or glulisine is preferred over regular insulin for prandial coverage in many patients. Insulin glargine is a long-acting biosynthetic human insulin that differs from normal insulin in that asparagine is replaced by glycine at amino acid 21, and two arginine residues are added to the C terminus of the B chain, leading to the formation of microprecipitates at physiologic pH in subcutaneous tissue. Compared to NPH insulin, the onset of insulin glargine action is later, the duration of action is longer (~24 h), and there is a less pronounced peak. A lower incidence of hypoglycemia, especially at night, has been reported with insulin glargine when compared to NPH insulin. A biosimilar version is now available. Insulin detemir has a fatty acid side chain that reversibly binds to albumin and prolongs its action by slowing absorption and catabolism, but its duration of action may only reach 12–20 h. Twice-daily injections of glargine, or especially detemir, are sometimes required to provide optimal 24-h coverage. Because of modification and extension of the carboxy-terminal terminus of the B chain, insulin degludec forms multihexamers in subcutaneous tissue and binds albumin, prolonging its duration of action (>42 h); it provides similar glycemic control as glargine but with less frequent nocturnal and severe hypoglycemia.

Basal insulin requirements are provided by long-acting insulin formulations (NPH insulin, insulin glargine, insulin detemir, or insulin degludec). These are usually prescribed with short-acting insulin in an attempt to mimic physiologic insulin release with meals. Although mixing of NPH and short-acting insulin formulations is common practice, this mixing may alter the insulin absorption profile (especially the short-acting insulins). For example, lispro absorption is delayed by mixing with NPH. The alteration in insulin absorption when the patient mixes different insulin formulations should not prevent mixing insulins. However, the following guidelines should be followed: (1) mix the different insulin formulations in the syringe immediately before injection (inject within 2 min after mixing); (2) do not store insulin as a mixture; (3) follow the same routine in terms of insulin mixing and administration to standardize the physiologic response to injected insulin; and (4) do not mix insulin glargine, detemir, or degludec with other insulins. The miscibility of some insulins allows for the production of combination insulins that contain 70% NPH and 30% regular (70/30), or equal mixtures of NPH and regular (50/50). By including the insulin analogue mixed with protamine, several additional combinations have a short-acting and long-acting profile (Table 397-4). Although more convenient for the patient (only two injections/day), combination insulin formulations do not allow independent adjustment of short-acting and long-acting activity. Several insulin formulations are available as insulin "pens," which are more convenient for some patients. Other novel insulins, such as one with a duration of action of several days, are in development. Insulin delivery by inhalation to provide meal-time insulin is approved, but not widely used. Prior to its use, the forced expiratory volume in one second (FEV_1) should be measured. Inhaled insulin can cause bronchospasm and cough and should be not be in individuals with lung disease or who smoke. Long-acting insulin/glucagon-like peptide-1 (GLP-1) receptor agonist combinations in fixed doses (degludec + liraglutide or glargine + lixisenatide) have recently become available, are effective, and are not associated with weight gain.

TABLE 397-4 Properties of Insulin Preparations[a]

PREPARATION	TIME OF ACTION		
	ONSET, h	PEAK, h	EFFECTIVE DURATION, h
Short-acting[b]			
Aspart	<0.25	0.5–1.5	2–4
Glulisine	<0.25	0.5–1.5	2–4
Lispro[f]	<0.25	0.5–1.5	2–4
Regular[g]	0.5–1.0	2–3	3–6
Inhaled human insulin	0.5–1.0	2–3	3
Long-acting[g]			
Degludec	1–9	—[c]	42[d]
Detemir	1–4	—[c]	12–24[d]
Glargine[f]	2–4	—[c]	20–24
NPH	2–4	4–10	10–16
Examples of insulin combinations[e]			
75/25–75% protamine lispro, 25% lispro	<0.25	Dual[f]	10–16
70/30–70% protamine aspart, 30% aspart	<0.25	Dual[f]	15-18
50/50–50% protamine lispro, 50% lispro	<0.25	Dual[f]	10–16
70/30–70% NPH, 30% regular	0.5–1	Dual[f]	10–16
Combination of long-acting insulin and GLP-1 receptor agonist	See text		

[a]Injectable insulin preparations (with exception of inhaled formulation) available in the United States; others are available in the United Kingdom and Europe. [b]Formulation with niacinamide has a slightly more rapid onset and offset. [c]Degludec, determir, and glargine have minimal peak activity. [d]Duration is dose-dependent. [e]Other insulin combinations are available. [f]Dual: two peaks—one at 2–3 h and the second one several hours later. [g]Also available in concentrations >U-100.

Insulin Regimens Representations of the various insulin regimens that may be used in type 1 DM are illustrated in **Fig. 397-1**. Although the insulin profiles are depicted as "smooth," symmetric curves, there is considerable patient-to-patient variation in the peak and duration. In all regimens, long-acting insulins (NPH, glargine, detemir, or degludec) supply basal insulin, whereas regular, insulin aspart, glulisine, or lispro provide prandial insulin. Short-acting insulin analogues should be injected just before (<10 min) and regular insulin 30–45 min prior to a meal. Sometimes short-acting insulin analogues are injected just after a meal (gastroparesis, unpredictable food intake).

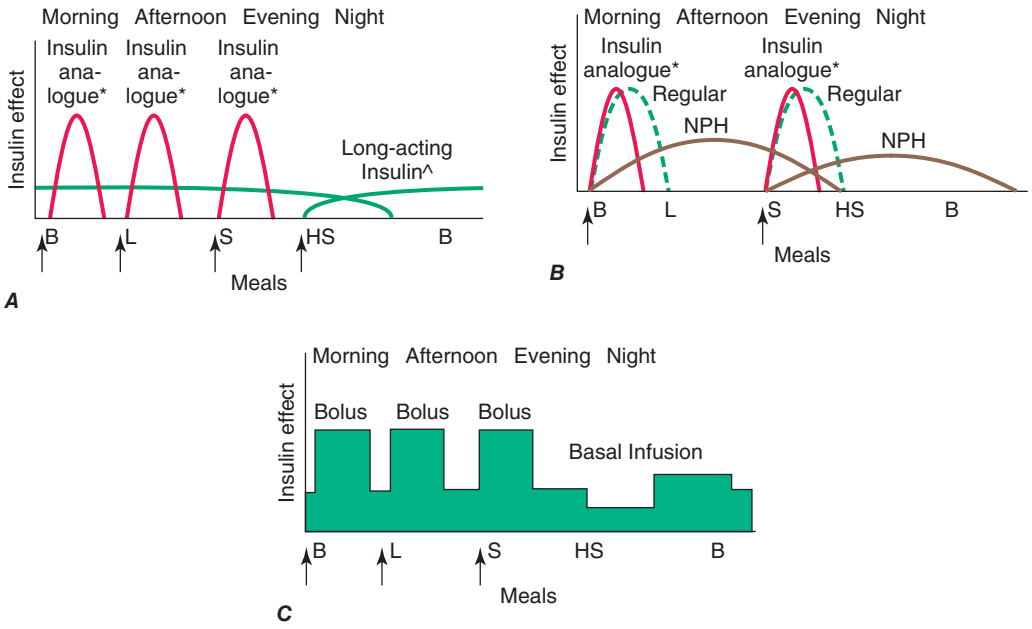


FIGURE 397-1 Representative insulin regimens for the treatment of diabetes. For each panel, the *y*-axis shows the amount of insulin effect and the *x*-axis shows the time of day. B, breakfast; HS, bedtime; L, lunch; S, supper. *Lispro, glulisine, or insulin aspart can be used. The time of insulin injection is shown with a *vertical arrow*. The type of insulin is noted above each insulin curve. ***A.*** Multiple-component insulin regimen consisting of long-acting insulin (degludec, determir, or glargine) to provide basal insulin coverage and three shots of glulisine, lispro, or insulin aspart to provide glycemic coverage for each meal. ***B.*** Injection of two shots of long-acting insulin (NPH) and short-acting insulin analogue (glulisine, lispro, insulin aspart [*solid red line*], or regular insulin [*green dashed line*]). Some deliver the second dose of NPH at bedtime or also use a short-acting insulin at lunch. Only one formulation of short-acting insulin is used. ***C.*** Insulin administration by insulin infusion device is shown with the basal insulin and a bolus injection at each meal. The basal insulin rate is decreased during the evening and increased slightly prior to the patient awakening in the morning. Glulisine, lispro, or insulin aspart is used in the insulin infusion device. *(Part C adapted from FR Kaufman: Medical Management of Type 1 Diabetes, 6th ed. American Diabetes Association, Alexandria, VA, 2012.)*

A shortcoming of current insulin regimens is that injected insulin immediately enters the systemic circulation, whereas endogenous insulin is secreted into the portal venous system. Thus, exogenous insulin administration exposes the liver to subphysiologic insulin levels. No insulin regimen reproduces the precise insulin secretory pattern of the pancreatic islet. However, the most physiologic regimens entail more frequent insulin injections, greater reliance on short-acting insulin, and more frequent capillary plasma glucose measurements (or by CGM). In general, individuals with type 1 DM require 0.4–1 units/kg per day of insulin divided into multiple doses, with ~50% of the insulin given as basal insulin.

MDI regimens refer to the combination of basal insulin and bolus insulin (preprandial short-acting insulin). The timing and dose of short-acting, preprandial insulin are altered to accommodate the SMBG results, anticipated food intake, and physical activity. Such regimens offer the patient with type 1 DM more flexibility in terms of lifestyle and the best chance for achieving near normoglycemia. One such regimen, shown in **Fig. 397-1*B***, consists of basal insulin with glargine, detemir, or degludec and preprandial lispro, glulisine, or insulin aspart. The insulin aspart, glulisine, or lispro dose is based on individualized algorithms that integrate the preprandial glucose and the anticipated carbohydrate intake. To determine the meal component of the preprandial insulin dose, the patient uses an insulin-to-carbohydrate ratio (a common ratio for type 1 DM is 1 unit/10–15 g of carbohydrate, but this must be determined for each individual). To this insulin, dose is added the supplemental or correcting insulin based on the preprandial blood glucose (one formula uses 1 unit of insulin for every 2.7 mmol/L [50 mg/dL] over the preprandial glucose target; another formula uses [body weight in kg] × [blood glucose – desired glucose in mg/dL]/1500). Such calculations must be adjusted based on each individual's sensitivity to insulin. An alternative multiple-component insulin regimen consists of bedtime NPH insulin, a small dose of NPH insulin at breakfast (20–30% of bedtime dose), and preprandial short-acting insulin. Other variations of this regimen are in use but have the disadvantage that NPH has a significant peak, making hypoglycemia more common. Frequent SMBG (≥4 times per day) is essential for these types of insulin regimens, although less frequent SMBG may be acceptable when used together with CGM.

In the past, one commonly used regimen consisted of twice-daily injections of NPH mixed with a short-acting insulin before the morning and evening meals (Fig. 397-1B). Such regimens usually prescribe two-thirds of the total daily insulin dose in the morning (with about two-thirds given as long-acting insulin and one-third as short-acting) and one-third before the evening meal (with approximately one-half given as long-acting insulin and one-half as short-acting). The drawback to such a regimen is that it forces a rigid schedule on the patient, in terms of daily activity and the content and timing of meals. Moreover, if the patient's meal pattern or content varies or if physical activity is increased, hyperglycemia or hypoglycemia may result. Moving the long-acting insulin from before the evening meal to bedtime may avoid nocturnal hypoglycemia and provide more insulin as glucose levels rise in the early morning as growth hormone and cortisol secretion peak (so-called dawn phenomenon). The insulin dose in such regimens should be adjusted based on SMBG results with the following general assumptions: (1) the fasting glucose is primarily determined by the prior evening long-acting insulin; (2) the prelunch glucose is a function of the morning short-acting insulin; (3) the presupper glucose is a function of the morning long-acting insulin; and (4) the bedtime glucose is a function of the presupper, short-acting insulin. This is not an optimal regimen for the patient with type 1 DM, but is sometimes used for patients with insulin-requiring type 2 DM.

CSII is a very effective insulin regimen for the patient with type 1 DM **(Fig. 397-1*C*)**. To the basal insulin infusion, a preprandial insulin ("bolus") is delivered by the insulin infusion device based on instructions from the patient, who uses an individualized algorithm incorporating the preprandial plasma glucose and anticipated carbohydrate intake. These sophisticated devices can accurately deliver small doses of insulin (microliters per hour) and have several advantages: (1) multiple basal infusion rates can be programmed to accommodate nocturnal versus daytime basal insulin requirement; (2) basal infusion rates can be altered during periods of exercise; (3) different waveforms of insulin infusion with meal-related bolus allow better matching of insulin depending on meal composition; and (4) programmed algorithms consider ongoing action of prior insulin administration and blood glucose values in calculating the insulin dose. These devices require instruction

by a health professional with considerable experience with insulin infusion devices and very frequent patient interactions with the diabetes management team. Insulin infusion devices present unique challenges, such as infection at the infusion site, unexplained hyperglycemia because the infusion set becomes obstructed, or diabetic ketoacidosis (DKA) if the insulin infusion device becomes disconnected. Because most physicians use lispro, glulisine, or insulin aspart in CSII, the extremely short half-life of these insulins quickly leads to insulin deficiency if the delivery system is interrupted. Essential to the safe use of infusion devices is thorough patient education, frequent SMBG (or by CGM), and a backup safety plan in the event of insulin infusion device failure. CGM sensor-augmented insulin infusion devices integrate the information from the CGM to inform insulin delivery. Currently, sensor communicating functions can interrupt basal insulin delivery during hypoglycemia (threshold suspension) or when hypoglycemia is anticipated (predictive suspension [not available in the United States]), which may be particularly useful for addressing nocturnal hypoglycemia. A partial closed-loop system has recently become available that combines patient-directed preprandial boluses with automated adjustment of between meal and basal insulin delivery based on CGM. Clinical experience with closed-loop systems is limited but increasing.

Other Agents That Improve Glucose Control The role of amylin, a 37-amino-acid peptide co-secreted with insulin from pancreatic beta cells, in normal glucose homeostasis is uncertain. However, based on the rationale that patients who are insulin deficient are also amylin deficient, an analogue of amylin (pramlintide) was created and found to reduce postprandial glycemic excursions in type 1 and type 2 diabetic patients taking insulin. Pramlintide injected just before a meal slows gastric emptying and suppresses glucagon but does not alter insulin levels. Pramlintide is approved for insulin-treated patients with type 1 and type 2 DM. Addition of pramlintide produces a modest reduction in the HbA_{1c} and seems to dampen meal-related glucose excursions. In type 1 DM, pramlintide is started as a 15-μg SC injection before each meal and titrated up to a maximum of 30–60 μg as tolerated. In type 2 DM, pramlintide is started as a 60-μg SC injection before each meal and may be titrated up to a maximum of 120 μg. The major side effects are nausea and vomiting, and dose escalations should be slow to limit these side effects. Because pramlintide slows gastric emptying, it may influence absorption of other medications and should not be used in combination with other drugs that slow gastrointestinal (GI) motility. The short-acting insulin given before the meal should initially be reduced to avoid hypoglycemia and then titrated as the effects of the pramlintide become evident. Because pramlintide suppresses glucagon, it may worsen hypoglycemia recovery and should not be used in patients with hypoglycemia unawareness.

TYPE 2 DIABETES MELLITUS

General Aspects The goals of glycemia-controlling therapy for type 2 DM are similar to those in type 1 DM. Whereas glycemic control tends to dominate the management of type 1 DM, the care of individuals with type 2 DM must also include attention to the treatment of conditions associated with type 2 DM (e.g., obesity, hypertension, dyslipidemia, CVD) and detection/management of DM-related complications **(Fig. 397-2; Chap. 398)**. Reduction in cardiovascular risk is of paramount importance because this is the leading cause of mortality in these individuals.

Type 2 DM management should begin with MNT (discussed above). An exercise regimen to increase insulin sensitivity and promote weight loss should also be instituted. Pharmacologic approaches to the management of type 2 DM include oral glucose-lowering agents, insulin, and other agents that improve glucose control; most physicians and patients prefer oral glucose-lowering agents as the initial choice. Any therapy that improves glycemic control reduces "glucose toxicity" to beta cells and may improve endogenous insulin secretion. However, type 2 DM is a progressive disorder and ultimately requires multiple therapeutic agents and often insulin in most patients.

Glucose-Lowering Agents Advances in the therapy of type 2 DM have generated oral glucose-lowering agents that target different pathophysiologic processes in type 2 DM. Based on their mechanisms of action, glucose-lowering agents are subdivided into agents that increase insulin secretion, reduce glucose production, increase insulin sensitivity, enhance GLP-1 action, or promote urinary excretion of glucose **(Table 397-5)**. Glucose-lowering agents other than insulin (with the exception of amylin analogue) are ineffective in type 1 DM and should not be used for glucose management of severely ill individuals with type 2 DM. Insulin is sometimes the initial glucose-lowering agent in type 2 DM.

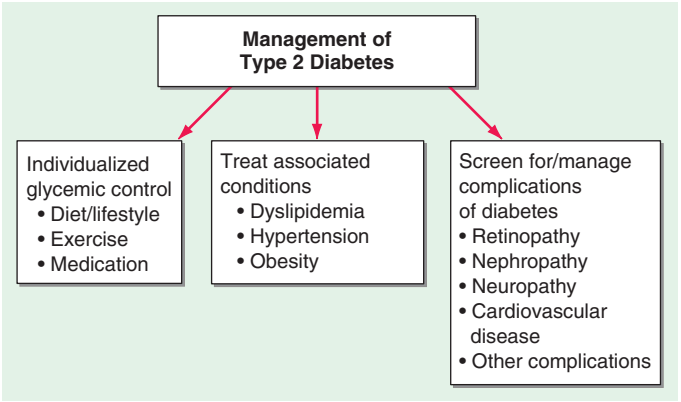


FIGURE 397-2 Essential elements in comprehensive care of type 2 diabetes.

BIGUANIDES Metformin, representative of this class of agents, reduces hepatic glucose production and improves peripheral glucose utilization slightly (Table 397-5). Metformin activates AMP-dependent protein kinase and enters cells through organic cation transporters (polymorphisms of these may influence the response to metformin). Recent evidence indicates that metformin's mechanism for reducing hepatic glucose production is to antagonize glucagon's ability to generate cAMP in hepatocytes. Metformin reduces fasting plasma glucose (FPG) and insulin levels, improves the lipid profile, and promotes modest weight loss. An extended-release form is available and may have fewer GI side effects (diarrhea, anorexia, nausea, metallic taste). Because of its relatively slow onset of action and GI symptoms with higher doses, the initial dose should be low and then escalated every 1–2 weeks based on SMBG measurements to a maximally tolerated dose of 2000 mg daily. Metformin is effective as monotherapy and can be used in combination with other oral agents or with insulin. Long-term use is associated with reduced micro- and probably macrovascular complications, but the data are less conclusive for macrovascular complications. The major toxicity of metformin, lactic acidosis, is very rare and can be prevented by careful patient selection. Vitamin B_{12} levels are lower during metformin treatment and should be monitored. Metformin should not be used in patients with moderate renal insufficiency (glomerular filtration rate [GFR] <45 mL/min), any form of acidosis, unstable congestive heart failure (CHF), liver disease, or severe hypoxemia. The National Institute for Health and Clinical Excellence in the United Kingdom suggests that metformin may be safe at a GFR >30 mL/min, with a reduced dose when the GFR is <45 mL/min. Metformin should be discontinued in hospitalized patients, in patients who can take nothing orally, and in those receiving radiographic contrast material. Insulin should be used until metformin can be restarted.

INSULIN SECRETAGOGUES—AGENTS THAT AFFECT THE ATP-SENSITIVE K^+ CHANNEL Insulin secretagogues stimulate insulin secretion by interacting with the ATP-sensitive potassium channel on the beta cell **(Chap. 396)**. These drugs are most effective in individuals with type 2 DM of relatively recent onset (<5 years) who have residual endogenous insulin production. First-generation sulfonylureas (chlorpropamide, tolazamide, tolbutamide) have a longer half-life, a greater incidence of hypoglycemia, and more frequent drug interactions, and are no longer used. Second-generation sulfonylureas have a more rapid onset of action and better coverage of the postprandial glucose rise, but the shorter half-life of some agents may require more than once-a-day dosing. Sulfonylureas reduce both fasting and postprandial glucose and

TABLE 397-5 Agents Used for Treatment of Type 1 or Type 2 Diabetes

	MECHANISM OF ACTION	EXAMPLES[a]	HBA_{1c} REDUCTION (%)[b]	AGENT-SPECIFIC ADVANTAGES	AGENT-SPECIFIC DISADVANTAGES	CONTRAINDICATIONS
Oral						
Biguanides[c*]	↓ Hepatic glucose production	Metformin	1–2	Weight neutral, do not cause hypoglycemia, inexpensive, extensive experience, ↓ CV events	Diarrhea, nausea, lactic acidosis, vitamin B12 deficiency	Renal insufficiency (see text for GFR <45 mL/min), CHF, radiographic contrast studies, hospitalized patients, acidosis
α-Glucosidase inhibitors[c**]	↓ GI glucose absorption	Acarbose, miglitol, voglibose	0.5–0.8	Reduce postprandial glycemia	GI flatulence, liver function tests	Renal/liver disease
Dipeptidyl peptidase IV inhibitors[c****]	Prolong endogenous GLP-1 action; ↑ Insulin, ↓ glucagon	Alogliptin, linagliptin, saxagliptin, sitagliptin, vildagliptin	0.5–0.8	Well tolerated, do not cause hypoglycemia	Angioedema/ urticarial and immune-mediated dermatologic effects	Reduced dose with renal disease
Insulin secretagogues: Sulfonylureas[c*]	↑ Insulin secretion	Glibornuride, gliclazide, glimepiride, glipizide, gliquidone, glyburide, glyclopyramide	1–2	Short onset of action, lower postprandial glucose, inexpensive	Hypoglycemia, weight gain	Renal/liver disease
Insulin secretagogues: Nonsulfonylureas[c***]	↑ Insulin secretion	Mitiglinide nateglinide, repaglinide	0.5–1.0	Short onset of action, lower postprandial glucose	Hypoglycemia	Renal/liver disease
Sodium-glucose cotransporter 2 inhibitors[***]	↑ renal glucose excretion	Canagliflozin, dapagliflozin, empagliflozin, ertugliflozin	0.5–1.0	do not cause hypoglycemia, ↓ weight and BP; see text for CVD effect	Urinary and genital infections, polyuria, dehydration, exacerbate tendency to hyperkalemia and DKA; see text	Moderate renal insufficiency, insulin-deficient DM
Thiazolidinediones[c***]	↓ Insulin resistance, ↑ glucose utilization	Pioglitazone, rosiglitazone	0.5–1.4	Lower insulin requirements	Peripheral edema, CHF, weight gain, fractures, macular edema	CHF, liver disease
Parenteral						
Amylin agonists[c,d***]	Slow gastric emptying, ↓ glucagon	Pramlintide	0.25–0.5	Reduce postprandial glycemia, weight loss	Injection, nausea, ↑ risk of hypoglycemia with insulin	Agents that also slow GI motility
GLP-1 receptor agonists[c***]	↑ Insulin, ↓ glucagon, slow gastric emptying, satiety	Albiglutide, dulaglutide, exenatide, liraglutide, lixisenatide, semaglutide	0.5–1.0	Weight loss, do not cause hypoglycemia; see text for CVD effect	Injection, nausea, ↑ risk of hypoglycemia with insulin secretagogues	Renal disease, agents that also slow GI motility; medullary carcinoma of thyroid, pancreatic disease
Insulin[c,d****]	↑ Glucose utilization, ↓ hepatic glucose production, and other anabolic actions	See text and Table 397-4	Not limited	Known safety profile	Injection, weight gain, hypoglycemia	
Medical nutrition therapy and physical activity[c*]	↓ Insulin resistance, ↑ insulin secretion	Low-calorie, low-fat diet, exercise	1–3	Other health benefits	Compliance difficult, long-term success low	

[a]Examples are approved for use in the United States; others are available in other countries. Examples may not include all agents in the class. [b]HbA*1c* reduction (absolute) depends partly on starting HbA_{1c}. [c]Used for treatment of type 2 diabetes. [d]Used in conjunction with insulin for treatment of type 1 diabetes. Cost of agent in the United States: *low, **moderate, ***high, ****variable.

Note: Some agents used to treat type 2 DM are not included in table (see text).

Abbreviations: ACE, angiotensin-converting enzyme; CHF, congestive heart failure; CV, cardiovascular; GI, gastrointestinal; HbA_{1c}, hemoglobin A_{1c}.

should be initiated at low doses and increased at 1- to 2-week intervals based on SMBG. In general, sulfonylureas increase insulin acutely and thus should be taken shortly before a meal; with chronic therapy, though, the insulin release is more sustained. Long-term use is associated with reduced micro- and macrovascular complications. Glimepiride and glipizide can be given in a single daily dose and are preferred over glyburide, especially in the elderly. Repaglinide, nateglinide, and mitiglinide are not sulfonylureas but also interact with the ATP-sensitive potassium channel. Because of their short half-life, these agents are given immediately before each meal to reduce meal-related glucose excursions.

Insulin secretagogues, especially the longer acting ones, have the potential to cause hypoglycemia, especially in elderly individuals. Hypoglycemia is usually related to delayed meals, increased physical activity, alcohol intake, or renal insufficiency. Individuals who ingest an overdose of some agents develop prolonged and serious hypoglycemia and should be monitored closely in the hospital **(Chap. 399)**. Most sulfonylureas are metabolized in the liver to compounds (some of which are active) that are cleared by the kidney. Thus, their use in individuals with significant hepatic or renal dysfunction is not advisable. Weight gain, a common side effect of sulfonylurea therapy, results from the increased insulin levels and improvement in glycemic control. Some sulfonylureas have significant drug interactions with alcohol and some medications including warfarin, aspirin, ketoconazole, α-glucosidase inhibitors, and fluconazole. A related isoform of ATP-sensitive potassium channels is present in the myocardium and the brain. All of these agents except glyburide have a low affinity for this isoform. Despite concerns that this agent might affect the myocardial response to ischemia and observational studies suggesting that sulfonylureas increase cardiovascular risk, studies have not shown an increased cardiac mortality with glyburide or other agents in this class.

INSULIN SECRETAGOGUES—AGENTS THAT ENHANCE GLP-1 RECEPTOR SIGNALING "Incretins" amplify glucose-stimulated insulin secretion (Chap. 396). Agents that either act as a GLP-1 receptor agonist or enhance endogenous GLP-1 activity are approved for the treatment of type 2 DM (Table 397-5). Agents in this class do not cause hypoglycemia because of the glucose-dependent nature of incretin-stimulated insulin secretion (unless there is concomitant use of an agent that can lead to hypoglycemia—sulfonylureas, etc.). GLP-1 receptor agonists increase glucose-stimulated insulin secretion, suppress glucagon, and slow gastric emptying. These agents do not promote weight gain; in fact, most patients experience modest weight loss and appetite suppression. Short-acting GLP-1 receptor agonists are exanatide and lixisenatide. Long-acting GLP-1 receptor agonists include liraglutide, exenatide, albiglutide, dulaglutide, and lixisenatide. Short-acting GLP-1 receptor agonists provide mostly postprandial coverage whereas the long-acting GLP-1 receptor agonists reduce both the postprandial and fasting glucose.

For example, exenatide, a synthetic version of a peptide initially identified in the saliva of the Gila monster (exendin-4), is an analogue of GLP-1. Unlike native GLP-1, which has a half-life of ~2 min, differences in the exenatide amino acid sequence render it resistant to the enzyme that degrades GLP-1 (dipeptidyl peptidase IV [DPP-IV]). Thus, exenatide has prolonged GLP-1-like action and binds to GLP-1 receptors found in islets, the GI tract, and the brain. Liraglutide, another GLP-1 receptor agonist, is almost identical to native GLP-1 except for an amino acid substitution and addition of a fatty acyl group (coupled with a γ-glutamic acid spacer) that promote binding to albumin and plasma proteins and prolong its half-life. Higher doses of liraglutide than used for glucose-lowering effects have been approved for weight loss therapy for obesity. Liraglutide treatment has also been associated with a decrease in CVD events in patients with type 2 DM and established CVD and with lower rates of diabetic kidney disease. In a similar patient population, semaglutide treatment was associated with fewer CVD events and reduced diabetic kidney disease, but with an increased rate of retinopathy-related complications. Whether the effect on CVD is a drug class effect is not clear as other GLP-1 receptor agonists have not reduced CVD events. Treatment with these agents should start at a low dose to minimize initial side effects (nausea being the limiting one). GLP-1 receptor agonists can be used as combination therapy with metformin, sulfonylureas, and thiazolidinediones. Some patients taking insulin secretagogues may require a reduction in those agents to prevent hypoglycemia. The major side effects are nausea, vomiting, and diarrhea. Some formulations carry a black box warning from the FDA because of an increased risk of thyroid C-cell tumors in rodents and are contraindicated in individuals with medullary carcinoma of the thyroid, multiple endocrine neoplasia, or pancreatic disease. Because GLP-1 receptor agonists slow gastric emptying, they may influence the absorption of other drugs. Whether GLP-1 receptor agonists enhance beta cell survival or promote beta cell proliferation in humans as in rodents is not known, but these agents do not appear to alter the natural history of type 2 DM.

DPP-IV inhibitors inhibit degradation of native GLP-1 and thus enhance the incretin effect. DPP-IV, which is widely expressed on the cell surface of endothelial cells and some lymphocytes, degrades a wide range of peptides (not GLP-1 specific). DPP-IV inhibitors promote insulin secretion in the absence of hypoglycemia or weight gain and appear to have a preferential effect on postprandial blood glucose. The levels of GLP-1 action in the patient are greater with the GLP-1 receptor agonists than with DPP-IV inhibitors. DPP-IV inhibitors are used either alone or in combination with other oral agents in type 2 DM. Reduced doses should be given to patients with renal insufficiency. There is conflicting evidence concerning a potentially increased risk for acute pancreatitis with GLP-1 receptor agonists and DPP-IV inhibitors, although initial concerns about possible premalignant lesions appear to be unfounded. For now, it is reasonable to avoid these agents in patients with pancreatic disease or with other significant risk factors for acute pancreatitis (e.g., heavy alcohol use, severely elevated serum triglycerides, hypercalcemia).

α-GLUCOSIDASE INHIBITORS α-Glucosidase inhibitors reduce postprandial hyperglycemia by delaying glucose absorption; they do not affect glucose utilization or insulin secretion (Table 397-5). Postprandial hyperglycemia, secondary to impaired hepatic and peripheral glucose disposal, contributes significantly to the hyperglycemic state in type 2 DM. These drugs, taken just before each meal, reduce glucose absorption by inhibiting the enzyme that cleaves oligosaccharides into simple sugars in the intestinal lumen. Therapy should be initiated at a low dose with the evening meal and increased to a maximal dose over weeks to months. The major side effects (diarrhea, flatulence, abdominal distention) are related to increased delivery of oligosaccharides to the large bowel and can be reduced somewhat by gradual upward dose titration. α-Glucosidase inhibitors may increase levels of sulfonylureas and increase the incidence of hypoglycemia. Simultaneous treatment with bile acid resins and antacids should be avoided. These agents should not be used in individuals with inflammatory bowel disease, gastroparesis, or a serum creatinine >177 μmol/L (2 mg/dL). This class of agents is not as potent as other oral agents in lowering the HbA_{1c} but is unique because it reduces the postprandial glucose rise. If hypoglycemia from other diabetes treatments occurs while taking these agents, the patient should consume glucose because the degradation and absorption of complex carbohydrates will be retarded.

THIAZOLIDINEDIONES Thiazolidinediones (Table 397-5) reduce insulin resistance by binding to the peroxisome proliferator-activated receptor γ (PPAR-γ) nuclear receptor (which forms a heterodimer with the retinoid X receptor). The PPAR-γ receptor is found at highest levels in adipocytes but is expressed at lower levels in many other tissues. Agonists of this receptor regulate a large number of genes, promote adipocyte differentiation, reduce hepatic fat accumulation, and promote fatty acid storage. Thiazolidinediones promote a redistribution of fat from central to peripheral locations. Circulating insulin levels decrease with use of the thiazolidinediones, indicating a reduction in insulin resistance. Although direct comparisons are not available, the two currently available thiazolidinediones appear to have similar efficacy. The prototype of this class of drugs, troglitazone, was withdrawn from the U.S. market after reports of hepatotoxicity and an association with an idiosyncratic liver reaction that sometimes led to hepatic failure. Although rosiglitazone and pioglitazone do not appear to induce the liver abnormalities seen with troglitazone, the FDA recommends measurement of liver function tests prior to initiating therapy. Modestly increased transaminase levels related to underlying fatty liver disease should not preclude treatment as these levels may improve with thiazolidinediones due to a reduction in hepatic fat content.

Rosiglitazone raises low-density lipoprotein (LDL), high-density lipoprotein (HDL), and triglycerides slightly. Pioglitazone raises HDL to a greater degree and LDL a lesser degree but lowers triglycerides. The clinical significance of the lipid changes with these agents is not known and may be difficult to ascertain because most patients with type 2 DM are also treated with a statin.

Thiazolidinediones are associated with weight gain (2–3 kg), a small reduction in the hematocrit, and a mild increase in plasma volume. Peripheral edema and CHF are more common in individuals treated with these agents. These agents are contraindicated in patients with liver disease or CHF (class III or IV). The FDA has issued an alert that rare patients taking these agents may experience a worsening of diabetic macular edema. An increased risk of fractures has been noted in postmenopausal women taking these agents. Thiazolidinediones have been shown to induce ovulation in premenopausal women with polycystic ovary syndrome. Women should be warned about the risk of pregnancy because the safety of thiazolidinediones in pregnancy is not established.

Concerns about increased cardiovascular risk associated with rosiglitazone led to considerable restrictions on its use and to the FDA issuing a "black box" warning in 2007. However, based on new information, the FDA has revised its guidelines and categorizes rosiglitazone similar to other drugs for type 2 DM. According to a recent FDA pronouncement, pioglitazone may be associated with an increased risk of bladder cancer. In one study, pioglitazone lowered the risk for recurrent stroke or myocardial infarction in insulin-resistant individuals without diabetes who had a prior stroke or transient ischemic attack.

Sodium-Glucose Co-Transporter 2 (SGLT2) Inhibitors These agents (Table 397-5) lower the blood glucose by selectively inhibiting this co-transporter, which is expressed almost exclusively in the proximal, convoluted tubule in the kidney. This inhibits glucose reabsorption, lowers the renal threshold for glucose, and leads to increased urinary glucose excretion. Thus, the glucose-lowering effect is insulin independent and not related to changes in insulin sensitivity or secretion. The loss of urinary glucose may promote modest weight reduction. Since these agents also impair proximal reabsorption of sodium, their use is associated with a diuretic effect and 3–6 mm Hg reduction in systolic blood pressure. Due to the increased urinary glucose, urinary and genital mycotic infections are more common in both men and women, and the diuretic effect can lead to reduced intravascular volume and acutely impaired kidney function. Inhibition of SGLT2 on the alpha cell may lead to increased glucagon and consequently liver production of glucose and ketones. Euglycemic DKA may occur during illness or when ongoing glucosuria masks stress-induced requirements for insulin. These agents should not be prescribed for patients with type 1 DM or pancreatogenic forms of DM associated with insulin deficiency. Empagliflozin and canagliflozin reduces CVD events and all cause cardiovascular mortality in patients with type 2 DM and established CVD, the risk for nephropathy, and the rate of hospitalization for CHF. A possible increased risk of bladder cancer has been seen with dapagliflozin; canagliflozin is associated with an increased risk of leg and foot amputation and bone fractures.

OTHER THERAPIES FOR TYPE 2 DM • Bile acid–binding resins Evidence indicates that bile acids, by signaling through nuclear receptors, may have a role in metabolism. Bile acid metabolism is abnormal in type 2 DM. The bile acid–binding resin colesevelam has been approved for the treatment of type 2 DM (already approved for treatment of hypercholesterolemia). Because bile acid–binding resins are minimally absorbed into the systemic circulation, how bile acid–binding resins lower blood glucose is not known. The most common side effects are GI (constipation, abdominal pain, and nausea). Bile acid–binding resins can increase plasma triglycerides and should be used cautiously in patients with a tendency for hypertriglyceridemia. The role of this class of drugs in the treatment of type 2 DM is not yet defined.

Bromocriptine A formulation of the dopamine receptor agonist bromocriptine (Cycloset) has been approved by the FDA for the treatment of type 2 DM. However, its role in the treatment of type 2 DM is uncertain.

INSULIN THERAPY IN TYPE 2 DM Insulin should be considered as part of the initial therapy in type 2 DM, particularly in lean individuals or those with severe weight loss, in individuals with underlying renal or hepatic disease that precludes oral glucose-lowering agents, or in individuals who are hospitalized or acutely ill. Insulin therapy is ultimately required by a substantial number of individuals with type 2 DM because of the progressive nature of the disorder and the relative insulin deficiency that develops in patients with long-standing diabetes. Both physician and patient reluctance often delay the initiation of insulin therapy, but glucose control and patient well-being are improved by insulin therapy in patients who have not reached glycemic targets.

Because endogenous insulin secretion continues and is capable of providing some coverage of mealtime caloric intake, insulin is usually initiated in a single dose of long-acting insulin (0.2–0.4 U/kg per day), given in the evening or just before bedtime (NPH, glargine, detemir, or degludec). Because fasting hyperglycemia and increased hepatic glucose production are prominent features of type 2 DM, bedtime insulin is more effective in clinical trials than a single dose of morning insulin. Glargine given at bedtime has less nocturnal hypoglycemia than NPH insulin. Some physicians prefer a relatively low, fixed starting dose of long-acting insulin (5–15 units) or a weight-based dose (0.1 units/kg). The insulin dose may then be adjusted in 10% increments as dictated by SMBG results. Both morning and bedtime long-acting insulin may be used in combination with oral glucose-lowering agents. Initially, basal insulin may be sufficient, but often prandial insulin coverage with multiple insulin injections is needed as diabetes progresses (see insulin regimens used for type 1 DM). Other insulin formulations that have a combination of short-acting and long-acting insulin (Table 397-4) are sometimes used in patients with type 2 DM because of convenience but do not allow independent adjustment of short-acting and long-acting insulin dose and often do not achieve the same degree of glycemic control as basal/bolus regimens. In selected patients with type 2 DM, insulin-infusion devices may be considered.

CHOICE OF INITIAL GLUCOSE-LOWERING AGENT The level of hyperglycemia and the patient's individualized goal (see "Establishment of Target Level of Glycemic Control") should influence the initial choice of therapy. Assuming that maximal benefit of MNT and increased physical activity has been realized, patients with mild hyperglycemia (FPG <7.0–11.0 mmol/L [126–199 mg/dL]) often respond well to a single, oral glucose-lowering agent, while those with moderate hyperglycemia (FPG 11.1–13.9 mmol/L [200–250 mg/dL]) will usually require more than one oral agent or insulin. Patients with more severe hyperglycemia (FPG >13.9 mmol/L [250 mg/dL]) may respond partially but are unlikely to achieve normoglycemia with oral therapy. Insulin can be used as initial therapy in individuals with severe hyperglycemia (FPG <13.9–16.7 mmol/L [250–300 mg/dL]) or in those who are symptomatic from the hyperglycemia. This approach is based on the rationale that more rapid glycemic control will reduce "glucose toxicity" to the islet cells, improve endogenous insulin secretion, and possibly allow oral glucose-lowering agents to be more effective. If this occurs, the insulin may be discontinued.

Insulin secretagogues, biguanides, α-glucosidase inhibitors, thiazolidinediones, GLP-1 receptor agonists, DPP-IV inhibitors, SLGT2 inhibitors, and insulin are approved for monotherapy of type 2 DM. Although each class of oral glucose-lowering agents has advantages and disadvantages (Table 397-5), certain generalizations apply: (1) insulin secretagogues, biguanides, GLP-1 receptor agonists, and thiazolidinediones improve glycemic control to a similar degree (1–2% reduction in HbA_{1c}) and are more effective than α-glucosidase inhibitors, DPP-IV inhibitors, and SLGT2 inhibitors; (2) assuming a similar degree of glycemic improvement, the clinical advantage of one class of drugs is not clear; any therapy that improves glycemic control is likely beneficial; (3) insulin secretagogues, GLP-1 receptor agonists, DPP-IV inhibitors, α-glucosidase inhibitors, and SLGT2 inhibitors begin to lower the plasma glucose immediately, whereas the glucose-lowering effects of the biguanides and thiazolidinediones are delayed by weeks; (4) not all agents are effective in all individuals with type 2 DM; (5) biguanides, α-glucosidase inhibitors, GLP-1 receptor agonists, DPP-IV inhibitors, thiazolidinediones, and SLGT2 inhibitors do not directly cause hypoglycemia; (6) most individuals will eventually require treatment with more than one class of oral glucose-lowering agents or insulin, reflecting the progressive nature of type 2 DM; and (7) durability of glycemic control is slightly less for sulfonylureas compared to metformin or thiazolidinediones.

Considerable clinical experience exists with metformin and sulfonylureas because they have been available for several decades. It is assumed that the α-glucosidase inhibitors, GLP-1 receptor agonists, DPP-IV inhibitors, thiazolidinediones, and SLGT2 inhibitors will reduce DM-related complications by improving glycemic control, but long-term data are not yet available. The thiazolidinediones are theoretically attractive because they target a fundamental abnormality in type 2 DM, namely insulin resistance.

Treatment algorithms by several professional societies (ADA/European Association for the Study of Diabetes [EASD], IDF, AACE) suggest metformin as initial therapy because of its efficacy, known side effect profile, and low cost **(Fig. 397-3)**. Metformin's advantages are that it promotes mild weight loss, lowers insulin levels, and improves the lipid profile slightly. Based on SMBG results and the HbA_{1c}, the dose of metformin should be increased until the glycemic target is achieved or maximum dose is reached.

COMBINATION THERAPY WITH GLUCOSE-LOWERING AGENTS A number of combinations of therapeutic agents are successful in type 2 DM metformin + second oral agent, metformin + GLP-1 receptor agonist, metformin + insulin, or combinations of a long-acting insulin and a GLP-1 receptor agonist. Because mechanisms of action of the first and

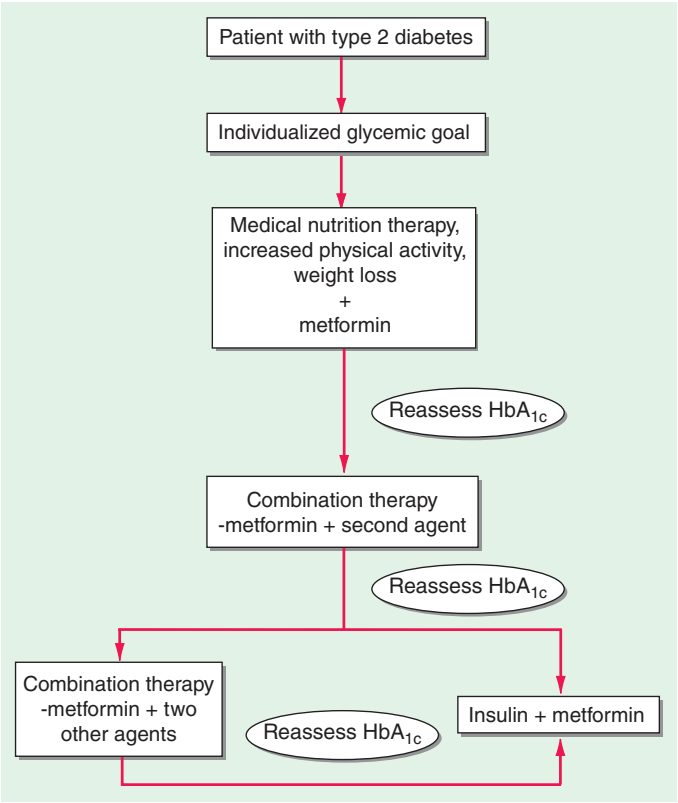


FIGURE 397-3 Glycemic management of type 2 diabetes. See text for discussion of treatment of severe hyperglycemia or symptomatic hyperglycemia. Agents that can be combined with metformin include insulin secretagogues, thiazolidinediones, α-glucosidase inhibitors, DPP-IV inhibitors, GLP-1 receptor agonists, SLGT2 inhibitors, and insulin. HbA_{1c}, hemoglobin HbA_{1c}.

second agents should be different, the effect on glycemic control is usually additive. There are little data to support the choice of one combination over another combination. Based on recent demonstrations of a beneficial cardiovascular effect in certain individuals with type 2 DM and CVD, or at high-risk of CVD, empagliflozin, canagliflozin, and liraglutide should now be considered in these populations. Medication costs vary considerably (Table 397-5), and this often factors into medication choice. Several fixed-dose combinations of oral agents are available, but evidence that they are superior to titration of single agent to a maximum dose and then addition of a second agent is lacking. If adequate control is not achieved with the combination of two agents (based on reassessment of the HbA_{1c} every 3 months), a third oral agent, GLP-1 receptor agonist, or basal insulin should be added (Fig. 397-3). Treatment approaches vary considerably from country to country. For example, α-glucosidase inhibitors are used commonly in South Asian patients (Indian), but infrequently in the United States or Europe. Whether this reflects an underlying difference in the disease or physician preference is not clear.

Treatment with insulin often becomes necessary as type 2 DM enters the phase of relative insulin deficiency and is signaled by inadequate glycemic control with one or two oral glucose-lowering agents. Insulin alone or in combination should be used in patients who fail to reach glycemic targets. For example, a single dose of long-acting insulin at bedtime is often effective in combination with metformin. As endogenous insulin production falls further, multiple injections of long-acting and short-acting insulin regimens are necessary to control postprandial glucose excursions. These insulin regimens are identical to the long-acting and short-acting combination regimens discussed above for type 1 DM, although usually at higher doses given insulin resistance. Weight gain and hypoglycemia are the major adverse effects of insulin therapy. The daily insulin dose required can become quite large (1–2 units/kg per day) as endogenous insulin production falls and insulin resistance persists. Individuals who require >1 unit/kg per day of long-acting insulin should be considered for combination therapy with metformin or a thiazolidinedione. The addition of metformin or a thiazolidinedione can reduce insulin requirements in some individuals with type 2 DM, while maintaining or even improving glycemic control. Insulin plus a thiazolidinedione promotes weight gain and is associated with peripheral edema. Addition of a thiazolidinedione to a patient's insulin regimen may necessitate a reduction in the insulin dose to avoid hypoglycemia. Patients requiring large doses of insulin (>200 units/day) can be treated with a more concentrated form of insulin.

SURGICAL THERAPIES

Whole pancreas transplantation can normalize glucose control in type 1 DM and when performed simultaneously with or after kidney transplantation can prolong the life of the kidney transplant by offering protection against recurrent diabetic nephropathy. Pancreatic islet transplantation is available as a less invasive form of beta cell replacement therapy for type 1 DM, but remains investigational in the United States. Due to the risks associated with chronic immunosuppression, whole pancreas and pancreatic islet transplantation may be considered for patients with severe metabolic instability or already requiring immunosuppression in support of a kidney or other organ transplant. Patients with chronic pancreatitis and preserved islet function who require pancreatectomy for pain relief may benefit from autologous islet transplantation as this may prevent or ameliorate postsurgical DM.

Metabolic (also referred to as bariatric) surgery for obese individuals with type 2 DM has shown considerable promise, sometimes with dramatic resolution of the diabetes or major reductions in the needed dose of glucose-lowering therapies **(Chap. 395)**. Several large, non-randomized clinical trials have demonstrated a much greater efficacy of metabolic surgery compared to medical management in the treatment of type 2 DM and may be considered in individuals with T2DM and a BMI >30 kg/m^2. The ADA clinical guidelines state that metabolic surgery should be considered in individuals with type 2 DM and a body mass index >30 kg/m^2 if hyperglycemia is inadequately controlled despite optimal medical therapy.

EMERGING THERAPIES

Many individuals with long-standing type 1 DM still produce very small amounts of insulin or have insulin-positive cells within the pancreas. This suggests that beta cells may slowly regenerate but are quickly destroyed by the autoimmune process. Particularly early in the disease course, efforts to suppress the autoimmune process and allow for beta cell regeneration are being tested at the time of diagnosis of type 1 DM, and for prevention in autoantibody-positive individuals at Stages 1 and 2 of type 1 DM (417-6). Closed-loop insulin infusion devices that infuse the appropriate amount of insulin in response to changing glucose levels are progressing rapidly. Bi-hormonal infusion devices that deliver both insulin and glucagon are under development. New therapies under evaluation or development for type 2 DM include activators of glucokinase, inhibitors of 11 β-hydroxysteroid dehydrogenase-1, GPR40 agonists, and agents to reduce inflammation.

Because whole pancreas and pancreatic islet transplantation are both limited by organ availability from deceased donors, stem cell-derived islet cells and xenogeneic sources of islets may eventually allow for a limitless supply of insulin-producing cells for transplantation.

ADVERSE EFFECTS OF THERAPY FOR DM

As with any therapy, the benefits of efforts directed toward glycemic control must be balanced against the risks of treatment (Table 397-5). Side effects of intensive treatment include an increased frequency of serious hypoglycemia, weight gain, increased economic costs, and greater demands on the patient. In the DCCT, quality of life was very similar in the intensive and standard therapy groups. The most serious complication of therapy for DM is hypoglycemia, and its treatment with oral glucose or glucagon injection is discussed in **Chap. 399**. Severe, recurrent hypoglycemia warrants examination of treatment regimen and glycemic goal for the individual patient. Weight gain occurs with most (insulin, insulin secretagogues, thiazolidinediones) but not all (metformin, α-glucosidase inhibitors, GLP-1 receptor

agonists, DPP-IV inhibitors) therapies. The weight gain is partially due to the anabolic effects of insulin and the reduction in glucosuria. As a result of concerns about CV safety of diabetes therapies, the FDA requires information about the cardiovascular safety profile as part of its evaluation of new medications for type 2 DM.

ACUTE DISORDERS RELATED TO SEVERE HYPERGLYCEMIA

Individuals with type 1 or type 2 DM and severe hyperglycemia (>13.9 mmol/L [250 mg/dL]) should be assessed for clinical stability, including mentation and hydration. Depending on the patient and the rapidity and duration of the severe hyperglycemia, an individual may require more intense and rapid therapy to lower the blood glucose. However, many patients with poorly controlled diabetes and hyperglycemia have few symptoms. The physician should assess if the patient is stable or if DKA or a hyperglycemic hyperosmolar state (HHS) should be considered. Ketones, an indicator of DKA, should be measured in individuals with type 1 DM when the plasma glucose is >13.9 mmol/L (250 mg/dL), during a concurrent illness, or with symptoms such as nausea, vomiting, or abdominal pain. Blood measurement of β-hydroxybutyrate is preferred over urine testing with nitroprusside-based assays that measure only acetoacetate and acetone.

DKA and HHS are acute, severe disorders directly related to diabetes. DKA was formerly considered a hallmark of type 1 DM, but also occurs in individuals with type 2 DM who can sometimes subsequently be treated with oral glucose-lowering agents (usually obese individuals of Hispanic or African-American descent). HHS is primarily seen in individuals with type 2 DM. Both disorders are associated with absolute or relative insulin deficiency, volume depletion, and acid-base abnormalities. DKA and HHS exist along a continuum of hyperglycemia, with or without ketosis. The metabolic similarities and differences in DKA and HHS are highlighted in Table 397-6. Both disorders are associated with potentially serious complications if not promptly diagnosed and carefully treated.

DIABETIC KETOACIDOSIS

Clinical Features The symptoms and physical signs of DKA are listed in Table 397-7 and usually develop over 24 h. DKA may be the initial symptom complex that leads to a diagnosis of type 1 DM, but more frequently, it occurs in individuals with established diabetes. Nausea and vomiting are often prominent, and their presence in an individual with diabetes warrants laboratory evaluation for DKA. Abdominal pain may be severe and can resemble acute pancreatitis or ruptured viscus. Hyperglycemia leads to glucosuria, volume depletion, and tachycardia. Hypotension can occur because of volume depletion in combination with peripheral vasodilatation. Kussmaul respirations and a fruity odor on the patient's breath (secondary to metabolic acidosis and increased acetone) are classic signs of the disorder. Lethargy and central nervous system depression may evolve into coma with severe DKA but should also prompt evaluation for other reasons for altered mental status (e.g., infection, hypoxemia). Cerebral edema, an extremely serious complication of DKA, is seen most frequently in children. Signs of infection, which may precipitate DKA, should be sought on physical examination, even in the absence of fever. Tissue ischemia (heart, brain) can also be a precipitating factor. Omission of insulin because of an eating disorder, mental health disorders, or an unstable psychosocial environment may sometimes be a factor precipitating DKA.

TABLE 397-7 Manifestations of Diabetic Ketoacidosis

Symptoms	Physical Findings
Nausea/vomiting	Tachycardia
Thirst/polyuria	Dehydration/hypotension
Abdominal pain	Tachypnea/Kussmaul respirations/ respiratory distress
Shortness of breath	Abdominal tenderness (may resemble acute pancreatitis or surgical abdomen)
Precipitating events	Lethargy/obtundation/cerebral edema/possibly coma
Inadequate insulin administration	
Infection (pneumonia/UTI/ gastroenteritis/sepsis)	
Infarction (cerebral, coronary, mesenteric, peripheral)	
Drugs (cocaine)	
Pregnancy	

Abbreviation: UTI, urinary tract infection.

TABLE 397-6 Laboratory Values in Diabetic Ketoacidosis (DKA) and Hyperglycemic Hyperosmolar State (HHS) (Representative Ranges at Presentation)

	DKA	HHS
Glucose,[a] mmol/L (mg/dL)	13.9–33.3 (250–600)	33.3–66.6 (600–1200)
Sodium, meq/L	125–135	135–145
Potassium[a,b]	Normal to ↑	Normal
Magnesium[a]	Normal	Normal
Chloride[a]	Normal	Normal
Phosphate[a,b]	Normal	Normal
Creatinine	Slightly ↑	Moderately ↑
Osmolality (mosm/mL)	300–320	330–380
Plasma ketones[a]	++++	+/−
Serum bicarbonate,[a] meq/L	<15	Normal to slightly ↓
Arterial pH	6.8–7.3	>7.3
Arterial Pco_2,[a] mmHg	20–30	Normal
Anion gap[a](Na − [Cl + HCO_3])	↑	Normal to slightly ↑

[a]Large changes occur during treatment of DKA. [b]Although plasma levels may be normal or high at presentation, total-body stores are usually depleted.

Pathophysiology DKA results from relative or absolute insulin deficiency combined with counterregulatory hormone excess (glucagon, catecholamines, cortisol, and growth hormone). Both insulin deficiency and glucagon excess, in particular, are necessary for DKA to develop. The decreased ratio of insulin to glucagon promotes gluconeogenesis, glycogenolysis, and ketone body formation in the liver, as well as increases in substrate delivery from fat and muscle (free fatty acids, amino acids) to the liver. Markers of inflammation (cytokines, C-reactive protein) are elevated in both DKA and HHS.

The combination of insulin deficiency and hyperglycemia reduces the hepatic level of fructose-2,6-bisphosphate, which alters the activity of phosphofructokinase and fructose-1,6-bisphosphatase. Glucagon excess decreases the activity of pyruvate kinase, whereas insulin deficiency increases the activity of phosphoenolpyruvate carboxykinase. These changes shift the handling of pyruvate toward glucose synthesis and away from glycolysis. The increased levels of glucagon and catecholamines in the face of low insulin levels promote glycogenolysis. Insulin deficiency also reduces levels of the GLUT4 glucose transporter, which impairs glucose uptake into skeletal muscle and fat and reduces intracellular glucose metabolism.

Ketosis results from a marked increase in free fatty acid release from adipocytes, with a resulting shift toward ketone body synthesis in the liver. Reduced insulin levels, in combination with elevations in catecholamines and growth hormone, increase lipolysis and the release of free fatty acids. Normally, these free fatty acids are converted to very low-density lipoprotein (VLDL) in the liver. However, in DKA, hyperglucagonemia alters hepatic metabolism to favor ketone body formation, through activation of the enzyme carnitine palmitoyltransferase I. This enzyme is crucial for regulating fatty acid transport into the mitochondria, where beta oxidation and conversion to ketone bodies occur. At physiologic pH, ketone bodies exist as ketoacids, which are neutralized by bicarbonate. As bicarbonate stores are depleted, metabolic acidosis ensues. Increased lactic acid production also contributes to the acidosis. The increased free fatty acids increase VLDL-triglyceride production. VLDL clearance is also reduced because the activity of insulin-sensitive lipoprotein lipase in muscle and fat is decreased.

DKA is often precipitated by increased insulin requirements, as occurs during a concurrent illness (Table 397-7). Failure to augment insulin therapy often compounds the problem. Complete omission or inadequate administration of insulin by the patient or health care team (in a hospitalized patient with type 1 DM) may precipitate DKA. Patients using insulin-infusion devices with short-acting insulin may develop DKA, because even a brief interruption in insulin delivery (e.g., mechanical malfunction) quickly leads to insulin deficiency.

Laboratory Abnormalities and Diagnosis The timely diagnosis of DKA is crucial and allows for prompt initiation of therapy. DKA is characterized by hyperglycemia (serum glucose > 13.9 mmol/L [250 mg/dL], ketosis, and metabolic acidosis (serum bicarbonate <15 mmol/L with increased anion gap) along with a number of secondary metabolic derangements (Table 397-6). Occasionally, the serum glucose is only minimally elevated, and may even be normal (euglycemic DKA). This has been noted in individuals treated with SGLT2 inhibitors. Arterial pH ranges between 6.8 and 7.3, depending on the severity of the acidosis. Despite a total-body potassium deficit, the serum potassium at presentation may be mildly elevated, secondary to the acidosis and volume depletion. Total-body stores of sodium, chloride, phosphorus, and magnesium are also reduced in DKA but are not accurately reflected by their levels in the serum because of hypovolemia and hyperglycemia. Elevated blood urea nitrogen (BUN) and serum creatinine levels reflect intravascular volume depletion. Leukocytosis, hypertriglyceridemia, and hyperlipoproteinemia are commonly found as well. Hyperamylasemia may suggest a diagnosis of pancreatitis, especially when accompanied by abdominal pain. However, in DKA the amylase is usually of salivary origin and thus is not diagnostic of pancreatitis. Serum lipase should be obtained if pancreatitis is suspected.

The measured serum sodium is reduced as a consequence of the hyperglycemia (1.6-mmol/L [1.6-meq] reduction in serum sodium for each 5.6-mmol/L [100-mg/dL] rise in the serum glucose). A normal serum sodium in the setting of DKA indicates a more profound water deficit. In "conventional" units, the calculated serum osmolality (2 × [serum sodium + serum potassium] + plasma glucose [mg/dL]/18 + BUN/2.8) is mildly to moderately elevated, although to a lesser degree than that found in HHS (see below).

In DKA, the ketone body, β-hydroxybutyrate, is synthesized at a threefold greater rate than acetoacetate; however, acetoacetate is preferentially detected by a commonly used ketosis detection reagent (nitroprusside). Serum ketones are present at significant levels (usually positive at serum dilution of ≥1:8). The nitroprusside tablet, or stick, is often used to detect urine ketones; certain medications such as captopril or penicillamine may cause false-positive reactions. Serum or plasma assays for β-hydroxybutyrate are preferred because they more accurately reflect the true ketone body level.

The metabolic derangements of DKA exist along a spectrum, beginning with mild acidosis with moderate hyperglycemia evolving into more severe findings. The degree of acidosis and hyperglycemia do not necessarily correlate closely because a variety of factors determine the level of hyperglycemia (oral intake, urinary glucose loss). Ketonemia is a consistent finding in DKA and distinguishes it from simple hyperglycemia. The differential diagnosis of DKA includes starvation ketosis, alcoholic ketoacidosis (bicarbonate usually >15 meq/L), and other forms of increased anion-gap acidosis (Chap. 51).

TREATMENT

Diabetic Ketoacidosis

The management of DKA is outlined in Table 397-8. After initiating IV fluid replacement and insulin therapy, the agent or event that precipitated the episode of DKA should be sought and aggressively treated. If the patient is vomiting or has altered mental status, a nasogastric tube should be inserted to prevent aspiration of gastric contents. Central to successful treatment of DKA is careful monitoring and frequent reassessment to ensure that the patient and the metabolic derangements are improving. A comprehensive flow sheet should record chronologic changes in vital signs, fluid intake and output, and laboratory values as a function of insulin administered.

TABLE 397-8 Management of Diabetic Ketoacidosis

1. Confirm diagnosis (↑ plasma glucose, positive serum ketones, metabolic acidosis).
2. Admit to hospital; intensive care setting may be necessary for frequent monitoring or if pH <7.00 or unconscious.
3. Assess:
   Serum electrolytes (K^+, Na^+, Mg^{2+}, Cl^-, bicarbonate, phosphate)
   Acid-base status—pH, HCO_3^-, Pco_2, β-hydroxybutyrate
   Renal function (creatinine, urine output)
4. Replace fluids: 2–3 L of 0.9% saline over first 1–3 h (10–20 mL/kg per hour); subsequently, 0.45% saline at 250–500 mL/h; change to 5% glucose and 0.45% saline at 150–250 mL/h when plasma glucose reaches 250 mg/dL (13.9 mmol/L).
5. Administer short-acting regular insulin: IV (0.1 units/kg), then 0.1 units/kg per hour by continuous IV infusion; increase two- to threefold if no response by 2–4 h. If the initial serum potassium is <3.3 mmol/L (3.3 meq/L), do not administer insulin until the potassium is corrected.
6. Assess patient: What precipitated the episode (noncompliance, infection, trauma, pregnancy, infarction, cocaine)? Initiate appropriate workup for precipitating event (cultures, CXR, ECG).
7. Measure capillary glucose every 1–2 h; measure electrolytes (especially K^+, bicarbonate, phosphate) and anion gap every 4 h for first 24 h.
8. Monitor blood pressure, pulse, respirations, mental status, fluid intake and output every 1–4 h.
9. Replace K^+: 10 meq/h when plasma K^+ <5.0–5.2 meq/L (or 20–30 meq/L of infusion fluid), ECG normal, urine flow and normal creatinine documented; administer 40–80 meq/h when plasma K^+ <3.5 meq/L or if bicarbonate is given. If initial serum potassium is >5.2 mmol/L (5.2 meq/L), do not supplement K^+ until the potassium is corrected.
10. See text about bicarbonate or phosphate supplementation.
11. Continue above until patient is stable, glucose goal is 8.3–11.1 mmol/L (150–200 mg/dL), and acidosis is resolved. Insulin infusion may be decreased to 0.02–0.1 units/kg per hour.
12. Administer long-acting insulin as soon as patient is eating. Allow for a 2–4 hour overlap in insulin infusion and SC long-acting insulin injection.

Abbreviations: CXR, chest x-ray; ECG, electrocardiogram.

Source: Adapted from M Sperling, in *Therapy for Diabetes Mellitus and Related Disorders*, American Diabetes Association, Alexandria, VA, 1998; and EA Nyenwe and AE Kitabchi: Metabolism 65:507, 2016.

After the initial bolus of normal saline, replacement of the sodium and free water deficit is carried out over the next 24 h (fluid deficit is often 3–5 L). When hemodynamic stability and adequate urine output are achieved, IV fluids should be switched to 0.45% saline depending on the calculated volume deficit. The change to 0.45% saline helps to reduce the trend toward hyperchloremia later in the course of DKA. Alternatively, initial use of lactated Ringer's IV solution may reduce the hyperchloremia that commonly occurs with normal saline.

A bolus of IV (0.1 units/kg) short-acting regular insulin is usually administered immediately (Table 397-8), and subsequent treatment should provide continuous and adequate levels of circulating insulin. IV administration is usually preferred (0.1 units/kg of regular insulin per hour) because it ensures rapid distribution and allows adjustment of the infusion rate as the patient responds to therapy. DKA can also be treated with SC short-acting insulin analogues. IV regular insulin should be continued until the acidosis resolves and the patient is metabolically stable. As the acidosis and insulin resistance associated with DKA resolve, the insulin infusion rate can be decreased (to 0.02–0.1 units/kg per hour). Long-acting insulin, in combination with SC short-acting insulin, should be administered as soon as the patient resumes eating, because this facilitates transition to an outpatient insulin regimen and reduces length of hospital stay. It is crucial to continue the insulin infusion until adequate insulin levels are achieved by administering long-acting insulin by the SC

route. Even relatively brief periods of inadequate insulin administration in this transition phase may result in DKA relapse.

Hyperglycemia usually improves at a rate of 4.2–5.6 mmol/L (50–100 mg/dL) per hour as a result of insulin-mediated glucose disposal, reduced hepatic glucose release, and rehydration. The latter reduces catecholamines, increases urinary glucose loss, and expands the intravascular volume. The decline in the plasma glucose within the first 1–2 h may be more rapid and is mostly related to volume expansion. When the plasma glucose reaches 11.1–13.9 mmol/L (200–250 mg/dL), glucose should be added to the 0.45% saline infusion to maintain the plasma glucose in the 8.3–11.1 mmol/L (150–200 mg/dL) range, and the insulin infusion should be continued at a lower rate to inhibit ketogenesis. More rapid correction of the serum glucose can precipitate the development of cerebral edema. Ketoacidosis begins to resolve as insulin reduces lipolysis, increases peripheral ketone body use, suppresses hepatic ketone body formation, and promotes bicarbonate regeneration. However, the acidosis and ketosis resolve more slowly than hyperglycemia. As ketoacidosis improves, β-hydroxybutyrate is converted to acetoacetate. Ketone body levels may appear to increase if measured by laboratory assays that use the nitroprusside reaction, which only detects acetoacetate and acetone. The improvement in acidosis and anion gap, a result of bicarbonate regeneration and decline in ketone bodies, is reflected by a rise in the serum bicarbonate level and the arterial pH. Depending on the rise of serum chloride, the anion gap (but not bicarbonate) will normalize. A hyperchloremic acidosis (serum bicarbonate of 15–18 mmol/L [15–18 meq/L]) often follows successful treatment and gradually resolves as the kidneys regenerate bicarbonate and excrete chloride.

Potassium stores are depleted in DKA (estimated deficit 3–5 mmol/kg [3–5 meq/kg]). During treatment with insulin and fluids, various factors contribute to the development of hypokalemia. These include insulin-mediated potassium transport into cells, resolution of the acidosis (which also promotes potassium entry into cells), and urinary loss of potassium salts of organic acids. Thus, potassium repletion should commence as soon as adequate urine output and a normal serum potassium are documented. If the initial serum potassium level is elevated, then potassium repletion should be delayed until the potassium falls into the normal range. Inclusion of 20–40 meq of potassium in each liter of IV fluid is reasonable, but additional potassium supplements may also be required. To reduce the amount of chloride administered, potassium phosphate or acetate can be substituted for the chloride salt. The goal is to maintain the serum potassium at >3.5 mmol/L (3.5 meq/L).

Despite a bicarbonate deficit, bicarbonate replacement is not usually necessary. In fact, theoretical arguments suggest that bicarbonate administration and rapid reversal of acidosis may impair cardiac function, reduce tissue oxygenation, and promote hypokalemia. The results of most clinical trials do not support the routine use of bicarbonate replacement, and one study in children found that bicarbonate use was associated with an increased risk of cerebral edema. However, in the presence of severe acidosis (arterial pH <7.0), the ADA advises bicarbonate (50 mmol [meq/L] of sodium bicarbonate in 200 mL of sterile water with 10 meq/L KCl per hour for 2 h until the pH is >7.0). Hypophosphatemia may result from increased glucose usage, but randomized clinical trials have not demonstrated that phosphate replacement is beneficial in DKA. If the serum phosphate is <0.32 mmol/L (1 mg/dL), then phosphate supplement should be considered and the serum calcium monitored. Hypomagnesemia may develop during DKA therapy and may also require supplementation.

With appropriate therapy, the mortality rate of DKA is low (<1%) and is related more to the underlying or precipitating event, such as infection or myocardial infarction. Venous thrombosis, upper GI bleeding, and acute respiratory distress syndrome occasionally complicate DKA. The major nonmetabolic complication of DKA therapy is cerebral edema, which most often develops in children as DKA is resolving. The etiology of and optimal therapy for cerebral edema are not well established, but overreplacement of free water and rapid normalization of serum glucose should be avoided.

Following treatment, the physician and patient should review the sequence of events that led to DKA to prevent future recurrences. Foremost is patient education about the symptoms of DKA, its precipitating factors, and the management of diabetes during a concurrent illness. During illness or when oral intake is compromised, patients should (1) frequently measure the capillary blood glucose; (2) measure urinary ketones when the serum glucose is >13.7 mmol/L (250 mg/dL); (3) drink fluids to maintain hydration; (4) continue or increase insulin; and (5) seek medical attention if dehydration, persistent vomiting, or uncontrolled hyperglycemia develop. Using these strategies, early DKA can be prevented or detected and treated appropriately on an outpatient basis.

HYPERGLYCEMIC HYPEROSMOLAR STATE

Clinical Features The prototypical patient with HHS is an elderly individual with type 2 DM, with a several-week history of polyuria, weight loss, and diminished oral intake that culminates in mental confusion, lethargy, or coma. The physical examination reflects profound dehydration and hyperosmolality and reveals hypotension, tachycardia, and altered mental status. Notably absent are symptoms of nausea, vomiting, and abdominal pain and the Kussmaul respirations characteristic of DKA. HHS is often precipitated by a serious, concurrent illness such as myocardial infarction or stroke. Sepsis, pneumonia, and other serious infections are frequent precipitants and should be sought. In addition, a debilitating condition (prior stroke or dementia) or social situation that compromises water intake usually contributes to the development of the disorder.

Pathophysiology Relative insulin deficiency and inadequate fluid intake are the underlying causes of HHS. Insulin deficiency increases hepatic glucose production (through glycogenolysis and gluconeogenesis) and impairs glucose utilization in skeletal muscle (see above discussion of DKA). Hyperglycemia induces an osmotic diuresis that leads to intravascular volume depletion, which is exacerbated by inadequate fluid replacement. The absence of ketosis in HHS is not understood. Presumably, the insulin deficiency is only relative and less severe than in DKA. Lower levels of counterregulatory hormones and free fatty acids have been found in HHS than in DKA in some studies. It is also possible that the liver is less capable of ketone body synthesis or that the insulin/glucagon ratio does not favor ketogenesis.

Laboratory Abnormalities and Diagnosis The laboratory features in HHS are summarized in Table 397-6. Most notable are the marked hyperglycemia (plasma glucose may be >55.5 mmol/L [1000 mg/dL]), hyperosmolality (>350 mosmol/L), and prerenal azotemia. The measured serum sodium may be normal or slightly low despite the marked hyperglycemia. The corrected serum sodium is usually increased (add 1.6 meq to measured sodium for each 5.6-mmol/L [100-mg/dL] rise in the serum glucose). In contrast to DKA, acidosis and ketonemia are absent or mild. A small anion-gap metabolic acidosis may be present secondary to increased lactic acid. Moderate ketonuria, if present, is secondary to starvation.

TREATMENT

Hyperglycemic Hyperosmolar State

Volume depletion and hyperglycemia are prominent features of both HHS and DKA. Consequently, therapy of these disorders shares several elements (Table 397-8). In both disorders, careful monitoring of the patient's fluid status, laboratory values, and insulin infusion rate is crucial. Underlying or precipitating problems should be aggressively sought and treated. In HHS, fluid losses and dehydration are usually more pronounced than in DKA due to the longer duration of the illness. The patient with HHS is usually older, more likely to have mental status changes, and more likely to have a life-threatening precipitating event with accompanying comorbidities. Even with proper treatment, HHS has a substantially higher mortality rate than DKA (up to 15% in some clinical series).

Fluid replacement should initially stabilize the hemodynamic status of the patient (1–3 L of 0.9% normal saline over the first 2–3 h). Because the fluid deficit in HHS is accumulated over a period of days to weeks, the rapidity of reversal of the hyperosmolar state must balance the need for free water repletion with the risk that too rapid a reversal may worsen neurologic function. If the serum sodium is >150 mmol/L (150 meq/L), 0.45% saline should be used. After hemodynamic stability is achieved, the IV fluid administration is directed at reversing the free water deficit using hypotonic fluids (0.45% saline initially, then 5% dextrose in water [D_5W]). The calculated free water deficit (which averages 9–10 L) should be reversed over the next 1–2 days (infusion rates of 200–300 mL/h of hypotonic solution). Potassium repletion is usually necessary and should be dictated by repeated measurements of the serum potassium. In patients taking diuretics, the potassium deficit can be quite large and may be accompanied by magnesium deficiency. Hypophosphatemia may occur during therapy and can be improved by using KPO_4 and beginning nutrition.

As in DKA, rehydration and volume expansion lower the plasma glucose initially, but insulin is also required. A reasonable regimen for HHS begins with an IV insulin bolus of 0.1 unit/kg followed by IV insulin at a constant infusion rate of 0.1 unit/kg per hour. If the serum glucose does not fall, increase the insulin infusion rate by twofold. As in DKA, glucose should be added to IV fluid when the plasma glucose falls to 11.1–13.9 mmol/L (200–250 mg/dL), and the insulin infusion rate should be decreased to 0.02–0.1 unit/kg per hour. The insulin infusion should be continued until the patient has resumed eating and can be transferred to a SC insulin regimen. The patient should be discharged from the hospital on insulin, although some patients can later switch to oral glucose-lowering agents.

MANAGEMENT OF DIABETES IN A HOSPITALIZED PATIENT

Virtually all medical and surgical subspecialties are involved in the care of hospitalized patients with diabetes. Hyperglycemia, whether in a patient with known diabetes or in someone without known diabetes, appears to be a predictor of poor outcome in hospitalized patients. General anesthesia, surgery, infection, or concurrent illness raises the levels of counterregulatory hormones (cortisol, growth hormone, catecholamines, and glucagon) and cytokines that may lead to transient insulin resistance and hyperglycemia. These factors increase insulin requirements by increasing glucose production and impairing glucose utilization and thus may worsen glycemic control. The concurrent illness or surgical procedure may lead to variable insulin absorption and also prevent the patient with DM from eating normally and, thus, may promote hypoglycemia. Glycemic control should be assessed on admission using the HbA_{1c}. Electrolytes, renal function, and intravascular volume status should be assessed as well. The high prevalence of CVD in individuals with DM (especially in type 2 DM) may necessitate preoperative cardiovascular evaluation **(Chap. 398)**.

The goals of diabetes management during hospitalization are near-normoglycemia, avoidance of hypoglycemia, and transition back to the outpatient diabetes treatment regimen. Upon hospital admission, frequent glycemic monitoring should begin, as should planning for diabetes management after discharge. Glycemic control appears to improve the clinical outcomes in a variety of settings, but optimal glycemic goals for the hospitalized patient are incompletely defined. In a number of cross-sectional studies of patients with diabetes, a greater degree of hyperglycemia was associated with worse cardiac, neurologic, and infectious outcomes. In some studies, patients who do not have preexisting diabetes but who develop modest blood glucose elevations during their hospitalization appear to benefit from achieving near-normoglycemia using insulin treatment. However, a large randomized clinical trial (Normoglycemia in Intensive Care Evaluation Survival Using Glucose Algorithm Regulation [NICE-SUGAR]) of individuals in the intensive care unit (ICU) (most of whom were receiving mechanical ventilation) found an increased mortality rate and a greater number of episodes of severe hypoglycemia with very strict glycemic control (target blood glucose of 4.5–6 mmol/L or 81–108 mg/dL) compared to individuals with a more moderate glycemic goal (target blood glucose of <10 mmol/L or 180 mg/dL). Currently, most data suggest that very strict blood glucose control in acutely ill patients likely worsens outcomes and increases the frequency of hypoglycemia. The ADA suggests the following glycemic goals for hospitalized patients: (1) in critically or non-critically ill patients: glucose of 7.8–10.0 mmol/L or 140–180 mg/dL; (2) in selected patients: glucose of 6.1–7.8 mmol/L or 110–140 mg/dL with avoidance of hypoglycemia.

Critical aspects for optimal diabetes care in the hospital include the following. (1) A hospital-wide system approach to treatment of hyperglycemia and prevention of hypoglycemia is needed. Inpatient diabetes management teams consisting of nurse practitioners and physicians are increasingly common. (2) Diabetes treatment plans should focus on the transition from the ICU and the transition from the inpatient to outpatient setting. (3) Adjustment of the discharge treatment regimen of patients whose diabetes was poorly controlled on admission (as reflected by the HbA_{1c}) is important.

The physician caring for an individual with diabetes in the perioperative period, during times of infection or serious physical illness, or simply when the patient is fasting for a diagnostic procedure must monitor the plasma glucose vigilantly, adjust the diabetes treatment regimen, and provide glucose infusion as needed. Hypoglycemia is frequent in hospitalized patients, and many of these episodes are avoidable. Hospital systems should have a diabetes management protocol to avoid inpatient hypoglycemia. Measures to reduce or prevent hypoglycemia include frequent glucose monitoring and anticipating potential modifications of insulin/glucose administration because of changes in the clinical situation or treatment (e.g., tapering of glucocorticoids) or interruption of enteral or parenteral infusions or PO intake.

Depending on the severity of the patient's illness and the hospital setting, the physician can use either an insulin infusion or SC insulin. Insulin infusions are preferred in the ICU or in a clinically unstable setting because the half-life of the infused insulin is quite short (minutes). The absorption of SC insulin may be variable in such situations. Insulin infusions can also effectively control plasma glucose in the perioperative period and when the patient is unable to take anything by mouth. Regular insulin is used rather than insulin analogues for IV insulin infusion because it is less expensive and equally effective. The physician must consider carefully the clinical setting in which an insulin infusion will be used, including whether adequate ancillary personnel are available to monitor the plasma glucose frequently and whether they can adjust the insulin infusion rate to maintain the plasma glucose within the optimal range. Insulin-infusion algorithms should integrate the insulin sensitivity of the patient, frequent blood glucose monitoring, and the trend of changes in the blood glucose to determine the insulin-infusion rate. Insulin-infusion algorithms jointly developed and implemented by nursing and physician staff are advised. Because of the short half-life of IV regular insulin, it is necessary to administer long-acting insulin prior to discontinuation of the insulin infusion (2–4 h before the infusion is stopped) to avoid a period of insulin deficiency.

In patients who are not critically ill or not in the ICU, basal or "scheduled" insulin is provided by SC, long-acting insulin supplemented by prandial and/or "corrective" insulin using a short-acting insulin (insulin analogues preferred). The use of "sliding scale," short-acting insulin alone, where no insulin is given unless the blood glucose is elevated, is inadequate for inpatient glucose management and should not be used. The short-acting, preprandial insulin dose should include coverage for food consumption (based on anticipated carbohydrate intake) plus a corrective or supplemental insulin based on the patient's insulin sensitivity and the blood glucose. For example, if the patient is thin (and likely insulin-sensitive), a corrective insulin supplement might be 1 unit for each 2.7 mmol/L (50 mg/dL) over the glucose target. If the patient is obese and insulin-resistant, then the insulin supplement might be 2 units for each 2.7 mmol/L (50 mg/dL) over the glucose target. It is critical to individualize the regimen and adjust the basal or "scheduled" insulin dose frequently, based on the corrective insulin required. A consistent carbohydrate-controlled diabetes meal plan for hospitalized patients provides a predictable

amount of carbohydrate for a particular meal each day (but not necessarily the same amount for breakfast, lunch, and supper) and avoids concentrated sweets. Individuals with type 1 DM who are undergoing general anesthesia and surgery or who are seriously ill should receive continuous insulin, either through an IV insulin infusion, their insulin infusion device, or by SC administration of a reduced dose of long-acting insulin. Short-acting insulin alone is insufficient. Prolongation of a surgical procedure or delay in the recovery room is not uncommon and may result in periods of insulin deficiency leading to DKA. Insulin infusion is the preferred method for managing patients with type 1 DM over a prolonged (several hours) perioperative period or when serious concurrent illness is present (0.5–1.0 units/h of regular insulin). If the diagnostic or surgical procedure is brief (<2 h), a reduced dose of SC insulin may suffice (20–50% basal reduction, with short-acting bolus insulin withheld or reduced). This approach prevents interruption of insulin infusion device therapy, or for MDI, facilitates the transition back to long-acting insulin after the procedure. The blood glucose should be monitored frequently during the illness or in the perioperative period.

Individuals with type 2 DM can be managed with either an insulin infusion or SC long-acting insulin (20–50% reduction depending on clinical setting) plus preprandial, short-acting insulin. Oral glucose-lowering agents should be discontinued upon admission and are not useful in regulating the plasma glucose in clinical situations where the insulin requirements and glucose intake are changing rapidly. Moreover, these oral agents may be dangerous if the patient is fasting (e.g., hypoglycemia with sulfonylureas) or at risk for declining kidney function due to, for example, radiographic contrast media or unstable CHF (lactic acidosis with metformin). Once clinically stable, oral glucose-lowering agents may be resumed in anticipation of discharge.

SPECIAL CONSIDERATIONS IN DM

TOTAL PARENTERAL NUTRITION (TPN)

(See also Chap. 328) TPN greatly increases insulin requirements. In addition, individuals not previously known to have DM may become hyperglycemic during TPN and require insulin treatment. IV insulin infusion is the preferred treatment for hyperglycemia, and rapid titration to the required insulin dose is done most efficiently using a separate insulin infusion. After the total insulin dose has been determined, a proportion of this insulin may be added directly to the TPN solution to cover the nutritional requirements for insulin, and adjusted based on the need for modified dosing of short-acting insulin. Total enteral nutrition (TEN) also increases insulin requirements and may lead to or worsen hyperglycemia. Hyperglycemia may be limited by using high protein formulations, but often requires insulin treatment. Short- or intermediate (i.e., NPH)-acting insulins should be used to cover bolus or continuous enteral feeding to minimize the risk for hypoglycemia should the TEN be interrupted or held. Patients with insulin deficiency (type 1 DM and pancreatogenic DM) should also receive long-acting insulin (0.1–0.2 units/kg per day) to cover basal insulin requirements should the TPN or TEN be interrupted or cycled.

GLUCOCORTICOIDS

Glucocorticoids increase insulin resistance, decrease glucose utilization, increase hepatic glucose production, and impair insulin secretion. These changes lead to a worsening of glycemic control in individuals with DM and may precipitate hyperglycemia in other individuals. If new-onset hyperglycemia remains during chronic treatment with supraphysiologic doses of glucocorticoid (>5 mg of prednisone or equivalent), the DM may be called "steroid-induced diabetes." The effects of glucocorticoids on glucose homeostasis are dose-related, usually reversible, and most pronounced in the postprandial period. If the FPG is near the normal range, oral diabetes agents (e.g., sulfonylureas, metformin) may be sufficient to reduce hyperglycemia. If the FPG is >11.1 mmol/L (200 mg/dL), oral agents are usually not efficacious, and insulin therapy is required. Short-acting insulin may be sufficient alone or together with long-acting insulin in order to control postprandial glucose excursions.

DIABETES MANAGEMENT IN OLDER ADULTS

Diabetes is very common in older adults, being present in ~25% of individuals over the age of 65. Increasingly, individuals with many years of type 1 DM are part of the patient population. As discussed above, individualized therapeutic goals and modalities in older adults should consider biologic age, other comorbidities and risk factors (hypertension, CV disease, etc.), neurocognitive and physical functional status, living arrangements, social support, and other medications. For example, the HbA1c goal for a highly functional 80-year-old should be different than that for an individual with diabetes in long-term care (skilled nursing facilities). In the former, the HbA1c goal (<7.5%) and selected therapies may be similar to younger individuals whereas in an individual with complex/poor health or cognitive impairment, an HbA1c goal of <8.5% would be reasonable. Critical to diabetes management in all older individuals is the avoidance of hypoglycemia, which can worsen underlying cognitive impairment or CV disease. Thus, medications that can cause hypoglycemia (insulin secretagogues, insulin) should be used carefully. In choosing medications for diabetes, the adverse effects (Table 397-5) should be considered (especially heart failure, renal insufficiency, etc.). Hypertension and dyslipidemia should be treated in elderly individuals with diabetes since there is clear benefit of blood pressure control with the benefit for lipid-lowering medications being less clearly demonstrated.

REPRODUCTIVE ISSUES

Reproductive capacity in either men or women with DM appears to be normal. Menstrual cycles may be associated with alterations in glycemic control in women with DM. Pregnancy is associated with marked insulin resistance; the increased insulin requirements often precipitate DM and lead to the diagnosis of gestational diabetes mellitus (GDM). Glucose, which at high levels is a teratogen to the developing fetus, readily crosses the placenta, but insulin does not. Thus, hyperglycemia from the maternal circulation may stimulate insulin secretion in the fetus. The anabolic and growth effects of insulin may result in macrosomia. GDM complicates ~7% (range 1–14%) of pregnancies. The incidence of GDM is greatly increased in certain ethnic groups, including African Americans and Latinas, consistent with a similar increased risk of type 2 DM. Current recommendations advise screening for glucose intolerance between weeks 24 and 28 of pregnancy in women not known to have diabetes. Therapy for GDM is similar to that for individuals with pregnancy-associated diabetes and involves MNT and insulin, if hyperglycemia persists. Oral glucose-lowering agents are not approved for use during pregnancy, but studies using metformin or glyburide have shown efficacy and have not found toxicity. With current practices, the morbidity and mortality rates of the mother with GDM and the fetus are not different from those in the nondiabetic population. Individuals who develop GDM are at marked increased risk for developing type 2 DM in the future and should be screened periodically for DM (see screening recommendations in Chap. 396). Most individuals with GDM revert to normal glucose tolerance after delivery, but some will continue to have overt diabetes or impairment of glucose tolerance after delivery. In addition, children of women with GDM appear to be at risk for obesity and glucose intolerance and have an increased risk of diabetes beginning in the later stages of adolescence.

Pregnancy in individuals with known DM requires meticulous planning and adherence to strict treatment regimens. Intensive insulin therapy and near-normalization of the HbA_{1c} (<6.5%) are essential for individuals with existing DM who are planning pregnancy. Consideration should be given to insulin infusion and CGM devices that may help to improve glycemic control prior to conception since the most crucial period of glycemic control is soon after fertilization. The risk of fetal malformations is increased 4–10 times in individuals with uncontrolled DM at the time of conception, and normal plasma glucose during the preconception period and throughout the periods of organ development in the fetus should be the goal, with more frequent monitoring of HbA1c every 2 months throughout gestation.

LIPODYSTROPHIC DM

Lipodystrophy, or the loss of subcutaneous fat tissue, may be generalized in certain genetic conditions such as leprechaunism, or acquired

as part of an autoimmune disorder. Generalized lipodystrophy is associated with leptin deficiency and severe insulin resistance and is often accompanied by acanthosis nigricans, hepatic steatosis, and severe hypertriglyceridemia. Recombinant human leptin (metreleptin) may allow for the achievement of metabolic control in generalized lipodystrophy, but is associated with the development of neutralizing antibodies and is available only through a restricted program under a Risk Evaluation and Mitigation Strategy (REMS).

Protease Inhibitors and Lipodystrophy Protease inhibitors used in the treatment of HIV disease (Chap. 197) have been associated with a centripetal accumulation of fat (visceral and abdominal area), accumulation of fat in the dorsocervical region, loss of extremity fat, decreased insulin sensitivity (elevations of the fasting insulin level and reduced glucose tolerance on IV glucose tolerance testing), and dyslipidemia. Although many aspects of the physical appearance of these individuals resemble Cushing's syndrome, increased cortisol levels do not account for this appearance. The possibility remains that this is related to HIV infection by some undefined mechanism, because some features of the syndrome were observed before the introduction of protease inhibitors. Therapy for HIV-related lipodystrophy is not well established.

FURTHER READING

American Diabetes Association: Comprehensive Medical Evaluation and Assessment of Comorbidities. Diabetes Care 41:S28, 2018.

Evert AB et al: Nutrition therapy recommendations for the management of adults with diabetes. Diabetes Care 37:S120, 2014.

Glycemic Targets. Diabetes care older adults. Diabetes Care 41:S119, 2018.

Hassanein M et al: Diabetes and Ramadan: Practical guidelines. Diabetes Res Clin Pract 126:303, 2017.

Lifestyle Management. Diabetes Care 41:S38, 2018.

Nyenwe EA, Kitabchi AE: The evolution of diabetic ketoacidosis: An update of its etiology, pathogenesis and management. Metabolism 65:507, 2016.

Qaseem A et al: Oral pharmacologic treatment of type 2 diabetes mellitus: A clinical practice guideline update from the American College of Physicians. Ann Intern Med 166:279, 2017.

Rubino F et al: Metabolic surgery in the treatment algorithm for type 2 diabetes: A joint statement by international diabetes organizations. Diabetes Care 39:86, 2016.

Thabit H, Hovorka R: Coming of age: the artificial pancreas for type 1 diabetes. Diabetologia 59:1795, 2016.

Young-Hyman D et al: Psychosocial care for people with diabetes: A position statement of the American Diabetes Association. Diabetes Care 39:2126, 2016.

398 Diabetes Mellitus: Complications

Alvin C. Powers, John M. Stafford, Michael R. Rickels

Diabetes-related complications affect many organ systems and are responsible for the majority of morbidity and mortality associated with the disease. For many years in the United States, diabetes has been the leading cause of new blindness in adults, renal failure, and nontraumatic lower extremity amputation. More recently, diabetes has also emerged as a leading contributor to coronary heart disease (CHD). Diabetes-associated complications related to hyperglycemia usually do not appear until the second decade of hyperglycemia. In contrast, diabetes-associated CHD risk, related in part to insulin resistance, may develop before hyperglycemia is established. Because type 2 diabetes mellitus (DM) often has a long asymptomatic period of hyperglycemia before diagnosis, many individuals with type 2 DM have both glucose-related and insulin resistance-related complications at the time of diagnosis. Fortunately, many of the diabetes-related complications can be prevented or delayed with a focus on diet, fitness, early detection, aggressive glycemic control, and efforts to minimize the risks of complications. Recent studies show a decline in diabetes-related complications in individuals, but this is tempered by the increase in the number of individuals with diabetes. For example, the rate of myocardial infarction (MI) associated with diabetes declined by 67% between 1990 and 2010.

TABLE 398-1 Diabetes-Related Complications

Microvascular
Eye disease
Retinopathy (nonproliferative/proliferative)
Macular edema
Neuropathy
Sensory and motor (mono- and polyneuropathy)
Autonomic
Nephropathy (albuminuria and declining renal function)
Macrovascular
Coronary heart disease
Peripheral arterial disease
Cerebrovascular disease
Other
Gastrointestinal (gastroparesis, diarrhea)
Genitourinary (uropathy/sexual dysfunction)
Dermatologic
Infectious
Cataracts
Glaucoma
Cheiroarthropathy[a]
Periodontal disease
Hearing loss
Other comorbid conditions associated with diabetes (relationship to hyperglycemia is uncertain): depression, obstructive sleep apnea, fatty liver disease, hip fracture, osteoporosis (in type 1 diabetes), cognitive impairment or dementia, low testosterone in men.

[a]Thickened skin and reduced joint mobility.

Diabetes-related complications can be divided into vascular and nonvascular complications and are similar for type 1 and type 2 DM (Table 398-1). The vascular complications of DM are further subdivided into microvascular (retinopathy, neuropathy, nephropathy) and macrovascular complications (CHD, peripheral arterial disease [PAD], cerebrovascular disease). Microvascular complications are diabetes-specific, whereas macrovascular complications have pathophysiologic features that are both shared with the general population and diabetes-specific. Nonvascular complications include infections, skin changes, and hearing loss. Some studies suggest that 2 DM increases the risk of dementia and impaired cognitive function.

GLYCEMIC CONTROL AND COMPLICATIONS

The microvascular complications of both type 1 and type 2 DM result from chronic hyperglycemia (Fig. 398-1). Evidence implicating a causative role for chronic hyperglycemia in the development of macrovascular complications is less conclusive. CHD events and mortality rate are two to four times greater in patients with type 2 DM and correlate with fasting and postprandial plasma glucose levels as well the hemoglobin A_{1c} (HbA_{1c}). Other factors such as dyslipidemia and hypertension also play important roles in macrovascular complications.

The Diabetes Control and Complications Trial (DCCT) provided definitive proof that reduction in chronic hyperglycemia can prevent many complications of type 1 DM (Fig. 398-1). This large multicenter clinical trial randomized >1400 individuals with type 1 DM to either intensive or conventional diabetes management and prospectively evaluated the development of diabetes-related complications during a

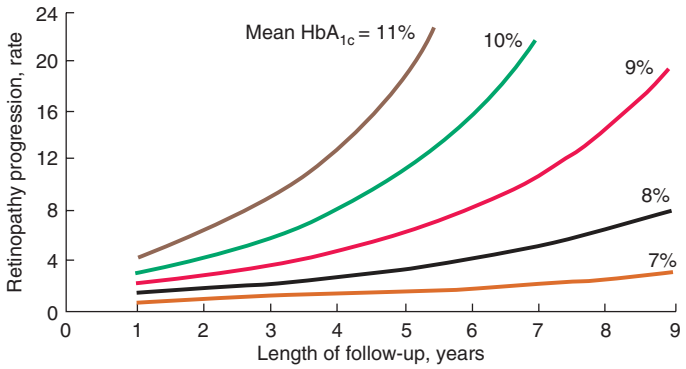


FIGURE 398-1 Relationship of glycemic control and diabetes duration to diabetic retinopathy. The progression of retinopathy in individuals in the Diabetes Control and Complications Trial is graphed as a function of the length of follow-up with different curves for different hemoglobin A_{1c} (HbA_{1c}) values. *(Adapted from The Diabetes Control and Complications Trial Research Group: Diabetes 44:968, 1995.)*

mean follow-up of 6.5 years. Individuals in the intensive diabetes management group received multiple administrations of insulin each day (injection or pump) along with extensive educational, psychological, and medical support. Individuals in the conventional diabetes management group received twice-daily insulin injections and quarterly nutritional, educational, and clinical evaluation. The goal in the former group was normoglycemia; the goal in the latter group was prevention of symptoms of diabetes. Individuals in the intensive diabetes management group achieved a substantially lower HbA_{1c} (7.3%) than individuals in the conventional diabetes management group (9.1%). After the DCCT results were reported in 1993, study participants continue to be followed in the Epidemiology of Diabetes Intervention and Complications (EDIC) trial, which recently completed 30 years of follow-up (DCCT + EDIC). At the end of the DCCT phase, study participants in both intensive and conventional arms were offered intensive therapy. However, during the subsequent follow-up of >18 years, the initial separation in glycemic control disappeared with both arms maintaining a mean HbA_{1c} of 8.0%.

The DCCT phase demonstrated that improvement of glycemic control reduced nonproliferative and proliferative retinopathy (47% reduction), albuminuria (39% reduction), clinical nephropathy (54% reduction), and neuropathy (60% reduction). Improved glycemic control also slowed the progression of early diabetic complications. During the DCCT phase, weight gain (4.6 kg) and severe hypoglycemia (requiring assistance of another person to treat) were more common in the intensive therapy group. The benefits of an improvement in glycemic control occurred over the entire range of HbA_{1c} values (Fig. 398-1), indicating that at any HbA_{1c} level, an improvement in glycemic control is beneficial. The results of the DCCT predicted that individuals in the intensive diabetes management group would gain 7.7 additional years of vision, 5.8 additional years free from end-stage renal disease (ESRD), and 5.6 years free from lower extremity amputations. If all complications of DM were combined, individuals in the intensive diabetes management group would experience >15.3 more years of life without significant microvascular complications of DM, compared to individuals who received standard therapy. This translates into an additional 5.1 years of life expectancy for individuals in the intensive diabetes management group. The 30-year follow-up data in the intensively treated group show a continued reduction in retinopathy, nephropathy, and cardiovascular disease. For example, individuals in the intensive therapy group had a 42–57% reduction in cardiovascular events (nonfatal MI, stroke, or death from a cardiovascular event) at a mean follow-up of 18 years, even though their subsequent glycemic control was the same as those in the conventional diabetes management group from years 6.5 to 17. During the EDIC phase, <1% of the cohort had become blind, lost a limb to amputation, or required dialysis. Other complications of diabetes, including autonomic neuropathy, bladder and sexual dysfunction, and cardiac autonomic neuropathy, were reduced in the intensive therapy group. Importantly, those in the intensive therapy group did not have a difference in cognitive function either during the DCCT phase (when the frequency of hypoglycemia was greater) or during the follow-up EDIC phase (when the frequency of hypoglycemia was similar).

The United Kingdom Prospective Diabetes Study (UKPDS) studied the course of >5000 individuals with type 2 DM for >10 years. This study used multiple treatment regimens and monitored the effect of intensive glycemic control and risk factor treatment on the development of diabetic complications. Newly diagnosed individuals with type 2 DM were randomized to (1) intensive management using various combinations of insulin, a sulfonylurea, or metformin or (2) conventional therapy using dietary modification and pharmacotherapy with the goal of symptom prevention. In addition, individuals were randomly assigned to different antihypertensive regimens. Individuals in the intensive treatment arm achieved an HbA_{1c} of 7%, compared to a 7.9% HbA_{1c} in the standard treatment group. The UKPDS demonstrated that each percentage point reduction in HbA_{1c} was associated with a 35% reduction in microvascular complications. As in the DCCT, there was a continuous relationship between glycemic control and development of complications. Improved glycemic control also reduced the cardiovascular event rate in the follow-up period of >10 years.

One of the major findings of the UKPDS was that strict blood pressure control significantly reduced both macro- and microvascular complications. In fact, the beneficial effects of blood pressure control were greater than the beneficial effects of glycemic control. Lowering blood pressure to moderate goals (144/82 mmHg) reduced the risk of DM-related death, stroke, microvascular endpoints, retinopathy, and heart failure (risk reductions between 32 and 56%).

Similar reductions in the risks of retinopathy and nephropathy were also seen in a small trial of lean Japanese individuals with type 2 DM randomized to either intensive glycemic control or standard therapy with insulin (Kumamoto study). These results demonstrate the effectiveness of improved glycemic control in individuals of different ethnicity and, presumably, a different etiology of DM (i.e., phenotypically different from those in the DCCT and UKPDS). The Action to Control Cardiovascular Risk in Diabetes (ACCORD) and Action in Diabetes and Vascular Disease: Preterax and Diamicron MR Controlled Evaluation (ADVANCE) trials also found that improved glycemic control reduced microvascular complications.

Thus, these large clinical trials in type 1 and type 2 DM indicate that chronic hyperglycemia plays a causative role in the pathogenesis of diabetic microvascular complications. In both the DCCT and the UKPDS, cardiovascular events were reduced at follow-up of >10 years, even though the improved glycemic control was not maintained. The positive impact of a period of improved glycemic control on later disease has been termed a *legacy effect* or *metabolic memory*.

A summary of the features of diabetes-related complications includes the following. (1) Duration and degree of hyperglycemia correlate with complications. (2) Intensive glycemic control is beneficial in all forms of DM. (3) Blood pressure control is critical, especially in type 2 DM. (4) Survival in patients with type 1 DM is improving, and diabetes-related complications are declining. (5) Not all individuals with diabetes develop diabetes-related complications. Other incompletely defined factors appear to modulate the development of complications. For example, despite long-standing DM, some individuals never develop nephropathy or retinopathy. Many of these patients have glycemic control that is indistinguishable from those who develop microvascular complications, suggesting a genetic susceptibility for developing particular complications, especially retinopathy and nephropathy.

MECHANISMS OF COMPLICATIONS

Chronic hyperglycemia is the important etiologic factor leading to complications of DM, but the mechanism(s) by which it leads to such diverse cellular and organ dysfunction is unknown. An emerging hypothesis is that hyperglycemia leads to epigenetic changes (Chap. 456) that influence gene expression in affected cells. Other hypotheses are that chronic hyperglycemia leads to formation of advanced glycosylation end products (AGEs; e.g., pentosidine, glucosepane, and carboxymethyllysine) which bind to specific cell surface receptor and/or the

nonenzymatic glycosylation of intra- and extracellular proteins, leading to cross-linking of proteins, accelerated atherosclerosis, glomerular dysfunction, endothelial dysfunction, and altered extracellular matrix composition. Other theories predict that hyperglycemia: (1) increases glucose metabolism via the sorbitol pathway related to the enzyme aldose reductase; (2) increases the formation of diacylglycerol, leading to activation of protein kinase C, which alters the transcription of genes for fibronectin, type IV collagen, contractile proteins, and extracellular matrix proteins in endothelial cells and neurons; and/or (3) increases the flux through the hexosamine pathway, which generates fructose-6-phosphate, a substrate for O-linked glycosylation and proteoglycan production, leading to altered function by glycosylation of proteins such as endothelial nitric oxide synthase.

Growth factors may play an important role in some diabetes-related microvascular complications, and their production is increased by most of these proposed pathways. For example, vascular endothelial growth factor A (VEGF-A) is increased locally in diabetic proliferative retinopathy and decreases after laser photocoagulation. A possible unifying mechanism is that hyperglycemia leads to increased production of reactive oxygen species or superoxide in the mitochondria and this may activate several of the pathways described above. Although hyperglycemia serves as the initial trigger for complications of diabetes, it is still unknown whether the same pathophysiologic processes are operative in all complications or whether some pathways predominate in certain organs.

The mechanisms of diabetes-related macrovascular complications including MI and stroke are glucose-related mechanisms, but also include traditional cardiovascular risk factors (dyslipidemia, hypertension), and insulin resistance. In type 2 diabetes, insulin resistance is present years prior to diagnosis and is associated with obesity and ectopic accumulation of lipids in muscle and liver. Additionally, insulin fails to appropriately suppress lipolysis from adipose tissue, which results in increased delivery of fatty acids to liver, muscle, endothelial cells, and cardiac tissues, leading to tissue accumulation of triglycerides, diacylglycerol, and ceramides.

OPHTHALMOLOGIC COMPLICATIONS OF DIABETES MELLITUS

DM is the leading cause of blindness between the ages of 20 and 74 in the United States. The gravity of this problem is highlighted by the finding that individuals with DM are 25 times more likely to become legally blind than individuals without DM. Severe vision loss is primarily the result of progressive diabetic retinopathy which leads significant macular edema and new blood vessel formation. Diabetic retinopathy is classified into two stages: nonproliferative and proliferative. Nonproliferative diabetic retinopathy usually appears late in the first decade or early in the second decade of the disease and is marked by retinal vascular microaneurysms, blot hemorrhages, and cotton-wool spots **(Fig. 398-2)**. Mild nonproliferative retinopathy may progress to more extensive disease, characterized by changes in venous vessel caliber, intraretinal microvascular abnormalities, and more numerous microaneurysms and hemorrhages. The pathophysiologic mechanisms invoked in nonproliferative retinopathy include loss of retinal pericytes, increased retinal vascular permeability, alterations in retinal blood flow, and abnormal retinal microvasculature, all of which can lead to retinal ischemia.

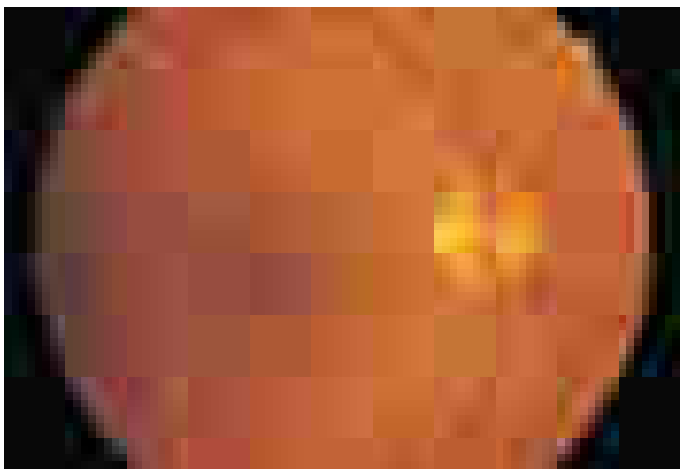


FIGURE 398-2 Diabetic retinopathy results in scattered hemorrhages, yellow exudates, and neovascularization. This patient has neovascular vessels proliferating from the optic disc, requiring urgent panretinal laser photocoagulation.

The appearance of neovascularization in response to retinal hypoxemia is the hallmark of proliferative diabetic retinopathy (Fig. 398-2). These newly formed vessels appear near the optic nerve and/or macula and rupture easily, leading to vitreous hemorrhage, fibrosis, and ultimately retinal detachment. Not all individuals with nonproliferative retinopathy go on to develop proliferative retinopathy, but the more severe the nonproliferative disease, the greater the chance of evolution to proliferative retinopathy within 5 years. This creates an important opportunity for early detection and treatment of diabetic retinopathy. Clinically significant macular edema can occur in the context of nonproliferative or proliferative retinopathy. Fluorescein angiography and optical coherence tomography are useful to detect macular edema, which is associated with a 25% chance of moderate visual loss over the next 3 years. Duration of DM and degree of glycemic control are the best predictors of the development of retinopathy; hypertension, nephropathy, and dyslipidemia are also risk factors. Nonproliferative retinopathy is found in many individuals who have had DM for >20 years. Although there is genetic susceptibility for retinopathy, it confers less influence than either the duration of DM or the degree of glycemic control.

TREATMENT

Diabetic Retinopathy

The most effective therapy for diabetic retinopathy is prevention. Intensive glycemic and blood pressure control will delay the development or slow the progression of retinopathy in individuals with either type 1 or type 2 DM. Paradoxically, during the first 6–12 months of improved glycemic control, established diabetic retinopathy may transiently worsen. Fortunately, this progression is temporary, and in the long term, improved glycemic control is associated with less diabetic retinopathy. Individuals with known retinopathy may be candidates for prophylactic laser photocoagulation when initiating intensive therapy. Once advanced retinopathy is present, improved glycemic control imparts less benefit, although adequate ophthalmologic care can prevent most blindness. Fenofibrate, while not reducing cardiovascular events in individuals with diabetes and dyslipidemia, does reduce the progression of retinopathy.

Regular, comprehensive eye examinations are essential for all individuals with DM **(see Table 397-1)**. Most diabetic eye disease can be successfully treated if detected early. Routine, nondilated eye examinations by the primary care provider or diabetes specialist are inadequate to detect diabetic eye disease, which requires a dilated eye exam performed by an optometrist or ophthalmologist, and subsequent management by a retinal specialist for optimal care of these disorders. Treatment of proliferative retinopathy or macular edema with laser photocoagulation and/or anti-VEGF therapy (ocular injection) usually is successful in preserving vision. Aspirin therapy (650 mg/d) does not appear to influence the natural history of diabetic retinopathy.

RENAL COMPLICATIONS OF DIABETES MELLITUS

Diabetic nephropathy is the leading cause of chronic kidney disease (CKD), ESRD, and CKD requiring renal replacement therapy. Furthermore, the prognosis of individuals with diabetes on dialysis is poor. Albuminuria in individuals with DM is associated with an increased risk of cardiovascular disease. Individuals with diabetic nephropathy commonly have diabetic retinopathy.

Like other microvascular complications, the pathogenesis of diabetic nephropathy is related to chronic hyperglycemia. The mechanisms by which chronic hyperglycemia leads to diabetic nephropathy, although

incompletely defined, involve the effects of soluble factors (growth factors, angiotensin II, endothelin, advanced glycation end products [AGEs]), hemodynamic alterations in the renal microcirculation (glomerular hyperfiltration or hyperperfusion, increased glomerular capillary pressure), and structural changes in the glomerulus (increased extracellular matrix, basement membrane thickening, mesangial expansion, fibrosis). Some of these effects may be mediated through angiotensin II receptors. Smoking accelerates the decline in renal function. Because only 20–40% of patients with diabetes develop diabetic nephropathy, additional genetic or environmental susceptibility factors remain unidentified. Known risk factors include race and a family history of diabetic nephropathy. Diabetic nephropathy and ESRD secondary to DM develop more commonly in African Americans, Native Americans, and Hispanic individuals with diabetes.

The natural history of diabetic nephropathy is characterized by a sequence of events that was initially defined for individuals with type 1 DM but appears similar in type 2 DM **(Fig. 398-3)**. Glomerular hyperperfusion and renal hypertrophy occur in the first years after the onset of DM and are associated with an increase of the estimated glomerular filtration rate (GFR). During the first 5 years of DM, thickening of the glomerular basement membrane, glomerular hypertrophy, and mesangial volume expansion occur as the GFR returns to normal. After 5–10 years of type 1 DM, many individuals begin to excrete small amounts of albumin in the urine. The American Diabetes Association (ADA) no longer uses the terms microalbuminuria or macroalbuminuria (previously used to refer to increased urinary protein of different levels) and instead uses the term albuminuria to refer to increased urinary protein excretion (spot urinary albumin-to-creatinine ratio >30 mg/g Cr). Likewise, this chapter uses the albuminuria, but emphasizes that this should be persistent and is a continuous variable. In some individuals with type 1 diabetes and albuminuria of short duration, the albuminuria regresses. Albuminuria is a risk factor for cardiovascular disease (CVD) and CKD. Diabetic kidney disease refers to albuminuria and reduced GFR (<60 mL/min/1.73 m^2); CKD related to diabetes, which may not be accompanied by albuminuria, is also discussed in **Chap. 305**. Most patients with CKD related to diabetes also have diabetic retinopathy. Once there is marked albuminuria and a reduction in GFR, the pathologic changes are likely irreversible.

The nephropathy that develops in type 2 DM differs from that of type 1 DM in the following respects: (1) albuminuria may be present when type 2 DM is diagnosed, reflecting its long asymptomatic period; (2) hypertension more commonly accompanies albuminuria; and (3) albuminuria may be less predictive of diabetic kidney disease. Finally, it should be noted that albuminuria in type 2 DM may be secondary to factors unrelated to DM, such as hypertension, congestive heart failure (CHF), prostate disease, or infection.

As part of comprehensive diabetes care **(Chap. 397)**, albuminuria should be detected at an early stage when effective therapies can be instituted. Because some individuals with type 1 or type 2 DM have a decline in GFR in the absence of albuminuria, assessment should include a spot urinary albumin-to-creatinine ratio and an estimated GFR **(Fig. 398-4)**. The urine protein measurement by routine urinalysis does not detect low levels of albumin excretion. Screening for albuminuria should commence 5 years after the onset of type 1 DM and at the time of diagnosis of type 2 DM.

Type IV renal tubular acidosis (hyporeninemic hypoaldosteronism) may occur in type 1 or 2 DM. These individuals develop a propensity to hyperkalemia and acidemia, which may be exacerbated by medications (especially angiotensin-converting enzyme [ACE] inhibitors, angiotensin receptor blockers [ARBs], and mineralocorticoid receptor antagonists). Patients with DM are predisposed to radiocontrast-induced nephrotoxicity. Risk factors for radiocontrast-induced nephrotoxicity are preexisting nephropathy and volume depletion. Individuals with DM undergoing radiographic procedures with contrast dye should be well hydrated before and after dye exposure, and the serum creatinine should be monitored for 24–48 h following the procedure. Metformin should be held until postintervention confirmation of preserved kidney function.

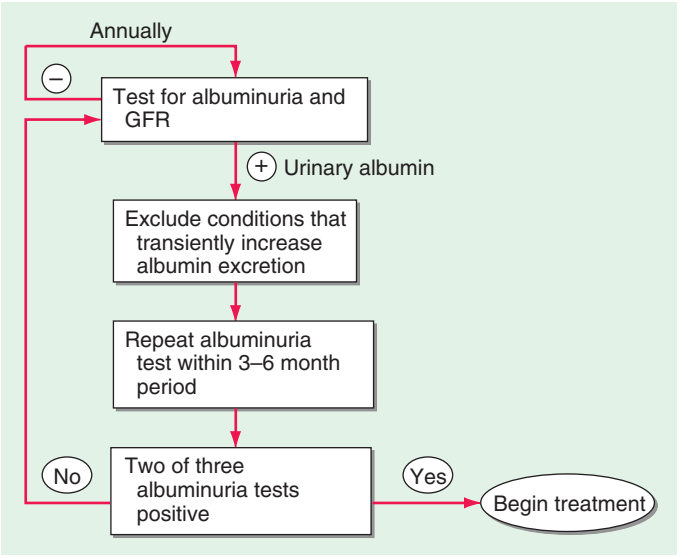


FIGURE 398-4 Screening for albuminuria should be performed annually in patients with type 1 diabetes for ≥5 years, in patients with type 2 diabetes, and during pregnancy. Albuminuria = spot urinary albumin-to-creatinine ratio >30 mg/g Cr; GRD = estimated glomerular filtration rate. Nondiabetes-related conditions that might increase cause albuminuria are urinary tract infection, hematuria, heart failure, febrile illness, severe hyperglycemia, severe hypertension, and vigorous exercise.

TREATMENT

Diabetic Nephropathy

The optimal therapy for diabetic nephropathy is prevention by control of glycemia **(Chap. 397 outlines glycemic goals and approaches)**. Interventions effective in slowing progression of albuminuria include (1) improved glycemic control, (2) strict blood pressure control, and (3) administration of an ACE inhibitor or ARB. Dyslipidemia should also be treated.

Improved glycemic control reduces the rate at which albuminuria appears and progresses in type 1 and type 2 DM. However, once there is a large amount of albuminuria, it is unclear whether improved glycemic control will slow progression of renal disease. During the later phase of declining renal function, insulin requirements may fall as the kidney is a site of insulin degradation. As the GFR decreases with progressive nephropathy, the use and dose of glucose-lowering agents should be reevaluated **(see Table 397-5)**. Some glucose-lowering medications (sulfonylureas and metformin) are contraindicated in advanced renal insufficiency.

Many individuals with type 1 or type 2 DM develop hypertension. Numerous studies in both type 1 and type 2 DM demonstrate the effectiveness of strict blood pressure control in reducing albumin excretion and slowing the decline in renal function. Blood pressure should be maintained at <140/90 mmHg in individuals with

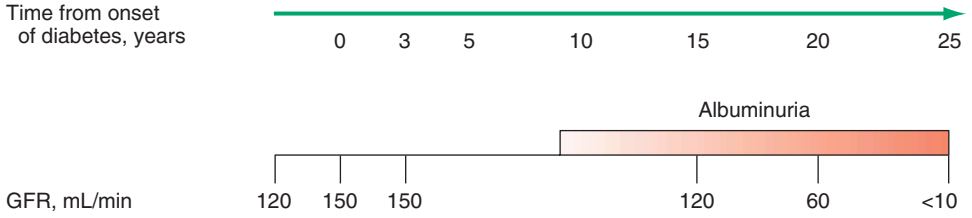


FIGURE 398-3 Time course of development of diabetic nephropathy. The relationship of time from onset of diabetes, albuminuria, and the glomerular filtration rate (GFR) are shown.

diabetes and possibly <130/80 in individuals at increased risk for CVD and CKD progression.

Either ACE inhibitors or ARBs should be used to reduce the albuminuria and the associated decline in GFR that accompanies it in individuals with type 1 or type 2 DM (see "Hypertension," below). Most experts believe that the two classes of drugs are equivalent in patient with diabetes. ARBs can be used as an alternative in patients who develop ACE inhibitor–associated cough or angioedema. After initiation of therapy, some increase the dose and monitor the urinary albumin. There is no benefit of intervention prior to onset of albuminuria or using a combination of an ACE inhibitor and an ARB. If use of either ACE inhibitors or ARBs is not possible or the blood pressure is not controlled, then, diuretics, calcium channel blockers (nondihydropyridine class), or beta blockers should be used. Some glucagon-like peptide-1 (GLP-1) receptor agonists and sodium-glucose co-transporter 2 Inhibitors improve glycemic control and reduce the progression of diabetic kidney disease in individuals with T2DM and established CVD **(Chap 397)**.

The ADA suggests a protein intake of 0.8 mg/kg of body weight/day in individuals with diabetic kidney disease.

Nephrology consultation should be considered when albuminuria appears and when the estimated GFR is <30 mL/min per 1.743 m^2. Complications of atherosclerosis are the leading cause of death in diabetic individuals with nephropathy and hyperlipidemia should be treated aggressively. Referral for transplant evaluation should be made when the GFR approaches 20 mL/min per 1.743 m^2. Preemptive (before dialysis) renal transplantation from a living kidney donor may be a preferred therapy. Kidney transplantation can be performed alone or as a combined pancreas-kidney transplant, which offers the promise of normoglycemia and freedom from both insulin and dialysis. As compared with nondiabetic individuals, hemodialysis in patients with DM is associated with more frequent complications, such as hypotension (due to autonomic neuropathy or loss of reflex tachycardia), more difficult vascular access, and accelerated progression of retinopathy.

NEUROPATHY AND DIABETES MELLITUS

Diabetic neuropathy, which occurs in ~50% of individuals with long-standing type 1 and type 2 DM, manifests as a diffuse neuropathy (distal symmetrical polyneuropathy and/or autonomic neuropathy), a mononeuropathy, and/or a radiculopathy/polyradiculopathy. As with other complications of DM, the development of neuropathy correlates with the duration of diabetes and glycemic control. Additional risk factors are body mass index (BMI) (the greater the BMI, the greater the risk of neuropathy) and smoking. The presence of CVD, elevated triglycerides, and hypertension is also associated with diabetic peripheral neuropathy. Both myelinated and unmyelinated nerve fibers are lost. Because the clinical features of diabetic neuropathy are similar to those of other neuropathies, the diagnosis of diabetic neuropathy should be made only after other possible etiologies are excluded **(Chap. 438)**.

Distal Symmetric Polyneuropathy (DSPN) DSPN, the most common form of diabetic neuropathy, most frequently presents with distal sensory loss and pain, but up to 50% of patients do not have symptoms of neuropathy. Symptoms may include a sensation of numbness, tingling, sharpness, or burning that begins in the feet and spreads proximally. Hyperesthesia, paresthesia, and dysesthesia also may occur. Pain typically involves the lower extremities, is usually present at rest, and worsens at night. Both an acute (lasting <12 months) and a chronic form of painful diabetic neuropathy may occur. The acute form is sometimes treatment-related, occurring in the context of improved glycemic control. As diabetic neuropathy progresses, the pain subsides and eventually disappears, but a sensory deficit persists and motor defects may develop. Physical examination **(Chap. 396)** often reveals sensory loss (to 10-g monofilament and/or vibration), loss of ankle deep-tendon reflexes, abnormal position sense, and muscular atrophy or foot drop. Annual screening for DSPN should begin 5 years after diagnosis of type 1 DM and at the time of diagnosis of type 2 DM and is aimed at detecting loss of protective sensation (LOPS). LOPS and DSPN are major risk factors for foot ulceration and falls due to small and large nerve fiber dysfunction.

Autonomic Neuropathy Individuals with long-standing type 1 or 2 DM may develop signs of autonomic dysfunction involving the cholinergic, noradrenergic, and peptidergic (peptides such as pancreatic polypeptide, substance P, etc.) systems. DM-related autonomic neuropathy can involve multiple systems, including the cardiovascular, gastrointestinal (GI), genitourinary, sudomotor, and metabolic systems. Cardiovascular autonomic neuropathy, reflected by decreased heart rate variability, resting tachycardia and orthostatic hypotension is associated with an increase in CVD. Reports of sudden death in DM have also been attributed to cardiovascular autonomic neuropathy. Gastroparesis and bladder-emptying abnormalities are often caused by the autonomic neuropathy seen in DM (discussed below). Hyperhidrosis of the upper extremities and anhidrosis of the lower extremities result from sympathetic nervous system dysfunction. Anhidrosis of the feet can promote dry skin with cracking, which increases the risk of foot ulcers. Autonomic neuropathy may reduce counterregulatory hormone release (especially catecholamines), leading to an inability to sense hypoglycemia appropriately (hypoglycemia unawareness; **Chap. 399**), thereby subjecting the patient to the risk of severe hypoglycemia and complicating efforts to improve glycemic control.

Mononeuropathy and/or Radiculopathy/Polyradiculopathy Mononeuropathy (dysfunction of isolated cranial or peripheral nerves) is less common than polyneuropathy in DM and presents with pain and motor weakness in the distribution of a single nerve. Mononeuropathies can occur at entrapment sites such as carpal tunnel or be noncompressive. Involvement of the third cranial nerve is most common and is heralded by diplopia. Physical examination reveals ptosis and ophthalmoplegia with normal pupillary constriction to light. Sometimes other cranial nerves, such as IV, VI, or VII (Bell's palsy), are affected. Peripheral mononeuropathies or simultaneous involvement of more than one nerve (mononeuropathy multiplex) may also occur. Diabetic radiculopathy or polyradiculopathy is a syndrome characterized by severe disabling pain in the distribution of one or more nerve roots. It may be accompanied by motor weakness. Intercostal or truncal radiculopathy causes pain over the thorax or abdomen. Involvement of the lumbar plexus or femoral nerve may cause severe pain in the thigh or hip and may be associated with muscle weakness in the hip flexors or extensors (diabetic amyotrophy). Fortunately, diabetic polyradiculopathies are usually self-limited and resolve over 6–12 months.

TREATMENT

Diabetic Neuropathy

Prevention of diabetic neuropathy is critical through improved glycemic control. Treatment of diabetic neuropathy is less than satisfactory. Lifestyle modifications (exercise, diet) has some efficacy in DSPN in type 2 DM and hypertension and hypertriglyceridemia should be treated. Efforts to improve glycemic control in long-standing diabetes may be confounded by hypoglycemia unawareness. Avoidance of neurotoxins (alcohol) and smoking, supplementation with vitamins for possible deficiencies (B_{12}, folate; **Chap. 326**). Patients should be educated that loss of sensation in the foot increases the risk for ulceration and its sequelae and that prevention of such problems is paramount. Patients with symptoms or signs of neuropathy or LOPS should check their feet daily and take precautions (footwear) aimed at preventing calluses or ulcerations. If foot deformities are present, a podiatrist should be involved.

Chronic, painful diabetic neuropathy is difficult to treat with only symptomatic treatment being available; evidence of the effectiveness of improved glycemic control in painful diabetic neuropathy is lacking. Two agents, duloxetine and pregabalin, have been approved by the U.S. Food and Drug Administration (FDA) for pain associated with diabetic neuropathy. Tapentadol, a centrally acting opioid, is also FDA-approved, but has only modest efficacy and poses addiction risk, making it and other opioids less desirable and not a first-line therapy. Diabetic neuropathy may respond to tricyclic

antidepressants, gabapentin, venlafaxine, carbamazepine, tramadol, and topical capsaicin, although none of these are FDA-approved for this indication. No direct comparisons of agents are available, and it is reasonable to switch agents if there is no response or if side effects develop. Referral to a pain management center may be necessary. Because the pain of acute diabetic neuropathy may resolve over time, medications may be discontinued as progressive neuronal damage from DM occurs.

Therapy of orthostatic hypotension secondary to autonomic neuropathy is also difficult. Nonpharmacologic maneuvers (adequate salt intake, avoidance of dehydration and diuretics, lower extremity support hose, and physical activity) may offer some benefit. A variety of agents have limited success (midodrine and droxidopa are FDA-approved for orthostatic hypotension of any etiology). Patients with resting tachycardia may be considered for beta-blocker therapy with caution exercised if there is hypoglycemia unawareness.

GASTROINTESTINAL/GENITOURINARY DYSFUNCTION

Long-standing type 1 and 2 DM may affect the motility and function of the GI and genitourinary systems. The most prominent GI symptoms are delayed gastric emptying (gastroparesis) and altered small- and large-bowel motility (constipation or diarrhea). Gastroparesis may present with symptoms of anorexia, nausea, vomiting, early satiety, and abdominal bloating. Microvascular complications (retinopathy and neuropathy) are usually present. Nuclear medicine scintigraphy after ingestion of a radiolabeled meal may document delayed gastric emptying, but may not correlate well with the patient's symptoms. Noninvasive "breath tests" following ingestion of a radiolabeled meal are emerging as a diagnostic tool. Although parasympathetic dysfunction secondary to chronic hyperglycemia is important in the development of gastroparesis, hyperglycemia itself also impairs gastric emptying. Nocturnal diarrhea, alternating with constipation, is a feature of DM-related GI autonomic neuropathy. In type 1 DM, these symptoms should also prompt evaluation for celiac sprue because of its increased frequency.

Diabetic autonomic neuropathy may lead to genitourinary dysfunction including cystopathy and female sexual dysfunction (reduced sexual desire, dyspareunia, reduced vaginal lubrication). Symptoms of diabetic cystopathy begin with an inability to sense a full bladder and a failure to void completely. As bladder contractility worsens, bladder capacity and the postvoid residual increase, leading to symptoms of urinary hesitancy, decreased voiding frequency, incontinence, and recurrent urinary tract infections.

Erectile dysfunction and retrograde ejaculation are very common in DM and may be one of the earliest signs of diabetic neuropathy **(Chap. 390)**. Erectile dysfunction, which increases in frequency with the age of the patient and the duration of diabetes, may occur in the absence of other signs of diabetic autonomic neuropathy.

TREATMENT

Gastrointestinal/Genitourinary Dysfunction

Current treatments for these complications of DM are inadequate and nonspecific. Improved glycemic control should be a goal, but has not clearly shown benefit. Smaller, more frequent meals that are easier to digest (liquid) and low in fat and fiber may minimize symptoms of gastroparesis. Medications that slow gastric emptying (opioids, GLP-1-receptor agonists) should be avoided. Metoclopramide may be used with severe symptoms but is restricted to short-term treatment in both the United States and Europe. Gastric electrical stimulatory devices are available but not approved. Diabetic diarrhea in the absence of bacterial overgrowth is treated symptomatically **(Chap. 318)**.

Diabetic cystopathy should be treated with scheduled voiding or self-catheterization. Drugs that inhibit type 5 phosphodiesterase are effective for erectile dysfunction, but their efficacy in individuals with DM is slightly lower than in the nondiabetic population **(Chap. 390)**.

CARDIOVASCULAR MORBIDITY AND MORTALITY

CVD is increased in individuals with type 1 or type 2 DM. The Framingham Heart Study revealed a marked increase in PAD, coronary artery disease, MI, and CHF (risk increase from one- to fivefold) in DM. In addition, the prognosis for individuals with diabetes who have coronary artery disease or MI is worse than for nondiabetics. CHD is more likely to involve multiple vessels in individuals with DM. In addition to CHD, cerebrovascular disease is increased in individuals with DM (threefold increase in stroke). Thus, after controlling for all known cardiovascular risk factors, type 2 DM increases the cardiovascular death rate twofold in men and fourfold in women. Congestive heart failure (CHF) is common in long-standing T2DM.

The American Heart Association considers DM as a controllable risk factor for cardiovascular disease; in some studies, type 2 DM patients without a prior MI have a similar risk for coronary artery-related events as nondiabetic individuals who have had a prior MI. Cardiovascular risk assessment in type 2 DM should encompass a more nuanced approach. Cardiovascular risk is lower and not equivalent in a younger individual with a brief duration of type 2 DM compared to an older individual with long-standing type 2 DM. In individuals without a known diagnosis of diabetes, elevated A1C is predictive not just of diabetes risk, but also risk of CHD, stroke, and all-cause mortality. Because of the extremely high prevalence of underlying CVD in individuals with diabetes (especially in type 2 DM), evidence of atherosclerotic vascular disease (e.g., cardiac stress test) should be sought in an individual with diabetes who has symptoms, even if atypical, suggestive of cardiac ischemia or peripheral or carotid arterial disease. The screening of asymptomatic individuals with diabetes for CHD is not recommended or cost-effective. The absence of chest pain ("silent ischemia") is common in individuals with diabetes, and a thorough cardiac evaluation should be considered prior to major surgical procedures.

The increase in cardiovascular morbidity and mortality rates in diabetes appears to relate to the synergism of hyperglycemia with other cardiovascular risk factors such as dyslipidemia (elevated triglycerides, low HDL-cholesterol and small-dense LDL), hypertension, obesity, reduced physical activity, and cigarette smoking. Additional risk factors prevalent include CKD (albuminuria, reduced GFR), abnormal platelet function, increased markers of inflammation, and endothelial dysfunction. The results of the ACCORD trial and VADT trial, which demonstrated that tight glucose control had limited benefit on cardiovascular outcomes in individuals with established cardiovascular disease, suggesting the importance of insulin resistance and dyslipidemia.

TREATMENT

Cardiovascular Disease

Treatment of coronary disease in the diabetic individual has substantial overlap with treatment of in individuals without diabetes **(Chap. 267)**. Revascularization procedures for CHD, including percutaneous coronary interventions (PCIs) and coronary artery bypass grafting (CABG), may be less efficacious in the diabetic individual. Initial success rates of PCI in diabetic individuals are similar to those in the nondiabetic population, but diabetic patients have higher rates of restenosis and lower long-term patency and survival rates in older studies. CABG plus optimal medical management likely has better outcomes than PCI for individuals with diabetes.

Aggressive cardiovascular risk modification in all individuals with DM and glycemic control should be individualized, as discussed in **Chap. 397**. In patients with known CHD and type 2 DM, an ACE inhibitor (or ARB), a statin, and acetylsalicylic acid (ASA; aspirin) should be considered. Beta blockers can be used in individuals with diabetes after MI. In patients with CHF, thiazolidinediones should not be used **(Chap. 397)**. However, metformin can be used in patients with stable CHF if the renal function is normal. Some newer glucose lowering therapies also have cardiovascular benefit, including the GLP-1 analogs semaglutide (SUSTAIN-6) and liraglutide (LEADER), and the SGLT2 inhibitors empagliflozin (EMPA-REG) and canagliflozin (CANVAS).

Antiplatelet therapy reduces cardiovascular events in individuals with DM who have CHD and is recommended. The ADA recommends considering the use of aspirin for primary prevention of coronary events in individuals with diabetes with an increased cardiovascular risk (>50 years with at least one risk factor such as hypertension, dyslipidemia, smoking, family history, or albuminuria). ASA is not recommended for primary prevention in those with a low cardiovascular risk (<50 years with no risk factors). The aspirin dose is the same as in nondiabetic individuals.

Cardiovascular Risk Factors • **DYSLIPIDEMIA** Individuals with DM may have several forms of dyslipidemia **(Chap. 400)**. Because of the additive cardiovascular risk of hyperglycemia and hyperlipidemia, lipid abnormalities should be assessed aggressively and treated as part of comprehensive diabetes care **(Chap. 397)**. The most common pattern of dyslipidemia is hypertriglyceridemia and reduced high-density lipoprotein (HDL) cholesterol levels. DM itself does not increase levels of low-density lipoprotein (LDL), but the small dense LDL particles found in type 2 DM are more atherogenic because they are more easily glycated and susceptible to oxidation.

Almost all treatment studies of diabetic dyslipidemia have been performed in individuals with type 2 DM because of the greater frequency of dyslipidemia in this form of diabetes. Interventional studies have shown that the beneficial effects of LDL reduction with statins are similar in the diabetic and nondiabetic populations. Large prospective trials of primary and secondary intervention for CHD have included some individuals with type 2 DM, and subset analyses have consistently found that reductions in LDL reduce cardiovascular events and morbidity in individuals with DM. No prospective studies have addressed similar questions in individuals with type 1 DM. Because the frequency of CVD is low in children and young adults with diabetes, assessment of cardiovascular risk should be incorporated into the guidelines discussed below.

Based on the guidelines provided by the ADA, all individuals with diabetes should be advised about lifestyle modification, including diet, weight loss, and increased physical activity **(Chap. 397)**. If individuals with diabetes have elevated triglyceride levels (>1.7 mmol/L [150 mg/dL]) or low HDL cholesterol (<1 mmol/L [40 mg/dL]) in men and (<1.3 mmol/L [50 mg/dL]) in women, lifestyle modification and improved glycemic control should be further emphasized. If triglycerides remain >5.7 mmol/L (500 mg/dL), treatment with fish oil and fibrate drugs may reduce the risk of pancreatitis.

In terms of the addition of pharmacologic therapy, the ADA recommends: (1) all patients with diabetes and atherosclerotic cardiovascular disease should receive high-intensity statin therapy; (2) in patients aged 40–75 years, consider using moderate-intensity statin therapy (without additional risk factors) intensity statin therapy (with additional risk factors); (3) in patients <40 years and additional risk factors, consider moderate-intensity statin therapy. The ADA recommendations for individuals with diabetes who are >75 years are similar to that for individuals aged 40–75 years. Combination therapy with a statin and a fibrate or niacin is not recommended with the exception of a statin and ezetimibe and a statin in patients with recent acute coronary syndrome. The choice of statin and dosing should be individualized based on response and side effects. If the LDL cholesterol remains >70 mg/dL in an individual with diabetes and CVD on a statin, consider the addition of ezetimibe or PCSDK9 inhibitor **(Chap. 400)**. Statin usage is associated with a mild increase in the risk of developing type 2 DM. This risk is greatest in individuals with other risk factors for type 2 DM **(Chap. 396)**. However, the cardiovascular benefits of statin use outweigh the mildly increased risk of diabetes.

HYPERTENSION Hypertension can accelerate other complications of DM, particularly CVD, nephropathy, and retinopathy. Blood pressure should be measured at every clinic visit. In targeting a goal of blood pressure of <140/90 mmHg, therapy should first emphasize lifestyle modifications such as weight loss, exercise, stress management, and sodium restriction. The BP goal should be individualized. In some younger individuals or those with increased CV risk, the provider may target a blood pressure of <130/80 mmHg. Realizing that more than one agent is usually required to reach the blood pressure goal, the ADA recommends that all patients with diabetes and hypertension be treated with an ACE inhibitor or an ARB initially. Subsequently, agents that reduce cardiovascular risk (beta blockers, thiazide diuretics, and calcium channel blockers) should be incorporated into the regimen. ACE inhibitors and ARBs are likely equivalent in most patients with diabetes and renal disease, but should not be combined. Serum potassium and renal function should be monitored.

Because of the high prevalence of atherosclerotic disease in individuals with type 2 DM, the possibility of renovascular hypertension should be considered when the blood pressure is not readily controlled.

LOWER EXTREMITY COMPLICATIONS

DM is the leading cause of nontraumatic lower extremity amputation in the United States. Foot ulcers and infections are also a major source of morbidity in individuals with DM. The reasons for the increased incidence of these disorders in DM involve the interaction of several pathogenic factors: neuropathy, abnormal foot biomechanics, PAD, and poor wound healing. The peripheral sensory neuropathy interferes with normal protective mechanisms and allows the patient to sustain major or repeated minor trauma to the foot, often without knowledge of the injury. Disordered proprioception causes abnormal weight bearing while walking and subsequent formation of callus or ulceration. Motor and sensory neuropathy lead to abnormal foot muscle mechanics and to structural changes in the foot (hammer toe, claw toe deformity, prominent metatarsal heads, Charcot joint). Autonomic neuropathy results in anhidrosis and altered superficial blood flow in the foot, which promote drying of the skin and fissure formation. PAD and poor wound healing impede resolution of minor breaks in the skin, allowing them to enlarge and to become infected.

Many individuals with type 2 DM develop a foot ulcer (great toe or metatarsophalangeal areas are most common), and a significant subset who develop an ulceration will ultimately undergo amputation (14–24% risk with that ulcer or subsequent ulceration). Risk factors for foot ulcers or amputation include male sex, diabetes for >10 years, peripheral neuropathy, abnormal structure of foot (bony abnormalities, callus, thickened nails), PAD, smoking, history of previous ulcer or amputation, visual impairment, poor glycemic control, and diabetic nephropathy, especially dialysis. Large calluses are often precursors to or overlie ulcerations.

TREATMENT

Lower Extremity Complications

The optimal therapy for foot ulcers and amputations is prevention through identification of high-risk patients, education of the patient, and institution of measures to prevent ulceration. High-risk patients should be identified during the routine, annual foot examination performed on all patients with DM (see "Ongoing Aspects of Comprehensive Diabetes Care" in **Chap. 397**). If the monofilament test or one of the other tests is abnormal, the patient is diagnosed with LOPS **(Chap. 396)**. Providers should consider screening for asymptomatic PAD in individuals >50 years of age who have diabetes and other risk factors using ankle-brachial index testing in high-risk individuals **(Chap. 275)**. Patient education should emphasize (1) careful selection of footwear, (2) daily inspection of the feet to detect early signs of poor-fitting footwear or minor trauma, (3) daily foot hygiene to keep the skin clean and moist, (4) avoidance of self-treatment of foot abnormalities and high-risk behavior (e.g., walking barefoot), and (5) prompt consultation with a health-care provider if an abnormality arises. Patients at high risk for ulceration or amputation may benefit from evaluation by a foot care specialist. Calluses and nail deformities should be treated by a podiatrist. Interventions directed at risk factor modification include orthotic shoes and devices, callus management, nail care, and prophylactic measures to reduce increased skin pressure from abnormal bony architecture. Attention to other risk factors for vascular disease (smoking, dyslipidemia, hypertension) and improved glycemic control are also important.

Despite preventive measures, foot ulceration and infection are common and represent a serious problem. Due to the multifactorial pathogenesis of lower extremity ulcers, management of these lesions is multidisciplinary and often demands expertise in orthopedics, vascular surgery, endocrinology, podiatry, and infectious diseases. The plantar surface of the foot is the most common site of ulceration. Ulcers may be primarily neuropathic (no accompanying infection) or may have surrounding cellulitis or osteomyelitis. Cellulitis without ulceration should be treated with antibiotics that provide broad-spectrum coverage, including anaerobes (see below).

An infected ulcer is a clinical diagnosis, because superficial culture of any ulceration will likely find multiple bacterial species of unknown significance. The infection surrounding the foot ulcer is often the result of multiple organisms, with aerobic gram-positive cocci (staphylococci including MRSA, Group A and B streptococci) being most common and with aerobic gram-negative bacilli and/or obligate anaerobes as co-pathogens.

Gas gangrene may develop in the absence of clostridial infection. Cultures should be obtained from the debrided ulcer base or from purulent drainage or aspiration of the wound. Wound depth should be determined by inspection and probing with a blunt-tipped sterile instrument. A wound that probes to the bone represents clinical evidence of osteomyelitis. Plain radiographs of the foot should be performed to assess the possibility of osteomyelitis in chronic ulcers that have not responded to therapy. Magnetic resonance imaging (MRI) is the most specific modality, with nuclear medicine scans and labeled white cell studies as alternatives. Surgical debridement is often necessary.

Osteomyelitis is best treated by a combination of prolonged antibiotics and debridement of infected bone when possible. The possible contribution of vascular insufficiency should be considered in all patients. Peripheral arterial bypass procedures are often effective in promoting wound healing and in decreasing the need for amputation of the ischemic limb **(Chap. 275)**.

Interventions with demonstrated efficacy in diabetic foot ulcers or wounds: (1) off-loading, (2) debridement, (3) wound dressings, (4) appropriate use of antibiotics, (5) revascularization, and (6) limited amputation. Off-loading is the complete avoidance of weight bearing on the ulcer, which removes the mechanical trauma that retards wound healing. Bed rest and a variety of orthotic devices or contact casting limit weight bearing on wounds or pressure points. Surgical debridement is important and effective, but the efficacy of other modalities for wound healing (enzymes, growth factors, cellular therapy, hyperbaric oxygen) is unclear. Dressings such as hydrocolloid dressings promote wound healing by creating a moist environment, controlling the exudate, and protecting the wound. Antiseptic agents should be avoided. Topical antibiotics are of limited value. Referral for physical therapy, orthotic evaluation, and rehabilitation should occur once the infection is controlled.

Mild or nonlimb-threatening infections can be treated with oral antibiotics directed predominantly at methicillin-susceptible staphylococci and streptococci (e.g., dicloxacillin, cephalosporin, amoxicillin/clavulanate). However, in patients with a prior history of MRSA or in locations with a high prevalence of MRSA, treatment with clindamycin, doxycycline, or trimethoprim-sulfamethoxazole is preferred. Trimethoprim-sulfamethoxazole exhibits less reliable coverage of streptococci than the β-lactams, and individuals with diabetes may develop adverse effects including acute kidney injury and hyperkalemia. Surgical debridement of necrotic tissue, local wound care (avoidance of weight bearing over the ulcer), and close surveillance for progression of infection are crucial. More severe infections require IV antibiotics as well as bed rest and local wound care. Urgent surgical debridement may be required. Optimization of glycemic control should be a goal. IV antibiotics should provide broad-spectrum coverage directed toward *Staphylococcus aureus*, including MRSA, streptococci, gram-negative aerobes, and anaerobic bacteria. Initial antimicrobial regimens include vancomycin plus a β-lactam/β-lactamase inhibitor or carbapenem or vancomycin plus a combination of a quinolone plus metronidazole. Daptomycin, ceftaroline, or linezolid may be substituted for vancomycin. If the infection surrounding the ulcer is not improving with IV antibiotics, reassessment of antibiotic coverage and reconsideration of the need for surgical debridement or revascularization are indicated. With clinical improvement, oral antibiotics and local wound care can be continued on an outpatient basis with close follow-up.

INFECTIONS

Individuals with DM have a greater frequency and severity of infection. The reasons for this include incompletely defined abnormalities in cell-mediated immunity and phagocyte function associated with hyperglycemia, as well as diminished vascularization. Hyperglycemia aids the colonization and growth of a variety of organisms (*Candida* and other fungal species). Many common infections are more frequent and severe in the diabetic population, whereas several rare infections are seen almost exclusively in the diabetic population. Examples of this latter category include rhinocerebral mucormycosis, emphysematous infections of the gallbladder and urinary tract, and "malignant" or invasive otitis externa. Invasive otitis externa is usually secondary to *Pseudomonas aeruginosa* infection in the soft tissue surrounding the external auditory canal, usually begins with pain and discharge, and may rapidly progress to osteomyelitis and meningitis. These infections should be sought, in particular, in patients presenting with severe hyperglycemia **(Chap. 397)**.

Pneumonia, urinary tract infections, and skin and soft tissue infections are all more common in the diabetic population. In general, the organisms that cause pulmonary infections are similar to those found in the nondiabetic population; however, gram-negative organisms, *S. aureus*, and *Mycobacterium tuberculosis* are more frequent pathogens. Urinary tract infections (either lower tract or pyelonephritis) are the result of common bacterial agents such as *Escherichia coli*, although several yeast species (*Candida albicans* and *Torulopsis glabrata*) are commonly observed. Complications of urinary tract infections include emphysematous pyelonephritis and emphysematous cystitis. Bacteriuria occurs frequently in individuals with diabetic cystopathy and does not require antibiotic therapy. Susceptibility to furunculosis, superficial candidal infections, and vulvovaginitis are increased. Poor glycemic control is a common denominator in individuals with these infections. Individuals with diabetes have an increased rate of colonization of *S. aureus* in the skinfolds and nares. Individuals with diabetes also have a greater risk of postoperative wound infections.

DERMATOLOGIC MANIFESTATIONS

The most common skin manifestations of DM are xerosis and pruritus and are usually relieved by skin moisturizers. Protracted wound healing and skin ulcerations are also frequent complications. Diabetic dermopathy, sometimes termed *pigmented pretibial papules*, or "diabetic skin spots," begins as an erythematous macule or papule that evolves into an area of circular hyperpigmentation. These lesions result from minor mechanical trauma in the pretibial region and are more common in elderly men with DM. Bullous diseases, such as bullosa diabeticorum (shallow ulcerations or erosions in the pretibial region), are also seen. *Necrobiosis lipoidica diabeticorum* is an uncommon disorder, accompanying diabetes in predominantly young women. This usually begins in the pretibial region as an erythematous plaque or papules that gradually enlarge, darken, and develop irregular margins, with atrophic centers and central ulceration. They are often painful. Vitiligo occurs at increased frequency in individuals with type 1 DM. *Acanthosis nigricans* (hyperpigmented velvety plaques seen on the neck, axilla, or extensor surfaces) is sometimes a feature of severe insulin resistance and accompanying diabetes. Generalized or localized *granuloma annulare* (erythematous plaques on the extremities or trunk), *lichen planus* (violaceous papules on the cutaneous surface +/− erosions in the mouth and genitalia), and *scleredema* (areas of skin thickening on the back or neck at the site of previous superficial infections) are more common in the diabetic population. *Lipoatrophy* and *lipohypertrophy* can occur at insulin injection sites but are now unusual with the use of human insulin, and avoided by rotating injection sites.

FURTHER READING

American Diabetes Association: Cardiovascular disease and risk management: Diabetes Care 41:S86, 2018.

Fox CS et al: Update on prevention of cardiovascular disease in adults with type 2 diabetes mellitus in light of recent evidence: A scientific statement from the American Heart Association and the American Diabetes Association. Diabetes Care 38:1777, 2015.

Hingorani A et al: The management of diabetic foot: A clinical practice guideline by the Society for Vascular Surgery in collaboration with the American Podiatric Medical Association and the Society for Vascular Medicine. J Vasc Surg 63:3S, 2016.

Holman RR et al: 10-Year follow-up of intensive glucose control in type 2 diabetes. N Engl J Med 359:1577, 2008.

Microvascular Complications and Foot Care. Diabetes Care 41:105, 2018.

Nathan DM et al: Diabetes control and complications trial/epidemiology of diabetes interventions and complications study at 30 years: Advances and contributions. Diabetes 62:3976, 2013.

Pop-Busui R et al: Diabetic neuropathy: A position statement by the American Diabetes Association. Diabetes Care 40:136, 2017.

Solomon SD et al: Diabetic retinopathy: A position statement by the American Diabetes Association. Diabetes Care 40:412, 2017.

399 Hypoglycemia

Philip E. Cryer, Stephen N. Davis

Hypoglycemia is most commonly caused by drugs used to treat diabetes mellitus or by exposure to other drugs, including alcohol. However, a number of other disorders, including critical organ failure, sepsis and inanition, hormone deficiencies, non-β-cell tumors, insulinoma, and prior gastric surgery, can cause hypoglycemia **(Table 399-1)**. Hypoglycemia may be documented by *Whipple's triad:* (1) symptoms consistent with hypoglycemia, (2) a low plasma glucose concentration measured with a precise method, and (3) relief of symptoms after the plasma glucose level is raised. The lower limit of the fasting plasma glucose concentration is normally ~70 mg/dL (~3.9 mmol/L), but lower venous glucose levels occur normally, late after a meal, during pregnancy, and during prolonged fasting (>24 h). Hypoglycemia can cause serious morbidity; if severe it can be fatal. It should be considered in any patient with episodes of confusion, an altered level of consciousness, or a seizure.

SYSTEMIC GLUCOSE BALANCE AND GLUCOSE COUNTERREGULATION

Glucose is an obligate metabolic fuel for the brain under physiologic conditions. The brain cannot synthesize glucose or store more than a few minutes' supply as glycogen and therefore requires a continuous supply of glucose from the arterial circulation. As the arterial plasma glucose concentration falls below the physiologic range, blood-to-brain glucose transport becomes insufficient to support brain energy metabolism and function. However, multiple integrated glucose counterregulatory mechanisms normally prevent or rapidly correct hypoglycemia.

Plasma glucose concentrations are normally maintained within a relatively narrow range—roughly 70–110 mg/dL (3.9–6.1 mmol/L) in the fasting state, with transient higher excursions after a meal—despite wide variations in exogenous glucose delivery from meals and in endogenous glucose utilization by, for example, exercising muscle. Between meals and during fasting, plasma glucose levels are maintained by endogenous glucose production, hepatic glycogenolysis, and hepatic (and renal) gluconeogenesis **(Fig. 399-1)**. Although hepatic glycogen stores are usually sufficient to maintain plasma glucose levels for ~8 h, this period can be shorter if glucose demand is increased by exercise or if glycogen stores are depleted by illness or starvation.

TABLE 399-1 Causes of Hypoglycemia in Adults

Ill or Medicated Individual
1. Drugs
Insulin or insulin secretagogue
Alcohol
Others
2. Critical illness
Hepatic, renal or cardiac failure
Sepsis
Inanition
3. Hormone deficiency
Cortisol
Growth hormone
Glucagon and epinephrine (in insulin-deficient diabetes)
4. Non–islet cell tumor (e.g., Mesenchymal tumors)
Seemingly Well Individual
5. Endogenous hyperinsulinism
Insulinoma
Functional β-cell disorders (nesidioblastosis)
Noninsulinoma pancreatogenous hypoglycemia
Post–gastric bypass hypoglycemia
Insulin autoimmune hypoglycemia
Antibody to insulin
Antibody to insulin receptor
Insulin secretagogue
Other
6. Disorders of Gluconeogenesis and fatty acid oxidation.
7. Exercise
8. Accidental, surreptitious, or malicious hypoglycemia

Source: Adapted from PE Cryer et al: J Clin Endocrinol Metab 94:709, 2009. ©The Endocrine Society, 2009.

Gluconeogenesis normally requires low insulin levels and the presence of anti-insulin (counterregulatory) hormones together with a coordinated supply of precursors from muscle and adipose tissue to the liver (and kidneys). Muscle provides lactate, pyruvate, alanine, glutamine, and other amino acids. Triglycerides in adipose tissue are broken down into fatty acids and glycerol, which is a gluconeogenic precursor. Fatty acids provide an alternative oxidative fuel to tissues other than the brain (which requires glucose).

Systemic glucose balance, maintenance of the normal plasma glucose concentration, is accomplished by a network of hormones, neural signals, and substrate effects that regulate endogenous glucose production and glucose utilization by tissues other than the brain **(Chap. 396)**. Among the regulatory factors, insulin plays a dominant role (**Table 399-2**; Fig. 399-1). As plasma glucose levels decline within the physiologic range, pancreatic β-cell insulin secretion decreases, thereby increasing hepatic glycogenolysis and hepatic (and renal) gluconeogenesis. Low insulin levels also reduce glucose utilization in peripheral tissues, inducing lipolysis and proteolysis and consequently releasing gluconeogenic precursors. Thus, a decrease in insulin secretion is the first defense against hypoglycemia.

As plasma glucose levels decline just below the physiologic range, glucose counterregulatory (plasma glucose–raising) hormones are released (Table 399-2; Fig. 399-1). Among these, pancreatic α-cell glucagon, and adrenomedullary epinephrine play a primary role. Glucagon stimulates hepatic glycogenolysis and gluconeogenesis. Adrenomedullary epinephrine also stimulates hepatic glycogenolysis and gluconeogenesis (and renal gluconeogenesis), but limits peripheral uptake of glucose and stimulates lipolysis with production of glycerol and fatty acids. Epinephrine becomes critical when glucagon is deficient. When hypoglycemia is prolonged beyond ~4 h, cortisol and growth hormone also support glucose production and restrict glucose utilization to a limited amount (~20% compared to epinephrine). Thus cortisol and growth hormone play no role in defense against acute hypoglycemia.

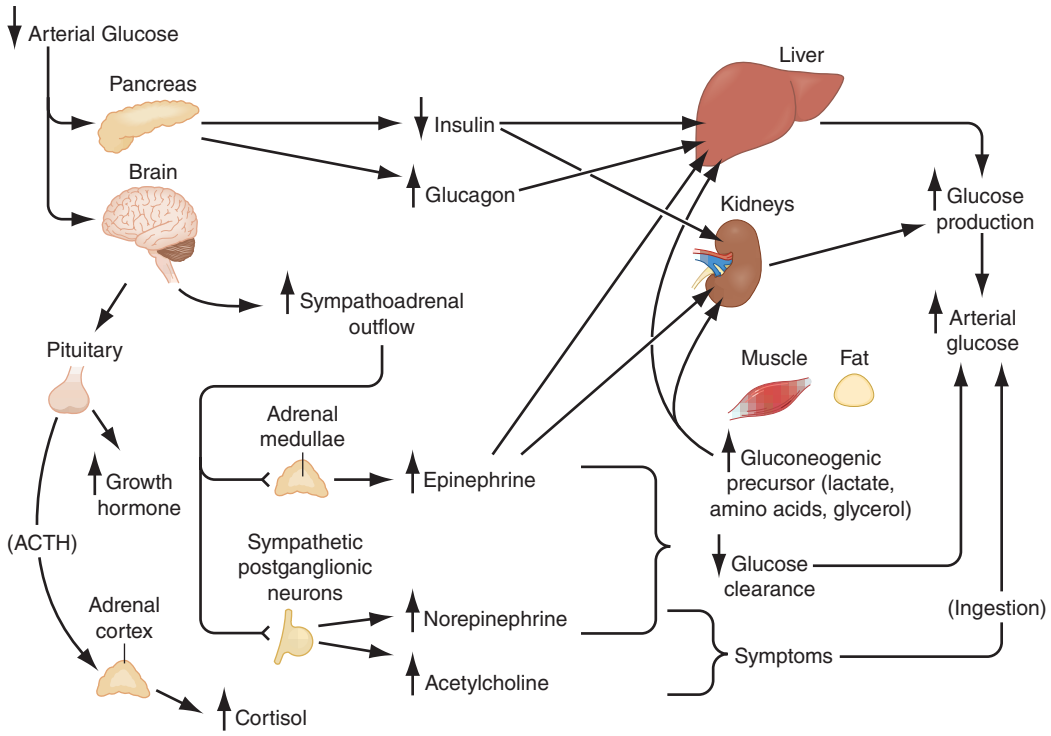


FIGURE 399-1 Physiology of glucose counterregulation: Mechanisms that normally prevent or rapidly correct hypoglycemia. In insulin-deficient diabetes, the key counterregulatory responses–suppression of insulin and increases in glucagon–are lost, and stimulation of sympathoadrenal outflow is attenuated. ACTH, adrenocorticotropic hormone.

As plasma glucose levels fall further, symptoms prompt behavioral defense against hypoglycemia, including the ingestion of food (Table 399-2; Fig. 399-1). The normal glycemic thresholds for these responses to decreasing plasma glucose concentrations are shown in Table 399-2. However, these thresholds are dynamic. They shift to higher-than-normal glucose levels in people with poorly controlled diabetes, who can experience symptoms of hypoglycemia when their glucose levels decline toward the normal range. On the other hand, thresholds shift to lower-than-normal glucose levels in people with recurrent hypoglycemia; i.e., patients with intensively treated diabetes or an insulinoma have symptoms at glucose levels lower than those that cause symptoms in healthy individuals.

Clinical Manifestations Neuroglycopenic manifestations of hypoglycemia are the direct result of central nervous system glucose deprivation. These features include behavioral changes, confusion, fatigue, seizure, loss of consciousness, cardiac arrhythmias, and, if hypoglycemia is severe, death. Neurogenic (or autonomic) manifestations of hypoglycemia result from the perception of physiologic changes caused by the central nervous system–mediated sympathoadrenal discharge that is triggered by hypoglycemia. They include *adrenergic* symptoms (mediated largely by norepinephrine released from sympathetic postganglionic neurons but perhaps also by epinephrine released from the adrenal medullae), such as palpitations, tremor, and anxiety, as well as *cholinergic* symptoms (mediated by acetylcholine released from sympathetic postganglionic neurons), such as sweating, hunger, and paresthesiae. Clearly, these are nonspecific symptoms. Their attribution to hypoglycemia requires that the corresponding plasma glucose concentration be low and that the symptoms resolve after the glucose level is raised (as delineated by Whipple's triad).

Common signs of hypoglycemia include diaphoresis and pallor. Heart rate and systolic blood pressure are typically increased, but may not be raised in an individual who has experienced repeated, recent episodes of hypoglycemia. Neuroglycopenic manifestations are often observable. Transient focal neurologic deficits occur occasionally. Permanent neurologic deficits are rare.

Etiology and Pathophysiology Hypoglycemia activates pro-inflammatory, pro-coagulant and pro-atherothrombotic responses in

TABLE 399-2 Physiologic Responses to Decreasing Plasma Glucose Concentrations

RESPONSE	GLYCEMIC THRESHOLD, mmol/L (mg/dL)	PHYSIOLOGIC ↓ EFFECTS	ROLE IN PREVENTION OR CORRECTION OF HYPOGLYCEMIA (GLUCOSE COUNTERREGULATION)
↓ Insulin	4.4–4.7 (80–85)	↑ R_a (↓ R_d), increased lipolysis; ↑ FFA ↑ Glycerol	Primary glucose regulatory factor/first defense against hypoglycemia
↑ Glucagon	3.6–3.9 (65–70)	↑ R_a	Primary glucose counterregulatory factor/second defense against hypoglycemia
↑ Epinephrine	3.6–3.9 (65–70)	↑ R_a, ↓ R_c, increased lipolysis; ↑ FFA & Glycerol	Third defense against hypoglycemia, critical when glucagon is deficient
↑ Cortisol and growth hormone	3.6–3.9 (65–70)	↑ R_a, ↓ R_c	Involved in defense against prolonged hypoglycemia; not critical
Symptoms	2.8–3.1 (50–55)	Recognition of hypoglycemia	Prompt behavioral defense against hypoglycemia (food ingestion)
↓ Cognition	<2.8 (<50)	—	Compromises behavioral defense against hypoglycemia

Note: R_a, rate of glucose appearance, glucose production by the liver and kidneys; R_c, rate of glucose clearance, glucose utilization relative to the ambient plasma glucose by insulin-sensitive tissues; R_d, rate of glucose disappearance, glucose utilization by insulin-sensitive tissues such as skeletal muscle. R_d by the brain is not altered by insulin, glucagon, epinephrine, cortisol, or growth hormone. FFA, free fatty acids.

Source: From PE Cryer, in S Melmed et al (eds): *Williams Textbook of Endocrinology,* 12th ed. New York, Elsevier, 2012.

T1DM, T2DM, and non-diabetic individuals. These responses increase platelet aggregation, reduce fibrinolytic balance (↑ plasminogen activator inhibitor-1), and increase intravascular coagulation. Hypoglycemia also reduces protective nitric oxide-mediated arterial vasodilator mechanisms in healthy, T1DM, and T2DM individuals.

HYPOGLYCEMIA IN DIABETES

Impact and Frequency Hypoglycemia is the limiting factor in the glycemic management of diabetes mellitus. First, it causes recurrent morbidity in most people with type 1 diabetes (T1DM) and in many with advanced type 2 diabetes (T2DM), and it is sometimes fatal. Second, it precludes maintenance of euglycemia over a lifetime of diabetes and thus full realization of the well-established microvascular benefits of glycemic control. Third, it causes a vicious cycle of recurrent hypoglycemia by producing hypoglycemia-associated autonomic failure—i.e., the clinical syndromes of defective glucose counterregulation and of hypoglycemia unawareness.

Hypoglycemia is a fact of life for people with T1DM if treated with insulin, sulfonylurea, or glinides. They suffer an average of two episodes of symptomatic hypoglycemia per week and at least one episode of severe, at least temporarily disabling hypoglycemia each year. An estimated 6–10% of people with T1DM die as a result of hypoglycemia. The incidence of hypoglycemia is lower in T2DM than in T1DM. However, its prevalence in insulin-requiring T2DM is surprisingly high. Recent studies have revealed a hypoglycemia prevalence approaching 70%. In fact, as patients with T2DM outnumber those with T1DM by ten- to twentyfold, the prevalence of hypoglycemia is now greater in T2DM. Insulin, sulfonylureas, or glinides can cause hypoglycemia in T2DM. Metformin, thiazolidinediones, α-glucosidase inhibitors, glucagon-like peptide 1 (GLP-1) receptor agonists, and dipeptidyl peptidase IV (DPP-IV) inhibitors do not cause hypoglycemia. However, they increase the risk when combined with a sulfonylurea, glinide, or insulin. Notably, the frequency of hypoglycemia approaches that in T1DM as persons with T2DM develop absolute insulin deficiency and require more complex treatment with insulin.

Conventional Risk Factors The conventional risk factors for hypoglycemia in diabetes are identified on the basis of the premise that relative or absolute insulin excess is the sole determinant of risk. Relative or absolute insulin excess occurs when (1) insulin (or insulin secretagogue) doses are excessive, ill-timed, or of the wrong type; (2) the influx of exogenous glucose is reduced (e.g., during an overnight fast, periods of temporary fasting, or after missed meals or snacks); (3) insulin-independent glucose utilization is increased (e.g., during exercise); (4) sensitivity to insulin is increased (e.g., with improved glycemic control, in the middle of the night, late after exercise, or with increased fitness or weight loss); (5) endogenous glucose production is reduced (e.g., after alcohol ingestion); and (6) insulin clearance is reduced (e.g., in renal failure). However, these conventional risk factors alone explain a minority of episodes; other factors are typically involved.

Hypoglycemia-Associated Autonomic Failure (HAAF) While marked insulin excess alone can cause hypoglycemia, iatrogenic hypoglycemia in diabetes (either T1DM and/or T2DM) is typically the result of the interplay of relative or absolute therapeutic insulin excess and compromised physiologic and behavioral defenses against falling plasma glucose concentrations (Table 399-2; Fig. 399-2). Defective glucose counterregulation compromises physiologic defense (particularly decrements in insulin and increments in glucagon and epinephrine), and hypoglycemia unawareness compromises behavioral defense (ingestion of carbohydrate).

DEFECTIVE GLUCOSE COUNTERREGULATION In the setting of absolute endogenous insulin deficiency, insulin levels do not decrease as plasma glucose levels fall; thus the first defense against hypoglycemia is lost. After a few years disease duration in T1DM, glucagon levels do not increase as plasma glucose levels fall; a second defense against hypoglycemia is lost. Reduced glucagon responses to hypoglycemia also occur in long duration T2DM. However, pancreatic alpha cells that produce glucagon are present in the same number and size in T1DM as compared to age matched non-diabetic individuals. Thus, the defect that restricts glucagon release during hypoglycemia in T1DM (and presumably in long-standing T2DM), appears to be a signaling defect, as glucagon responses to other physiologic stress in T1DM (e.g., exercise) are preserved. Finally, the increase in epinephrine levels, the third critical defense against acute hypoglycemia, is typically attenuated. The glycemic threshold for the sympathoadrenal (adrenomedullary epinephrine and sympathetic neural norepinephrine) response is shifted to lower plasma glucose concentrations. That shift is typically the result of recent antecedent iatrogenic hypoglycemia. In the setting of absent decrements in insulin and of absent increments in glucagon, the attenuated increment in epinephrine causes the clinical syndrome of defective glucose counterregulation. Affected patients are at ≥25-fold greater risk of severe iatrogenic hypoglycemia during intensive glycemic therapy for their diabetes than are patients with normal epinephrine responses. This functional—and potentially reversible—disorder is distinct from classic diabetic autonomic neuropathy—a structural and irreversible disorder.

HYPOGLYCEMIA UNAWARENESS The attenuated sympathoadrenal response (largely the reduced sympathetic neural response) to hypoglycemia causes the clinical syndrome of *hypoglycemia unawareness*—i.e., loss of the warning adrenergic and cholinergic symptoms that previously allowed the patient to recognize developing hypoglycemia and therefore to abort the episode by ingesting carbohydrates. Affected

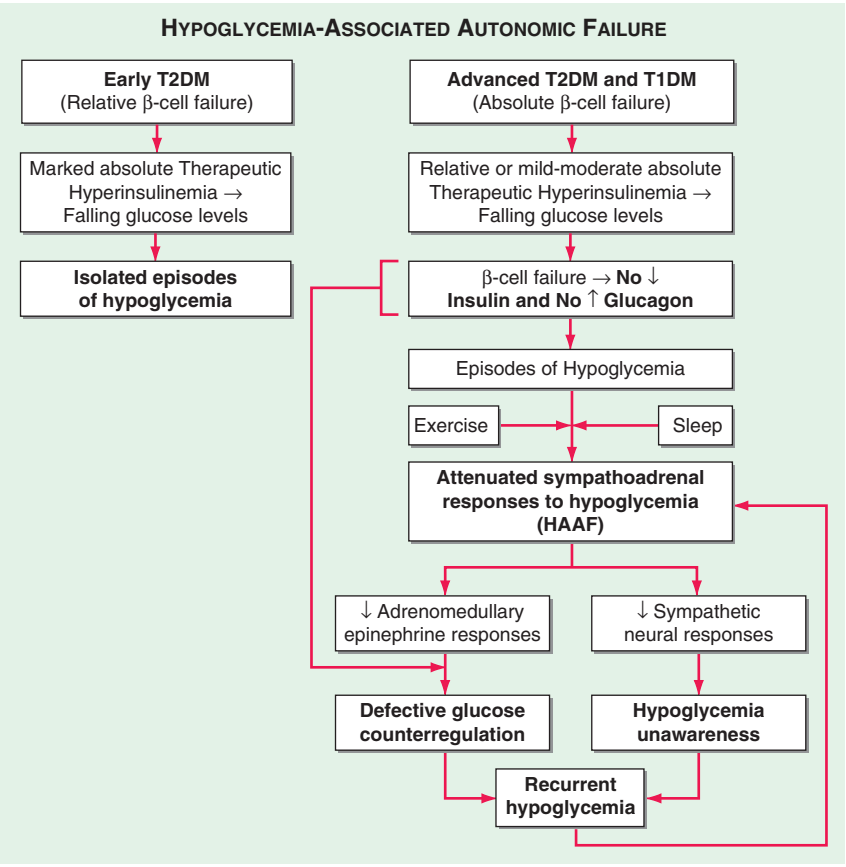


FIGURE 399-2 Hypoglycemia-associated autonomic failure (HAAF) in insulin-deficient diabetes. T1DM, type 1 diabetes mellitus; T2DM, type 2 diabetes mellitus. *(Modified from PE Cryer: Hypoglycemia in Diabetes. Pathophysiology, Prevalence, and Prevention, 2nd ed. © American Diabetes Association, 2012.)*

patients are at a sixfold increased risk of severe iatrogenic hypoglycemia during intensive glycemic therapy of their diabetes.

HAAF IN DIABETES The concept of HAAF in diabetes posits that recent antecedent iatrogenic hypoglycemia (or sleep or prior exercise) causes both defective glucose counterregulation (by reducing the epinephrine response to a given level of subsequent hypoglycemia in the setting of absent insulin and glucagon responses) and hypoglycemia unawareness (by reducing the sympathoadrenal response to a given level of subsequent hypoglycemia). These impaired responses create a vicious cycle of recurrent iatrogenic hypoglycemia (Fig. 399-2). Hypoglycemia unawareness and, to some extent, the reduced epinephrine component of defective glucose counterregulation are reversible by as little as 2–3 weeks of scrupulous avoidance of hypoglycemia in most affected patients.

On the basis of this pathophysiology, additional risk factors for hypoglycemia in diabetes include (1) absolute insulin deficiency, indicating that insulin levels will not decrease and glucagon levels will not increase as plasma glucose levels fall; (2) a history of severe hypoglycemia or of hypoglycemia unawareness, implying recent antecedent hypoglycemia, as well as prior exercise or sleep, indicating that the sympathoadrenal response will be attenuated; (3) impaired renal function resulting in reduced clearance of exogenous and endogenous insulin; (4) classical diabetic autonomic neuropathy, and (5) lower hemoglobin A_{1C} (HBA_{1C}), or lower glycemic goals even at elevated HBA_{1C} levels (8–10%), as they represent an increased probability of recent antecedent hypoglycemia.

Hypoglycemia Risk Factor Reduction Several recent multicenter, randomized, controlled trials investigating the potential benefits of tight glucose control in either inpatient or outpatient settings have reported a high prevalence of severe hypoglycemia. In the NICE-SUGAR study, attempts to control in-hospital plasma glucose values towards physiologic levels resulted in increased mortality risk. The ADVANCE and ACCORD studies and the Veterans Affairs Diabetes Trial (VADT) also found a significant incidence of severe hypoglycemia among T2DM patients. Severe hypoglycemia with accompanying serious cardiovascular morbidity and mortality also occurred in the standard (e.g., not receiving intensified treatment) control group in both the ACCORD study and the VADT. Thus, severe hypoglycemia can and does occur at HbA_{1c} values of 8–10% in both T1DM and T2DM. Somewhat surprisingly, all three studies found little or no benefit of intensive glucose control to reduce macrovascular events in T2DM. In fact, the ACCORD study was ended early because of the increased mortality rate in the intensive glucose control arm. Whether iatrogenic hypoglycemia was the cause of the increased mortality risk is not known. In light of these findings, some new recommendations and paradigms have been formulated. Whereas there is little debate regarding the need to reduce hyperglycemia in the hospital, the glycemic maintenance goals have been modified to lie between 140 and 180 mg/dL. Accordingly, the benefits of insulin therapy and reduced hyperglycemia can be obtained while the prevalence of hypoglycemia is reduced.

Similarly, evidence exists that intensive glucose control can reduce the prevalence of microvascular disease in both T1DM and T2DM. These benefits need to be weighed against the increased prevalence of hypoglycemia. Certainly, the level of glucose control (i.e., the HbA_{1c} level) should be evaluated for each patient. Multicenter trials have demonstrated that individuals with recently diagnosed T1DM or T2DM can have better glycemic control with less hypoglycemia. In addition, there is still long-term benefit in reducing HbA_{1c} values from higher to lower, albeit still above recommended levels. Perhaps a reasonable therapeutic goal is the lowest HbA_{1c} level that does not cause severe hypoglycemia and that preserves awareness of hypoglycemia.

Pancreatic transplantation (both whole-organ and islet-cell) has been used in part as a treatment for severe hypoglycemia. Generally, rates of hypoglycemia are reduced after transplantation. This decrease appears to be due to increased physiologic insulin and glucagon responses during hypoglycemia.

The use of continuous glucose monitors,either alone or in combination with continuous subcutaneous infusion via a wearable pump, offers promise as a method of reducing hypoglycemia while improving HbA_{1c}. Specifically, continuous glucose monitoring coupled with temporary discontinuation of subcutaneous insulin infusion when the monitor predicts a low glucose concentration is particularly promising. Furthermore, ongoing progress utilizing a portable wearable "artificial pancreas" incorporating continuous glucose sensor modulation of either insulin alone or bi-hormonal delivery of both insulin and glucagon has been established. Additionally, stem cell-derived β-cells also offer promise of novel therapeutic interventions to reduce hypoglycemia.

Other interventions to stimulate counterregulatory responses, such as selective serotonin-reuptake inhibitors, β-adrenergic receptor antagonists, opiate receptor antagonists, and fructose, remain experimental and have not been assessed in large-scale clinical trials.

Thus, intensive glycemic therapy **(Chap. 397)** needs to be applied along with the patient's education and empowerment, frequent self-monitoring of blood glucose, flexible insulin (and other drug) regimens (including the use of insulin analogues, both short- and longer-acting), individualized glycemic goals, and ongoing professional guidance, support, and consideration of both the conventional risk factors and those indicative of compromised glucose counterregulation. Given a history of hypoglycemia unawareness, a 2- to 3-week period of scrupulous avoidance of hypoglycemia is indicated.

HYPOGLYCEMIA WITHOUT DIABETES

There are many causes of hypoglycemia (Table 399-1). Because hypoglycemia is common in insulin- or insulin secretagogue–treated diabetes, it is often reasonable to assume that a clinically suspicious episode is the result of hypoglycemia. On the other hand, because hypoglycemia is rare in the absence of relevant drug-treated diabetes, it is reasonable to conclude that a hypoglycemic disorder is present only in patients in whom Whipple's triad can be demonstrated.

Particularly when patients are ill or medicated, the initial diagnostic focus should be on the possibility of drug involvement and then on critical illnesses, hormone deficiency, or non–islet cell tumor hypoglycemia. In the absence of any of these etiologic factors and in a seemingly well individual, the focus should shift to possible endogenous hyperinsulinism or accidental, surreptitious, or even malicious hypoglycemia.

Drugs Insulin and insulin secretagogues suppress glucose production and stimulate glucose utilization. Ethanol blocks gluconeogenesis but not glycogenolysis. Thus, alcohol-induced hypoglycemia typically occurs after a several-day ethanol binge during which the person eats little food, with consequent glycogen depletion. Ethanol is usually measurable in blood at the time of presentation, but its levels correlate poorly with plasma glucose concentrations. Because gluconeogenesis becomes the predominant route of glucose production during prolonged hypoglycemia, alcohol can contribute to the progression of hypoglycemia in patients with insulin-treated diabetes.

Many other drugs have been associated with hypoglycemia. These include commonly used drugs such as angiotensin-converting enzyme inhibitors and angiotensin receptor antagonists, β-adrenergic receptor antagonists, quinolone antibiotics, indomethacin, quinine, and sulfonamides.

Critical Illness Among hospitalized patients, serious illnesses such as renal, hepatic, or cardiac failure; sepsis; and inanition are second only to drugs as causes of hypoglycemia.

Rapid and extensive hepatic destruction (e.g., toxic hepatitis) causes fasting hypoglycemia because the liver is the major site of endogenous glucose production. The mechanism of hypoglycemia in patients with cardiac failure is unknown. Hepatic congestion and hypoxia may be involved. Although the kidneys are a source of glucose production, hypoglycemia in patients with renal failure is also caused by the reduced clearance of insulin (thereby inappropriately increasing insulin relative to the prevailing glucose levels), and the reduced mobilization of gluconeogenic precursors in renal failure.

Sepsis is a relatively common cause of hypoglycemia. Increased glucose utilization is induced by cytokine production in macrophage-rich

tissues such as the liver, spleen, and lung. Hypoglycemia develops if glucose production fails to keep pace. Cytokine-induced inhibition of gluconeogenesis in the setting of nutritional glycogen depletion, in combination with hepatic and renal hypoperfusion, may also contribute to hypoglycemia.

Hypoglycemia can be seen with starvation. Due to brain conversion and utilization of alternative substrates, such as lactate, pyruvate, and ketone bodies, there is only a modest counterregulatory neuroendocrine and autonomic nervous system response. During periods of prolonged starvation (fasting) plasma glucose levels are lower in women as compared to men; perhaps because of loss of whole-body fat stores and subsequent depletion of gluconeogenic precursors (e.g., amino acids), necessitating increased glucose utilization.

Hormone Deficiencies Neither cortisol nor growth hormone is critical to the prevention of hypoglycemia, at least in adults. Nonetheless, hypoglycemia can occur with prolonged fasting in patients with primary adrenocortical failure (Addison's disease) or hypopituitarism. Anorexia and weight loss are typical features of chronic cortisol deficiency and likely result in glycogen depletion. Cortisol deficiency is associated with impaired gluconeogenesis and low levels of gluconeogenic precursors; these associations suggest that substrate-limited gluconeogenesis, in the setting of glycogen depletion, is the cause of hypoglycemia. Growth hormone deficiency can cause hypoglycemia in young children. In addition to extended fasting, high rates of glucose utilization (e.g., during exercise or in pregnancy) or low rates of glucose production (e.g., after alcohol ingestion) can precipitate hypoglycemia in adults with previously unrecognized hypopituitarism.

Hypoglycemia is not a feature of the epinephrine-deficient state that results from bilateral adrenalectomy when glucocorticoid replacement is adequate, nor does it occur during pharmacologic adrenergic blockade when other glucoregulatory systems are intact. Combined deficiencies of glucagon and epinephrine play a key role in the pathogenesis of iatrogenic hypoglycemia in people with insulin-deficient diabetes, as discussed earlier. Otherwise, deficiencies of these hormones are not usually considered in the differential diagnosis of a hypoglycemic disorder.

Non-β-Cell Tumors Fasting hypoglycemia, often termed *non–islet cell tumor hypoglycemia*, occurs occasionally in patients with large mesenchymal or epithelial tumors (e.g., hepatomas, adrenocortical carcinomas, carcinoids). The glucose kinetic patterns resemble those of hyperinsulinism (see next), but insulin secretion is suppressed appropriately during hypoglycemia. In most instances, hypoglycemia is due to overproduction of an incompletely processed form of insulin-like growth factor II ("big IGF-II") that does not complex normally with circulating binding proteins and thus more readily gains access to target tissues. The tumors are usually apparent clinically, plasma ratios of IGF-II to IGF-I are high, and free IGF-II levels (and levels of pro-IGF-II [1–21]) are elevated. Curative surgery is seldom possible, but reduction of tumor bulk may ameliorate hypoglycemia. Therapy with a glucocorticoid, growth hormone, or both has also been reported to alleviate hypoglycemia. Hypoglycemia attributed to ectopic IGF-I production has been reported but is rare.

Endogenous Hyperinsulinism Hypoglycemia due to endogenous hyperinsulinism can be caused by (1) a primary β-cell disorder—typically a β-cell tumor (*insulinoma*), sometimes multiple insulinomas, or a functional β-cell disorder with β-cell hypertrophy or hyperplasia; (2) an antibody to insulin or to the insulin receptor; (3) a β-cell secretagogue such as a sulfonylurea; or perhaps (4) ectopic insulin secretion, among other very rare mechanisms. None of these causes is common.

The fundamental pathophysiologic feature of endogenous hyperinsulinism caused by a primary β-cell disorder or an insulin secretagogue is the failure of insulin secretion to fall to very low levels during hypoglycemia. This feature is assessed by measurement of plasma insulin, C-peptide (the connecting peptide that is cleaved from proinsulin to produce insulin), proinsulin, and glucose concentrations during hypoglycemia. Insulin, C-peptide, and proinsulin levels need not be high relative to normal, euglycemic values; rather, they are inappropriately high in the setting of a low plasma glucose concentration. Critical diagnostic findings are a plasma insulin concentration ≥3 μU/mL (≥18 pmol/L), a plasma C-peptide concentration ≥0.6 ng/mL (≥0.2 nmol/L), and a plasma proinsulin concentration ≥5.0 pmol/L when the plasma glucose concentration is <55 mg/dL (<3.0 mmol/L) with symptoms of hypoglycemia. A low plasma β-hydroxybutyrate concentration (≤2.7 mmol/L) and an increment in plasma glucose level of >25 mg/dL (>1.4 mmol/L) after IV administration of glucagon (1.0 mg) indicate increased insulin (or IGF) actions.

The diagnostic strategy is (1) to measure plasma glucose, insulin, C-peptide, proinsulin, and β-hydroxybutyrate concentrations and to screen for circulating oral hypoglycemic agents during an episode of hypoglycemia and (2) to assess symptoms during the episode and seek their resolution following correction of hypoglycemia by IV injection of glucagon (i.e., to document Whipple's triad). This is straightforward if the patient is hypoglycemic when seen. Since endogenous hyperinsulinemic disorders usually, but not invariably, cause fasting hypoglycemia, a diagnostic episode may develop after a relatively short outpatient fast. Serial sampling during an inpatient diagnostic fast of up to 72 h or after a mixed meal is more problematic. An alternative is to give patients a detailed list of the required measurements and ask them to present to an emergency room, with the list, during a symptomatic episode. Obviously, a normal plasma glucose concentration during a symptomatic episode indicates that the symptoms are not the result of hypoglycemia.

An *insulinoma*—an insulin-secreting pancreatic islet β-cell tumor—is the prototypical cause of endogenous hyperinsulinism and therefore should be sought in patients with a compatible clinical syndrome. However, insulinoma is not the only cause of endogenous hyperinsulinism. Some patients with fasting endogenous hyperinsulinemic hypoglycemia have diffuse islet involvement with β-cell hypertrophy and sometimes hyperplasia. This pattern is commonly referred to as *nesidioblastosis*, although β-cells budding from ducts are not invariably found. Other patients have a similar islet pattern but with postprandial hypoglycemia, a disorder termed *noninsulinoma pancreatogenous hypoglycemia*. Postgastric bypass postprandial hypoglycemia, which most often follows Roux-en-Y gastric bypass, is also characterized by diffuse islet involvement and endogenous hyperinsulinism. Some have suggested that exaggerated GLP-1 responses to meals cause hyperinsulinemia and hypoglycemia, but the relevant pathogenesis has not been clearly established. If medical treatments with agents such as an α-glucosidase inhibitor, diazoxide, or octreotide fail, partial pancreatectomy may be required. Autoimmune hypoglycemias include those caused by an antibody to insulin that binds post-meal insulin and then gradually disassociates, with consequent late postprandial hypoglycemia. Alternatively, an insulin receptor antibody can function as an agonist. The presence of an insulin secretagogue, such as a sulfonylurea or a glinide, results in a clinical and biochemical pattern similar to that of an insulinoma but can be distinguished by the presence of the circulating secretagogue. Finally, there are reports of very rare phenomena such as ectopic insulin secretion, a gain-of-function insulin receptor mutation, and exercise-induced hyperinsulinemia.

Insulinomas are uncommon, with an estimated yearly incidence of 1 in 250,000. Because >90% of insulinomas are benign, they are a treatable cause of potentially fatal hypoglycemia. The median age at presentation is 50 years in sporadic cases, but the tumor usually presents in the third decade when it is a component of multiple endocrine neoplasia type 1 (Chap. 381). More than 99% of insulinomas are within the substance of the pancreas, and the tumors are usually small (<2.0 cm in diameter in 90% of cases). Therefore, they come to clinical attention because of hypoglycemia rather than mass effects. CT or MRI detects ~70–80% of insulinomas. These methods detect metastases in the roughly 10% of patients with a malignant insulinoma. Transabdominal ultrasound often identifies insulinomas, and endoscopic ultrasound has a sensitivity of ~90%. Somatostatin receptor scintigraphy is thought to detect insulinomas in about half of patients. Selective pancreatic arterial calcium injections, with the endpoint of a sharp increase in hepatic venous insulin levels, regionalize insulinomas with high sensitivity, but this invasive procedure is seldom necessary except to confirm endogenous hyperinsulinism in the diffuse islet disorders. Intraoperative pancreatic

ultrasonography almost invariably localizes insulinomas that are not readily palpable by the surgeon. Surgical resection of a solitary insulinoma is generally curative. Diazoxide, which inhibits insulin secretion, or the somatostatin analogue octreotide can be used to treat hypoglycemia in patients with unresectable tumors; everolimus, an mTOR (mammalian target of rapamycin) inhibitor, is promising.

ACCIDENTAL, SURREPTITIOUS, OR MALICIOUS HYPOGLYCEMIA

Accidental ingestion of an insulin secretagogue (e.g., as the result of a pharmacy or other medical error) or even accidental administration of insulin can occur. Factitious hypoglycemia, caused by surreptitious or even malicious administration of insulin or an insulin secretagogue, shares many clinical and laboratory features with insulinoma. It is most common among health care workers, patients with diabetes or their relatives, and people with a history of other factitious illnesses. However, it should be considered in all patients being evaluated for hypoglycemia of obscure cause. Ingestion of an insulin secretagogue causes hypoglycemia with increased C-peptide levels, whereas exogenous insulin causes hypoglycemia with low C-peptide levels reflecting suppression of insulin secretion.

Analytical error in the measurement of plasma glucose concentrations is rare. On the other hand, glucose monitors used to guide treatment of diabetes are not quantitative instruments, particularly at low glucose levels, and should not be used for the definitive diagnosis of hypoglycemia. Even with a quantitative method, low measured glucose concentrations can be artifactual—e.g., the result of continued glucose metabolism by the formed elements of the blood ex vivo, particularly in the presence of leukocytosis, erythrocytosis, or thrombocytosis or with delayed separation of the serum from the formed elements (pseudohypoglycemia).

INBORN ERRORS OF METABOLISM CAUSING HYPOGLYCEMIA

Nondiabetic hypoglycemia also results from inborn errors of metabolism. Such hypoglycemia most commonly occurs in infancy but can also occur in adulthood. Cases in adults can be classified into those resulting in fasting hypoglycemia, postprandial hypoglycemia, and exercise-induced hypoglycemia.

Fasting Hypoglycemia Although rare, disorders of glycogenolysis can result in fasting hypoglycemia. These disorders include glycogen storage disease (GSD) of types 0, I, III, and IV and Fanconi-Bickel syndrome (Chap. 412). Patients with GSD types I and III characteristically have high blood lactate levels before and after meals, respectively. Both groups have hypertriglyceridemia, but ketones are high in GSD type III. Defects in fatty acid oxidation also result in fasting hypoglycemia. These defects can include (1) defects in the carnitine cycle; (2) fatty-acid β-oxidation disorders; (3) electron transfer disturbances; and (4) ketogenesis disorders. Finally, defects in gluconeogenesis (fructose-1, 6-biphosphatase) have been reported to result in recurrent hypoglycemia and lactic acidosis.

Postprandial Hypoglycemia Inborn errors of metabolism resulting in postprandial hypoglycemia are also rare. These errors include (1) glucokinase, SUR1, and Kir6.2 potassium channel mutations; (2) congenital disorders of glycosylation; and (3) inherited fructose intolerance.

Exercise-Induced Hypoglycemia Exercise-induced hypoglycemia, by definition, follows exercise. It results in hyperinsulinemia caused by increased activity of monocarboxylate transporter 1 in β cells.

APPROACH TO THE PATIENT

Hypoglycemia

In addition to the recognition and documentation of hypoglycemia as well as its treatment (often on an urgent basis), diagnosis of the hypoglycemic mechanism is critical for the selection of therapy that prevents, or at least minimizes, recurrent hypoglycemia.

RECOGNITION AND DOCUMENTATION

Hypoglycemia is suspected in patients with typical symptoms; in the presence of confusion, an altered level of consciousness, or a seizure; or in a clinical setting in which hypoglycemia is known to occur. Blood should be drawn, whenever possible, before the administration of glucose to allow documentation of a low plasma glucose concentration. Convincing documentation of hypoglycemia requires the fulfillment of Whipple's triad. Thus, the ideal time to measure the plasma glucose level is during a symptomatic episode. A normal glucose level excludes hypoglycemia as the cause of the symptoms. A low glucose level confirms that hypoglycemia is the cause of the symptoms, provided the latter resolve after the glucose level is raised. When the cause of the hypoglycemic episode is obscure, additional measurements—made while the glucose level is low and before treatment—should include plasma insulin, C-peptide, proinsulin, and β-hydroxybutyrate levels; also critical are screening for circulating oral hypoglycemic agents and assessment of symptoms before and after the plasma glucose concentration is raised.

When the history suggests prior hypoglycemia and no potential mechanism is apparent, the diagnostic strategy is to evaluate the patient as just described and assess for Whipple's triad during and after an episode of hypoglycemia. On the other hand, while it cannot be ignored, a distinctly low plasma glucose concentration measured in a patient without corresponding symptoms raises the possibility of an artifact (pseudohypoglycemia).

DIAGNOSIS OF THE HYPOGLYCEMIC MECHANISM

In a patient with documented hypoglycemia, a plausible hypoglycemic mechanism can often be deduced from the history, physical examination, and available laboratory data (Table 399-1). Drugs, particularly alcohol or agents used to treat diabetes, should be the first consideration—even in the absence of known use of a relevant drug—given the possibility of surreptitious, accidental, or malicious drug administration. Other considerations include evidence of a relevant critical illness, hormone deficiencies (less commonly), and a non-β-cell tumor that can be pursued diagnostically (rarely). Absent one of these mechanisms in an otherwise seemingly well individual, the physician should consider endogenous hyperinsulinism and proceed with measurements and assessment of symptoms during spontaneous hypoglycemia or under conditions that might elicit hypoglycemia.

URGENT TREATMENT

If the patient is able and willing, oral treatment with glucose tablets or glucose-containing fluids, candy, or food is appropriate. A reasonable initial dose is 15–20 g of glucose. If the patient is unable or unwilling (because of neuroglycopenia) to take carbohydrates orally, parenteral therapy is necessary. IV administration of glucose (25 g) should be followed by a glucose infusion guided by serial plasma glucose measurements. If IV therapy is not practical, SC or IM glucagon (1.0 mg in adults) can be used, particularly in patients with T1DM. Because it acts by stimulating glycogenolysis, glucagon is ineffective in glycogen-depleted individuals (e.g., those with alcohol-induced hypoglycemia). Glucagon also stimulates insulin secretion and is therefore less useful in T2DM. The somatostatin analogue octreotide can be used to suppress insulin secretion in sulfonylurea-induced hypoglycemia. These treatments raise plasma glucose concentrations only transiently, and patients should therefore be urged to eat as soon as is practical to replete glycogen stores.

PREVENTION OF RECURRENT HYPOGLYCEMIA

Prevention of recurrent hypoglycemia requires an understanding of the hypoglycemic mechanism. Offending drugs can be discontinued or their doses reduced. Hypoglycemia caused by a sulfonylurea can persist for hours or even days. Underlying critical illnesses can often be treated. Cortisol and growth hormone can be replaced if levels are deficient. Surgical, radiotherapeutic, or chemotherapeutic reduction of a non–islet cell tumor can alleviate hypoglycemia even if the tumor cannot be cured; glucocorticoid or growth hormone

administration also may reduce hypoglycemic episodes in such patients. Surgical resection of an insulinoma is curative; medical therapy with diazoxide or octreotide can be used if resection is not possible and in patients with a nontumor β-cell disorder. Partial pancreatectomy may be necessary in the latter patients. The treatment of autoimmune hypoglycemia (e.g., with glucocorticoid or immunosuppressive drugs) is problematic, but these disorders are sometimes self-limited. Failing these treatments, frequent feedings and avoidance of fasting may be required. Administration of uncooked cornstarch at bedtime or even an overnight intragastric infusion of glucose may be necessary for some patients.

FURTHER READING

BONDS DE et al: The association between symptomatic, severe hypoglycaemia and mortality in type 2 diabetes: Retrospective epidemiological analysis of the ACCORD study. BMJ 340:b4909, 2010.

CRYER PE: *Hypoglycemia in Diabetes*, 3rd ed. Alexandria, VA, American Diabetes Association, 2016.

CRYER PE: Hypoglycemia, in *Williams Textbook of Endocrinology*, 13th ed, S Melmed et al (eds). Philadelphia, Saunders, 2016, pp. 1582–1607.

JOY NG et al: Effects of acute and antecedent hypoglycemia on endothelial function and markers of atherothrombotic balance in health humans. Diabetes 64:2571, 2015.

RUSSELL SJ et al: Outpatient glycemic control with a bionic pancreas in type 1 diabetes. N Engl J Med 371:313, 2014.

400 Disorders of Lipoprotein Metabolism

Daniel J. Rader, Sekar Kathiresan

Lipoproteins are complexes of lipids and proteins that are essential for transport of cholesterol, triglycerides (TGs), and fat-soluble vitamins in the blood. Disorders of lipoprotein metabolism include primary and secondary conditions that substantially increase or decrease specific circulating lipids (e.g., cholesterol or TGs) or lipoproteins (e.g., low density or high density lipoproteins, see below). The demonstration that cholesterol-lowering therapy significantly reduces the clinical complications of atherosclerotic cardiovascular disease (ASCVD) makes it important for clinicians to be familiar with the diagnosis and treatment of lipoprotein disorders. This chapter reviews normal lipoprotein physiology, the pathophysiology of disorders of lipoprotein metabolism, the effects of genetic and environmental factors on lipoprotein metabolism, and the clinical approaches to the diagnosis and management of lipoprotein disorders.

LIPOPROTEIN METABOLISM

LIPOPROTEIN CLASSIFICATION AND COMPOSITION

Lipoproteins are large macromolecular complexes composed of lipids and proteins that transport poorly soluble lipids (primarily TGs, cholesterol, and fat-soluble vitamins) through body fluids (plasma, interstitial fluid, and lymph) to and from tissues. Lipoproteins play an essential role in the absorption of dietary cholesterol, long-chain fatty acids, and fat-soluble vitamins; the transport of TGs, cholesterol, and fat-soluble vitamins from the liver to peripheral tissues; and the transport of cholesterol from peripheral tissues to the liver and intestine.

Lipoproteins contain a core of hydrophobic lipids (TGs and cholesteryl esters) surrounded by a shell of hydrophilic lipids (phospholipids, unesterified cholesterol) and proteins (called apolipoproteins) that interact with body fluids. The plasma lipoproteins are divided into five major classes based on their relative density (Fig. 400-1 and Table 400-1): chylomicrons, very-low-density lipoproteins (VLDLs), intermediate-density lipoproteins (IDLs), low-density lipoproteins (LDLs), and high-density lipoproteins (HDLs). Each lipoprotein class comprises a family of particles that vary in density, size, and protein composition. Because lipid is less dense than water, the density of a lipoprotein particle is primarily determined by the amount of lipid per particle. Chylomicrons are the most lipid-rich and therefore least dense lipoprotein particles, whereas HDLs have the least lipid and are therefore the most dense lipoproteins. In addition to their density, lipoprotein particles can be classified according to their size, determined either by nondenaturing gel electrophoresis or by nuclear magnetic resonance profiling. There is a strong inverse relationship between density and size, with the largest particles being the most buoyant (chylomicrons) and the smallest particles being the most dense (HDL).

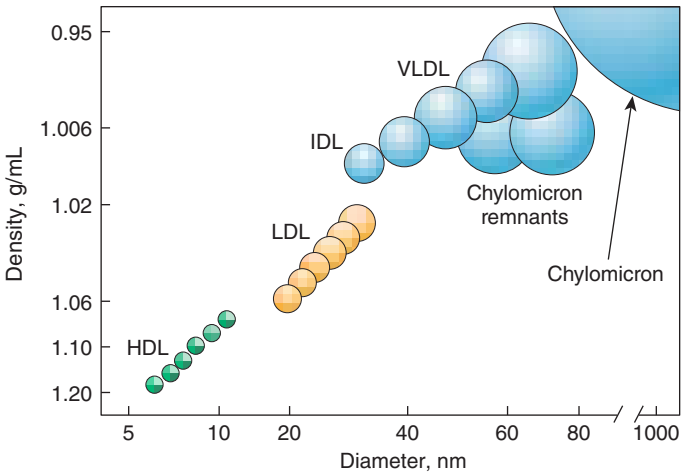


FIGURE 400-1 The density and size distribution of the major classes of lipoprotein particles. Lipoproteins are classified by density and size, which are inversely related. HDL, high-density lipoprotein; IDL, intermediate-density lipoprotein; LDL, low-density lipoprotein; VLDL, very-low-density lipoprotein.

The proteins associated with lipoproteins, called *apolipoproteins* (Table 400-2), are required for the assembly, structure, function, and metabolism of lipoproteins. Apolipoproteins provide a structural basis for lipoproteins, activate enzymes important in lipoprotein metabolism, and act as ligands for cell surface receptors. ApoB is a very large protein and is the major structural protein of chylomicrons, VLDLs, IDLs, and LDLs; one molecule of apoB, either apoB-48 (chylomicron) or apoB-100 (VLDL, IDL, or LDL), is present on each lipoprotein particle. The human liver synthesizes the full-length apoB-100, whereas the intestine makes the shorter apoB-48, which is derived from the same *APOB* gene by post-transcriptional mRNA editing. HDLs have different apolipoproteins that define this lipoprotein class, most importantly apoA-I, which is synthesized in both the liver and intestine and is found on virtually all HDL particles. ApoA-II is the second most abundant HDL apolipoprotein and is on approximately two-thirds of the HDL particles. ApoC-II, apoC-III, and apoA-V regulate the metabolism of TGs-rich lipoproteins. ApoE plays a critical role in the metabolism and clearance of TG-rich particles. Most apolipoproteins, other than apoB, exchange actively among lipoprotein particles in the blood. Apolipoprotein(a) [apo(a)] is a distinctive apolipoprotein that results in the formation of a lipoprotein known as lipoprotein(a) [Lp(a)], and is discussed more below.

TRANSPORT OF INTESTINALLY DERIVED DIETARY LIPIDS BY CHYLOMICRONS

One critical role of lipoproteins is the efficient transport of dietary lipids from the intestine to tissues that require fatty acids for energy or store and metabolize lipids and of intestinal cholesterol to the liver (Fig. 400-2). Dietary lipids are hydrolyzed by lipases within the intestinal lumen and emulsified with bile acids to form micelles. Dietary cholesterol, fatty acids, and fat-soluble vitamins are absorbed in the proximal small intestine. Cholesterol and retinol are esterified (by the

TABLE 400-1 Major Lipoprotein Classes

LIPOPROTEIN	DENSITY, g/mL[a]	SIZE, nm[b]	ELECTROPHORETIC MOBILITY[c]	APOLIPOPROTEINS		OTHER CONSTITUENTS
				MAJOR	OTHER	
Chylomicrons	0.930	75–1200	Origin	ApoB-48	A-I, A-V, C-I, C-II, C-III, E	Retinyl esters
Chylomicron remnants	0.930–1.006	30–80	Slow pre-β	ApoB-48	A-I, A-V, C-I, C-II, C-III, E	Retinyl esters
VLDL	0.930–1.006	30–80	Pre-β	ApoB-100	A-I, A-II, A-V, C-I, C-II, C-III, E	Vitamin E
IDL	1.006–1.019	25–35	Slow pre-β	ApoB-100	C-I, C-II, C-III, E	Vitamin E
LDL	1.019–1.063	18–25	β	ApoB-100		Vitamin E
HDL	1.063–1.210	5–12	α	ApoA-I	A-II, A-IV, A-V, C-III, E	LCAT, CETP, paroxonase
Lp(a)	1.050–1.120	25	Pre-β	ApoB-100	Apo(a)	Oxidized phospholipids

[a]The density of the particle is determined by ultracentrifugation. [b]The size of the particle is measured using gel electrophoresis. [c]The electrophoretic mobility of the particle on agarose gel electrophores reflects the size and surface charge of the particle, with β being the position of LDL and α being the position of HDL.

Note: All of the lipoprotein classes contain phospholipids, esterified and unesterified cholesterol, and triglycerides to varying degrees.

Abbreviations: CETP, cholesteryl ester transfer protein; HDL, high-density lipoprotein; IDL, intermediate-density lipoprotein; LCAT, lecithin-cholesterol acyltransferase; LDL, low-density lipoprotein; Lp(a), lipoprotein A; VLDL, very-low-density lipoprotein.

addition of a fatty acid) in the enterocyte to form cholesteryl esters and retinyl esters, respectively. Longer-chain fatty acids (>12 carbons) are incorporated into TGs and packaged with apoB-48, cholesteryl esters, retinyl esters, phospholipids, in a process that requires the action of the microsomal TGs transfer protein (MTP), to form chylomicrons. Nascent chylomicrons are secreted into the intestinal lymph and delivered via the thoracic duct directly to the systemic circulation, where they are extensively processed by peripheral tissues before reaching the liver. The particles encounter lipoprotein lipase (LPL), which is anchored to a glycosylphosphatidylinositol-anchored protein, GPIHBP1, that is attached to the endothelial surfaces of capillaries in adipose tissue, heart, and skeletal muscle (Fig. 400-2). The TGs of chylomicrons are hydrolyzed by LPL, and free fatty acids are released. ApoC-II and apoA-V are apolipoproteins that are transferred to circulating chylomicrons from HDL in the post-prandial state; apoC-II acts as a required cofactor for LPL activation and apoA-V serves as a facilitator of LPL activity. The released free fatty acids are taken up by adjacent myocytes or adipocytes and either oxidized to generate energy or reesterified and stored as TG. Some of the released free fatty acids bind albumin before entering cells and are transported to other tissues, especially the liver. The chylomicron particle progressively shrinks in size as the hydrophobic TG core is hydrolyzed and the hydrophilic lipids (cholesterol and phospholipids) and apolipoproteins on the particle surface are transferred to HDL, ultimately creating chylomicron remnants.

Chylomicron remnants are rapidly removed from the circulation by the liver through a process that requires apoE as a ligand for receptors in the liver. Consequently, few, if any, chylomicrons or chylomicron remnants are generally present in the blood after a 12-h fast, except in patients with certain disorders of lipoprotein metabolism.

TRANSPORT OF HEPATICALLY DERIVED LIPIDS BY VLDL AND LDL

Another key role of lipoproteins is the transport of hepatic lipids from the liver to the periphery (Fig. 400-2) to provide an energy source during fasting. During the fasting state, lipolysis of adipose TGs generates fatty acids that are transported to the liver, and the liver is also capable of synthesizing fatty acids through de novo lipogenesis. These fatty acids are esterified by the liver into TGs, which are packaged into VLDL particles along with apoB-100, cholesteryl esters, phospholipids, and vitamin E in a process that, like for chylomicron assembly, requires MTP. VLDL thus resemble chylomicrons in that they are "triglyceride-rich lipoproteins," but they contain apoB-100 rather than apoB-48, are smaller and less buoyant, and have a higher ratio of cholesterol to TG (~1 mg of cholesterol for every 5 mg of TG). After secretion by the liver into the plasma, as with chylomicrons, the TGs of VLDL are hydrolyzed by LPL, especially in muscle, heart, and adipose tissue. After the relatively TG-depleted VLDL remnants dissociate from LPL, they are referred to as IDLs, which contain roughly similar amounts of cholesterol and TG. The liver removes ~40–60% of IDL by receptor–mediated endocytosis via binding to apoE, which is acquired through transfer of this protein from HDL. The remainder of IDL is further remodeled by hepatic lipase (HL) to form LDL. During this process, phospholipids and TG in the particle are hydrolyzed, and most of the remaining apolipoproteins except apoB-100 are transferred to other lipoproteins. Approximately 70% of LDL is removed from the circulation by receptor–mediated endocytosis (primarily the LDL receptor) in the liver with apoB-100 serving as the ligand for the LDL receptor. It should be noted that apoB-48 does not contain the LDL receptor-binding ligand region and, therefore, clearance of apoB-48-containing

TABLE 400-2 Major Apolipoproteins

APOLIPOPROTEIN	PRIMARY SOURCE	LIPOPROTEIN ASSOCIATION	FUNCTION
ApoA-I	Intestine, liver	HDL, chylomicrons	Structural protein for HDL Activates LCAT
ApoA-II	Liver	HDL, chylomicrons	Structural protein for HDL
ApoA-IV	Intestine, liver	HDL, chylomicrons	Unknown
ApoA-V	Liver	VLDL, chylomicrons	Promotes LPL-mediated triglyceride lipolysis
Apo(a)	Liver	Lp(a)	Structural protein for Lp(a)
ApoB-48	Intestine	Chylomicrons, chylomicron remnants	Structural protein for chylomicrons
ApoB-100	Liver	VLDL, IDL, LDL, Lp(a)	Structural protein for VLDL, LDL, IDL, Lp(a) Ligand for binding to LDL receptor
ApoC-I	Liver	Chylomicrons, VLDL, HDL	Unknown
ApoC-II	Liver	Chylomicrons, VLDL, HDL	Cofactor for LPL
ApoC-III	Liver, intestine	Chylomicrons, VLDL, HDL	Inhibits LPL activity and lipoprotein binding to receptors
ApoE	Liver	Chylomicron remnants, IDL, HDL	Ligand for binding to LDL receptor and other receptors

Abbreviations: HDL, high-density lipoprotein; IDL, intermediate-density lipoprotein; LCAT, lecithin-cholesterol acyltransferase; LDL, low-density lipoprotein; Lp(a), lipoprotein A; LPL, lipoprotein lipase; VLDL, very-low-density lipoprotein.

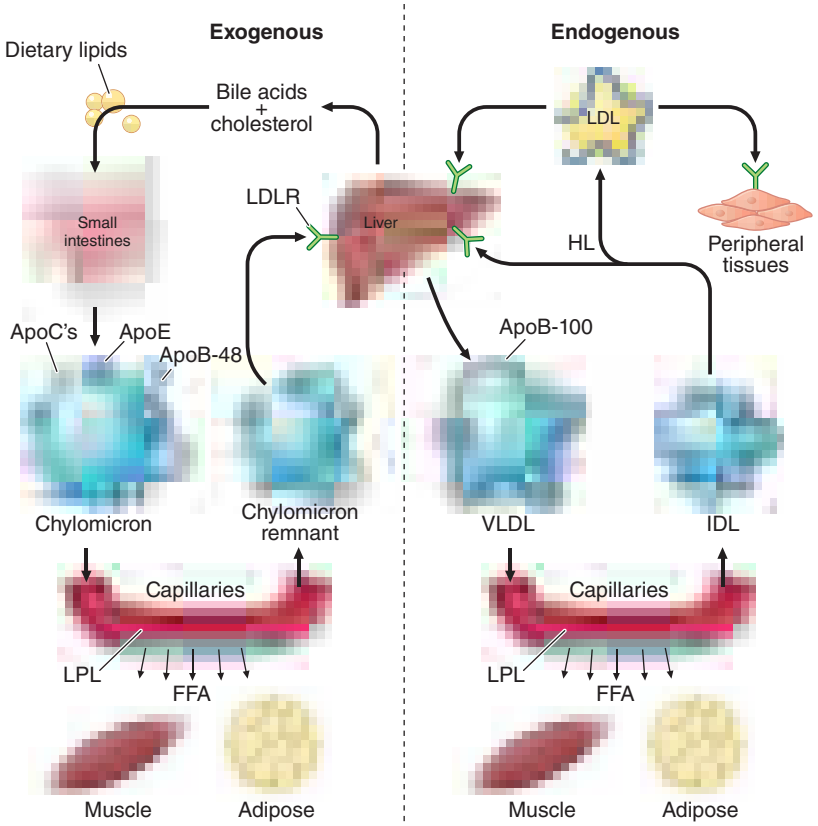


FIGURE 400-2 The exogenous and endogenous lipoprotein metabolic pathways. The exogenous pathway transports dietary lipids to the periphery and the liver. The endogenous pathway transports hepatic lipids to the periphery. FFA, free fatty acid; HL, hepatic lipase; IDL, intermediate-density lipoprotein; LDL, low-density lipoprotein; LDLR, low-density lipoprotein receptor; LPL, lipoprotein lipase; VLDL, very-low-density lipoprotein.

chylomicron remnants is dependent on apoE-mediated clearance as noted above. Some LDL particles are lipolytically processed to "small dense LDL" particles.

Lp(a) is a lipoprotein similar to LDL in lipid and protein composition, but it contains an additional distinctive protein called apolipoprotein(a) [apo(a)]. Apo(a) is synthesized in the liver and attached to apoB-100 by a disulfide linkage. The major site of clearance of Lp(a) is the liver, but the uptake pathway is not known.

HDL METABOLISM AND REVERSE CHOLESTEROL TRANSPORT

All nucleated cells synthesize cholesterol, but only hepatocytes and enterocytes can effectively excrete cholesterol from the body, into either the bile or the gut lumen, respectively. In the liver, cholesterol is secreted into the bile, either directly or after conversion to bile acids. Cholesterol in peripheral cells is transported from the plasma membranes of peripheral cells to the liver and intestine by a process termed "reverse cholesterol transport" that is facilitated by HDL (Fig. 400-3).

Nascent HDL particles are synthesized by the intestine and the liver. Newly secreted apoA-I rapidly acquires phospholipids and unesterified cholesterol from its site of synthesis (intestine or liver) via cellular efflux promoted by the membrane protein ATP-binding cassette protein A1 (ABCA1). This process results in the formation of discoidal HDL particles, which then recruit additional unesterified cholesterol from cells or circulating lipoproteins. Within the HDL particle, the cholesterol is esterified to cholesteryl ester (CE) through the addition of a fatty acid by lecithin-cholesterol acyltransferase (LCAT), a plasma enzyme associated with HDL; the hydrophobic CE forms the core of the mature HDL particle. As HDL acquires more CE, it becomes spherical, and additional apolipoproteins and lipids are transferred to the particles from the surfaces of chylomicrons and VLDLs during lipolysis.

HDL cholesterol is transported to hepatocytes by two major pathways. HDL CE can be "selectively" taken up by hepatocytes via the scavenger receptor class B1 (SR-B1), a cell surface HDL receptor that mediates the selective transfer of CE from HDL with subsequent dissociation and "recycling" of the HDL particle. In addition, HDL CE can be transferred to apoB-containing lipoproteins in exchange for TG by the cholesteryl ester transfer protein (CETP). The CE esters are then removed from the circulation by LDL receptor–mediated endocytosis. HDL-derived CE taken up by the hepatocyte through these pathways is hydrolyzed and much of the cholesterol is ultimately excreted directly into the bile or converted to bile acids with excretion to bile, providing a biliary route into the intestinal lumen. There is also evidence that, under certain conditions, HDL cholesterol can be transported directly into the intestinal lumen without requiring a transhepatobiliary route, a process known as "'transintenstinal cholesterol excretion."

HDL particles undergo extensive remodeling within the plasma compartment by a variety of lipid transfer proteins and lipases. The phospholipid transfer protein (PLTP) transfers phospholipids from other lipoproteins to HDL or among different classes of HDL particles and is a regulator of HDL metabolism. After CETP- and PLTP-mediated lipid exchange, the TG-enriched HDL becomes a much better substrate for HL, which hydrolyzes the TGs and phospholipids to generate smaller HDL particles. A related enzyme called endothelial lipase (EL) hydrolyzes HDL phospholipids, generating smaller HDL particles that are catabolized faster. Remodeling of HDL influences the metabolism, function, and plasma concentrations of HDL.

DISORDERS OF ELEVATED CHOLESTEROL AND TGS

Disorders of lipoprotein metabolism are collectively referred to as "dyslipidemias." Dyslipidemias are generally characterized clinically by increased plasma levels of cholesterol, TGs, or both, variably

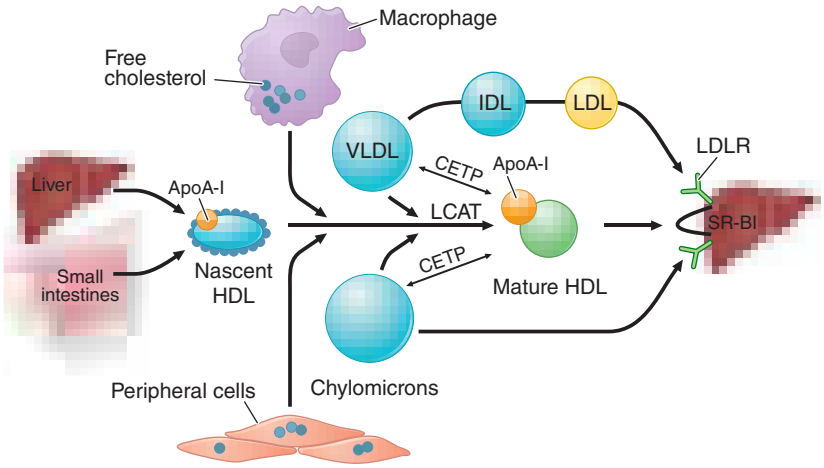


FIGURE 400-3 High-density lipoprotein (HDL) metabolism and reverse cholesterol transport. This pathway transports excess cholesterol from the periphery back to the liver for excretion in the bile. The liver and the intestine produce nascent HDLs. Free cholesterol is acquired from macrophages and other peripheral cells and esterified by lecithin-cholesterol acyltransferase (LCAT), forming mature HDLs. HDL cholesterol can be selectively taken up by the liver via SR-BI (scavenger receptor class BI). Alternatively, HDL cholesteryl ester can be transferred by cholesteryl ester transfer protein (CETP) from HDLs to very-low-density lipoproteins (VLDLs) and chylomicrons, which can then be taken up by the liver. IDL, intermediate-density lipoprotein; LDL, low-density lipoprotein; LDLR, low-density lipoprotein receptor.

accompanied by reduced levels of HDL cholesterol. Unusually low levels of cholesterol also fall within the broad scope of lipoprotein disorders. Because plasma lipids are commonly screened (see below), dyslipidemia is frequently seen in clinical practice. The majority of patients with dyslipidemia have some combination of genetic predisposition (often polygenic) and environmental contribution (diet, lifestyle, medical condition, or drug). Many, but not all, patients with dyslipidemia are at increased risk for ASCVD, which is the primary reason for making the diagnosis, as intervention can substantially reduce this risk. In addition, patients with markedly elevated levels of TGs may be at risk for acute pancreatitis and require intervention to reduce this risk.

Although literally hundreds of proteins influence lipoprotein metabolism and may interact to produce dyslipidemia in an individual patient, there are a limited number of discrete "nodes" or pathways that regulate lipoprotein metabolism and are dysfunctional in specific dyslipidemias. These include: (1) assembly and secretion of TG-rich VLDLs by the liver; (2) lipolysis of TG-rich lipoproteins by LPL; (3) receptor-mediated uptake of apoB-containing lipoproteins by the liver; (4) cellular cholesterol metabolism in the hepatocyte and the enterocyte; and (5) neutral lipid transfer and phospholipid hydrolysis in the plasma. The following discussion will focus on these regulatory nodes, recognizing that in many cases these nodes interact with and influence each other.

DYSLIPIDEMIA CAUSED BY EXCESSIVE HEPATIC SECRETION OF VLDL

Excessive production of VLDL by the liver is one of the most common causes of dyslipidemia. Individuals with excessive hepatic VLDL production usually have elevated fasting TGs and low levels of HDL cholesterol (HDL-C), with variable elevations in LDL cholesterol (LDL-C). A cluster of other metabolic risk factors are often found in association with VLDL overproduction, including obesity, glucose intolerance, insulin resistance, and hypertension (the so-called "metabolic syndrome," Chap. 401). Some of the major factors that drive hepatic VLDL secretion include a high-carbohydrate diet, excessive alcohol use, obesity and insulin resistance, nephrotic syndrome, and genetic factors.

Secondary Causes of VLDL Overproduction

• **HIGH-CARBOHYDRATE DIET** Dietary carbohydrates are utilized as a substrate for fatty acid synthesis in the liver. Some of the newly synthesized fatty acids are esterified, forming TGs, and secreted in VLDL. Thus, excessive intake of calories as carbohydrates, which is frequent in Western societies, leads to increased hepatic VLDL-TG secretion.

ALCOHOL Excessive alcohol consumption inhibits hepatic oxidation of free fatty acids, thus promoting hepatic TG synthesis and VLDL secretion. Regular alcohol use also raises plasma levels of HDL-C and should be considered in patients with the relatively unusual combination of elevated TGs and elevated HDL-C.

OBESITY AND INSULIN RESISTANCE (See also Chaps. 395 and 396) Obesity and insulin resistance are frequently accompanied by dyslipidemia characterized by elevated plasma levels of TG, low HDL-C, variable levels of LDL-C, and increased levels of small dense LDL. The increase in adipocyte mass and accompanying decreased insulin sensitivity associated with obesity have multiple effects on lipid metabolism, with one of the major effects being excessive hepatic VLDL production. More free fatty acids are delivered from the expanded and insulin-resistant adipose tissue to the liver, where they are re-esterified in hepatocytes to form TGs, which are packaged into VLDLs for secretion into the circulation. In addition, the increased insulin levels promote increased fatty acid synthesis in the liver. In insulin-resistant patients who progress to type 2 diabetes mellitus, dyslipidemia remains common, even when the patient is under relatively good glycemic control. In addition to increased VLDL production, insulin resistance can also result in decreased LPL activity, resulting in reduced catabolism of chylomicrons and VLDLs and more severe hypertriglyceridemia (see below).

NEPHROTIC SYNDROME (See also Chap. 305) Nephrotic syndrome is a classic cause of excessive VLDL production. The molecular mechanism of VLDL overproduction remains poorly understood but has been attributed to the effects of hypoalbuminemia leading to increased hepatic protein synthesis. Effective treatment of the underlying renal disease often normalizes the lipid profile, but most patients with chronic nephrotic syndrome require lipid-lowering drug therapy.

CUSHING'S SYNDROME (See also Chap. 379) Endogenous or exogenous glucocorticoid excess is associated with increased VLDL synthesis and secretion and hypertriglyceridemia. Patients with Cushing's syndrome frequently have dyslipidemia especially characterized by hypertriglyceridemia and low HDL-C, although elevations in plasma levels of LDL-C can also be seen.

Primary (Genetic) Causes of VLDL Overproduction

Genetic variation influences hepatic VLDL production. A number of genes have been identified in which common and low-frequency variants probably contribute to increased VLDL production, likely involving interactions with diet and other environmental factors. The best recognized inherited condition associated with VLDL overproduction is familial combined hyperlipidemia.

FAMILIAL COMBINED HYPERLIPIDEMIA (FCHL) FCHL is generally characterized by elevations in plasma levels of TGs (VLDL) and LDL-C (including small dense LDL) and reduced plasma levels of HDL-C. It is estimated to occur in ~1 in 100–200 individuals and is an important contributor to premature coronary heart disease (CHD); ~20% of patients who develop CHD under age 60 have FCHL. FCHL can manifest in childhood but is usually not fully expressed until adulthood. The disease clusters in families, and affected family members typically have one of three possible phenotypes: (1) elevated plasma levels of LDL-C, (2) elevated plasma levels of TGs due to elevation in VLDL, or (3) elevated plasma levels of both LDL-C and TG. The lipoprotein profile can switch among these three phenotypes in the same individual over time and may depend on factors such as diet, exercise, weight, and insulin sensitivity. Patients with FCHL have substantially elevated plasma levels of apoB, often disproportionately high relative to the plasma LDL-C concentration, indicating the presence of small dense LDL particles, which are characteristic of this syndrome.

Individuals with this phenotype generally share the same metabolic defect, namely overproduction of VLDL by the liver. The molecular etiology of this condition remains poorly understood, and no single gene has been identified in which mutations cause this disorder in a simple Mendelian fashion. It is likely that defects in a combination of genes can cause the condition, suggesting that a more appropriate term for the disorder might be *polygenic combined hyperlipidemia.*

The presence of a mixed dyslipidemia (plasma TG levels between 200 and 600 mg/dL and total cholesterol levels between 200 and 400 mg/dL, usually with HDL-C levels <40 mg/dL in men and <50 mg/dL in women) and a family history of dyslipidemia and/or premature CHD suggests the diagnosis. Measurement of apoB levels can help support the diagnosis if they are substantially elevated relative to the LDL-C level. Individuals with this phenotype should be treated aggressively due to significantly increased risk of premature CHD. Decreased dietary intake of simple carbohydrates, increase aerobic exercise, and weight loss can all have beneficial effects on the lipid profile. Patients with type 2 diabetes should be aggressively treated to maintain good glucose control. Most patients with FCHL require lipid-lowering drug therapy, starting with statins, to reduce apoB-containing lipoprotein levels and lower the risk of cardiovascular disease.

LIPODYSTROPHY Lipodystrophy is a condition in which the generation of adipose tissue generally or in certain fat depots is impaired. Lipodystrophies are often associated with insulin resistance and elevated plasma levels of VLDL and chylomicrons due to increased fatty acid synthesis and VLDL production, as well as reduced clearance of TG-rich particles. Patients with congenital generalized lipodystrophy—a recessive disorder caused by mutations in the *AGPAT2* and *BSCL2* genes—are very rare. These patients have nearly complete absence of subcutaneous fat, accompanied by profound insulin resistance and leptin deficiency, severe hypertriglyceridemia, and accumulation of TGs in multiple tissues including the liver. Patients with generalized

lipodystrophy can often be effectively treated with recombinant leptin administration. Partial lipodystrophy is a dominantly inherited disorder that is somewhat more common than the generalized form. It is caused by mutations in several different genes, including lamin A/C (*LMNA*), PPAR gamma (*PPARG*), perilipin (*PLIN1*), and *AKT2*. Partial lipodystrophy is characterized by markedly reduced subcutaneous fat in the extremities and buttocks, accompanied by increased facial, neck, and truncal fat. These patients generally have insulin resistance, often quite severe, accompanied by type 2 diabetes, hepatosteatosis, and dyslipidemia. The dyslipidemia, attributed mostly to increased VLDL production but also possibly due to other factors, is usually characterized by substantially elevated TGs and cholesterol and can be difficult to manage clinically. Patients with partial lipodystrophy are at substantially increased risk of atherosclerotic vascular disease and should therefore be treated aggressively for their dyslipidemia with statins and, if necessary, additional lipid-lowering therapies.

DYSLIPIDEMIA CAUSED BY IMPAIRED LIPOLYSIS OF TG-RICH LIPOPROTEINS

Impaired lipolysis of the TGs in TG-rich lipoproteins (TRLs) also commonly contributes to dyslipidemia. As noted above, LPL is the key enzyme responsible for hydrolyzing the TGs in chylomicrons and VLDL. LPL is synthesized and secreted into the extracellular space from adipocytes, skeletal myocytes, and cardiomyocytes. It is then transported from the subendothelial to the vascular endothelial surfaces by GPIHPB1, which helps dock it to the endothelial surface. Individuals with impaired LPL activity, whether secondary or due to a primary genetic disorder, have elevated fasting TGs and low levels of HDL-C, usually without elevation in LDL-C or apoB. Insulin resistance, in addition to causing excessive VLDL production, can also cause impaired LPL activity and lipolysis. A number of common, low-frequency, and rare genetic variants have been described that influence LPL activity, and single-gene Mendelian disorders that reduce LPL activity have also been described (Table 400-3).

Secondary Causes of Impaired Lipolysis of TRLs • **OBESITY AND INSULIN RESISTANCE** (See also Chaps. 394, 395, and 396) In addition to hepatic overproduction of VLDL, as discussed above, obesity, insulin resistance, and type 2 diabetes have been reported to be associated with variably reduced LPL activity. This may be due in part to the effects of tissue insulin resistance leading to reduced transcription of LPL in skeletal muscle and adipose, as well as to increased production of the LPL inhibitor apoC-III by the liver. This reduction in LPL activity often exacerbates the effects of increased VLDL production and contributes to the dyslipidemia seen in these patients.

Primary (Genetic) Causes and Genetic Predisposition to Impaired Lipolysis of TRLs • **FAMILIAL CHYLOMICRONEMIA SYNDROME** As noted above, LPL is required for the hydrolysis of TGs in chylomicrons and VLDLs. Genetic deficiency or inactivity of LPL results in impaired lipolysis and profound elevations in plasma chylomicrons, causing *familial chylomicronemia syndrome*. While chylomicronemia predominates, in fact these patients often have elevated plasma levels of VLDL as well. The fasting plasma is turbid, and if left undisturbed for several hours, the chylomicrons float to the top and form a creamy supernatant layer. Fasting TG levels are almost invariably >1000 mg/dL. Fasting cholesterol levels are also elevated but to a lesser degree. The most common cause of FCS involves mutations in the *LPL* gene. *LPL deficiency* has autosomal recessive inheritance (loss of function mutations in both alleles) and has an estimated frequency of ~1 in 1 million, though its true prevalence is unknown. Heterozygotes with LPL mutations often have moderate elevations in plasma TG levels and increased risk for CHD.

Familial chylomicronemia syndrome can be caused by mutations in genes other than LPL. For example, apoC-II is a required cofactor for LPL. *ApoC-II deficiency* due to loss of function mutations in both *APOC2* alleles results in functional lack of LPL activity and severe hyperchylomicronemia that is indistinguishable from LPL deficiency. It is also recessive in inheritance pattern and much rarer than LPL deficiency. Individuals heterozygous for a mutation in *APOC2* do not generally have hypertriglyceridemia. Another apolipoprotein, apoA-V, facilitates the association of VLDL and chylomicrons with LPL and promotes hydrolysis of the TGs. Individuals harboring loss-of-function mutations in both *APOA5* alleles causing *ApoA-V deficiency* develop a form of familial chylomicronemia syndrome. Heterozygosity for variants in *APOA5* that reduce its function contributes to the polygenic basis of hypertriglyceridemia. GPIHBP1 is required for transport and tethering of LPL to the endothelial luminal surface. Homozygosity for mutations

TABLE 400-3 Primary Hyperlipoproteinemias Caused by Known Single-Gene Mutations

GENETIC DISORDER	PROTEIN (GENE) DEFECT	LIPOPROTEINS ELEVATED	CLINICAL FINDINGS	GENETIC TRANSMISSION	ESTIMATED INCIDENCE
Hypertriglyceridemia					
Lipoprotein lipase deficiency	LPL (*LPL*)	Chylomicrons, VLDL	Eruptive xanthomas, hepatosplenomegaly, pancreatitis	AR	~1/1,000,000
Familial apoC-II deficiency	ApoC-II (*APOC2*)	Chylomicrons, VLDL	Eruptive xanthomas, hepatosplenomegaly, pancreatitis	AR	<1/1,000,000
ApoA-V deficiency	ApoA-V (*APOA5*)	Chylomicrons, VLDL	Eruptive xanthomas, hepatosplenomegaly, pancreatitis	AR	<1/1,000,000
GPIHBP1 deficiency	*GPIHBP1*	Chylomicrons	Eruptive xanthomas, pancreatitis	AR	<1/1,000,000
Combined Hyperlipidemia					
Familial hepatic lipase deficiency	Hepatic lipase (*LIPC*)	VLDL remnants, HDL	Pancreatitis, CHD	AR	<1/1,000,000
Familial dysbetalipoproteinemia	ApoE (*APOE*)	Chylomicron remnants, VLDL remnants	Palmar and tuberoeruptive xanthomas, CHD, PVD	AR	~1/10,000
Hypercholesterolemia					
Familial hypercholesterolemia	LDL receptor (*LDLR*)	LDL	Tendon xanthomas, CHD	AD	~1/250 to 1/500
Familial defective apoB-100	ApoB-100 (*APOB*)	LDL	Tendon xanthomas, CHD	AD	<~1/1500
Autosomal dominant hypercholesterolemia, type 3	PCSK9 (*PCSK9*)	LDL	Tendon xanthomas, CHD	AD	<1/1,000,000
Autosomal recessive hypercholesterolemia	ARH (*LDLRAP*)	LDL	Tendon xanthomas, CHD	AR	<1/1,000,000
Sitosterolemia	*ABCG5* or *ABCG8*	LDL	Tendon xanthomas, CHD	AR	<1/1,000,000

Abbreviations: AD, autosomal dominant; apo, apolipoprotein; AR, autosomal recessive; ARH, autosomal recessive hypercholesterolemia; CHD, coronary heart disease; LDL, low-density lipoprotein; LPL, lipoprotein lipase; PVD, peripheral vascular disease; VLDL, very-low density lipoprotein.

in *GPIHBP1* that interfere with its synthesis or folding cause familial chylomicronemia syndrome. Autoantibodies to GPIHBP1 have also been reported to cause severe hyperchylomicronemia.

Familial chylomicronemia syndrome can present in childhood or adulthood with recurrent episodes of severe abdominal pain due to acute pancreatitis. In this setting, the diagnosis should be suspected if a fasting TG level is >750 mg/dL. Eruptive xanthomas, which are small, yellowish-white papules, may appear in clusters on the back, buttocks, and extensor surfaces of the arms and legs. On funduscopic examination, the retinal blood vessels may be opalescent (lipemia retinalis). Hepatosplenomegaly is sometimes noted as a result of uptake of circulating chylomicrons by reticuloendothelial cells in the liver and spleen. Premature CHD is not generally a feature of familial chylomicronemia syndromes.

The diagnosis of familial chylomicronemia syndrome is a clinical diagnosis based on persistence and severity of hypertriglyceridemia in the setting of a history of proven or suspected acute pancreatitis. While LPL activity can be measured in "postheparin plasma" obtained after an IV heparin injection to release the endothelial-bound LPL, this assay is not widely available. Molecular sequencing of the candidate FCS genes can be used to confirm the diagnosis, but is not required for making the clinical diagnosis.

Because of the risk of pancreatitis, it is important to consider the diagnosis and institute therapeutic interventions in familial chylomicronemia syndrome. The goal is to prevent pancreatitis by reducing fasting TG levels to <500 mg/dL. Dietary fat intake should be markedly restricted (to as little as 15 gm/day), often with fat-soluble vitamin supplementation. Consultation with a registered dietician familiar with this disorder is essential. Usually dietary fat restriction alone is not successful in resolving the chylomicronemia, in which case fish oils have been modestly effective in some patients; fibrates (such as fenofibrate) may be tried but are also unlikely to be effective. A new therapeutic approach involving the suppression of *APOC3* with an antisense oligonucleotide is a promising approach for patients with FCS. In patients with apoC-II deficiency, apoC-II can be provided by infusing fresh-frozen plasma to resolve the chylomicronemia in the acute setting. Management of patients with familial chylomicronemia syndrome is particularly challenging during pregnancy when VLDL production is increased.

FAMILIAL HYPERTRIGLYCERIDEMIA (FHTG) FHTG is characterized by elevated fasting TGs without a clear secondary cause, average to below average LDL-C levels, low HDL-C levels, and a family history of hypertriglyceridemia. Plasma LDL-C levels are often reduced due to defective conversion of TG-rich lipoproteins to LDL. In contrast to FCHL, apoB levels are not elevated. The identification of other first-degree relatives with hypertriglyceridemia is useful in making the diagnosis. Unlike in FCHL, this condition is not generally associated with a significantly increased risk of CHD. However, if the hypertriglyceridemia is exacerbated by environmental factors, medical conditions, or drugs, the TGs can rise to a level at which acute pancreatitis is a risk. Indeed, management of patients with this condition is mostly focused on reduction of TGs to prevent pancreatitis.

Individuals with this phenotype generally have reduced lipolysis of TRLs, although overproduction of VLDL by the liver can also contribute. While this disorder runs in families, often with a dominant pattern of inheritance, a molecular etiology has not been established. Combinations of gene variants have been shown to cause this phenotype and therefore a more appropriate term for this condition might be *polygenic hypertriglyceridemia.*

It is important to consider and rule out secondary causes of the hypertriglyceridemia as discussed above. Increased intake of simple carbohydrates, obesity, insulin resistance, alcohol use, estrogen treatment, and certain medications can exacerbate this phenotype. Patients who are at high risk for CHD due to other risk factors should be treated with statin therapy. In patients who are otherwise not at high risk for CHD, lipid-lowering drug therapy can frequently be avoided with appropriate dietary and lifestyle changes. Patients with plasma TG levels >500 mg/dL after a trial of diet and exercise should be considered for drug therapy with a fibrate or fish oil to reduce TGs in order to prevent pancreatitis.

DYSLIPIDEMIA CAUSED BY IMPAIRED HEPATIC UPTAKE OF APOB-CONTAINING LIPOPROTEINS

Impaired uptake of LDL and remnant lipoproteins by the liver is another common cause of dyslipidemia. As discussed above, the LDL receptor is the major receptor responsible for uptake of LDL and remnant particles by the liver. Down-regulation of LDL receptor activity or genetic variation that reduces the activity of the LDL receptor pathway leads to elevations in LDL-C. One major factor that reduces LDL receptor activity is a diet high in saturated and *trans* fats. Other medical conditions that reduce LDL receptor activity include hypothyroidism and estrogen deficiency. In addition, genetic variation in a number of genes influences LDL clearance, and mutations in some of these genes cause several discrete Mendelian disorders of elevated LDL-C (Table 400-3).

Secondary Causes of Impaired Hepatic Uptake of Lipoproteins • **HYPOTHYROIDISM** (See also Chap. 375) Hypothyroidism is associated with elevated plasma LDL-C levels due primarily to a reduction in hepatic LDL receptor function and delayed clearance of LDL. Thyroid hormone increases hepatic expression of the LDL receptor. Hypothyroid patients also frequently have increased levels of circulating IDL, and some patients with hypothyroidism also have mild hypertriglyceridemia. Because hypothyroidism is often subtle and therefore easily overlooked, all patients presenting with elevated plasma levels of LDL-C, especially if there has been an unexplained increase in LDL-C, should be screened for hypothyroidism. Thyroid replacement therapy usually ameliorates the hypercholesterolemia; if not, the patient probably has a primary lipoprotein disorder and may require lipid-lowering drug therapy with a statin.

CHRONIC KIDNEY DISEASE (See also Chap. 305) Chronic kidney disease (CKD) is often associated with mild hypertriglyceridemia (150–400 mg/dL) due to the accumulation of VLDLs and remnant lipoproteins in the circulation. TG lipolysis and remnant clearance are both reduced in patients with renal failure. Because the risk of ASCVD is increased in CKD, patients should usually be treated with lipid-lowering agents, particularly statins.

Patients with solid organ transplants often have increased lipid levels due to the effect of the drugs required for immunosuppression. These patients can present a difficult clinical management problem, but statins are often indicated in these patients, with careful attention to the potential for untoward muscle-related side effects.

Primary (Genetic) Causes of Impaired Hepatic Uptake of Lipoproteins Genetic variation contributes substantially to elevated LDL-C levels in the general population. It has been estimated that at least 50% of variation in LDL-C is genetically determined. Many patients with elevated LDL-C have *polygenic hypercholesterolemia* due to multiple genetic variants exerting modest LDL-raising effects. In patients who are genetically predisposed to higher LDL-C levels, diet plays a key role; indeed increased saturated and *trans* fats in the diet shifts the entire distribution of LDL levels in the population to the right. Importantly, single-gene (Mendelian) causes of elevated LDL-C are relatively common and should be considered in the differential diagnosis of elevated LDL-C (Table 400-3).

FAMILIAL HYPERCHOLESTEROLEMIA (FH) FH, also known as autosomal dominant hypercholesterolemia (ADH), is an autosomal co-dominant disorder characterized by elevated plasma levels of LDL-C in the absence of hypertriglyceridemia. FH is caused by mutations that lead to reduced function of the LDL receptor, with the most common being mutations in the *LDLR* gene itself. The reduction in LDL receptor activity in the liver results in a reduced rate of clearance of LDL from the circulation. The plasma level of LDL increases to a level such that the rate of LDL production equals the rate of LDL clearance by residual LDL receptor as well as non-LDL receptor mechanisms. The elevated levels of LDL-C in FH are primarily due to delayed removal of LDL from the blood; in addition, because the removal of IDL is also delayed, the production of LDL from IDL is also increased. Individuals with two

mutated *LDLR* alleles (FH homozygotes, or compound heterozygotes) have much higher LDL-C levels than those with one mutant allele (FH heterozygotes).

Although mutations in the *LDLR* are the most common cause of FH, mutations in at least two other genes, *APOB* and *PCSK9*, can also cause ADH. ApoB-100 is the critical structural protein in LDL and contains a ligand for binding to the LDL receptor. Mutations in the LDL receptor–binding domain of apoB-100 cause a form of FH, also known as ADH type 2 or familial defective apoB (FDB). The mutations reduce the affinity of LDL binding to the LDL receptor, such that LDL is removed from the circulation at a reduced rate. Of note, truncating mutations in *APOB* cause low LDL-C levels (see below). The proprotein convertase subtilisin/kexin type 9 (PCSK9) is a secreted protein that binds to the LDL receptor and targets it for lysosomal degradation. Normally, after LDL binds to the LDL receptor, it is internalized along with the receptor, and in the low pH of the endosome, the LDL receptor dissociates from the LDL and recycles to the cell surface. When circulating PCSK9 binds the receptor, the complex is internalized and the receptor is directed to the lysosome, rather than to the cell surface, reducing the number of active LDL receptors. Gain-of-function mutations in *PCSK9* that enhance the activity of PCSK9 cause a form of FH, also known as ADH type 3. Of note, loss-of-function mutations in *PCSK9* markedly lower LDL-C levels (see below).

The population frequency of heterozygous FH was originally estimated to be 1 in 500 individuals, but recent data suggest it is ~1 in 250 individuals, making it one of the most common single-gene disorders in humans. FH has a much higher prevalence in certain founder populations, such as South African Afrikaners, Christian Lebanese, French Canadians, and Lancaster County Amish. Heterozygous FH is characterized by elevated plasma levels of LDL-C (~190–400 mg/dL) and usually relatively normal levels of TGs. Patients with heterozygous FH have hypercholesterolemia from birth, and disease recognition is often based on detection of hypercholesterolemia on routine screening, or a notable family history of hypercholesterolemia, or premature coronary heart disease. Inheritance of FH is dominant, meaning that the condition is inherited from one parent, and ~50% of the patient's siblings and children can be expected to have FH. For this reason, family-based "cascade screening" can be very effective in identifying additional persons with FH. The family history is frequently positive for premature CHD on the side of the family from which the mutation was inherited. Physical findings in some, but not all, patients with heterozygous FH include corneal arcus and/or tendon xanthomas, particularly involving the dorsum of the hands and the Achilles tendons. Untreated heterozygous FH is associated with a markedly increased risk of cardiovascular disease; untreated men with heterozygous FH have an ~50% chance of having a myocardial infarction before age 60 years, and women with heterozygous FH are at substantially increased risk as well. The age of onset of cardiovascular disease is highly variable and depends on the specific molecular defect, the level of LDL-C, and coexisting cardiovascular risk factors.

The diagnosis of FH is generally a clinical diagnosis based on substantial hypercholesterolemia with LDL-C >190 mg/dL in the absence of a secondary etiology, and a family history of hypercholesterolemia and/or premature coronary disease. Secondary causes of significant hypercholesterolemia such as hypothyroidism, nephrotic syndrome, and obstructive liver disease should be excluded. Sequencing of the FH genes (*LDLR*, *APOB*, *PCSK9*) to confirm the diagnosis is available and worthy of consideration; persons with confirmed FH are at higher risk of CVD than those with similar LDL-C levels who don't have FH and therefore may benefit from more aggressive treatment of hypercholesterolemia.

FH patients should always be actively treated to lower plasma levels of LDL-C, preferably starting in childhood. Initiation of a diet low in saturated and *trans* fats is recommended, but heterozygous FH patients require pharmacologic therapy for effective control of their LDL-C levels. Statins are the initial drug class of choice, and usually a more potent member of the class. Many heterozygous FH patients cannot achieve adequate control of their LDL-C levels even with high-intensity statin therapy, and a cholesterol absorption inhibitor (ezetimibe), a PCSK9 inhibitor, or a bile acid sequestrant are the next-line classes of drugs.

Homozygous FH (hoFH) is caused by mutations in both alleles of the LDL receptor or double heterozygosity for mutations in two FH genes. Patients with homozygous FH have been classified into those with virtually no detectable LDL receptor activity (*receptor negative*) and those patients with markedly reduced but detectable LDL receptor activity (*receptor defective*). LDL-C levels in patients with homozygous FH range from about 400 to >1000 mg/dL, with receptor-defective patients at the lower end and receptor-negative patients at the higher end of the range. TGs are usually normal. Some patients with homozygous FH, particularly receptor-negative patients, present in childhood with cutaneous xanthomas on the hands, wrists, elbows, knees, heels, or buttocks. The devastating consequence of homozygous FH is accelerated ASCVD, which often presents in childhood or early adulthood. Atherosclerosis often develops first in the aortic root, where it can cause aortic valvular or supravalvular stenosis, and typically extends into the coronary ostia, which become stenotic. Symptoms can be atypical, and sudden death is not uncommon. Untreated, receptor-negative patients with homozygous FH rarely survive beyond the second decade; patients with receptor-defective LDL receptor defects have a better prognosis but almost invariably develop clinically apparent atherosclerotic vascular disease by age 30, and often much sooner. Carotid and femoral disease develops later in life and is usually not clinically significant.

Homozygous FH should be suspected in a child or young adult with LDL >400 mg/dL without secondary cause. Cutaneous xanthomas, evidence of CVD, and hypercholesterolemia in both parents all are supportive of the diagnosis. While the diagnosis is usually made on clinical grounds, specific mutations can usually be identified by DNA sequencing. Patients with homozygous FH must be treated aggressively to delay the onset and progression of CVD. Although receptor negative patients have no response to statins and PCSK9 inhibitors, receptor defective patients can have modest responses to these medicines and they should be tried in patients with hoFH. Two drugs that reduce the hepatic production of VLDL and thus LDL, a small-molecule inhibitor of the microsomal TG transfer protein (MTP) and an antisense oligonucleotide to apoB, are approved in the United States for the treatment of patients with homozygous FH and should be considered in patients who have insufficient response to statins and PCSK9 inhibitors. LDL apheresis, a physical method of purging the blood of LDL in which the LDL particles are selectively removed from the circulation, should be considered in hoFH patients who have persistently elevated LDL-C levels despite attempts at drug therapy. Liver transplantation is effective in decreasing plasma LDL-C levels in this disorder and is sometimes used as a last resort. Liver-directed gene therapy is under development for hoFH.

FH is an autosomal dominant disorder. There are a few rare conditions that cause an FH-like phenotype in an autosomal recessive manner and should be considered in patients with substantial hypercholesterolemia who do not report a dominant family history of hypercholesterolemia or premature CHD.

AUTOSOMAL RECESSIVE HYPERCHOLESTEROLEMIA (ARH) ARH is a very rare autosomal recessive disorder that was originally reported in individuals of Sardinian descent. The disease is caused by mutations in the gene *LDLRAP1* encoding the protein LDLR adaptor protein (also called the ARH protein) which is required for LDL receptor–mediated endocytosis in the liver. LDLRAP1 binds to the cytoplasmic domain of the LDL receptor and links the receptor to the endocytic machinery. In the absence of LDLRAP1, LDL binds to the extracellular domain of the LDL receptor, but the lipoprotein-receptor complex fails to be internalized. ARH, like homozygous FH, is characterized by hypercholesterolemia, tendon xanthomas, and premature coronary artery disease (CAD). The levels of plasma LDL-C tend to be intermediate between the levels present in FH homozygotes and FH heterozygotes, and CAD is not usually symptomatic until the third decade. LDL receptor function in cultured fibroblasts is normal or only modestly reduced

in ARH, whereas LDL receptor function in lymphocytes and the liver is negligible. Unlike FH homozygotes, the hyperlipidemia responds to treatment with statins, but these patients often require additional therapy to lower plasma LDL-C to acceptable levels.

SITOSTEROLEMIA Sitosterolemia is a rare autosomal recessive disease that is caused by biallelic loss-of-function mutations in either of two members of the ATP-binding cassette (ABC) half transporter family, *ABCG5* and *ABCG8*. These genes are expressed in both enterocytes and hepatocytes. The proteins heterodimerize to form a functional complex that transports plant sterols such as sitosterol and campesterol, and animal sterols, predominantly cholesterol, across the biliary membrane of hepatocytes into the bile and across the intestinal luminal surface of enterocytes into the gut lumen, reducing their absorption and promoting their excretion. In normal individuals, <5% of dietary plant sterols are absorbed by the proximal small intestine. The small amounts of plant sterols that enter the circulation are preferentially excreted into the bile and thus levels of plant sterols are kept very low in tissues. In sitosterolemia, the intestinal absorption of sterols is increased and biliary and fecal excretion of the sterols is reduced, resulting in increased plasma and tissue levels of both plant sterols and cholesterol. The increase in hepatic sterol levels results in transcriptional suppression of the expression of the LDL receptor, resulting in reduced uptake of LDL and substantially increased LDL-C levels. In addition to the clinical picture of severe hypercholesterolemia, often accompanied by tendon xanthomas and premature ASCVD, these patients also have anisocytosis and poikilocytosis of erythrocytes and megathrombocytes due to the incorporation of plant sterols into cell membranes. Episodes of hemolysis and splenomegaly are a distinctive clinical feature of this disease compared to other genetic forms of hypercholesterolemia and can be a clue to the diagnosis. Sitosterolemia should be suspected in a patient with severe hypercholesterolemia without a family history of such or who fails to respond to statin therapy. Sitosterolemia can be diagnosed by a laboratory finding of a substantial increase in plasma sitosterol and/or other plant sterols, and should be confirmed by gene sequencing of *ABCG5* and *ABCG8*. It is important to make the diagnosis, because diet, bile acid sequestrants, and cholesterol-absorption inhibitors are the most effective agents to reduce LDL-C and plasma plant sterol levels in these patients. Of note, heterozygosity for mutations in *ABCG5* or *ABCG8* is now recognized to cause a moderate form of hypercholesterolemia.

LYSOSOMAL ACID LIPASE DEFICIENCY (LALD) LALD, also known as *cholesteryl ester storage disease*, is an autosomal recessive disorder caused by loss-of-function variants in both alleles of the gene *LIPA* encoding the enzyme lysosomal acid lipase (LAL). LAL is responsible for hydrolyzing neutral lipids, particularly TGs and cholesteryl esters, after delivery to the lysosome by cell-surface receptors such as the LDL receptor. It is particularly important in the liver, which clears large amounts of lipoproteins from the circulation. LALD is characterized by elevated LDL-C, usually in association with low HDL-C and with variably elevated TG levels, together with progressive fatty liver ultimately leading to hepatic fibrosis. Genetic deficiency of LAL results in accumulation of neutral lipid in the hepatocytes, leading to hepatosplenomegaly, microvesicular steatosis, and ultimately fibrosis and end-stage liver disease. The most severe form of this disorder, Wolman's disease, presents in infancy and is rapidly fatal. The etiology of the elevated LDL-C levels is primarily due to impaired LDL receptor–mediated clearance of LDL. LALD should be suspected in nonobese patients with elevated LDL-C, low HDL-C, and evidence of fatty liver in the absence of overt insulin resistance. The diagnosis can be made with a dried blood spot assay of LAL activity and confirmed by DNA genotyping for the most common mutation, followed if necessary by sequencing of the gene to find the second mutation. Liver biopsy is required to assess the degree of inflammation and fibrosis. LALD is underdiagnosed; it is critically important to suspect it and make the diagnosis because enzyme replacement therapy is now available and is highly effective in treating this condition.

The above conditions primarily cause elevations in LDL due to impaired catabolism of LDL from the blood. There are a few forms of primary dyslipidemia that impair the catabolism of "remnant" TG-rich lipoproteins (after their processing by LPL) and therefore cause elevations in both cholesterol and TGs due to remnant accumulation.

FAMILIAL DYSBETALIPOPROTEINEMIA (FDBL) FDBL (also known as *type III hyperlipoproteinemia*) is usually a recessive disorder characterized by a mixed hyperlipidemia (elevated cholesterol and TGs) due to the accumulation of remnant lipoprotein particles (chylomicron remnants and VLDL remnants, or IDL). ApoE is present in multiple copies on chylomicron remnants and IDL, and mediates their removal via hepatic lipoprotein receptors (Fig. 400-2). FDBL is due to genetic variants of apoE, most commonly apoE2, that result in an apoE protein with reduced ability to bind lipoprotein receptors. The *APOE* gene is polymorphic in sequence, resulting in the expression of three common isoforms: apoE3, which is the most common; and apoE2 and apoE4, which both differ from apoE3 by a single amino acid. Although associated with slightly higher LDL-C levels and increased CHD risk, the apoE4 allele is not associated with FDBL. Individuals who carry one or two apoE4 alleles have an increased risk of Alzheimer's disease. ApoE2 has a lower affinity for the LDL receptor; therefore, chylomicron remnants and IDL containing apoE2 are removed from plasma at a slower rate. Individuals who are homozygous for the E2 allele (the E2/E2 genotype) comprise the most common subset of patients with FDBL.

Approximately 0.5% of the general population are apoE2/E2 homozygotes, but only a small minority of these individuals actually develop hyperlipidemia characteristic of FDBL. In most cases, an additional, sometimes identifiable, factor precipitates the development of hyperlipoproteinemia. The most common precipitating factors are a high-fat diet, diabetes mellitus, obesity, hypothyroidism, renal disease, HIV infection, estrogen deficiency, alcohol use, or certain drugs. The disease seldom presents in women before menopause. Certain "dominant negative" mutations in apoE can cause a dominant form of FDBL where the hyperlipidemia is fully manifest in the heterozygous state, but these mutations are very rare.

Patients with FDBL usually present in adulthood with hyperlipidemia, xanthomas, or premature coronary or peripheral vascular disease. In FDBL, in contrast to other disorders of elevated TGs, the plasma levels of cholesterol and TG are often elevated to a similar degree, and the level of HDL-C is usually normal or reduced. Two distinctive types of xanthomas, tuberoeruptive and palmar, are seen in FDBL patients. Tuberoeruptive xanthomas begin as clusters of small papules on the elbows, knees, or buttocks and can grow to the size of small grapes. Palmar xanthomas (alternatively called *xanthomata striata palmaris*) are orange-yellow discolorations of the creases in the palms and wrists. Both of these xanthoma types are virtually pathognomonic for FDBL. Subjects with FDBL have premature ASCVD and tend to have more peripheral vascular disease than is typically seen in FH.

The definitive diagnosis of FDBL can be made either by the documentation of very high levels of remnant lipoproteins or by identification of the apoE2/E2 genotype. A variety of methods are used to identify remnant lipoproteins in the plasma, including "β-quantification" by ultracentrifugation (ratio of directly measured VLDL-C to total plasma TG >0.30), lipoprotein electrophoresis (broad β band), or nuclear magnetic resonance lipoprotein profiling. The Friedewald formula for calculation of LDL-C is not valid in FDBL because the VLDL particles are depleted in TG and enriched in cholesterol. The plasma levels of LDL-C are actually low in this disorder due to defective metabolism of VLDL to LDL. DNA-based apoE genotyping can be performed to confirm homozygosity for apoE2. However, absence of the apoE2/E2 genotype does not strictly rule out the diagnosis of FDBL, because other mutations in apoE can (rarely) cause this condition.

Because FDBL is associated with increased risk of premature ASCVD, it should be treated aggressively. Other metabolic conditions that can worsen the hyperlipidemia (see above) should be managed. Patients with FDBL are typically diet-responsive and can respond favorably to weight reduction and to low-cholesterol, low-fat diets. Alcohol intake should be curtailed. Pharmacologic therapy is often required, and statins are the first line in management. In the event of

statin intolerance or insufficient control of hyperlipidemia, cholesterol absorption inhibitors, fibrates, and PCSK9 inhibitors are also effective in the treatment of FDBL.

HEPATIC LIPASE DEFICIENCY Hepatic lipase (HL; gene name *LIPC*) is a member of the same gene family as LPL and hydrolyzes TGs and phospholipids in remnant lipoproteins and HDL. Hydrolysis of lipids in remnant particles by HL contributes to their hepatic uptake via an apoE-mediated process. HL deficiency is a very rare autosomal recessive disorder characterized by elevated plasma levels of cholesterol and TGs (mixed hyperlipidemia) due to the accumulation of lipoprotein remnants, accompanied by elevated plasma level of HDL-C. The diagnosis is confirmed by measuring HL activity in postheparin plasma and/or confirmation of loss-of-function mutations in both alleles of HL/*LIPC*. Due to the small number of patients with HL deficiency, the association of this genetic defect with ASCVD is not entirely clear, although anecdotally patients with HL deficiency who have premature CVD have been described. As with FDBL, statin therapy is recommended to reduce remnant lipoproteins and cardiovascular risk.

Additional Secondary Causes of Dyslipidemia

Many of the secondary causes of dyslipidemia **(Table 400-4)** have been described above. Additional considerations are discussed here.

LIVER DISORDERS **(See also Chap. 329)** Because the liver is the principal site of formation and clearance of lipoproteins, liver disorders can affect plasma lipid levels in a variety of ways. Hepatitis due to infection, drugs, or alcohol is often associated with increased VLDL synthesis and mild to moderate hypertriglyceridemia. Severe hepatitis and liver failure are associated with dramatic reductions in plasma cholesterol and TGs due to reduced lipoprotein biosynthetic capacity. Cholestasis is often associated with hypercholesterolemia. A major pathway by which cholesterol is excreted from the body is via secretion into bile, either directly or after conversion to bile acids, and cholestasis blocks this critical excretory pathway. In cholestasis, free cholesterol, coupled with phospholipids, is secreted into the plasma as a constituent of a lamellar particle called *LP-X*. The particles can deposit in skinfolds, producing lesions resembling those seen in patients with FDBL (xanthomata strata palmaris). Planar and eruptive xanthomas can also be seen in patients with cholestasis.

DRUGS Many drugs have an impact on lipid metabolism and can result in significant alterations in the lipoprotein profile (Table 400-4). Estrogen administration is associated with increased VLDL and HDL synthesis, resulting in elevated plasma levels of both TGs and HDL-C. This lipoprotein pattern is distinctive because the levels of plasma TG and HDL-C are typically inversely related. Plasma TG levels should be monitored when birth control pills or postmenopausal estrogen therapy is initiated to ensure that the increase in VLDL production does not lead to severe hypertriglyceridemia. Use of low-dose preparations of estrogen or the estrogen patch can minimize the effect of exogenous estrogen on lipids.

INHERITED CAUSES OF LOW LEVELS OF APoB-CONTAINING LIPOPROTEINS

Plasma concentrations of LDL-C <60 mg/dL are unusual. Although in some cases LDL-C levels in this range may be reflective of malnutrition or serious chronic illness, LDL-C <60 mg/dL in an otherwise healthy individual suggests an inherited condition. The major inherited causes of low LDL-C are reviewed here.

Abetalipoproteinemia

The synthesis and secretion of apoB-containing lipoproteins in the enterocytes of the proximal small bowel and in the hepatocytes of the liver involve a complex series of events that coordinate the coupling of various lipids with apoB-48 and apoB-100, respectively. Abetalipoproteinemia is a rare autosomal recessive disease caused by loss-of-function mutations in the gene encoding microsomal TG transfer protein (MTP; gene name *MTTP*), a protein that transfers lipids to nascent chylomicrons and VLDLs in the intestine and liver, respectively. Plasma levels of cholesterol and TG are extremely low in this disorder, and chylomicrons, VLDLs, LDLs, and apoB are undetectable in plasma. The parents of patients with abetalipoproteinemia (obligate heterozygotes) have normal plasma

TABLE 400-4 Secondary Causes of Dyslipidemia

LDL		HDL					
ELEVATED	REDUCED	ELEVATED	REDUCED	VLDL ELEVATED	IDL ELEVATED	CHYLOMICRONS ELEVATED	LP(a) ELEVATED
Hypothyroidism	Severe liver disease	Alcohol	Smoking	Obesity	Multiple myeloma	Autoimmune disease	Chronic kidney disease
Nephrotic syndrome	Malabsorption	Exercise	DM type 2	DM type 2	Monoclonal gammopathy	DM type 2	Nephrotic syndrome
Cholestasis	Malnutrition	Exposure to chlorinated hydrocarbons	Obesity	Glycogen storage disease	Autoimmune disease		Inflammation
Acute intermittent porphyria	Gaucher's disease	Drugs: estrogen	Malnutrition	Nephrotic syndrome	Hypothyroidism		Menopause
Anorexia nervosa	Chronic infectious disease		Gaucher's disease	Hepatitis			Orchidectomy
Hepatoma	Hyperthyroidism		Cholesteryl ester storage disease	Alcohol			Hypothyroidism
Drugs: thiazides, cyclosporin, carbamazepine	Drugs: niacin toxicity		Drugs: anabolic steroids, beta blockers	Renal failure			Acromegaly
				Sepsis			Drugs: growth hormone, isotretinoin
				Stress			
				Cushing's syndrome			
				Pregnancy			
				Acromegaly			
				Lipodystrophy			
				Drugs: estrogen, beta blockers, glucocorticoids, bile acid binding resins, retinoic acid			

Abbreviations: DM, diabetes mellitus; HDL, high-density lipoprotein; IDL, intermediate-density lipoprotein; LDL, low-density lipoprotein; Lp(a), lipoprotein A; VLDL, very-low-density lipoprotein.

lipid and apoB levels. Abetalipoproteinemia usually presents in early childhood with diarrhea and failure to thrive due to fat malabsorption. The initial neurologic manifestations are loss of deep tendon reflexes, followed by decreased distal lower extremity vibratory and proprioceptive sense, dysmetria, ataxia, and the development of a spastic gait, often by the third or fourth decade. Patients with abetalipoproteinemia also develop a progressive pigmented retinopathy presenting with decreased night and color vision, followed by reductions in daytime visual acuity and ultimately progressing to near-blindness. The presence of spinocerebellar degeneration and pigmented retinopathy in this disease has resulted in some patients with abetalipoproteinemia being misdiagnosed as having Friedreich's ataxia.

Most of the clinical manifestations of abetalipoproteinemia result from defects in the absorption and transport of fat-soluble vitamins. Vitamin E and retinyl esters are normally transported from enterocytes to the liver by chylomicrons, and vitamin E is dependent on VLDL for transport out of the liver and into the circulation. As a consequence of the inability of these patients to secrete apoB-containing particles, patients with abetalipoproteinemia are markedly deficient in vitamin E and are also mildly to moderately deficient in vitamins A and K. Patients with abetalipoproteinemia should be referred to specialized centers for confirmation of the diagnosis and appropriate therapy. Treatment consists of a low-fat, high-caloric, vitamin-enriched diet accompanied by large supplemental doses of vitamin E. It is imperative that treatment be initiated as soon as possible to prevent development of neurologic sequelae, which can progress even with appropriate therapy. New therapies for this serious disease are needed.

Familial Hypobetalipoproteinemia (FHBL) FHBL generally refers to a condition of low total cholesterol, LDL-C, and apoB due to mutations in the *APOB* gene. Most of the mutations causing FHBL result in a truncated apoB protein, resulting in impaired assembly and secretion of chylomicrons from enterocytes and VLDL from the liver. Mutations that result in VLDL particles containing a truncated apoB protein are cleared from the circulation at an accelerated rate, which also contributes to patients with this disorder having low levels of LDL-C and apoB. Individuals heterozygous for these mutations usually have LDL-C levels <60–80 mg/dL and also tend to have lower levels of plasma TG. Many FHBL patients have elevated levels of hepatic fat (due to reduced VLDL export) and sometimes have increased levels of liver transaminases, although it appears that these patients infrequently develop associated inflammation and fibrosis.

Mutations in both apoB alleles cause homozygous FHBL, an extremely rare disorder resembling abetalipoproteinemia with nearly undetectable LDL-C and apoB. The neurologic defects in this form of hypobetalipoproteinemia tend to be less severe than is typically seen in abetalipoproteinemia. Homozygous hypobetalipoproteinemia can be distinguished from abetalipoproteinemia by examining the inheritance pattern of the plasma LDL-C level. The levels of LDL-C and apoB are normal in the parents of patients with abetalipoproteinemia and low in those of patients with homozygous hypobetalipoproteinemia.

Familial Combined Hypolipidemia Nonsense mutations in both alleles of the gene Angiopoietin-like 3 (*ANGPTL3*) lead to low plasma levels of all three major lipid fractions—TG, LDL-C, and HDL-C, a phenotype termed familial combined hypolipidemia. ANGPTL3 is a protein synthesized by the liver and secreted into the bloodstream. It inhibits LPL, thus delaying clearance of TRLs from the blood and increasing TRL blood concentrations. Deficiency of ANGPTL3, therefore, raises LPL activity and predominantly lowers blood TG. ANGPTL3 deficiency is associated with a reduced risk for CHD. Therapies to antagonize ANGPTL3 are in development and initial human studies show that inhibition of ANGPTL3 by either an antisense oligonucleotide or a monoclonal antibodies lower blood levels of TG and LDL-C.

PCSK9 Deficiency Another inherited cause of low LDL-C results from loss-of-function mutations in *PCSK9*. PCSK9 is a secreted protein that binds to the extracellular domain of the LDL receptor in the liver and promotes the degradation of the receptor. Heterozygosity for nonsense mutations in PCSK9 that interfere with the synthesis of the protein are associated with increased hepatic LDL receptor activity and reduced plasma levels of LDL-C. Such mutations are more frequent in individuals of African descent. Individuals who are heterozygous for a loss-of-function mutation in PCSK9 have an ~30–40% reduction in plasma levels of LDL-C and have a substantial protection from CHD relative to those without a PCSK9 mutation, presumably due to having lower plasma cholesterol levels since birth. Homozygotes for these nonsense mutations have been reported and have extremely low LDL-C levels (<20 mg/dL) but appear otherwise healthy. A sequence variation of somewhat higher frequency (R46L) is found predominantly in individuals of European descent. This mutation impairs, but does not completely destroy, PCSK9 function. As a consequence, the plasma levels of LDL-C in individuals carrying this mutation are more modestly reduced (~15–20%); individuals with these mutations have a 45% reduction in CHD risk. The discovery of this condition led to the development of therapies that antagonize PCSK9, thus reducing LDL-C levels and risk of CHD. Two antibodies against PCSK9 are currently on the market **(Table 400-5)**.

DISORDERS OF REDUCED HDL CHOLESTEROL

Low levels of HDL-C are very commonly encountered in clinical practice. Low HDL-C is an important independent predictor of increased cardiovascular risk and has been used regularly in standardized risk calculators. However, it is doubtful that low HDL-C is directly causal for the development of ASCVD. HDL metabolism is strongly influenced by TG metabolism, insulin resistance, and inflammation, among other environmental and medical factors. Thus the HDL-C measurement integrates a number of cardiovascular risk factors, potentially explaining its strong inverse association with ASCVD.

The majority of patients with low HDL-C have some combination of genetic predisposition and secondary factors. Variants in dozens of genes have been shown to influence HDL-C levels. Even more important quantitatively, obesity and insulin resistance have strong suppressive effects on HDL-C, and low HDL-C in these conditions is widely observed. Furthermore, the vast majority of patients with elevated TGs have reduced levels of HDL-C. Most patients with low HDL-C who have been studied in detail have accelerated catabolism of HDL and its associated apoA-I as the physiologic basis for the low HDL-C. Importantly, although HDL-C remains an important biomarker for assessing cardiovascular risk, it is not currently a direct target of intervention for raising the level in order to reduce cardiovascular risk.

INHERITED CAUSES OF VERY LOW LEVELS OF HDL-C

Mutations in genes encoding proteins that play critical roles in HDL synthesis and catabolism can result in reductions in plasma levels of HDL-C. Unlike the genetic forms of hypercholesterolemia, which are invariably associated with premature coronary atherosclerosis, genetic forms of hypoalphalipoproteinemia (low HDL-C) are often not associated with clearly increased risk of ASCVD.

Gene Deletions in the *APOA5-A1-C3-A4* Locus and Coding Mutations in *APOA1* Complete genetic deficiency of apoA-I due to a complete deletion of the *APOA1* gene results in the virtual absence of circulating HDL and appears to increase the risk of premature ASCVD. The genes encoding *APOA5*, *APOA1*, *APOC3*, and *APOA4* are clustered together on chromosome 11. Some patients with no apoA-I have genomic deletions that include other genes in the cluster. ApoA-I is required for LCAT activity. In the absence of LCAT, free cholesterol levels increase in both plasma (not HDL) and in tissues. The free cholesterol can form deposits in the cornea and in the skin, resulting in corneal opacities and planar xanthomas. Premature CHD is associated with apoA-I deficiency.

Missense and nonsense mutations in the apoA-I gene are present in some patients with low plasma levels of HDL-C (usually 15–30 mg/dL),

TABLE 400-5 Summary of the Major Approved Drugs Used for the Treatment of Dyslipidemia

DRUG	MAJOR INDICATIONS	STARTING DOSE	MAXIMAL DOSE	MECHANISM	COMMON SIDE EFFECTS
HMG-CoA reductase inhibitors (statins)	Elevated LDL-C; increased CV risk			↓ Cholesterol synthesis, ↑ Hepatic LDL receptors, ↓ VLDL production	Myalgias and myopathy ↑ transaminases, ↑ diabetes risk
Lovastatin		20–40 mg daily	80 mg daily		
Pravastatin		40–80 mg daily	80 mg daily		
Simvastatin		20–40 mg daily	80 mg daily		
Fluvastatin		20–40 mg daily	80 mg daily		
Atorvastatin		20–40 mg daily	80 mg daily		
Rosuvastatin		5–20 mg daily	40 mg daily		
Pitavastatin		1–2 mg daily	4 mg daily		
Cholesterol absorption inhibitor	Elevated LDL-C			↓ Cholesterol absorption, ↑ LDL receptors	Elevated transaminases
Ezetimibe		10 mg daily	10 mg daily		
Bile acid sequestrants	Elevated LDL-C			↑ Bile acid excretion and ↑ LDL receptors	Bloating, constipation, elevated triglycerides
Cholestyramine		4 g daily	32 g daily		
Colestipol		5 g daily	40 g daily		
Colesevelam		3750 mg daily	4375 mg daily		
PCSK9 inhibitors	Elevated LDL-C			↓ PCSK9 activity, ↑ LDL receptors	Injection site reactions
Evolocumab		140 mg SQ q 2 weeks	420 mg SQ q 1 month (hoFH)		
Alirocumab		75 mg SQ q 2 weeks	150 mg SQ q 2 weeks		
MTP inhibitor	HoFH			↓ VLDL production	Nausea, diarrhea, increased hepatic fat
Lomitapide		5 mg daily	60 mg daily		
ApoB inhibitor	HoFH			↓ VLDL production	Injection site reactions, flu-like symptoms, increased hepatic fat
Mipomersen		200 mg SC weekly	200 mg SC weekly		
Fibric acid derivatives	Elevated TG			↑ LPL, ↓ VLDL synthesis	Dyspepsia, myalgia, gallstones, elevated transaminases
Gemfibrozil		600 mg bid	600 mg bid		
Fenofibrate		145 mg qd	145 mg qd		
Omega-3 fatty acids	Elevated TG	4 g daily	4 g daily	↑ TG catabolism	Dyspepsia, fishy odor to breath
Omega-3 acid ethyl esters					
Icosapent ethyl		4 g daily	4 g daily		

Abbreviations: GI, gastrointestinal; HDL-C, high-density lipoprotein cholesterol; HoFH, homozygous familial hypercholesterolemia; LDL, low-density lipoprotein; LDL-C, LDL-cholesterol; LPL, lipoprotein lipase; TG, triglyceride; VLDL, very-low-density lipoprotein.

but are a rare cause of low plasma HDL-C levels. Most individuals with low plasma HDL-C levels due to missense mutations in apoA-I do not appear to have premature CHD. Patients who are heterozygous for an Arg173Cys substitution in apoA-I (so-called apoA-I_{Milano}) have very low plasma levels of HDL-C due to impaired LCAT activation and accelerated clearance of the HDL particles containing the abnormal apoA-I. Despite having very low plasma levels of HDL-C, these individuals do not have an increased risk of premature CHD. A few selected missense mutations in apoA-I and apoA-II promote the formation of amyloid fibrils, which can cause systemic amyloidosis.

Tangier Disease (*ABCA1* Deficiency) Tangier disease is a rare autosomal co-dominant form of extremely low plasma HDL-C levels that is caused by mutations in the gene encoding ABCA1, a cellular transporter that facilitates efflux of unesterified cholesterol and phospholipids from cells to apoA-I (Fig. 400-3). ABCA1 in the liver and intestine rapidly lipidates the apoA-I secreted from the basolateral membranes of these tissues. In the absence of ABCA1, the nascent, poorly lipidated apoA-I is immediately cleared from the circulation. Thus, patients with Tangier disease have extremely low circulating plasma levels of HDL-C (<5 mg/dL) and apoA-I (<5 mg/dL). Cholesterol accumulates in the reticuloendothelial system of these patients, resulting in hepatosplenomegaly and pathognomonic enlarged, grayish yellow or orange tonsils. An intermittent peripheral neuropathy (mononeuritis multiplex) or a sphingomyelia-like neurologic disorder can also be seen in this disorder. Tangier disease is probably associated with some increased risk of premature atherosclerotic disease, although the association is not as robust as might be anticipated, given the very low levels of HDL-C and apoA-I in these patients. Patients with Tangier disease also have low plasma levels of LDL-C, which may attenuate the atherosclerotic risk. Obligate heterozygotes for ABCA1 mutations have moderately reduced plasma HDL-C levels (15–30 mg/dL), and their risk of premature CHD remains uncertain.

Familial LCAT Deficiency This rare autosomal recessive disorder is caused by mutations in LCAT, an enzyme synthesized in the liver and secreted into the plasma, where it circulates associated with lipoproteins (Fig. 400-3). As reviewed above, the enzyme is activated by apoA-I and mediates the esterification of cholesterol to form cholesteryl esters. Consequently, in familial LCAT deficiency, the proportion of free cholesterol in circulating lipoproteins is greatly increased (from

~25% to >70% of total plasma cholesterol). Deficiency in this enzyme interferes with the maturation of HDL particles and results in rapid catabolism of circulating apoA-I.

Two genetic forms of familial LCAT deficiency have been described in humans: complete deficiency (also called *classic LCAT deficiency*) and partial deficiency (also called *fish-eye disease*). Progressive corneal opacification due to the deposition of free cholesterol in the cornea, very low plasma levels of HDL-C (usually <10 mg/dL), and variable hypertriglyceridemia are characteristic of both disorders. In partial LCAT deficiency, there are no other known clinical sequelae. In contrast, patients with complete LCAT deficiency have hemolytic anemia and progressive renal insufficiency that eventually leads to end-stage renal disease. Remarkably, despite the extremely low plasma levels of HDL-C and apoA-I, premature ASCVD is not a consistent feature of either LCAT deficiency or fish eye disease. The diagnosis can be confirmed in a specialized laboratory by assaying plasma LCAT activity or by sequencing the *LCAT* gene.

Primary Hypoalphalipoproteinemia The condition of low plasma levels of HDL-C (the "alpha lipoprotein") is referred to as *hypoalphalipoproteinemia*. Primary hypoalphalipoproteinemia is defined as a plasma HDL-C level below the tenth percentile in the setting of relatively normal cholesterol and TG levels, no apparent secondary causes of low plasma HDL-C, and no clinical signs of LCAT deficiency or Tangier disease. This syndrome is often referred to as *isolated low HDL*. A family history of low HDL-C facilitates the diagnosis of an inherited condition, which may follow an autosomal dominant pattern. The metabolic etiology of this disease appears to be primarily accelerated catabolism of HDL and its apolipoproteins. Some of these patients may have ABCA1 mutations and therefore have heterozygous Tangier disease. Several kindreds with primary hypoalphalipoproteinemia and an increased incidence of premature CHD have been described, although it is not clear if the low HDL-C level is the cause of the accelerated atherosclerosis in these families. Association of hypoalphalipoproteinemia with premature CHD may depend on the specific nature of the gene defect or the underlying metabolic defect that either directly or indirectly causes the low plasma HDL-C level.

SCREENING, DIAGNOSIS, AND MANAGEMENT OF DISORDERS OF LIPOPROTEIN METABOLISM

SCREENING

Hypercholesterolemia is a cause of premature CHD that is highly treatable and therefore persons should be actively screened. Plasma lipid levels should be measured, preferably after a 12-h overnight fast, in all adults; guidelines suggest that screening of children between 9 and 11 years of age is also recommended. In most clinical laboratories, the total cholesterol and TGs in the plasma are measured enzymatically, and then the cholesterol in the supernatant is measured after precipitation of apoB-containing lipoproteins to determine the HDL-C. The LDL-C is then estimated using the following equation (the Friedewald formula):

$$\text{LDL-C} = \text{total cholesterol} - (\text{TG}/5) - \text{HDL-C}$$

(The VLDL cholesterol content is estimated by dividing the plasma TG by 5, reflecting the ratio of TG to cholesterol in VLDL particles.) This formula is reasonably accurate if test results are obtained on fasting plasma and if the TG level does not exceed ~200 mg/dL; by convention it cannot be used if the TG level is >400 mg/dL. LDL-C can be directly measured by a number of methods. Further evaluation and treatment are based primarily on the clinical assessment of absolute cardiovascular risk using risk calculators such as the AHA/ACC risk calculator based on a large amount of observational data.

DIAGNOSIS

A critical first step in managing a lipoprotein disorder is to attempt to determine the class or classes of lipoproteins that are increased or decreased in the patient. Once the dyslipidemia is accurately classified, efforts should be directed to rule out any possible secondary causes (Table 400-4). A careful social, medical, and family history should be obtained. A fasting glucose should be obtained in the initial workup of all subjects with an elevated TG level. Nephrotic syndrome and chronic renal insufficiency should be excluded by obtaining urine protein and serum creatinine. Liver function tests should be performed to rule out hepatitis and cholestasis. Hypothyroidism should be ruled out by measuring serum thyroid-stimulating hormone.

Once secondary causes have been ruled out, attempts should be made to diagnose a primary lipid disorder, because the underlying genetic defect can provide important prognostic information regarding the risk of developing CHD, the response to drug therapy, and the management of other family members. Obtaining the correct diagnosis often requires a detailed family medical history, lipid analyses in family members, and sometimes specialized testing.

Severe Hypertriglyceridemia If the fasting plasma TG level is >750 mg/dL, the patient may have chylomicronemia. If the elevated TG levels are persistent and the total cholesterol-to-TG ratio is >8, particularly in the setting of a history of pancreatitis, familial chylomicronemia syndrome should be considered. While LPL activity measured in postheparin plasma can support diagnosis, genetic testing for FCS genes may be indicated. Most individuals with persistent severe hypertriglyceridemia do not have a single gene disorder (FCS) but instead are genetically predisposed and have secondary factors (diet, obesity, glucose intolerance, alcohol ingestion, estrogen therapy) that contribute to the hyperlipidemia. Such patients are still at risk for acute pancreatitis and should be treated to reduce their TG levels and thus their risk of pancreatitis.

Severe Hypercholesterolemia If the levels of LDL-C are very high (greater than a ninety-fifth percentile for age and sex) in absence of secondary causes, familial hypercholesterolemia should be considered, particularly if there is a family history of hypercholesterolemia and/or premature CHD. While FH is a clinical diagnosis, genetic sequencing is now widely available as may be considered in order to determine the molecular diagnosis; a finding of a causal mutation may appropriately result in earlier and more aggressive therapy to lower LDL-C and could also promote family-based cascade screening. Recessive forms of severe hypercholesterolemia are rare, but if a patient with severe hypercholesterolemia has parents with normal cholesterol levels, ARH, sitosterolemia, and LALD should be considered. Patients with more moderate hypercholesterolemia that does not segregate in families as a monogenic trait are likely to have polygenic hypercholesterolemia.

Combined Hyperlipidemia Elevations in the plasma levels of both cholesterol and TGs are seen in patients with increased plasma levels of both VLDL and LDL or of remnant lipoproteins. A β-quantification to determine the VLDL cholesterol/TG ratio in plasma (see discussion of FDBL), an NMR lipoprotein profile, or a direct measurement of the plasma LDL-C should be performed at least once prior to initiation of lipid-lowering therapy to determine if the hyperlipidemia is due to the accumulation of remnants or to an increase in both LDL and VLDL. Measurement of plasma apoB levels can help identify patients with FCHL who may require more aggressive treatment.

Given the prevalence of primary and secondary dyslipidemias and the clinical benefits of early diagnosis and initiation of therapy, it is essential that physicians screen lipids systematically, rule out secondary causes of dyslipidemia, suspect inherited disorders of lipoprotein metabolism where appropriate, actively promote family-based cascade screening, and be knowledgeable about the existing therapeutic options, including PCSK9 inhibitors. The field of "clinical lipidology" has matured and is moving toward a more systematic clinical application of genomic medicine. Diagnostic DNA sequencing or genotyping in patients with suspected FH, FCS, and FDBL has the potential to enhance molecular diagnosis, facilitate appropriate therapeutic interventions, and promote family-based cascade screening.

APPROACH TO THE PATIENT

Lipoprotein Disorders

The major goals in the clinical management of lipoprotein disorders are: (1) prevention of acute pancreatitis in patients with severe hypertriglyceridemia; and (2) prevention of CVD and related cardiovascular events.

MANAGEMENT OF SEVERE HYPERTRIGLYCERIDEMIA TO PREVENT PANCREATITIS

Although the observational relationship between severe hypertriglyceridemia, particularly chylomicronemia, and acute pancreatitis is well-established, there has never been a clinical trial designed or powered to prove that intervention to reduce TGs reduces the risk of pancreatitis. Nevertheless, it is generally considered appropriate medical practice to intervene in patients with TGs >500 mg/dL in order to reduce the risk of pancreatitis. It remains controversial whether individuals with severe hypertriglyceridemia are at increased risk for ASCVD.

Lifestyle Modifying the lifestyle of the patient with severe hypertriglyceridemia often is associated with a significant reduction in plasma TG level. Patients who drink alcohol should be encouraged to decrease or preferably eliminate their intake. Patients with severe hypertriglyceridemia often benefit from a formal dietary consultation with a dietician intimately familiar with counseling patients on the dietary management of high TGs. Dietary fat intake should be restricted to reduce the formation of chylomicrons in the intestine. The excessive intake of simple carbohydrates should be discouraged because insulin drives TG production in the liver. Aerobic exercise and even increase in regular physical activity can have a positive effect in reducing TG levels and should be strongly encouraged. For patients who are overweight, weight loss can help to reduce TG levels. In extreme cases, bariatric surgery has been shown to not only produce effective weight loss but also substantially reduce plasma TG levels.

Pharmacologic Therapy for Severe Hypertriglyceridemia Despite the above interventions, however, many patients with severe hypertriglyceridemia require pharmacologic therapy (Table 400-5). Patients who persist in having fasting TG >500 mg/dL despite active lifestyle management are candidates for pharmacologic therapy. The two major classes of drugs used for management of these patients are fibrates and omega-3 fatty acids (fish oils). In addition, statins can reduce plasma TG levels and also reduce CVD risk, and should be used in patients with severe hypertriglyceridemia who are at increased risk of CVD.

Fibrates Fibric acid derivatives, or fibrates, are agonists of PPARα, a nuclear receptor involved in the regulation of lipid metabolism. Fibrates stimulate LPL activity (enhancing TG hydrolysis), reduce apoC-III synthesis (enhancing lipoprotein remnant clearance), promote β-oxidation of fatty acids, and may reduce VLDL TG production. This class of therapeutic agents sometimes lowers but more often raises the plasma level of LDL-C in individuals with severe hypertriglyceridemia. Fibrates are generally well tolerated, but are associated with an increase in the incidence of gallstones. Fibrates can cause myopathy, especially when combined with other lipid-lowering therapy (statins, niacin), and can raise creatinine. Fibrates should be used with caution in patients with CKD. Importantly, fibrates can potentiate the effect of warfarin and certain oral hypoglycemic agents, so the anticoagulation status and plasma glucose levels should be closely monitored in patients on these agents.

Omega 3 Fatty Acids (Fish Oils) Omega-3 fatty acids, or omega-3 polyunsaturated fatty acids (n-3 PUFAs), commonly known as fish oils, are present in high concentration in fish and in flaxseed. The most widely used n-3 PUFAs for the treatment of hyperlipidemias are the two active molecules in fish oil: eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA). n-3 PUFAs have been concentrated into tablets and in doses of 3–4 g/d are effective at lowering fasting TG levels. Fish oils can cause an increase in plasma LDL-C levels in some patients. In general, fish oils are well tolerated, with the major side effect being dyspepsia. They appear to be safe, at least at doses up to 3–4 g, but can be associated with a prolongation in the bleeding time.

MANAGEMENT OF CHOLESTEROL TO PREVENT CARDIOVASCULAR DISEASE

In contrast to hypertriglyceridemia and pancreatitis, there are abundant and compelling data that intervention to reduce LDL-C substantially reduces the risk of CVD, including myocardial infarction and stroke, as well as total mortality. Thus, it is imperative that patients with hypercholesterolemia be assessed for cardiovascular risk and for the need for intervention. It is also worth noting that patients at high risk for CVD who have plasma LDL-C levels in the "normal" or average range also benefit from intervention to reduce LDL-C levels.

Lifestyle The first approach to a patient with hypercholesterolemia and high cardiovascular risk is to make any necessary lifestyle changes. In obese patients, efforts should be made to reduce body weight to the ideal level. Patients should receive dietary counseling to reduce the content of saturated fats, *trans* fats, and cholesterol in the diet. Regular aerobic exercise has relatively little impact on reducing plasma LDL-C levels, although it has cardiovascular benefits independent of LDL lowering.

Pharmacologic Therapy for Hypercholesterolemia The decision to use LDL-lowering drug therapy (Table 400-5)—with a statin being first-line therapy—depends on the level of LDL-C as well as the level of cardiovascular risk. In general, patients with a Mendelian disorder of elevated LDL-C such as FH must be treated to reduce the very high lifetime risk of CVD, and treatment should be initiated as early as possible in adulthood or, in some cases, during childhood.

Otherwise, the decision to initiate LDL-lowering drug therapy is generally determined by the level of cardiovascular risk. In patients with established CVD, statin therapy is well supported by clinical trial data and should be used regardless of the LDL-C level. For patients >40 years old without clinical CVD, the AHA/ACC risk calculator *(http://my.americanheart.org/professional/StatementsGuidelines/PreventionGuidelines/Prevention-Guidelines_UCM_457698_SubHomePage.jsp)* can be used to determine the 10-year absolute risk for CVD, and current guidelines suggest that a 10-year risk >7.5% merits consideration of statin therapy regardless of plasma LDL-C level. For younger patients, the assessment of lifetime risk of CVD may help inform the decision to start a statin.

HMG-CoA Reductase Inhibitors (Statins) Statins inhibit HMG-CoA reductase, a key enzyme in cholesterol biosynthesis. By inhibiting cholesterol biosynthesis, statins lead to increased hepatic LDL receptor activity and accelerated clearance of circulating LDL, resulting in a dose-dependent reduction in plasma levels of LDL-C. The magnitude of LDL lowering associated with statin treatment varies widely among individuals, but once a patient is on a statin, the doubling of the statin dose produces an ~6% further reduction in the level of plasma LDL-C. The statins currently available differ in their LDL-C–reducing potency (Table 400-5). Current recommendations are to use "high-intensity" statin therapy in most patients deemed at high risk of CVD. Currently, there is no convincing evidence that any of the different statins confer an advantage that is independent of the effect on LDL-C. Statins also reduce plasma TGs in a dose-dependent fashion, which is roughly proportional to their LDL-C–lowering effects (if the TGs are <400 mg/dL). Statins have a modest HDL-raising effect (5–10%) that is not generally dose-dependent.

Statins are well tolerated and can be taken in tablet form once a day. Potential side effects include dyspepsia, headaches, fatigue, and muscle or joint pains. Severe myopathy and even rhabdomyolysis

occur rarely with statin treatment. The risk of statin-associated myopathy is increased by the presence of older age, frailty, renal insufficiency, and coadministration of drugs that interfere with the metabolism of statins, such as erythromycin and related antibiotics, antifungal agents, immunosuppressive drugs, and fibric acid derivatives (particularly gemfibrozil). Severe myopathy can usually be avoided by careful patient selection, avoidance of interacting drugs, and instructing the patient to contact the physician immediately in the event of unexplained muscle pain. In the event of muscle symptoms, the plasma creatine kinase (CK) level should be obtained to differentiate myopathy from myalgia. Serum CK levels need not be monitored on a routine basis in patients taking statins, because an elevated CK in the absence of symptoms does not predict the development of myopathy and does not necessarily suggest the need for discontinuing the drug.

Another consequence of statin therapy can be elevation in liver transaminases (alanine aminotransferase [ALT] and aspartate aminotransferase [AST]). They should be checked before starting therapy, at 2–3 months, and then annually. Substantial (greater than three times the upper limit of normal) elevation in transaminases is relatively rare, and mild-to-moderate (one to three times normal) elevation in transaminases in the absence of symptoms need not mandate discontinuing the medication. Severe clinical hepatitis associated with statins is exceedingly rare, and the trend is toward less frequent monitoring of transaminases in patients taking statins. The statin-associated elevation in liver enzymes resolves upon discontinuation of the medication.

Statins appear to be remarkably safe. Meta-analyses of large randomized controlled clinical trials with statins do not suggest an increase in any major noncardiac diseases except type 2 diabetes. A small excess percentage of those taking statins will develop diabetes but the benefits associated with the reduction in cardiovascular events outweigh the increase in incidence of diabetes. Statins are the drug class of choice for LDL-C reduction and are by far the most widely used class of lipid-lowering drugs.

Cholesterol Absorption Inhibitors Cholesterol within the lumen of the small intestine is derived from the diet (about one-third) and the bile (about two-thirds) and is actively absorbed by the enterocyte through a process that involves the protein NPC1L1. Ezetimibe (Table 400-5) is a cholesterol absorption inhibitor that binds directly to and inhibits NPC1L1 and blocks the intestinal absorption of cholesterol. Ezetimibe (10 mg) inhibits cholesterol absorption by almost 60%, resulting in a reduction in delivery of dietary sterols in the liver and an increase in hepatic LDL receptor expression. The mean reduction in plasma LDL-C on ezetimibe (10 mg) is 18%, and the effect is additive when used in combination with a statin. Effects on TG and HDL-C levels are negligible. When used in combination with a statin, monitoring of liver transaminases is recommended. The only roles for ezetimibe in monotherapy are in patients who do not tolerate statins and in sitosterolemia.

Bile Acid Sequestrants (Resins) Bile acid sequestrants bind bile acids in the intestine and promote their excretion rather than reabsorption in the ileum. To maintain the bile acid pool size, the liver diverts cholesterol to bile acid synthesis. The decreased hepatic intracellular cholesterol content results in upregulation of the LDL receptor and enhanced LDL clearance from the plasma. Bile acid sequestrants, including cholestyramine, colestipol, and colesevelam (Table 400-5), primarily reduce plasma LDL-C levels but can cause an increase in plasma TGs. Therefore, patients with hypertriglyceridemia generally should not be treated with bile acid–binding resins. Cholestyramine and colestipol are insoluble resins that must be suspended in liquids. Colesevelam is available as tablets but generally requires up to six to seven tablets per day for effective LDL-C lowering. Most side effects of resins are limited to the gastrointestinal tract and include bloating and constipation. Because bile acid sequestrants are not systemically absorbed, they are very safe and the cholesterol-lowering drug of choice in children and in women of childbearing age who are lactating, pregnant, or could become pregnant. They are effective in combination with statins and in combination with ezetimibe and are particularly useful with one or both of these drugs for patients with severe hypercholesterolemia or those with statin intolerance.

PCSK9 Inhibitors PCSK9 inhibitors are antibodies that bind to circulating PCSK9 and prevent its interaction with the LDL receptor. This permits more LDL receptors to recycle back to the cell surface and functionally increases the number of LDL receptors available to remove LDL from the blood. They are highly effective in lowering LDL-C, with a mean 50–60% reduction in LDL-C. They also reduce plasma levels of Lp(a) modestly. PCSK9 inhibition has been proven to reduce cardiovascular events in patients with existing CHD. These antibodies are administered subcutaneously every 2–4 weeks. They are generally well-tolerated, with the major side effect being injection site reactions. They are indicated as second line (after statin) or third line (after statin + ezetimibe) therapy in patients with FH or CHD in whom LDL-C is not reduced to acceptable levels with statin (+/- ezetimibe) alone.

Specialized Drugs For Homozygous FH Two "orphan" drugs are approved specifically for the management of homozygous FH. They include a small-molecule inhibitor of MTP, called lomitapide, and an antisense oligonucleotide against apoB, called mipomersen. These drugs reduce VLDL production and LDL-C levels in homozygous FH patients. Due to their mechanism of action, each drug causes an increase in hepatic fat, the long-term consequences of which are unknown. In addition, lomitapide is associated with gastrointestinal-related side effects, and mipomersen is associated with skin reactions and flu-like symptoms. One of these drugs should be considered in hoFH patients after a trial of a statin plus PCSK9 inhibitor is shown to be insufficient to reduce LDL-C levels.

LDL Apheresis Patients who cannot reduce their LDL-C to acceptable levels despite optimally tolerated drug therapy are candidates for LDL apheresis. In this process, the patient's plasma is passed over a column that selectively removes the LDL, and the LDL-depleted plasma is returned to the patient. LDL apheresis is indicated for patients on maximally tolerated combination drug therapy (including a PCSK9 inhibitor) who have CHD and a plasma LDL-C level >200 mg/dL or no CHD and a plasma LDL-C level >300 mg/dL; LDL apheresis could be considered in high-risk patients who have an LDL-C > 160 mg/dL on maximal therapy.

FURTHER READING

Abul-Husn NS et al: Genetic identification of familial hypercholesterolemia within a single U.S. health care system. Science 354:6319, 2016.

Cholesterol Treatment Trialists' (CTT) Collaboration et al: Efficacy and safety of LDL-lowering therapy among men and women: Meta-analysis of individual data from 174,000 participants in 27 randomised trials. Lancet 385:1397, 2015.

Dron JS, Hegele RA: Genetics of triglycerides and the risk of atherosclerosis. Curr Atheroscler Rep 19:31, 2017.

Garg A (ed): *Dyslipidemias: Pathophysiology, Evaluation and Management*. Humana Press, 2015.

Gidding SS et al: The agenda for familial hypercholesterolemia: A scientific statement from the American Heart Association. Circulation 132:2167, 2015.

Khavandi M et al: Treatment of dyslipidemias to prevent cardiovascular disease in patients with type 2 diabetes. Curr Cardiol Rep 19:7, 2017.

Khera AV et al: Association of rare and common variation in the lipoprotein lipase gene with coronary artery disease. JAMA 317:937, 2017.

Rader DJ: New Therapeutic approaches to the treatment of dyslipidemia. Cell Metab 23:405, 2016.

Sabatine MS et al: Evolocumab and clinical outcomes in patients with cardiovascular disease. N Engl J Med 376:1713, 2017.

401 The Metabolic Syndrome

Robert H. Eckel

The metabolic syndrome (syndrome X, insulin resistance syndrome) consists of a constellation of metabolic abnormalities that confer increased risk of cardiovascular disease (CVD) and diabetes mellitus. Evolution of the criteria for the metabolic syndrome since the original definition by the World Health Organization in 1998 reflects growing clinical evidence and analysis by a variety of consensus conferences and professional organizations. The major features of the metabolic syndrome include central obesity, hypertriglyceridemia, low levels of high-density lipoprotein (HDL) cholesterol, hyperglycemia, and hypertension **(Table 401-1)**.

GLOBAL HEALTH/EPIDEMIOLOGY

The most challenging feature of the metabolic syndrome to define is waist circumference. Intraabdominal circumference (visceral adipose tissue) is the most strongly related to insulin resistance and risk of diabetes and CVD, and for any given waist circumference the distribution of adipose tissue between subcutaneous (SC) and visceral depots varies substantially. Thus, within and between populations, there is a lesser vs greater risk at the same waist circumference. These differences in populations reflect the range of waist circumferences considered to confer risk in different geographic locations (Table 401-1).

The prevalence of the metabolic syndrome varies around the world, in part reflecting the age and ethnicity of the populations studied and the diagnostic criteria applied. In general, the prevalence of the metabolic syndrome increases with age. The highest recorded prevalence worldwide is among Native Americans, with an age-adjusted 53% of women and 45% of men meeting the criteria of the National Cholesterol Education Program and Adult Treatment Panel III (NCEP:ATPIII). Greater global industrialization is associated with rising rates of obesity, and expected increase in the prevalence of the metabolic syndrome, especially as the population ages. Moreover, the rising prevalence and severity of obesity among children reflects features of the metabolic syndrome in a younger population, now estimated to be up to 23% and >60% amongst obese and overweight children.

In 2012, the overall prevalence of the metabolic syndrome in the United States was 33% with a higher prevalence in women than men (36% vs 30%, respectively). When stratified by race/ethnicity, the highest prevalence of the metabolic syndrome was 35% in Hispanics followed by 33% in non-Hispanic Caucasians and blacks. From 2003–2004 to 2011–2012, overall prevalence of the metabolic syndrome increased from 33% in 2003–2004 to 35% in 2011–2012. In France, studies of a cohort of >18-year-old adults revealed a gradual increase from <10% prevalence for each sex to 19% of men and ~10% for women. However, greater compliance with the French Nutrition and Health Program-Guideline Score was associated with a lower prevalence in both genders. A similar heart healthy nutritional impact on the prevalence of metabolic syndrome is now apparent in the United States.

The frequency distribution of the five components of the syndrome for the U.S. population (NHANES III) is summarized in **Fig. 401-1**. Increases in waist circumference predominate among women, whereas increases in fasting plasma triglyceride levels (i.e., >150 mg/dL), reductions in HDL cholesterol levels, and hyperglycemia are more likely in men.

RISK FACTORS

Overweight/Obesity The metabolic syndrome was first described in the early twentieth century; however, the worldwide overweight/obesity epidemic has recently been the force driving its increasing recognition. Central adiposity is a key feature of the syndrome, and the syndrome's prevalence reflects the strong relationship between waist circumference and increasing adiposity. However, despite the importance of obesity, patients who are of normal weight may also be insulin-resistant and may have the metabolic syndrome. This phenotype is particularly evident for populations in India, Southeast Asia, and Central America.

Sedentary Lifestyle Physical inactivity and less cardiorespiratory fitness are a predictor of CVD events and the related risk of death. Many components of the metabolic syndrome are associated with a sedentary lifestyle, including increased adipose tissue (predominantly central), reduced HDL cholesterol, and increased triglycerides, blood pressure, and glucose in genetically susceptible persons. Compared with individuals who watch television or videos or use the computer <1 h daily, those who do so for >4 h daily have a twofold increased risk of the metabolic syndrome.

Genetics No single gene explains the complex phenotype called the metabolic syndrome. However, using genome wide association and candidate gene approaches, a number of genetic variants are associated with the metabolic syndrome. Although many of the loci have unknown function, many others relate to body weight and composition, insulin resistance, and lipid and lipoprotein metabolism.

Aging The metabolic syndrome affects nearly 50% of the U.S. population aged >60, and at >60 years of age women are more often affected. The age dependency of the syndrome's prevalence is seen in most populations around the world.

Diabetes Mellitus Diabetes mellitus can be included in both the NCEP and the harmonizing definitions of the metabolic syndrome, but the greatest value of the metabolic syndrome, and especially fasting glucose, is predicting type 2 diabetes. The great majority (~75%) of patients with type 2 diabetes or impaired glucose tolerance have the metabolic

TABLE 401-1 NCEP:ATPIII[a] 2001 and Harmonizing Definition Criteria for the Metabolic Syndrome

<table>
<tr><th>NCEP:ATPIII 2001</th><th colspan="3">HARMONIZING DEFINITION[b]</th></tr>
<tr><td rowspan="6">Three or more of the following:
• Central obesity: waist circumference >102 cm (M), >88 cm (F)
• Hypertriglyceridemia: triglyceride level ≥150 mg/dL or specific medication
• Low HDL[c] cholesterol: <40 mg/dL and <50 mg/dL for men and women, respectively, or specific medication
• Hypertension: blood pressure ≥130 mmHg systolic or ≥85 mmHg diastolic or specific medication
• Fasting plasma glucose level ≥100 mg/dL or specific medication or previously diagnosed type 2 diabetes</td><td colspan="3">Three of the following:
Waist circumference (cm)</td></tr>
<tr><td>Men</td><td>Women</td><td>Ethnicity</td></tr>
<tr><td>≥94</td><td>≥80</td><td>Europid, sub-Saharan African, Eastern and Middle Eastern</td></tr>
<tr><td>≥90</td><td>≥80</td><td>South Asian, Chinese, and ethnic South and Central American</td></tr>
<tr><td>≥85</td><td>≥90</td><td>Japanese</td></tr>
<tr><td colspan="3">• Fasting triglyceride level >150 mg/dL or specific medication
• HDL cholesterol level <40 mg/dL and <50 mg/dL for men and women, respectively, or specific medication
• Blood pressure >130 mm systolic or >85 mm diastolic or previous diagnosis or specific medication
• Fasting plasma glucose level ≥100 mg/dL (alternative indication: drug treatment of elevated glucose levels)</td></tr>
</table>

[a]National Cholesterol Education Program and Adult Treatment Panel III. [b]In this analysis, the following thresholds for waist circumference were used: white men, ≥94 cm; African-American men, ≥94 cm; Mexican-American men, ≥90 cm; white women, ≥80 cm; African-American women, ≥80 cm; Mexican-American women, ≥80 cm. For participants whose designation was "other race—including multiracial," thresholds that were once based on Europid cutoffs (≥94 cm for men and ≥80 cm for women) and on South Asian cutoffs (≥90 cm for men and ≥80 cm for women) were used. For participants who were considered "other Hispanic," the International Diabetes Federation thresholds for ethnic South and Central Americans were used. [c]High-density lipoprotein.

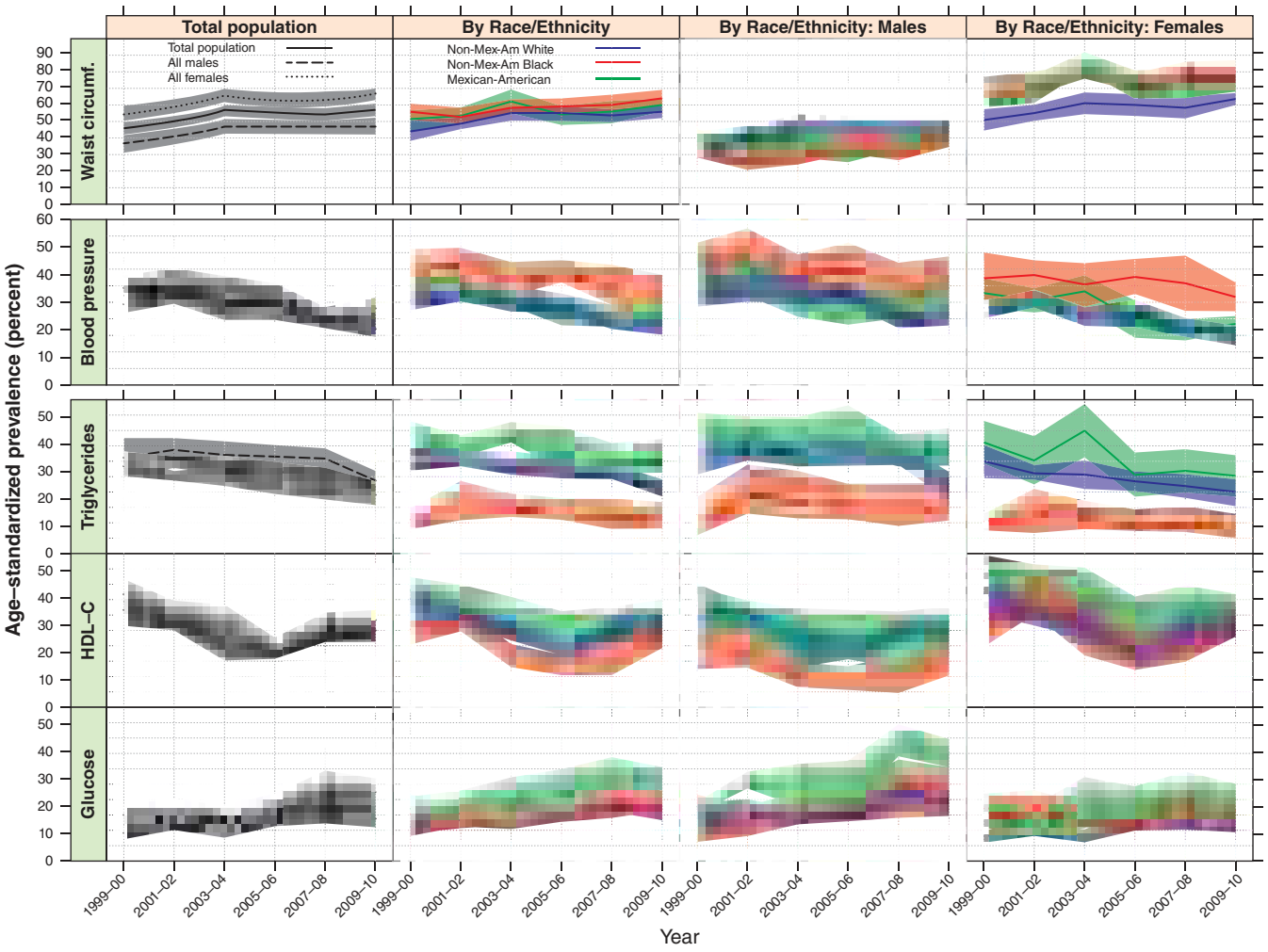


FIGURE 401-1 Prevalence and trends of the five components of metabolic syndrome in the adult US population, 1999–2010. Prevalence and trends of the components of Metabolic Syndrome for US adults aged 20 or older for the total population by sex (first column), by race (second column), and by race and sex (third and fourth columns). Shaded areas represent 95% confidence intervals. HDL-C, high-density lipoprotein cholesterol; Non-Mex-Am, Non-Mexican American; Waist circumf, waist circumference. *(From H Beltrán-Sánchez et al: Prevalence and trends of metabolic syndrome in the adult U.S. population, 1999-2010. J Am Coll Cardiol 62:697, 2013.)*

syndrome. The presence of the metabolic syndrome in these populations relates to a higher prevalence of CVD than in patients who have type 2 diabetes or impaired glucose tolerance but do not have the syndrome.

Cardiovascular Disease Individuals with the metabolic syndrome are twice as likely to die of CVD as those who do not, and their risk of an acute myocardial infarction or stroke is threefold higher. The approximate prevalence of the metabolic syndrome among patients with coronary heart disease (CHD) is 60%, with a prevalence of ~35% among patients with premature coronary artery disease (≤age 45) and a particularly high prevalence among women. With appropriate cardiac rehabilitation and changes in lifestyle (e.g., nutrition, physical activity, weight reduction, and—in some cases—pharmacologic therapy), the prevalence of the syndrome can be reduced.

Lipodystrophy Lipodystrophic disorders in general are associated with the metabolic syndrome. Moreover, it is quite common for such patients to present with the metabolic syndrome. Both genetic lipodystrophy (e.g., Berardinelli-Seip congenital lipodystrophy, Dunnigan familial partial lipodystrophy) and acquired lipodystrophy (e.g., HIV-related lipodystrophy and in HIV patients receiving certain antiretroviral therapies) may give rise to severe insulin resistance and many of the components of the metabolic syndrome.

ETIOLOGY

Insulin Resistance The most accepted and unifying hypothesis to describe the pathophysiology of the metabolic syndrome is insulin resistance, caused systemically by an incompletely understood defect in insulin action **(Chap. 396)**. The onset of insulin resistance is heralded by postprandial hyperinsulinemia, which is followed by fasting hyperinsulinemia and ultimately by hyperglycemia.

An early major contributor to the development of insulin resistance is an overabundance of circulating fatty acids **(Fig. 401-2)**. Plasma albumin-bound free fatty acids are derived predominantly from adipose-tissue triglyceride stores released by intracellular lipolytic enzymes. The lipolysis of triglyceride-rich lipoproteins in tissues by lipoprotein lipase also produces free fatty acids. Insulin mediates both anti-lipolysis and the stimulation of lipoprotein lipase in adipose tissue. Of note, the inhibition of lipolysis in adipose tissue is the most sensitive pathway of insulin action. Thus, when insulin resistance develops, increased lipolysis produces more fatty acids, which further decrease the anti-lipolytic effect of insulin. Excessive fatty acids enhance substrate availability and create insulin resistance by modifying downstream signaling. Fatty acids impair insulin-mediated glucose uptake and accumulate as triglycerides in both skeletal and cardiac muscle, whereas increased fatty acid flux increases glucose production and triglyceride production and accumulation in the liver.

Leptin resistance also may be a pathophysiologic mechanism to explain the metabolic syndrome. Physiologically, leptin reduces appetite, promotes energy expenditure, and enhances insulin sensitivity. In addition, leptin may regulate cardiac and vascular function through a nitric oxide–dependent mechanism. However, when obesity develops, hyperleptinemia ensues, with evidence of leptin resistance in the brain and other tissues resulting in inflammation, insulin resistance,

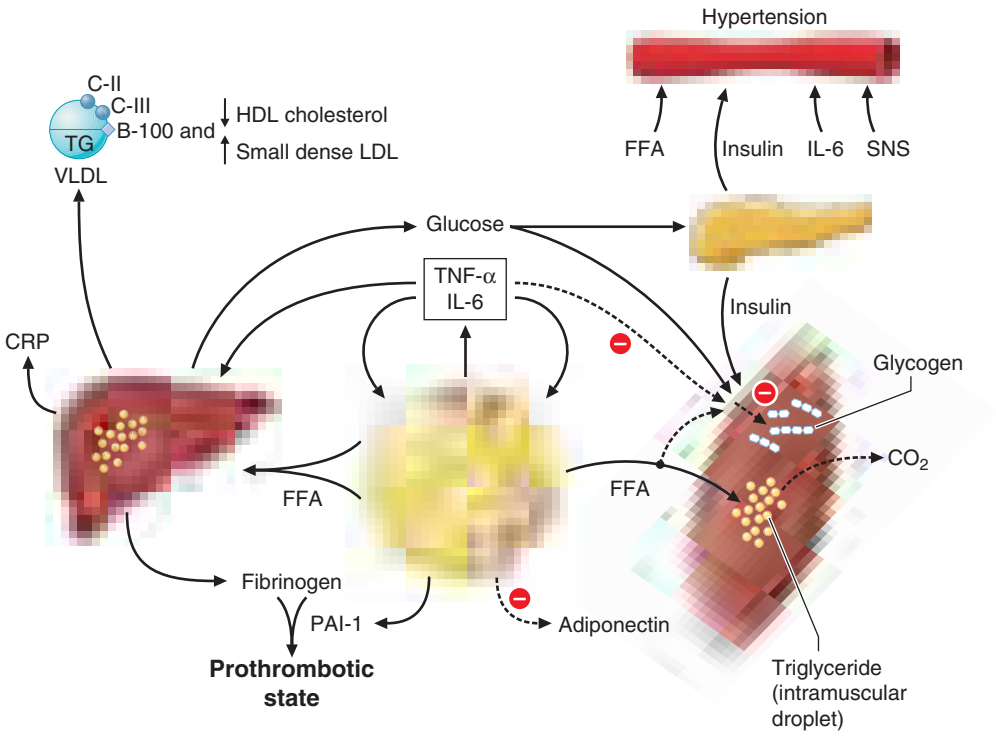


FIGURE 401-2 Pathophysiology of the metabolic syndrome. Free fatty acids (FFAs) are released in abundance from an expanded adipose tissue mass. In the liver, FFAs result in increased production of glucose and triglycerides and secretion of very low-density lipoproteins (VLDLs). Associated lipid/lipoprotein abnormalities include reductions in high-density lipoprotein (HDL) cholesterol and an increased low-density lipoprotein (LDL) particle number (no.). FFAs also reduce insulin sensitivity in muscle by inhibiting insulin-mediated glucose uptake. Associated defects include a reduction in glucose partitioning to glycogen and increased lipid accumulation in triglyceride (TG). The increase in circulating glucose, and to some extent FFAs, increases pancreatic insulin secretion, resulting in hyperinsulinemia. Hyperinsulinemia may result in enhanced sodium reabsorption and increased sympathetic nervous system (SNS) activity and contribute to hypertension, as might higher levels of circulating FFAs. The pro-inflammatory state is superimposed and contributory to the insulin resistance produced by excessive FFAs. The enhanced secretion of interleukin 6 (IL-6) and tumor necrosis factor α (TNF-α) produced by adipocytes and monocyte-derived macrophages results in more insulin resistance and lipolysis of adipose tissue triglyceride stores to circulating FFAs. IL-6 and other cytokines also enhance hepatic glucose production, VLDL production by the liver, hypertension and insulin resistance in muscle. Insulin resistance also contributes to increased triglyceride accumulation in the liver. Cytokines and FFAs also increase hepatic production of fibrinogen and adipocyte production of plasminogen activator inhibitor 1 (PAI-1), resulting in a pro-thrombotic state. Higher levels of circulating cytokines stimulate hepatic production of C-reactive protein (CRP). Reduced production of the anti-inflammatory and insulin-sensitizing cytokine adiponectin is also associated with the metabolic syndrome. *(Modified from RH Eckel et al: Lancet 365:1415, 2005.)*

hyperlipidemia, and a plethora of cardiovascular disorders, such as hypertension, atherosclerosis, CHD, and heart failure.

The oxidative stress hypothesis provides a unifying theory for aging and the predisposition to the metabolic syndrome. In studies of insulin-resistant individuals with obesity, type 2 diabetes, the offspring of patients with type 2 diabetes, and the elderly, a defect in mitochondrial oxidative phosphorylation leads to the accumulation of triglycerides and related lipid molecules in muscle, liver and perhaps other tissues, i.e., β-cells.

Recently, the gut microbiome has emerged as an important contributor to the development of obesity and related metabolic disorders, including the metabolic syndrome. Although the mechanisms remain uncertain, interaction among genetic predisposition, diet, bile acid metabolism, and the intestinal flora is important.

Increased Waist Circumference Waist circumference is an important component of the most recent and frequently applied diagnostic criteria for the metabolic syndrome. However, measuring waist circumference does not reliably distinguish increases in adipose tissue from that in visceral fat; this distinction requires CT or MRI. With increases in visceral adipose tissue, adipose tissue–derived free fatty acids reach the liver. In contrast, increases in abdominal SC fat release lipolysis products into the systemic circulation and therefore have fewer effects on hepatic metabolism. Relative increases in visceral versus SC adipose tissue with increasing waist circumference in Asians and Asian Indians may explain the greater prevalence of the syndrome in those populations than in African-American men, in whom SC fat predominates. It is also possible that visceral fat is a marker for—but not the source of—excess postprandial free fatty acids in obesity.

Dyslipidemia (See also Chap. 400) In general, free fatty acid flux to the liver results in increased production of apoB-containing, triglyceride-rich, very low-density lipoproteins (VLDLs). The effect of insulin on this process is complex, but *hypertriglyceridemia* is an excellent marker of the insulin-resistant condition. Not only is hypertriglyceridemia a feature of the metabolic syndrome, but patients with the metabolic syndrome have elevated levels of apoC-III carried on VLDLs and other lipoproteins. This increase in apoC-III is inhibitory to lipoprotein lipase, further contributing to hypertriglyceridemia, and confers more risk for atherosclerotic cardiovascular disease (ASCVD).

The other major lipoprotein disturbance in the metabolic syndrome is a *reduction in HDL cholesterol*. This reduction is a consequence of changes in HDL composition and metabolism. In the presence of hypertriglyceridemia, a decrease in the cholesterol content of HDL is a consequence of reduced cholesteryl ester content of the lipoprotein core in combination with cholesteryl ester transfer protein–mediated alterations in triglycerides that make the particle small and dense. This change in lipoprotein composition also results in increased clearance of HDL from the circulation. These changes in HDL have a relationship to insulin resistance that is probably indirect, occurring in concert with the changes in triglyceride-rich lipoprotein metabolism.

In addition to HDLs, low-density lipoproteins (LDLs) have alterations in composition in the metabolic syndrome. With fasting serum triglycerides at >2.0 mM (~180 mg/dL), there is usually a predominance of small, dense LDLs, which are thought to be more atherogenic although their association with hypertriglyceridemia and low HDLs make their independent contribution to ASCVD events difficult to assess. Individuals with hypertriglyceridemia often have increases in cholesterol content of both VLDL1 and VLDL2 sub-fractions and in

LDL particle number. Both of these lipoprotein changes may contribute to atherogenic risk in patients with the metabolic syndrome.

Glucose Intolerance (See also Chap. 396) Defects in insulin action in the metabolic syndrome lead to impaired suppression of glucose production by the liver (and kidney) and reduced glucose uptake and metabolism in insulin-sensitive tissues—i.e., muscle and adipose tissue. There is a strong relationship between impaired fasting glucose or impaired glucose tolerance and insulin resistance in studies of humans, nonhuman primates, and rodents. To compensate for defects in insulin action, insulin secretion and/or clearance increases or decreases, respectively, so that euglycemia remains. Ultimately, this compensatory mechanism fails because of defects in insulin secretion, resulting in progression from impaired fasting glucose and/or impaired glucose tolerance to type 2 diabetes mellitus.

Hypertension The relationship between insulin resistance and hypertension is well established. Paradoxically, under normal physiologic conditions, insulin is a vasodilator with secondary effects on sodium reabsorption in the kidney. However, in the setting of insulin resistance, the vasodilatory effect of insulin is lost but the renal effect on sodium reabsorption is preserved. Sodium reabsorption is increased in Caucasians with the metabolic syndrome but not in Africans or Asians. Insulin also increases the activity of the sympathetic nervous system, an effect that is preserved in the setting of insulin resistance. Insulin resistance is also associated with pathway-specific impairment in phosphatidylinositol-3-kinase signaling. In the endothelium, this impairment may cause an imbalance between the production of nitric oxide and the secretion of endothelin 1, with a consequent decrease in blood flow. In addition, increases in angiotensinogen gene expression in adipose tissue of obese subjects results in increases in circulating angiotensin II and vasoconstriction. Although these mechanisms are provocative, the inadequate evaluation of insulin action by measurement of fasting insulin levels or by homeostasis model assessment shows that insulin resistance contributes only partially to the increased prevalence of hypertension in the metabolic syndrome.

Another possible mechanism underlying hypertension in the metabolic syndrome is the vasoactive role of perivascular adipose tissue. Reactive oxygen species released by NADPH oxidase impair endothelial function and result in local vasoconstriction. Other paracrine effects such as leptin or other pro-inflammatory cytokines released from adipose tissue, such as TNF-α may also be important.

Hyperuricemia is another consequence of insulin resistance in the metabolic syndrome. There is growing evidence not only that uric acid is associated with hypertension but also that reduction of uric acid normalizes blood pressure in hyperuricemic adolescents with hypertension. The mechanism appears to be in part related to an adverse effect of uric acid on nitric acid synthase in the macula densa of the kidney and stimulation of the renin-angiotensin aldosterone system.

Pro-Inflammatory Cytokines The increases in pro-inflammatory cytokines—including interleukins 1, 6, and 18; resistin; TNF-α; and the systemic biomarker C-reactive protein—reflect overproduction by the expanded adipose tissue mass (Fig. 401-2). Adipose tissue–derived macrophages may be the primary source of pro-inflammatory cytokines locally and in the systemic circulation. It remains unclear, however, how much of the insulin resistance is caused by the paracrine effects of these cytokines and how much by the endocrine effects.

Adiponectin Adiponectin is an anti-inflammatory cytokine produced exclusively by adipocytes. Adiponectin enhances insulin sensitivity and inhibits many steps in the inflammatory process. In the liver, adiponectin inhibits the expression of gluconeogenic enzymes and the rate of glucose production. In muscle, adiponectin increases glucose transport and enhances fatty acid oxidation, partially through the activation of AMP kinase. Reductions in adiponectin levels are common in the metabolic syndrome. The relative contributions of adiponectin deficiency and overabundance of the pro-inflammatory cytokines are unclear.

CLINICAL FEATURES

Symptoms and Signs The metabolic syndrome typically is not associated with symptoms. On physical examination, waist circumference and blood pressure are often elevated. The presence of either or both of these signs should prompt the clinician to search for other biochemical abnormalities that may be associated with the metabolic syndrome. Less frequently, lipoatrophy or acanthosis nigricans are present on examination. Because these physical findings characteristically are associated with severe insulin resistance, other components of the metabolic syndrome are much more common.

Associated Diseases • **CARDIOVASCULAR DISEASE** The relative risk for new-onset CVD in patients with the metabolic syndrome who do not have diabetes averages 1.5–3 fold. However, in INTERHEART, a study of 26,903 subjects from 52 countries, the risk for acute myocardial infarction in subjects with the metabolic syndrome (WHO or IDF definition) is comparable to that conferred by some, but not all, of the component risk factors. Diabetes mellitus (OR: 2.72) and hypertension (OR: 2.60) are stronger than other risk factors. Although congestive heart failure and the metabolic syndrome can occur together, typically this consequence is secondary to metabolic syndrome-related ASCVD or hypertension. Metabolic syndrome is also associated with increases in the risk for stroke, peripheral vascular disease, and Alzheimer's disease. However, as for myocardial infarction, the risk beyond the additive role of the components of the metabolic syndrome remains debatable. In the Reasons for Geographic and Racial Differences in Stroke (REGARDS) cohort, an observational study of black and white adults ≥45 years old across the United States, there were 9741 participants and 41% had the metabolic syndrome. After adjustment for multiple confounders, metabolic syndrome was associated with increases in high sensitivity C-reactive protein (hsCRP), and this relationship was associated with a 1.34 relative risk for all-cause mortality; but <50% of deaths were from CVD. The population attributable risk was 9.5% for the metabolic syndrome alone and 14.7% for both metabolic syndrome and increased hsCRP. The relationship metabolic syndrome and hsCRP to mortality was greater for whites than blacks.

TYPE 2 DIABETES Overall, the risk for type 2 diabetes among patients with the metabolic syndrome is increased three- to fivefold. In the Framingham Offspring Study's 8-year follow-up of middle-aged participants, the population-attributable risk of the metabolic syndrome for developing type 2 diabetes was 62% among men and 47% among women, yet increases in fasting plasma glucose explained most, if not all, of this increased risk.

Other Associated Conditions In addition to the features specifically used to define the metabolic syndrome, other metabolic alterations are secondary to, or accompany insulin resistance. Those alterations include increases in apoB and apoC-III, uric acid, pro-thrombotic factors (fibrinogen, plasminogen activator inhibitor 1), serum viscosity, asymmetric dimethylarginine, homocysteine, white blood cell count, pro-inflammatory cytokines, C-reactive protein, increased urine albumin/creatinine ratio, non-alcoholic fatty liver disease (NAFLD), and/or non-alcoholic steatohepatitis NASH), polycystic ovary syndrome, and obstructive sleep apnea.

NONALCOHOLIC FATTY LIVER DISEASE NAFLD has become the most common liver disease, in part a consequence of the insulin resistance of the metabolic syndrome. The mechanism relates to increases in free fatty acid flux, reductions in intrahepatic fatty acid oxidation with resultant increases in triglyceride biosynthesis and hepatocellular accumulation, with variable inflammation and oxidative stress. The more serious NASH, a consequence of NAFLD in some patients and precursor of cirrhosis and end stage liver disease, includes a more substantial pro-inflammatory contribution. NASH is now present in 3–5% of the U.S. population and other Western countries. Of patients with the metabolic syndrome, ~25–60% have NAFLD and up to 35% have NASH. As the prevalence of overweight/obesity and the metabolic syndrome increases, NASH may become one of the more common causes of end-stage liver disease and hepatocellular carcinoma.

HYPERURICEMIA (See also Chap. 410) Hyperuricemia reflects defects in insulin action on the renal tubular reabsorption of uric acid and may contribute to hypertension through its effect on the endothelium. An increase in asymmetric dimethylarginine, an endogenous inhibitor of nitric oxide synthase, also relates to endothelial dysfunction. In addition, increases in urine albumin/creatinine ratio may relate to altered endothelial pathophysiology in the insulin-resistant state.

POLYCYSTIC OVARY SYNDROME (See also Chap. 385) Polycystic ovary syndrome is highly associated with insulin resistance (50–80%) and the metabolic syndrome, with a prevalence of the syndrome between 40 and 50%. Women with polycystic ovary syndrome are 2–4 times more likely to have the metabolic syndrome than are women without polycystic ovary syndrome.

OBSTRUCTIVE SLEEP APNEA (See also Chap. 27) Obstructive sleep apnea is commonly associated with obesity, hypertension, increased circulating cytokines, impaired glucose tolerance, and insulin resistance. In fact, obstructive sleep apnea may predict metabolic syndrome, even in the absence of excess adiposity. Moreover, when biomarkers of insulin resistance are compared between patients with obstructive sleep apnea and weight-matched controls, insulin resistance is found to be more severe in those with apnea. Continuous positive airway pressure treatment improves insulin sensitivity in patients with obstructive sleep apnea.

DIAGNOSIS

The diagnosis of the metabolic syndrome relies on fulfillment of the criteria listed in Table 401-1, as assessed using tools at the bedside and in the laboratory. The medical history should include evaluation of symptoms for obstructive sleep apnea in all patients and polycystic ovary syndrome in premenopausal women. Family history will help determine risk for CVD and diabetes mellitus. Blood pressure and waist circumference measurements provide information necessary for the diagnosis.

Laboratory Tests Measurement of fasting lipids and glucose is needed in determining whether the metabolic syndrome is present. The measurement of additional biomarkers associated with insulin resistance can be individualized. Such tests might include those for apoB, hsCRP, fibrinogen, uric acid, urinary albumin/creatinine ratio, and liver function. A sleep study should be performed if symptoms of obstructive sleep apnea are present. If polycystic ovary syndrome is suspected based on clinical features and anovulation, testosterone, luteinizing hormone, and follicle-stimulating hormone should be measured.

TREATMENT

The Metabolic Syndrome

LIFESTYLE (SEE ALSO CHAP. 395)

Obesity, particularly abdominal, is the driving force behind the metabolic syndrome. Thus, weight reduction is the primary approach to the disorder. With at least a 5% and more so with 10% weight reduction, improvement in insulin sensitivity results in favorable modifications in many components of the metabolic syndrome. In general, recommendations for weight loss include a combination of caloric restriction, increased physical activity, and behavior modification. Caloric restriction is the most important component, whereas increases in physical activity are important for maintenance of weight loss. Some but not all evidence suggests that the addition of exercise to caloric restriction may promote greater weight loss from the visceral depot. The tendency for weight regain after successful weight reduction underscores the need for long-lasting behavioral changes.

Diet Before prescribing a weight-loss diet, it is important to emphasize that it has taken the patient a long time to develop an expanded fat mass; thus, the correction need not occur quickly. Given that ~3500 kcal = 1 lb. of fat, ~500-kcal restriction daily equates to weight reduction of 1 lb. per week. Diets restricted in carbohydrate typically provide a more rapid initial weight loss. However, after 1 year, the amount of weight reduction is minimally more reduced or no different from that with caloric restriction alone. Thus, adherence to the diet is more important than the chosen diet. Moreover, there is concern about low-carbohydrate diets enriched in saturated fat, particularly for patients at risk for ASCVD. Therefore, a high-quality dietary pattern—i.e., a diet enriched in fruits, vegetables, whole grains, lean poultry, and fish—should be encouraged to maximize overall health benefit.

Physical Activity Before prescribing a physical activity program to patients with the metabolic syndrome, it is important to ensure that the increased activity does not incur risk. Some high-risk patients should undergo formal cardiovascular evaluation before initiating an exercise program. For an inactive participant, gradual increases in physical activity should be encouraged to enhance adherence and avoid injury. Although increases in physical activity can lead to modest weight reduction, 60–90 min of daily activity is required to achieve this goal. Even if an overweight or obese adult is unable to undertake this level of activity, a significant health benefit will follow from at least 30 min of moderate-intensity activity daily. The caloric value of 30 min of a variety of activities can be found at *www.heart.org/HEARTORG/GettingHealthy/WeightManagement/LosingWeight/Losing-Weight_UCM_307904_Article.jsp*. Of note, a variety of routine activities, such as gardening, walking, and housecleaning, require moderate caloric expenditure. Thus, physical activity should not be defined solely in terms of formal exercise such as jogging, swimming, or tennis.

Behavior Modification Behavioral treatment typically includes recommendations for dietary restriction and more physical activity that predicts sufficient weight loss that benefits metabolic health. The subsequent challenge is the duration of the program because weight regain so often follows successful weight reduction. Improved long-term outcomes often follow a variety of methods, such as a personal or group counselor, the Internet, social media, and telephone follow-up to maintain contact between providers and patients.

Obesity (See also Chap. 395) In some patients with the metabolic syndrome, treatment options need to extend beyond lifestyle intervention. Weight-loss drugs come in two major classes: appetite suppressants and absorption inhibitors. Appetite suppressants approved by the U.S. Food and Drug Administration include phentermine (for short-term use [3 months] only) as well as the more recent additions phentermine/topiramate, lorcaserin, naltrexone/bupropion and high dose (3.0 mg) liraglutide (rather than 1.8 mg, maximum for treatment of type 2 diabetes), which are approved without restrictions on the duration of therapy. In clinical trials, the phentermine/topiramate extended release combination has resulted in ~8% weight loss relative to placebo in 50% of patients. Side effects include palpitations, headache, paresthesias, constipation, and insomnia. Lorcaserin results in less weight loss—typically ~5% beyond placebo—but can cause headache and nasopharyngitis. Naltrexone/bupropion extended release reduces body weight ≥10% in ~20% of patients, however, the drug combination is contraindicated in patients with seizure disorders or any condition that predisposes to seizures. Naltrexone/bupropion also increases pulse and blood pressure and should not be given to patients with uncontrolled hypertension. High dose liraglutide results in ~6% weight loss relative to placebo with ~33% of patients with >10% weight loss. Common side effects are limited to the upper gastrointestinal tract, including nausea, and less frequently, emesis.

Orlistat inhibits fat absorption by ~30% and is moderately effective compared with placebo (~4% more weight loss). Moreover, orlistat reduced the incidence of type 2 diabetes, an effect that was especially evident among patients with impaired glucose tolerance at baseline. This drug is often difficult to take because of oily leakage per rectum. In general, for all weight loss drugs, greater weight

reduction leads to greater improvement in metabolic syndrome components, including the conversion from prediabetes to type 2 diabetes.

Metabolic or bariatric surgery is an option for patients with the metabolic syndrome who have a body mass index >40 kg/m^2, or >35 kg/m^2 with comorbidities. An evolving application for metabolic surgery includes patients with a body mass index as low as 30 kg/m^2 and type 2 diabetes. Gastric bypass or vertical sleeve gastrectomy results in dramatic weight reduction and improvement in most features of the metabolic syndrome. A survival benefit with gastric bypass has also been realized.

LDL CHOLESTEROL **(SEE ALSO CHAP. 400)**

The rationale for the NCEP: ATPIII's development of criteria for the metabolic syndrome was to go beyond LDL cholesterol in identifying and reducing the risk of ASCVD. The working assumption by the panel was that LDL cholesterol goals had already been achieved and that increasing evidence supports a linear reduction in ASCVD events because of progressive lowering of LDL cholesterol with statins. The 2013 ACC/AHA Cholesterol Guidelines have no specific recommendations for patients with the metabolic syndrome; however, a statin should be prescribed in all patients with diabetes age 40–79 with an LDL cholesterol between 60 and 189 mg/dL. For those patients with diabetes and known ASCVD, the current evidence supports a high intensity-statin dose (e.g., atorvastatin 40–80 mg or rosuvastatin 20–40 mg daily). For those patients with the metabolic syndrome but without diabetes, the 10–year ASCVD risk estimator should be employed and a risk ≥7.5% should lead to a discussion between provider and patient about initiating statin therapy for primary prevention of ASCVD.

Diets restricted in saturated fats (<6% of calories) and *trans*-fats (as few as possible) should be applied aggressively. Although evidence is controversial, dietary cholesterol can also be restricted. If LDL cholesterol remains elevated, pharmacologic intervention is needed. Based on substantial evidence, treatment with statins, which lower LDL cholesterol by 15–60%, is the first-choice medication intervention. Of note, for each doubling of the statin dose, LDL cholesterol is further lowered by only ~6%. Hepatotoxicity (more than a threefold increase in hepatic aminotransferases) is rare, but myopathy occurs in ~10–20% of patients. The cholesterol absorption inhibitor ezetimibe is well tolerated and should be the second-choice medication intervention. Ezetimibe typically reduces LDL cholesterol by 15–20%. Proprotein convertase subtilisin/kexin type 9 (PCSK9) inhibitors are potent LDL cholesterol lowering drugs (~45–60%) but are not needed for most patients with the metabolic syndrome. Of course, if these patients also have familial hypercholesterolemia, or insufficient LDL cholesterol lowering on statins ± ezetimibe, a PCSK9 inhibitor should be considered. The bile acid sequestrants cholestyramine, colestipol, and colesevalam may be more effective than ezetimibe but, because they can increase triglyceride levels, must be used with caution in patients with the metabolic syndrome when fasting triglycerides are >300 mg/dL. Side effects include gastrointestinal symptoms (palatability, bloating, belching, constipation, anal irritation). Nicotinic acid has similar LDL cholesterol–lowering capabilities (<20%). Fibrates are best employed to lower LDL cholesterol when triglycerides are not elevated. Fenofibrate may be more effective than gemfibrozil in this setting.

TRIGLYCERIDES **(SEE ALSO CHAP. 400)**

The NCEP: ATPIII focused on non-HDL cholesterol rather than on triglycerides whereas the 2013 ACA/AHA Cholesterol Guidelines stated that fasting triglycerides >500 mg/dL should be treated to prevent more serious hypertriglyceridemia and pancreatitis. Although a fasting triglyceride value of >150 mg/dL is a component of the metabolic syndrome, post hoc analyses of multiple fibrate trials have suggested reduction in the primary ASCVD outcome in patients (with or without concomitant statin therapy) with fasting triglycerides >200 mg/dL, often in the setting of reduced levels of HDL cholesterol.

A fibrate (gemfibrozil or fenofibrate) is the drug of choice to lower fasting triglyceride levels, which are typically reduced by 30–45%. Concomitant administration with drugs metabolized by the 3A4 cytochrome P450 system (including some statins) increases the risk of myopathy. In these cases, fenofibrate may be preferable to gemfibrozil. In the Veterans Affairs HDL Intervention Trial, gemfibrozil was administered to men with known CHD and levels of HDL cholesterol <40 mg/dL. A coronary disease event and mortality rate benefit was experienced predominantly among men with hyperinsulinemia and/or diabetes, many of whom were identified retrospectively as having the metabolic syndrome. Of note, the degree of triglyceride lowering in this trial did not predict benefit. Although levels of LDL cholesterol did not change, a decrease in LDL particle number correlated with benefit.

Other drugs that lower triglyceride levels include statins, nicotinic acid, and prescription omega-3 fatty acids. For this purpose, an intermediate or high dose of the "more potent" statins (atorvastatin, rosuvastatin) is needed. The effect of nicotinic acid on fasting triglycerides is dose related and ~20–35%, an effect that is less pronounced than that of fibrates. In patients with the metabolic syndrome and diabetes, nicotinic acid may increase fasting glucose levels and clinical trials with nicotinic acid + statin have failed to reduce ASCVD events. Prescriptions of omega-3 fatty acid preparations that include high doses of eicosapentaenoic acid ± docosahexaenoic acid (~1.5–4.5 g/d) lower fasting triglyceride levels by ~25–40%. Here, no drug interactions with fibrates or statins occur, and the main side effect of their use is eructation with a fishy taste. Freezing the nutraceutical can partially block this unpleasant side effect. Clinical trials of high-dose omega-3 fatty acids in patients with and without the metabolic syndrome are ongoing.

HDL CHOLESTEROL **(SEE ALSO CHAP. 400)**

Very few lipid-modifying compounds increase HDL cholesterol levels. Statins, fibrates, and bile acid sequestrants have modest effects (5–10%), whereas ezetimibe and omega-3 fatty acids have no effect. Nicotinic acid is the only currently available drug with predictable HDL cholesterol-raising properties. The response is dose related, and nicotinic acid can increase HDL cholesterol by up to 30% above baseline. After several trials of nicotinic acid versus placebo in statin-treated patients, there is no evidence that raising HDL cholesterol with nicotinic acid beneficially affects ASCVD events in patients with or without the metabolic syndrome.

BLOOD PRESSURE **(SEE ALSO CHAP. 271)**

The direct relationship between blood pressure and all-cause mortality rate has been well established in studies comparing patients with hypertension (>140/90 mmHg), patients with pre-hypertension (>120/80 mmHg but <140/90 mmHg), and individuals with normal blood pressure (<120/80 mmHg). In patients who have the metabolic syndrome without diabetes, the best choice for the initial antihypertensive medication is an angiotensin-converting enzyme (ACE) inhibitor or an angiotensin II receptor blocker, as these two classes of drugs are effective and well tolerated. In all patients with hypertension, a sodium-restricted dietary pattern enriched in fruits and vegetables, whole grains, and low-fat dairy products should be advocated. Home monitoring of blood pressure may assist in maintaining good blood-pressure control.

IMPAIRED FASTING GLUCOSE **(SEE ALSO CHAP. 396)**

In patients with the metabolic syndrome and type 2 diabetes, aggressive glycemic control may favorably modify fasting levels of triglycerides and/or HDL cholesterol. In patients with impaired fasting glucose who do not have diabetes, a lifestyle intervention that includes weight reduction, dietary saturated fat restriction, and increased physical activity has been shown to reduce the incidence of type 2 diabetes. Metformin also reduces the incidence of diabetes, although the effect is less pronounced than that of lifestyle intervention.

INSULIN RESISTANCE (SEE ALSO CHAP. 397)

Several drug classes (biguanides, thiazolidinediones [TZDs]) increase insulin sensitivity. Because insulin resistance is the primary pathophysiologic mechanism for the metabolic syndrome, representative drugs in these classes reduce its prevalence. Both metformin and TZDs enhance insulin action in the liver and suppress endogenous glucose production. TZDs, but not metformin, also improve insulin-mediated glucose uptake in muscle and adipose tissue. In a meta-analysis of nine trials involving 12,026 participants, the TZD pioglitazone versus placebo was associated with reduction in ASCVD events in patients with insulin resistance (metabolic syndrome), prediabetes and type 2 diabetes. However, adverse effects including weight gain, bone fracture, and congestive heart failure with/or without edema were seen. Benefit of TZDs has been seen in patients with NAFLD, and with metformin in women with polycystic ovary syndrome, and both drug classes have been shown to reduce markers of inflammation.

FURTHER READING

ALBERTI KG et al: Harmonizing the metabolic syndrome: A joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. Circulation 120:1640, 2009.

BELTRAN-SANCHEZ H et al: Prevalence and trends of metabolic syndrome in the adult U.S. population, 1999–2010. J Am Coll Cardiol 62:693, 2013.

BROWN AE, WALKER M: Genetics of insulin resistance and the metabolic syndrome. Curr Cardiol Rep 18:75, 2016.

ECKEL RH et al: The metabolic syndrome. Lancet 365:1415, 2005.

FERLAND A, ECKEL RH: Does sustained weight loss reverse the metabolic syndrome? Curr Hypertens Rep. 13:456, 2011.

GENSER L et al: Obesity, type 2 diabetes, and the metabolic syndrome: Pathophysiologic relationships and guidelines for surgical intervention. Surg Clin North Am 96:681, 2016.

GRUNDY SM et al: Diagnosis and management of the metabolic syndrome: An American Heart Association/National Heart, Lung, and Blood Institute Scientific Statement. Circulation 112:2735, 2005.

LIM S, ECKEL RH: Pharmacological treatment and therapeutic perspective of metabolic syndrome. Rev Endocr Metab Disord 15:329, 2014.

O'NEILL S, O'DRISCOLL L: Metabolic syndrome: A closer look at the growing epidemic and its associated pathologies. Obes Rev 16:1, 2015.

TUNE JD et al: Cardiovascular consequences of metabolic syndrome. Transl Res 183:57, 2017.

Section 4 Disorders of Bone and Mineral Metabolism

402 Bone and Mineral Metabolism in Health and Disease

F. Richard Bringhurst, Marie B. Demay, Henry M. Kronenberg

BONE STRUCTURE AND METABOLISM

Bone is a dynamic tissue that is remodeled constantly throughout life. The arrangement of compact and cancellous bone provides strength and density suitable for both mobility and protection. Compact or cortical bone forms the roughly cylindrical shell of long bones; cancellous or trabecular bone forms the plate-like meshwork that internally supports the cortical shell. In addition, bone provides a reservoir for calcium, magnesium, phosphorus, sodium, and other ions necessary for homeostatic functions. Bone also hosts and regulates hematopoiesis by providing niches for hematopoietic cell proliferation and differentiation. The skeleton is highly vascular and receives about 10% of the cardiac output. Remodeling of bone is accomplished by two distinct cell types: osteoblasts produce bone matrix, and osteoclasts resorb the matrix. The activities of these cells are coordinated by osteocytes, long-lived regulatory cells embedded within bone matrix.

The extracellular components of bone consist of a solid mineral phase in close association with an organic matrix, of which 90–95% is type I collagen (Chap. 406). The noncollagenous portion of the organic matrix is heterogeneous and contains serum proteins such as albumin as well as many locally produced proteins, whose functions are incompletely understood. Those proteins include cell attachment/signaling proteins such as thrombospondin, osteopontin, and fibronectin; calcium-binding proteins such as matrix gla protein and osteocalcin; and proteoglycans such as biglycan and decorin. Some of the proteins organize collagen fibrils; others influence mineralization and binding of the mineral phase to the matrix.

The mineral phase is made up of calcium and phosphate and is best characterized as a poorly crystalline hydroxyapatite. The mineral phase of bone is deposited initially in intimate relation to the collagen fibrils and is found in specific locations in the "holes" between the collagen fibrils. This architectural arrangement of mineral and matrix results in a two-phase material well suited to withstand mechanical stresses. The organization of collagen influences the amount and type of mineral phase formed in bone. Although the primary structures of type I collagen in skin and bone tissues are similar, there are differences in posttranslational modifications and distribution of intermolecular cross-links. The holes in the packing structure of the collagen are larger in mineralized collagen of bone and dentin than in unmineralized collagens such as those in tendon. Single amino acid substitutions in the helical portion of either the α1 (*COL1A1*) or α2 (*COL1A2*) chains of type I collagen disrupt the organization of bone in osteogenesis imperfecta. The severe skeletal fragility associated with this group of disorders highlights the importance of the fibrillar matrix in the structure of bone (Chap. 406).

Osteoblasts synthesize and secrete the organic matrix and regulate its mineralization. They are derived from cells of mesenchymal origin (Fig. 402-1A). Active osteoblasts are found on the surface of newly forming bone. As an osteoblast secretes matrix, which then is mineralized, the cell may become an *osteocyte*, still connected with its blood supply through a series of canaliculi. Osteocytes account for the vast majority of the cells in bone. They are thought to be the mechanosensors in bone that communicate signals to surface osteoblasts and osteoclasts and their progenitors through the canalicular network and thereby serve as master regulators of bone formation and resorption. Remarkably, osteocytes also secrete fibroblast growth factor 23 (FGF23), a major hormonal regulator of phosphate metabolism (see below). Mineralization of the matrix, both in trabecular bone and in osteones of compact cortical bone (*Haversian systems*), begins soon after the matrix is secreted (primary mineralization) but is not completed for several weeks or even longer (secondary mineralization). Although this mineralization takes advantage of the high concentrations of calcium and phosphate, already near saturation in serum, mineralization is a carefully regulated process that is dependent on the activity of osteoblast-derived alkaline phosphatase, which probably works by hydrolyzing inhibitors of mineralization.

Genetic studies in humans and mice have identified several key genes that control osteoblast development. *Runx2* is a transcription factor expressed specifically in chondrocyte (cartilage cells) and osteoblast progenitors as well as in hypertrophic chondrocytes and mature osteoblasts. *Runx2* regulates the expression of several important osteoblast proteins, including osterix (another transcription factor needed for osteoblast maturation), osteopontin, bone sialoprotein, type I collagen, osteocalcin, and receptor-activator of NFκB (RANK) ligand. *Runx2* expression is regulated in part by bone morphogenic proteins (BMPs). *Runx2*-deficient mice are devoid of osteoblasts, whereas mice with a deletion

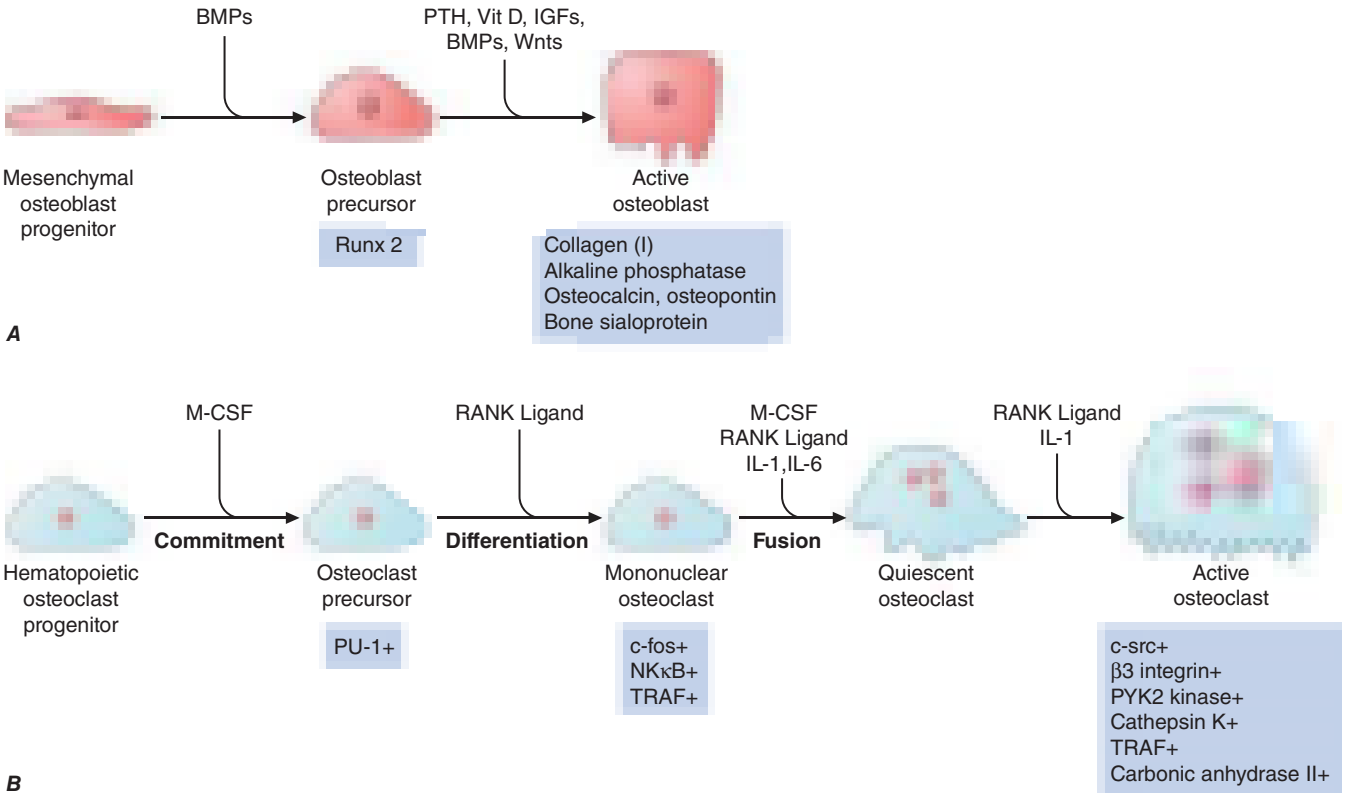


FIGURE 402-1 Pathways regulating development of (A) osteoblasts and (B) osteoclasts. Hormones, cytokines, and growth factors that control cell proliferation and differentiation are shown above the arrows. Transcription factors and other markers specific for various stages of development are depicted below the arrows. BMPs, bone morphogenic proteins; wnts, wingless-type mouse mammary tumor virus integration site; PTH, parathyroid hormone; Vit D, vitamin D; IGFs, insulin-like growth factors; Runx2, Runt-related transcription factor 2; M-CSF, macrophage colony-stimulating factor; PU-1, a monocyte- and B lymphocyte–specific ets family transcription factor; NFκB, nuclear factor κB; TRAF, tumor necrosis factor receptor–associated factor; RANK ligand, receptor activator of NFκB ligand; IL-1, interleukin 1; IL-6, interleukin 6. *(Modified from T Suda et al: Endocr Rev 20:345, 1999, with permission.)*

of only one allele (*Runx2* +/–) exhibit a delay in formation of the clavicles and some cranial bones. The latter abnormalities are similar to those in the human disorder *cleidocranial dysplasia*, which is also caused by heterozygous inactivating mutations in *Runx2*.

The paracrine signaling molecule, Indian hedgehog (Ihh), also plays a critical role in osteoblast development, as evidenced by Ihh-deficient mice that lack osteoblasts in the type of bone formed on a cartilage mold (endochondral ossification). Signals originating from members of the wnt (wingless-type mouse mammary tumor virus integration site) family of paracrine factors are also important for osteoblast proliferation and differentiation. Osteocytes regulate osteoblasts partly by secreting a potent inhibitor of wnt signaling called sclerostin. Numerous other growth-regulatory factors affect osteoblast function, including the three closely related transforming growth factor βs, fibroblast growth factors (FGFs) 2 and 18, platelet-derived growth factor, and insulin-like growth factors (IGFs) I and II. Hormones such as parathyroid hormone (PTH) and 1,25-dihydroxyvitamin D [1,25$(OH)_2$D] activate receptors expressed by osteoblasts to assure mineral homeostasis and influence a variety of bone cell functions.

Resorption of bone is carried out mainly by *osteoclasts*, multinucleated cells that are formed by fusion of cells derived from the common precursor of macrophages and osteoclasts. Thus, these cells derive from the hematopoietic lineage, quite different from the mesenchymal cells that become osteoblasts. Multiple factors that regulate osteoclast development have been identified **(Fig. 402-1*B*)**. Factors produced by osteocytes, osteoblasts and marrow stromal cells allow cells of the osteoblast lineage to control osteoclast development and activity. Macrophage colony-stimulating factor (M-CSF) plays a critical role during several steps in the pathway and ultimately leads to fusion of osteoclast progenitor cells to form multinucleated, active osteoclasts. RANK ligand, a member of the tumor necrosis factor (TNF) family, is expressed on the surface of osteocytes, osteoblasts, and stromal fibroblasts. In a process involving cell-cell interactions, RANK ligand binds to the RANK receptor on osteoclast progenitors, stimulating osteoclast differentiation and activation. Alternatively, a soluble decoy receptor, referred to as osteoprotegerin, can bind RANK ligand and inhibit osteoclast differentiation. Several growth factors and cytokines (including interleukins 1, 6, and 11; TNF; and interferon γ) modulate osteoclast differentiation and function. Most hormones that influence osteoclast function do not target these cells directly but instead target cells of the osteoblast lineage to increase production of M-CSF and RANK. Both PTH and 1,25$(OH)_2$D increase osteoclast number and activity by this indirect mechanism. Calcitonin, in contrast, binds to its receptor on the basal surface of osteoclasts and directly inhibits osteoclast function. Estradiol has multiple cellular targets in bone, including osteoclasts, immune cells, and osteoblasts; actions on all these cells serve to decrease osteoclast number and decreased bone resorption.

Osteoclast-mediated resorption of bone takes place in scalloped spaces (*Howship's lacunae*) where the osteoclasts are attached through a specific αvβ3 integrin to components of the bone matrix such as osteopontin. The osteoclast forms a tight seal to the underlying matrix and secretes protons, chloride, and proteinases into a confined space that has been likened to an extracellular lysosome. The active osteoclast surface forms a ruffled border that contains a specialized proton-pump ATPase that secretes acid that solubilizes the mineral phase. Carbonic anhydrase (type II isoenzyme) within the osteoclast generates the needed protons. The bone matrix is resorbed in the acid environment adjacent to the ruffled border by proteases, such as cathepsin K, that act at low pH.

In the embryo and the growing child, bone develops mostly by replacing previously calcified cartilage (endochondral bone formation) with subsequent remodeling, or, in a few bones, is formed without a cartilage matrix (intramembranous bone formation). During endochondral bone formation, chondrocytes proliferate, secrete and mineralize a matrix, enlarge (hypertrophy), and then die, enlarging bone and providing the matrix and factors that stimulate endochondral bone

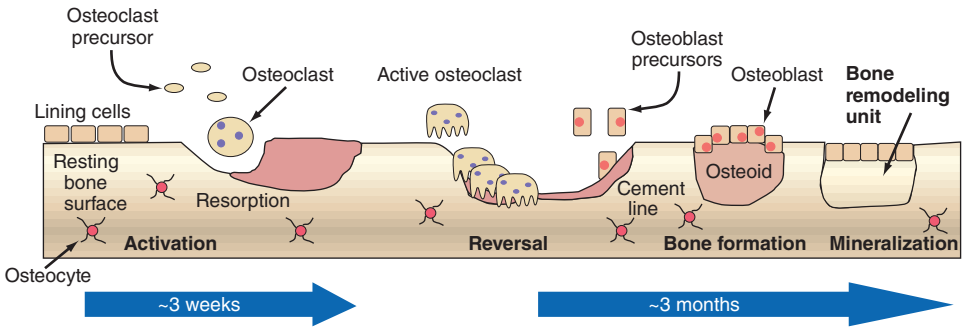


FIGURE 402-2 Schematic representation of bone remodeling. The cycle of bone remodeling is carried out by the basic multicellular unit (BMU), which consists of a group of osteoclasts and osteoblasts. In cortical bone, the BMUs tunnel through the tissue, whereas in cancellous bone, they move across the trabecular surface. The process of bone remodeling is initiated by contraction of the lining cells and the recruitment of osteoclast precursors. These precursors fuse to form multinucleated, active osteoclasts that mediate bone resorption. Osteoclasts adhere to bone and subsequently remove it by acidification and proteolytic digestion. As the BMU advances, osteoclasts leave the resorption site and osteoblasts move in to cover the excavated area and begin the process of new bone formation by secreting osteoid, which eventually is mineralized into new bone. After osteoid mineralization, osteoblasts flatten and form a layer of lining cells over new bone.

formation. This program is regulated by both local factors such as IGF-I and -II, Ihh, parathyroid hormone–related peptide (PTHrP), BMPs, and FGFs and by systemic hormones such as growth hormone, glucocorticoids, and estrogen.

New bone, whether formed in infants or in adults during repair, has a relatively high ratio of cells to matrix and is characterized by coarse fiber bundles of collagen that are interlaced and randomly dispersed (woven bone). In adults, the more mature bone is organized with fiber bundles regularly arranged in parallel or concentric sheets (lamellar bone). In long bones, deposition of lamellar bone in a concentric arrangement around blood vessels forms the Haversian systems. Growth in length of bones is dependent on proliferation of cartilage cells and the endochondral sequence at the growth plate. Growth in width and thickness is accomplished by formation of bone at the periosteal surface and by resorption at the endosteal surface, with the rate of formation exceeding that of resorption. In adults, after the growth plates of cartilage close through the actions of estrogen, growth in length and endochondral bone formation cease. Even in adults, however, remodeling of bone (within Haversian systems as well as along the surfaces of trabecular bone) continues throughout life. In adults, ~4% of the surface of trabecular bone (such as iliac crest) is involved in active resorption, whereas 10–15% of trabecular surfaces are covered with osteoid, unmineralized new bone formed by osteoblasts. Radioisotope studies indicate that as much as 18% of the total skeletal calcium is deposited and removed each year. Thus, bone is an active metabolizing tissue that requires an intact blood supply. The cycle of bone resorption and formation is a highly orchestrated process, directed by osteocytes and carried out by the basic multicellular unit, which is composed of a group of osteoclasts and osteoblasts **(Fig. 402-2)**.

The response of bone to fractures, infection, and interruption of blood supply and to expanding lesions is relatively limited. Dead bone must be resorbed, and new bone must be formed, a process carried out in association with growth of new blood vessels into the involved area. In injuries that disrupt the organization of the tissue such as a fracture in which apposition of fragments is poor or when motion exists at the fracture site, progenitor stromal cells recapitulate the endochondral bone formation of early development and form cartilage that is replaced by bone and, variably, fibrous tissue. When there is good apposition with fixation and little motion at the fracture site, repair occurs predominantly by formation of new bone without other mediating tissue.

Remodeling of bone occurs along lines of force generated by mechanical stress. The signals from these mechanical stresses are sensed by osteocytes, which transmit signals to osteoclasts and osteoblasts or their precursors. One such signal made by osteocytes is sclerostin, an inhibitor of wnt signaling. Mechanical forces suppress sclerostin production and thus increase bone formation by osteoblasts. Expanding lesions in bone such as tumors induce resorption at the surface in contact with the tumor by producing ligands such as PTHrP that stimulate osteoclast differentiation and function. Thus, bone plasticity reflects the interaction of cells with each other and with the environment.

Measurement of the products of osteoblast and osteoclast activity can assist in the diagnosis and management of bone diseases. Osteoblast activity can be assessed by measuring serum bone-specific alkaline phosphatase. Similarly, osteocalcin, a protein secreted from osteoblasts, is made virtually only by osteoblasts. Osteoclast activity can be assessed by measurement of products of collagen degradation. Collagen molecules are covalently linked to each other in the extracellular matrix through the formation of hydroxypyridinium cross-links **(Chap. 406)**. After digestion by osteoclasts, these cross-linked peptides can be measured both in urine and in blood.

CALCIUM METABOLISM

Over 99% of the 1–2 kg of calcium present normally in the adult human body resides in the skeleton, where it provides mechanical stability and serves as a reservoir sometimes needed to maintain extracellular fluid (ECF) calcium concentration **(Fig. 402-3)**. Skeletal calcium accretion first becomes significant during the third trimester of fetal life, accelerates throughout childhood and adolescence, reaches a peak in early adulthood, and gradually declines thereafter at rates that rarely exceed 1–2% per year. These slow changes in total skeletal calcium content contrast with relatively high daily rates of closely matched fluxes of calcium into and out of bone (~250–500 mg each), a process mediated by coupled osteoblastic and osteoclastic activity. Another 0.5–1% of skeletal calcium is freely exchangeable (e.g., in chemical equilibrium) with that in the ECF.

The concentration of ionized calcium in the ECF must be maintained within a narrow range because of the critical role calcium plays in

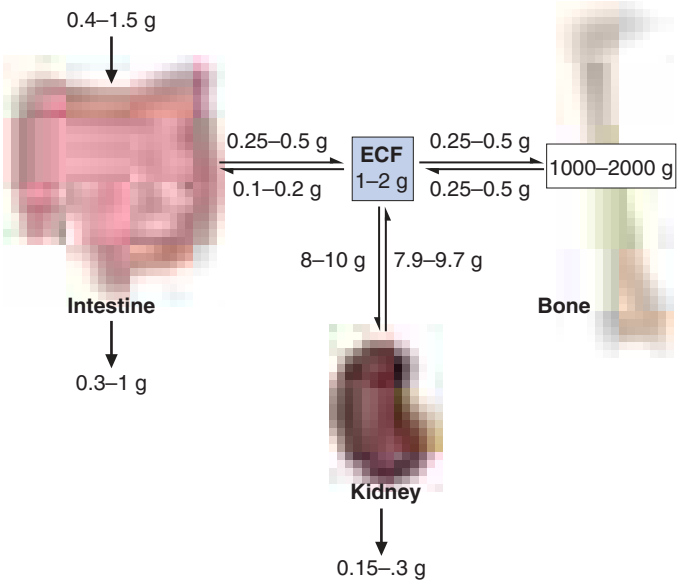


FIGURE 402-3 Calcium homeostasis. Schematic illustration of calcium content of extracellular fluid (ECF) and bone as well as of diet and feces; magnitude of calcium flux per day as calculated by various methods is shown at sites of transport in intestine, kidney, and bone. Ranges of values shown are approximate and were chosen to illustrate certain points discussed in the text. In conditions of calcium balance, rates of calcium release from and uptake into bone are equal.

a wide array of cellular functions, especially those involved in neuromuscular activity, secretion, and signal transduction. Intracellular cytosolic free calcium levels are ~100 nmol/L and are 10,000-fold lower than ionized calcium concentration in the blood and ECF (1.1–1.3 mmol/L). Cytosolic calcium does not play the structural role played by extracellular calcium; instead, it serves a signaling function. The steep chemical gradient of calcium from outside to inside the cell promotes rapid calcium influx through various membrane calcium channels that can be activated by hormones, metabolites, or neurotransmitters, swiftly changing cellular function. In blood, total calcium concentration is normally 2.2–2.6 m*M* (8.5–10.5 mg/dL), of which ~50% is ionized. The remainder is bound ionically to negatively charged proteins (predominantly albumin and immunoglobulins) or loosely complexed with phosphate, citrate, sulfate, or other anions. Alterations in serum protein concentrations directly affect the total blood calcium concentration even if the ionized calcium concentration remains normal. An algorithm to correct for protein changes adjusts the total serum calcium (in mg/dL) upward by 0.8 times the deficit in serum albumin (g/dL) or by 0.5 times the deficit in serum immunoglobulin (in g/dL). Such corrections provide only rough approximations of actual free calcium concentrations, however, and may be misleading, particularly during acute illness. Acidosis also alters ionized calcium by reducing its association with proteins. The best practice is to measure blood ionized calcium directly by a method that employs calcium-selective electrodes in acute settings during which calcium abnormalities might occur.

Control of the ionized calcium concentration in the ECF ordinarily is accomplished by adjusting the rates of calcium movement across intestinal and renal epithelia. These adjustments are mediated mainly via changes in blood levels of the hormones, PTH and $1,25(OH)_2D$. Blood ionized calcium directly suppresses PTH secretion by activating calcium-sensing receptors (CaSRs) in parathyroid cells. Also, ionized calcium indirectly affects PTH secretion by lowering $1,25(OH)_2D$ production. This active vitamin D metabolite inhibits PTH production by an incompletely understood mechanism of negative feedback **(Chap. 403)**.

Normal dietary calcium intake in the United States varies widely, ranging from 10–37 mmol/d (400–1500 mg/d). A National Academy of Medicine (formerly, Institute of Medicine) analysis recommends a daily allowance of 25–30 mmol (1000–1200 mg) for most adults. Intestinal absorption of ingested calcium involves both active (transcellular) and passive (paracellular) mechanisms. Passive calcium absorption is nonsaturable and approximates 5% of daily calcium intake, whereas active absorption involves apical calcium entry via specific ion channels (TRPV5 and TRPV6), whose expression is controlled principally by $1,25(OH)_2D$, and normally ranges from 20 to 70%. Active calcium transport occurs mainly in the proximal small bowel (duodenum and proximal jejunum), although some active calcium absorption occurs in most segments of the small intestine. Optimal rates of calcium absorption require gastric acid. This is especially true for weakly dissociable calcium supplements such as calcium carbonate. In fact, large boluses of calcium carbonate are poorly absorbed because of their neutralizing effect on gastric acid. In achlorhydric subjects and for those taking drugs that inhibit gastric acid secretion, supplements should be taken with meals to optimize their absorption. Use of calcium citrate may be preferable in these circumstances. Calcium absorption may also be blunted in disease states such as pancreatic or biliary insufficiency, in which ingested calcium remains bound to unabsorbed fatty acids or other food constituents. At high levels of calcium intake, synthesis of $1,25(OH)_2D$ is reduced; this decreases the rate of active intestinal calcium absorption. The opposite occurs with dietary calcium restriction. Some calcium, ~2.5–5 mmol/d (100–200 mg/d), is excreted as an obligate component of intestinal secretions and is not regulated by calciotropic hormones.

The feedback-controlled hormonal regulation of intestinal absorptive efficiency results in a relatively constant daily net calcium absorption of ~5–7.5 mmol/d (200–400 mg/d) despite large changes in daily dietary calcium intake. This daily load of absorbed calcium is excreted by the kidneys in a manner that is also tightly regulated by the concentration of ionized calcium in the blood. Approximately 8–10 g/d of calcium is filtered by the glomeruli, of which only 2–3% appears in the urine. Most filtered calcium (65%) is reabsorbed in the proximal tubules via a passive, paracellular route that is coupled to concomitant NaCl reabsorption and not specifically regulated. The cortical thick ascending limb of Henle's loop (cTAL) reabsorbs roughly another 20% of filtered calcium, also via a paracellular mechanism. Calcium reabsorption in the cTAL requires a tight-junctional protein called paracellin-1 and is inhibited by increased blood concentrations of calcium or magnesium, acting via the CaSR, which is highly expressed on basolateral membranes in this nephron segment. Operation of the renal CaSR provides a mechanism, independent of those engaged directly by PTH or $1,25(OH)_2D$, by which serum ionized calcium can control renal calcium reabsorption. Finally, ~10% of filtered calcium is reabsorbed in the distal convoluted tubules (DCTs) by a transcellular mechanism. Calcium enters the luminal surface of the cell through specific apical calcium channels (TRPV5), whose number is regulated. It then moves across the cell in association with a specific calcium-binding protein (calbindin-D28k) that buffers cytosolic calcium concentrations from the large mass of transported calcium. Ca^{2+}-ATPases and Na^+/Ca^{2+} exchangers actively extrude calcium across the basolateral surface and thereby maintain the transcellular calcium gradient. All these processes are stimulated directly or indirectly by PTH. The DCT is also the site of action of thiazide diuretics, which lower urinary calcium excretion by inducing sodium depletion and thereby augmenting proximal calcium reabsorption. Conversely, dietary sodium loads, or increased distal sodium delivery caused by loop diuretics or saline infusion, induce calciuresis.

The homeostatic mechanisms that normally maintain a constant serum ionized calcium concentration may fail at extremes of calcium intake or when the hormonal systems or organs involved are compromised. Thus, even with maximal activity of the vitamin D–dependent intestinal active transport system, sustained calcium intakes <5 mmol/d (<200 mg/d) cannot provide enough net calcium absorption to replace obligate losses via the intestine, the kidney, sweat, and other secretions. In this case, increased blood levels of PTH and $1,25(OH)_2D$ activate osteoclastic bone resorption to obtain needed calcium from bone, which leads to progressive bone loss and negative calcium balance. Increased PTH and $1,25(OH)_2D$ also enhance renal calcium reabsorption, and $1,25(OH)_2D$ enhances calcium absorption in the gut. At very high calcium intakes (>100 mmol/d [>4 g/d]), passive intestinal absorption continues to deliver calcium into the ECF despite maximally downregulated intestinal active transport and renal tubular calcium reabsorption. This can cause severe hypercalciuria, nephrocalcinosis, progressive renal failure, and hypercalcemia (e.g., "milk-alkali syndrome"). Deficiency or excess of PTH or vitamin D, intestinal disease, and renal failure represent other commonly encountered challenges to normal calcium homeostasis **(Chap. 403)**.

PHOSPHORUS METABOLISM

Although 85% of the ~600 g of body phosphorus is present in bone mineral, phosphorus is also a major intracellular constituent both as the free anion(s) and as a component of numerous organophosphate compounds, including structural proteins, enzymes, transcription factors, carbohydrate and lipid intermediates, high-energy stores (ATP [adenosine triphosphate], creatine phosphate), and nucleic acids. Unlike calcium, phosphorus exists intracellularly at concentrations close to those present in ECF (e.g., 1–2 mmol/L). In cells and in the ECF, phosphorus exists in several forms, predominantly as $H_2PO_4^-$ or $NaHPO_4^-$, with perhaps 10% as HPO_4^{2-}. This mixture of anions will be referred to here as "phosphate." In serum, about 12% of phosphorus is bound to proteins. Concentrations of phosphates in blood and ECF generally are expressed in terms of elemental phosphorus, with the normal range in adults being 0.75–1.45 mmol/L (2.5–4.5 mg/dL). Because the volume of the intracellular fluid compartment is twice that of the ECF, measurements of ECF phosphate may not accurately reflect phosphate availability within cells that follows even modest shifts of phosphate from one compartment to the other.

Phosphate is widely available in foods and is absorbed efficiently (65%) by the small intestine even in the absence of vitamin D. However, phosphate absorptive efficiency may be enhanced (to 85–90%) via active transport mechanisms that are stimulated by $1,25(OH)_2D$. These mechanisms involve activation of Na^+/PO_4^{2-} co-transporters that move phosphate into

intestinal cells against an unfavorable electrochemical gradient. Daily net intestinal phosphate absorption varies widely with the composition of the diet but is generally in the range of 500–1000 mg/d. Phosphate absorption can be inhibited by large doses of calcium salts or by sevelamer hydrochloride (Renagel), strategies commonly used to control levels of serum phosphate in renal failure. Aluminum hydroxide antacids also reduce phosphate absorption but are used less commonly because of the potential for aluminum toxicity. Low serum phosphate stimulates renal proximal tubular synthesis of $1,25(OH)_2D$, perhaps by suppressing blood levels of FGF23 (see below).

Serum phosphate levels vary by as much as 50% on a normal day. This reflects the effect of food intake but also an underlying circadian rhythm that produces a nadir between 7 and 10 A.M. Carbohydrate administration, especially as IV dextrose solutions in fasting subjects, can decrease serum phosphate by >0.7 mmol/L (2 mg/dL) due to rapid uptake into and utilization by cells. A similar response is observed in the treatment of diabetic ketoacidosis and during metabolic or respiratory alkalosis. Because of this wide variation in serum phosphate, it is best to perform measurements in the basal, fasting state.

Control of serum phosphate is determined mainly by the rate of renal tubular reabsorption of the filtered load, which is ~4–6 g/d. Because intestinal phosphate absorption is highly efficient, urinary excretion is not constant but varies directly with dietary intake. The fractional excretion of phosphate (ratio of phosphate to creatinine clearance) is generally in the range of 10–15%. The proximal tubule is the principal site at which renal phosphate reabsorption is regulated. This is accomplished by changes in the levels of apical expression and activity of specific Na^+/PO_4^{2-} co-transporters (NaPi-2a and NaPi-2c) in the proximal tubule. Levels of these transporters at the apical surface of these cells are reduced rapidly by PTH, a major hormonal regulator of renal phosphate excretion. FGF23 can impair phosphate reabsorption dramatically by a similar mechanism. Activating *FGF23* mutations cause the rare disorder autosomal dominant hypophosphatemic rickets (ADHR). In contrast to PTH, FGF23 also leads to reduced synthesis of $1,25(OH)_2D$, which may worsen the resulting hypophosphatemia by lowering intestinal phosphate absorption. Renal reabsorption of phosphate is responsive to changes in dietary intake such that experimental dietary phosphate restriction leads to a dramatic lowering of urinary phosphate within hours, preceding any decline in serum phosphate (e.g., filtered load). This physiologic renal adaptation to changes in dietary phosphate availability occurs independently of PTH and may be mediated in part by changes in levels of serum FGF23. Findings in *FGF23*-knockout mice suggest that FGF23 normally acts to lower blood phosphate and $1,25(OH)_2D$ levels. In turn, elevation of blood phosphate increases blood levels of FGF23.

Renal phosphate reabsorption is impaired by hypocalcemia, hypomagnesemia, and severe hypophosphatemia. Phosphate clearance is enhanced by ECF volume expansion and impaired by dehydration. Phosphate retention is an important pathophysiologic feature of renal insufficiency **(Chap. 305)**.

HYPOPHOSPHATEMIA

Causes Hypophosphatemia can occur by one or more of three primary mechanisms: (1) inadequate intestinal phosphate absorption, (2) excessive renal phosphate excretion, and (3) rapid redistribution of phosphate from the ECF into bone or soft tissue **(Table 402-1)**. Because phosphate is so abundant in foods, inadequate intestinal absorption is almost never observed now that aluminum hydroxide antacids, which bind phosphate in the gut, are no longer widely used. Fasting or starvation, however, may result in depletion of body phosphate and predispose to subsequent hypophosphatemia during refeeding, especially if this is accomplished with IV glucose alone.

Chronic hypophosphatemia usually signifies a persistent renal tubular phosphate-wasting disorder. Excessive activation of PTH/PTHrP receptors in the proximal tubule as a result of primary or secondary hyperparathyroidism or because of the PTHrP-mediated hypercalcemia syndrome in malignancy **(Chap. 403)** is among the more common causes of renal hypophosphatemia, especially because of the high prevalence of vitamin D deficiency in older Americans. Familial hypocalciuric hypercalcemia and Jansen's chondrodystrophy are rare examples of genetic disorders in this category **(Chap. 403)**.

TABLE 402-1 Causes of Hypophosphatemia

- I. Reduced renal tubular phosphate reabsorption
  - A. PTH/PTHrP-dependent
    - 1. Primary hyperparathyroidism
    - 2. Secondary hyperparathyroidism
      - a. Vitamin D deficiency/resistance
      - b. Calcium starvation/malabsorption
      - c. Bartter's syndrome
      - d. Autosomal recessive renal hypercalciuria with hypomagnesemia
    - 3. PTHrP-dependent hypercalcemia of malignancy
    - 4. Familial hypocalciuric hypercalcemia
  - B. PTH/PTHrP-independent
    - 1. Excess FGF23 or other "phosphatonins"
      - a. X-linked hypophosphatemic rickets (XLH)
      - b. Autosomal recessive hypophosphatemia (ARHP)
      - c. Autosomal dominant hypophosphatemic rickets (ADHR) (DMP1, ENPP1 deficiency)
      - d. Tumor-induced osteomalacia syndrome (TIO)
      - e. McCune-Albright syndrome (fibrous dysplasia)
      - f. Epidermal nevus syndrome
    - 2. Intrinsic renal disease
      - a. Fanconi's syndrome(s)
      - b. Cystinosis
      - c. Wilson's disease
      - d. NaPi-2a or NaPi-2c mutations
    - 3. Other systemic disorders
      - a. Poorly controlled diabetes mellitus
      - b. Alcoholism
      - c. Hyperaldosteronism
      - d. Hypomagnesemia
      - e. Amyloidosis
      - f. Hemolytic-uremic syndrome
      - g. Renal transplantation or partial liver resection
      - h. Rewarming or induced hyperthermia
    - 4. Drugs or toxins
      - a. Ethanol
      - b. Acetazolamide, other diuretics
      - c. High-dose estrogens or glucocorticoids
      - d. Heavy metals (lead, cadmium, saccharated ferric oxide)
      - e. Toluene, *N*-methyl formamide
      - f. Cisplatin, ifosfamide, foscarnet, rapamycin
- II. Impaired intestinal phosphate absorption
  - A. Aluminum-containing antacids
  - B. Sevalamer
- III. Shifts of extracellular phosphate into cells
  - A. Intravenous glucose
  - B. Insulin therapy for prolonged hyperglycemia or diabetic ketoacidosis
  - C. Catecholamines (epinephrine, dopamine, albuterol)
  - D. Acute respiratory alkalosis
  - E. Gram-negative sepsis, toxic shock syndrome
  - F. Recovery from starvation or acidosis
  - G. Rapid cellular proliferation
    - 1. Leukemic blast crisis
    - 2. Intensive erythropoietin, other growth factor therapy
- IV. Accelerated net bone formation
  - A. After parathyroidectomy
  - B. Treatment of vitamin D deficiency, Paget's disease
  - C. Osteoblastic metastases

Abbreviations: PTH, parathyroid hormone; PTHrP, parathyroid hormone–related peptide.

Several genetic and acquired diseases cause PTH/PTHrP-independent tubular phosphate wasting with associated rickets and osteomalacia. All these diseases manifest severe hypophosphatemia; renal phosphate wasting, sometimes accompanied by aminoaciduria; inappropriately low blood levels of $1,25(OH)_2D$; low-normal serum levels of calcium; and evidence of impaired cartilage or bone mineralization. Analysis of these diseases led to the discovery of the hormone FGF23, which is an important physiologic regulator of phosphate metabolism. FGF23 decreases phosphate reabsorption in the proximal tubule and also suppresses the 1α-hydroxylase responsible for synthesis of $1,25(OH)_2D$. FGF23 is synthesized by cells of the osteoblast lineage, primarily osteocytes. High-phosphate diets increase FGF23 levels, and low-phosphate diets decrease them. ADHR was the first disease linked to abnormalities in FGF23. ADHR results from activating mutations in the gene that encodes FGF23. These mutations alter a cleavage site that ordinarily allows for inactivation of intact FGF23. Several other genetic disorders feature elevated FGF23 and hypophosphatemia. The most common of these is X-linked hypophosphatemic rickets (XLH), which results from inactivating mutations in an endopeptidase termed *PHEX* (*p*hosphate-regulating gene with *h*omologies to *e*ndopeptidases on the *X* chromosome) that is expressed most abundantly on the surface of osteocytes and mature osteoblasts. Patients with XLH usually have high FGF23 levels, and ablation of the *FGF23* gene reverses the hypophosphatemia found in the mouse version of XLH. How inactivation of PHEX leads to increased levels of FGF23 has not been determined. Two rare autosomal recessive hypophosphatemic syndromes associated with elevated FGF23 are due to inactivating mutations of dentin matrix protein-1 (DMP1) and ectonucleotide pyrophosphatase/phosphodiesterase 1 (ENPP1), respectively, both of which normally are highly expressed in bone and presumably regulate FGF23 production. An unusual hypophosphatemic disorder, tumor-induced osteomalacia (TIO), is an acquired disorder in which tumors, usually of mesenchymal origin and generally histologically benign, secrete FGF23 and/or other molecules that induce renal phosphate wasting. The hypophosphatemic syndrome resolves completely within hours to days after successful resection of the responsible tumor. Such tumors typically express large amounts of FGF23 mRNA, and patients with TIO usually exhibit elevations of FGF23 in their blood.

Dent's disease is an X-linked recessive disorder caused by inactivating mutations in CLCN5, a chloride transporter expressed in endosomes of the proximal tubule; features include hypercalciuria, hypophosphatemia, and recurrent kidney stones. Renal phosphate wasting is common among poorly controlled diabetic patients and alcoholics, who therefore are at risk for iatrogenic hypophosphatemia when treated with insulin or IV glucose, respectively. Diuretics and certain other drugs and toxins can cause defective renal tubular phosphate reabsorption (Table 402-1).

In hospitalized patients, hypophosphatemia is often attributable to massive redistribution of phosphate from the ECF into cells. Insulin therapy for diabetic ketoacidosis is a paradigm for this phenomenon, in which the severity of the hypophosphatemia is related to the extent of antecedent depletion of phosphate and other electrolytes **(Chap. 397)**. The hypophosphatemia is usually greatest at a point many hours after initiation of insulin therapy and is difficult to predict from baseline measurements of serum phosphate at the time of presentation, when prerenal azotemia can obscure significant phosphate depletion. Other factors that may contribute to such acute redistributive hypophosphatemia include antecedent starvation or malnutrition, administration of IV glucose without other nutrients, elevated blood catecholamines (endogenous or exogenous), respiratory alkalosis, and recovery from metabolic acidosis.

Hypophosphatemia also can occur transiently (over weeks to months) during the phase of accelerated net bone formation that follows parathyroidectomy for severe primary hyperparathyroidism or during treatment of vitamin D deficiency or lytic Paget's disease. This is usually most prominent in patients who preoperatively have evidence of high bone turnover (e.g., high serum levels of alkaline phosphatase). Osteoblastic metastases can also lead to this syndrome.

Clinical and Laboratory Findings The clinical manifestations of severe hypophosphatemia reflect a generalized defect in cellular energy metabolism because of ATP depletion, a shift from oxidative phosphorylation toward glycolysis, and associated tissue or organ dysfunction. Acute, severe hypophosphatemia occurs mainly or exclusively in hospitalized patients with underlying serious medical or surgical illness and preexisting phosphate depletion due to excessive urinary losses, severe malabsorption, or malnutrition. Chronic hypophosphatemia tends to be less severe, with a clinical presentation dominated by musculoskeletal complaints such as bone pain, osteomalacia, pseudofractures, and proximal muscle weakness or, in children, rickets and short stature.

Neuromuscular manifestations of severe hypophosphatemia are variable but may include muscle weakness, lethargy, confusion, disorientation, hallucinations, dysarthria, dysphagia, oculomotor palsies, anisocoria, nystagmus, ataxia, cerebellar tremor, ballismus, hyporeflexia, impaired sphincter control, distal sensory deficits, paresthesia, hyperesthesia, generalized or Guillain-Barré–like ascending paralysis, seizures, coma, and even death. Serious sequelae such as paralysis, confusion, and seizures are likely only at phosphate concentrations <0.25 mmol/L (<0.8 mg/dL). Rhabdomyolysis may develop during rapidly progressive hypophosphatemia. The diagnosis of hypophosphatemia-induced rhabdomyolysis may be overlooked, as up to 30% of patients with acute hypophosphatemia (<0.7 m*M*) have creatine phosphokinase elevations that peak 1–2 days after the nadir in serum phosphate, when the release of phosphate from injured myocytes may have led to a near normalization of circulating levels of phosphate.

Respiratory failure and cardiac dysfunction, which are reversible with phosphate treatment, may occur at serum phosphate levels of 0.5–0.8 mmol/L (1.5–2.5 mg/dL). Renal tubular defects, including tubular acidosis, glycosuria, and impaired reabsorption of sodium and calcium, may occur. Hematologic abnormalities correlate with reductions in intracellular ATP and 2,3-diphosphoglycerate and may include erythrocyte microspherocytosis and hemolysis; impaired oxyhemoglobin dissociation; defective leukocyte chemotaxis, phagocytosis, and bacterial killing; and platelet dysfunction with spontaneous gastrointestinal hemorrhage.

TREATMENT

Hypophosphatemia

Severe hypophosphatemia (<0.75 mmol/L [<2 mg/dL]), particularly in the setting of underlying phosphate depletion, constitutes a dangerous electrolyte abnormality that should be corrected promptly. Unfortunately, the cumulative deficit in body phosphate cannot be predicted easily from knowledge of the circulating level of phosphate, and therapy must be approached empirically. The threshold for IV phosphate therapy and the dose administered should reflect consideration of renal function, the likely severity and duration of the underlying phosphate depletion, and the presence and severity of symptoms consistent with those of hypophosphatemia. In adults, phosphate may be safely administered IV as neutral mixtures of sodium or potassium phosphate salts at initial doses of 0.2–0.8 mmol/kg of elemental phosphorus over 6 h (e.g., 10–50 mmol over 6 h), with doses >20 mmol/6 h reserved for those who have serum levels <0.5 mmol/L (1.5 mg/dL) and normal renal function. A suggested approach is presented in **Table 402-2**. Serum levels of phosphate and calcium must be monitored closely (every 6–12 h) throughout treatment. It is necessary to avoid a serum calcium-phosphorus product >50 to reduce the risk of heterotopic calcification. Hypocalcemia, if present, should be corrected before administering IV phosphate. Less severe hypophosphatemia, in the range of 0.5–0.8 mmol/L (1.5–2.5 mg/dL), usually can be treated with oral phosphate in divided doses of 750–2000 mg/d as elemental phosphorus; higher doses can cause bloating and diarrhea.

Management of chronic hypophosphatemia requires knowledge of the cause(s) of the disorder. Hypophosphatemia related to the secondary hyperparathyroidism of vitamin D deficiency usually

TABLE 402-2 Intravenous Therapy for Hypophosphatemia

Consider
Likely severity of underlying phosphate depletion
Concurrent parenteral glucose administration
Presence of neuromuscular, cardiopulmonary, or hematologic complications of hypophosphatemia
Renal function (reduce dose by 50% if serum creatinine >220 μmol/L [>2.5 mg/dL])
Serum calcium level (correct hypocalcemia first; reduce dose by 50% in hypercalcemia)
Guidelines

SERUM PHOSPHORUS, MM (MG/DL)	RATE OF INFUSION, MMOL/H	DURATION, H	TOTAL ADMINISTERED, MMOL
<0.8 (<2.5)	2	6	12
<0.5 (<1.5)	4	6	24
<0.3 (<1)	8	6	48

Rates shown are calculated for a 70-kg person; levels of serum calcium and phosphorus must be measured every 6–12 h during therapy; infusions can be repeated to achieve stable serum phosphorus levels >0.8 mmol/L (>2.5 mg/dL); most formulations available in the United States provide 3 mmol/mL of sodium or potassium phosphate.

TABLE 402-3 Causes of Hyperphosphatemia

I. Impaired renal phosphate excretion
- A. Renal insufficiency
- B. Hypoparathyroidism
  1. Developmental
  2. Autoimmune
  3. After neck surgery or radiation
  4. Activating mutations of the calcium-sensing receptor
- C. Parathyroid suppression
  1. Parathyroid-independent hypercalcemia
     a. Vitamin D or vitamin A intoxication
     b. Sarcoidosis, other granulomatous diseases
     c. Immobilization, osteolytic metastases
     d. Milk-alkali syndrome
  2. Severe hypermagnesemia or hypomagnesemia
- D. Pseudohypoparathyroidism
- E. Acromegaly
- F. Tumoral calcinosis
- G. Heparin therapy

II. Massive extracellular fluid phosphate loads
- A. Rapid administration of exogenous phosphate (intravenous, oral, rectal)
- B. Extensive cellular injury or necrosis
  1. Crush injuries
  2. Rhabdomyolysis
  3. Hyperthermia
  4. Fulminant hepatitis
  5. Cytotoxic therapy
  6. Severe hemolytic anemia
- C. Transcellular phosphate shifts
  1. Metabolic acidosis
  2. Respiratory acidosis

responds to treatment with vitamin D and calcium alone. XLH, ADHR, TIO, and related renal tubular disorders usually are managed with divided oral doses of phosphate, often with calcium and $1,25(OH)_2D$ supplements to bypass the block in renal $1,25(OH)_2D$ synthesis and prevent secondary hyperparathyroidism caused by suppression of ECF calcium levels. Thiazide diuretics may be used to prevent nephrocalcinosis in patients who are managed this way. Complete normalization of hypophosphatemia is generally not possible in these conditions. Optimal therapy for TIO is extirpation of the responsible tumor, which may be localized by radiographic skeletal survey or bone scan (many are located in bone) or by radionuclide scanning using sestamibi or labeled octreotide. Successful treatment of TIO-induced hypophosphatemia with octreotide has been reported in a small number of patients.

HYPERPHOSPHATEMIA

Causes When the filtered load of phosphate and glomerular filtration rate (GFR) are normal, control of serum phosphate levels is achieved by adjusting the rate at which phosphate is reabsorbed by the proximal tubular NaPi-2 co-transporters. The principal hormonal regulators of NaPi-2 activity are PTH and FGF23. Hyperphosphatemia, defined in adults as a fasting serum phosphate concentration >1.8 mmol/L (5.5 mg/dL), usually results from impaired glomerular filtration, hypoparathyroidism, excessive delivery of phosphate into the ECF (from bone, gut, or parenteral phosphate therapy), or a combination of these factors **(Table 402-3)**. The upper limit of normal serum phosphate concentrations is higher in children and neonates (2.4 mmol/L [7 mg/dL]). It is useful to distinguish hyperphosphatemia caused by impaired renal phosphate excretion from that which results from excessive delivery of phosphate into the ECF (Table 402-3).

In chronic renal insufficiency, reduced GFR leads to phosphate retention. Hyperphosphatemia in turn further impairs renal synthesis of $1,25(OH)_2D$, increases FGF23 levels, and stimulates PTH secretion and hypertrophy both directly and indirectly (by lowering blood ionized calcium levels). Thus, hyperphosphatemia is a major cause of the secondary hyperparathyroidism of renal failure and must be addressed early in the course of the disease **(Chaps. 305 and 403)**.

Hypoparathyroidism leads to hyperphosphatemia via increased expression of NaPi-2 co-transporters in the proximal tubule. Hypoparathyroidism, or parathyroid suppression, has multiple potential causes, including autoimmune disease; developmental, surgical, or radiation-induced absence of functional parathyroid tissue; vitamin D intoxication or other causes of PTH-independent hypercalcemia; cellular PTH resistance (pseudohypoparathyroidism or hypomagnesemia); infiltrative disorders such as Wilson's disease and hemochromatosis; and impaired PTH secretion caused by hypermagnesemia, severe hypomagnesemia, or activating mutations in the CaSR. Hypocalcemia may also contribute directly to impaired phosphate clearance, as calcium infusion can induce phosphaturia in hypoparathyroid subjects. Increased tubular phosphate reabsorption also occurs in acromegaly, during heparin administration, and in tumoral calcinosis. Tumoral calcinosis is caused by a rare group of genetic disorders in which FGF23 is processed in a way that leads to low levels of active FGF23 in the bloodstream. This may result from mutations in the FGF23 sequence or via inactivating mutations in the *GALNT3* gene, which encodes a galactosaminyl transferase that normally adds sugar residues to FGF23 that slow its proteolysis. A similar syndrome results from FGF23 resistance due to inactivating mutations of the FGF23 co-receptor Klotho. These abnormalities cause elevated serum $1,25(OH)_2D$, parathyroid suppression, increased intestinal calcium absorption, and focal hyperostosis with large, lobulated periarticular heterotopic ossifications (especially at shoulders or hips) and are accompanied by hyperphosphatemia. In some forms of tumoral calcinosis serum phosphorus levels are normal.

When large amounts of phosphate are delivered rapidly into the ECF, hyperphosphatemia can occur despite normal renal function. Examples include overzealous IV phosphate therapy, oral or rectal administration of large amounts of phosphate-containing laxatives or enemas (especially in children), extensive soft tissue injury or necrosis (crush injuries, rhabdomyolysis, hyperthermia, fulminant hepatitis, cytotoxic chemotherapy), extensive hemolytic anemia, and transcellular phosphate shifts induced by severe metabolic or respiratory acidosis.

Clinical Findings The clinical consequences of acute, severe hyperphosphatemia are due mainly to the formation of widespread calcium phosphate precipitates and resulting hypocalcemia. Thus, tetany, seizures, accelerated nephrocalcinosis (with renal failure,

hyperkalemia, hyperuricemia, and metabolic acidosis), and pulmonary or cardiac calcifications (including development of acute heart block) may occur. The severity of these complications relates to the elevation of serum phosphate levels, which can reach concentrations as high as 7 mmol/L (20 mg/dL) in instances of massive soft tissue injury or tumor lysis syndrome.

TREATMENT

Hyperphosphatemia

Therapeutic options for management of severe hyperphosphatemia are limited. Volume expansion may enhance renal phosphate clearance. Aluminum hydroxide antacids or sevalamer may be helpful in chelating and limiting absorption of offending phosphate salts present in the intestine. Hemodialysis is the most effective therapeutic strategy and should be considered early in the course of severe hyperphosphatemia, especially in the setting of renal failure and symptomatic hypocalcemia.

MAGNESIUM METABOLISM

Magnesium is the major intracellular divalent cation. Normal concentrations of extracellular magnesium and calcium are crucial for normal neuromuscular activity. Intracellular magnesium forms a key complex with ATP and is an important cofactor for a wide range of enzymes, transporters, and nucleic acids required for normal cellular function, replication, and energy metabolism. The concentration of magnesium in serum is closely regulated within the range of 0.7–1 mmol/L (1.5–2 meq/L; 1.7–2.4 mg/dL), of which 30% is protein-bound and another 15% is loosely complexed to phosphate and other anions. One-half of the 25 g (1000 mmol) of total body magnesium is located in bone, only one-half of which is insoluble in the mineral phase. Almost all extraskeletal magnesium is present within cells, where the total concentration is 5 m*M*, 95% of which is bound to proteins and other macromolecules. Because only 1% of body magnesium resides in the ECF, measurements of serum magnesium levels may not accurately reflect the level of total body magnesium stores.

Dietary magnesium content normally ranges from 6 to 15 mmol/d (140–360 mg/d), of which 30–40% is absorbed, mainly in the jejunum and ileum. Intestinal magnesium absorptive efficiency is stimulated by $1,25(OH)_2D$ and can reach 70% during magnesium deprivation. Urinary magnesium excretion normally matches net intestinal absorption and is ~4 mmol/d (100 mg/d). Regulation of serum magnesium concentrations is achieved mainly by control of renal magnesium reabsorption. Only 20% of filtered magnesium is reabsorbed in the proximal tubule, whereas 60% is reclaimed in the cTAL and another 5–10% in the DCT. Magnesium reabsorption in the cTAL occurs via a paracellular route that requires both a lumen-positive potential, created by NaCl reabsorption, and tight-junction proteins encoded by members of the Claudin gene family. Magnesium reabsorption in the cTAL is increased by PTH but inhibited by hypercalcemia or hypermagnesemia, both of which activate the CaSR in this nephron segment.

HYPOMAGNESEMIA

Causes Hypomagnesemia usually signifies substantial depletion of body magnesium stores (0.5–1 mmol/kg). Hypomagnesemia can result from intestinal malabsorption; protracted vomiting, diarrhea, or intestinal drainage; defective renal tubular magnesium reabsorption; or rapid shifts of magnesium from the ECF into cells, bone, or third spaces **(Table 402-4)**. Dietary magnesium deficiency is unlikely except possibly in the setting of alcoholism. A rare genetic disorder that causes selective intestinal magnesium malabsorption has been described (primary infantile hypomagnesemia). Another rare inherited disorder (hypomagnesemia with secondary hypocalcemia) is caused by mutations in the gene encoding TRPM6, a protein that, along with TRPM7, forms a channel important for both intestinal and distal-tubular renal transcellular magnesium transport. Malabsorptive states, often compounded by vitamin D deficiency, can critically limit magnesium absorption and produce hypomagnesemia despite the compensatory effects of secondary hyperparathyroidism and of hypocalcemia and hypomagnesemia to enhance cTAL magnesium reabsorption. Diarrhea or surgical drainage fluid may contain ≥5 mmol/L of magnesium. Proton pump inhibitors (omeprazole and others) may produce hypomagnesemia by an unknown mechanism that does not involve renal wasting of magnesium.

Several genetic magnesium-wasting syndromes have been described, including inactivating mutations of genes encoding the DCT NaCl co-transporter (Gitelman's syndrome), proteins required for cTAL Na-K-2Cl transport (Bartter's syndrome), claudin 16 or claudin

TABLE 402-4 Causes of Hypomagnesemia

- I. Impaired intestinal absorption
  - A. Hypomagnesemia with secondary hypocalcemia (TRPM6 mutations)
  - B. Malabsorption syndromes
  - C. Vitamin D deficiency
  - D. Proton pump inhibitors
- II. Increased intestinal losses
  - A. Protracted vomiting/diarrhea
  - B. Intestinal drainage, fistulas
- III. Impaired renal tubular reabsorption
  - A. Genetic magnesium-wasting syndromes
    1. Gitelman's syndrome
    2. Bartter's syndrome
    3. Claudin 16 or 19 mutations
    4. Potassium channel mutations (Kv1.1, Kir4.1)
    5. Na^+,K^+-ATPase γ-subunit mutations (FXYD2)
  - B. Acquired renal disease
    1. Tubulointerstitial disease
    2. Postobstruction, ATN (diuretic phase)
    3. Renal transplantation
  - C. Drugs and toxins
    1. Ethanol
    2. Diuretics (loop, thiazide, osmotic)
    3. Cisplatin
    4. Pentamidine, foscarnet
    5. Cyclosporine
    6. Aminoglycosides, amphotericin B
    7. Cetuximab
  - D. Other
    1. Extracellular fluid volume expansion
    2. Hyperaldosteronism
    3. SIADH
    4. Diabetes mellitus
    5. Hypercalcemia
    6. Phosphate depletion
    7. Metabolic acidosis
    8. Hyperthyroidism
- IV. Rapid shifts from extracellular fluid
  - A. Intracellular redistribution
    1. Recovery from diabetic ketoacidosis
    2. Refeeding syndrome
    3. Correction of respiratory acidosis
    4. Catecholamines
  - B. Accelerated bone formation
    1. Post-parathyroidectomy
    2. Treatment of vitamin D deficiency
    3. Osteoblastic metastases
  - C. Other
    1. Pancreatitis, burns, excessive sweating
    2. Pregnancy (third trimester) and lactation

Abbreviations: ATN, acute tubular necrosis; SIADH, syndrome of inappropriate antidiuretic hormone.

19 (autosomal recessive renal hypomagnesemia with hypercalciuria), a DCT Na^+,K^+-ATPase γ-subunit (autosomal dominant renal hypomagnesemia with hypocalciuria), DCT K+ channels (Kv1.1, Kir4.1) and a mitochondrial gene encoding a tRNA. Activating mutations of the CaSR can cause hypomagnesemia as well as hypocalcemia. ECF expansion, hypercalcemia, and severe phosphate depletion may impair magnesium reabsorption, as can various forms of renal injury, including those caused by drugs such as cisplatin, cyclosporine, aminoglycosides, and pentamidine as well as the EGF receptor inhibitory antibody, cetuximab (EGF action is required for normal DCT apical expression of TRPM6) (Table 402-4). A rising blood concentration of ethanol directly impairs tubular magnesium reabsorption, and persistent glycosuria with osmotic diuresis leads to magnesium wasting and probably contributes to the high frequency of hypomagnesemia in poorly controlled diabetic patients. Magnesium depletion is aggravated by metabolic acidosis, which causes intracellular losses as well.

Hypomagnesemia due to rapid shifts of magnesium from ECF into the intracellular compartment can occur during recovery from diabetic ketoacidosis, starvation, or respiratory acidosis. Less acute shifts may be seen during rapid bone formation after parathyroidectomy, with treatment of vitamin D deficiency, or with osteoblastic metastases. Large amounts of magnesium may be lost with acute pancreatitis, extensive burns, or protracted and severe sweating and during pregnancy and lactation.

Clinical and Laboratory Findings Hypomagnesemia may cause generalized alterations in neuromuscular function, including tetany, tremor, seizures, muscle weakness, ataxia, nystagmus, vertigo, apathy, depression, irritability, delirium, and psychosis. Patients are usually asymptomatic when serum magnesium concentrations are >0.5 mmol/L (1 meq/L; 1.2 mg/dL), although the severity of symptoms may not correlate with serum magnesium levels. Cardiac arrhythmias may occur, including sinus tachycardia, other supraventricular tachycardias, and ventricular arrhythmias. Electrocardiographic abnormalities may include prolonged PR or QT intervals, T-wave flattening or inversion, and ST straightening. Sensitivity to digitalis toxicity may be enhanced.

Other electrolyte abnormalities often seen with hypomagnesemia, including hypocalcemia (with hypocalciuria) and hypokalemia, may not be easily corrected unless magnesium is administered as well. The hypocalcemia may be a result of concurrent vitamin D deficiency, although hypomagnesemia can cause impaired synthesis of $1,25(OH)_2D$, cellular resistance to PTH, and, at very low serum magnesium (<0.4 mmol/L [<0.8 meq/L; <1 mg/dL]), a defect in PTH secretion; these abnormalities are reversible with therapy.

TREATMENT

Hypomagnesemia

Mild, asymptomatic hypomagnesemia may be treated with oral magnesium salts ($MgCl_2$, MgO, $Mg[OH]_2$) in divided doses totaling 20–30 mmol/d (40–60 meq/d). Diarrhea may occur with larger doses. More severe hypomagnesemia should be treated parenterally, preferably with IV $MgCl_2$, which can be administered safely as a continuous infusion of 50 mmol/d (100 meq Mg^{2+}/d) if renal function is normal. If GFR is reduced, the infusion rate should be lowered by 50–75%. Use of IM $MgSO_4$ is discouraged; the injections are painful and provide relatively little magnesium (2 mL of 50% $MgSO_4$ supplies only 4 mmol). $MgSO_4$ may be given IV instead of $MgCl_2$, although the sulfate anions may bind calcium in serum and urine and aggravate hypocalcemia. Serum magnesium should be monitored at intervals of 12–24 h during therapy, which may continue for several days because of impaired renal conservation of magnesium (only 50–70% of the daily IV magnesium dose is retained) and delayed repletion of intracellular deficits, which may be as high as 1–1.5 mmol/kg (2–3 meq/kg).

It is important to consider the need for calcium, potassium, and phosphate supplementation in patients with hypomagnesemia. Vitamin D deficiency frequently coexists and should be treated with oral or parenteral vitamin D or 25(OH)D (but not with $1,25(OH)_2D$, which may impair tubular magnesium reabsorption, possibly via PTH suppression). In severely hypomagnesemic patients with concomitant hypocalcemia and hypophosphatemia, administration of IV magnesium alone may worsen hypophosphatemia, provoking neuromuscular symptoms or rhabdomyolysis, due to rapid stimulation of PTH secretion. This is avoided by administering both calcium and magnesium.

HYPERMAGNESEMIA

Causes Hypermagnesemia is rarely seen in the absence of renal insufficiency, as normal kidneys can excrete large amounts (250 mmol/d) of magnesium. Mild hypermagnesemia due to excessive reabsorption in the cTAL occurs with CaSR mutations in familial hypocalciuric hypercalcemia and has been described in some patients with adrenal insufficiency, hypothyroidism, or hypothermia. Massive exogenous magnesium exposures, usually via the gastrointestinal tract, can overwhelm renal excretory capacity and cause life-threatening hypermagnesemia (Table 402-5). A notable example of this is prolonged retention of even normal amounts of magnesium-containing cathartics in patients with intestinal ileus, obstruction, or perforation. Extensive soft tissue injury or necrosis can also deliver large amounts of magnesium into the ECF in patients who have suffered trauma, shock, sepsis, cardiac arrest, or severe burns.

Clinical and Laboratory Findings The most prominent clinical manifestations of hypermagnesemia are vasodilation and neuromuscular blockade, which may appear at serum magnesium concentrations >2 mmol/L (>4 meq/L; >4.8 mg/dL). Hypotension that is refractory to vasopressors or volume expansion may be an early sign. Nausea, lethargy, and weakness may progress to respiratory failure, paralysis, and coma, with hypoactive tendon reflexes, at serum magnesium levels >4 mmol/L. Other findings may include gastrointestinal hypomotility or ileus; facial flushing; pupillary dilation; paradoxical bradycardia; prolongation of PR, QRS, and QT intervals; heart block; and, at serum magnesium levels approaching 10 mmol/L, asystole.

Hypermagnesemia, acting via the CaSR, causes hypocalcemia and hypercalciuria due to both parathyroid suppression and impaired cTAL calcium reabsorption.

TREATMENT

Hypermagnesemia

Successful treatment of hypermagnesemia generally involves identifying and interrupting the source of magnesium and employing measures to increase magnesium clearance from the ECF. Use of magnesium-free cathartics or enemas may be helpful in clearing ingested magnesium from the gastrointestinal tract. Vigorous IV

TABLE 402-5 Causes of Hypermagnesemia

I.	Excessive magnesium intake
	A. Cathartics, urologic irrigants
	B. Parenteral magnesium administration
II.	Rapid mobilization from soft tissues
	A. Trauma, shock, sepsis
	B. Cardiac arrest
	C. Burns
III.	Impaired magnesium excretion
	A. Renal failure
	B. Familial hypocalciuric hypercalcemia
IV.	Other
	A. Adrenal insufficiency
	B. Hypothyroidism
	C. Hypothermia

hydration should be attempted, if appropriate. Hemodialysis is effective and may be required in patients with significant renal insufficiency. Calcium, administered IV in doses of 100–200 mg over 1–2 h, has been reported to provide temporary improvement in signs and symptoms of hypermagnesemia.

VITAMIN D

SYNTHESIS AND METABOLISM

1,25-dihydroxyvitamin D [1,25(OH)$_2$D] is the major steroid hormone involved in regulation of mineral ion homeostasis. Vitamin D and its metabolites are hormones and hormone precursors rather than vitamins, since in the proper biologic setting, they can be synthesized endogenously (Fig. 402-4). In response to ultraviolet radiation of the skin, a photochemical cleavage results in the formation of vitamin D from 7-dehydrocholesterol. Cutaneous production of vitamin D is decreased by melanin and high solar protection factor sunblocks, which effectively impair skin penetration by ultraviolet light. The increased use of sunblocks in North America and Western Europe and a reduction in the magnitude of solar exposure of the general population over the last several decades has led to an increased reliance on dietary sources of vitamin D. In the United States and Canada, these sources largely consist of fortified cereals and dairy products, in addition to fish oils and egg yolks. Vitamin D from plant sources is in the form of vitamin D$_2$, whereas that from animal sources is vitamin D$_3$. These two forms have equivalent biologic activity and are activated equally well by the vitamin D hydroxylases in humans. Vitamin D enters the circulation, whether absorbed from the intestine or synthesized cutaneously, bound to vitamin D–binding protein, an α-globulin synthesized in the liver. Vitamin D is subsequently 25-hydroxylated in the liver by a cytochrome P450 oxidase in the mitochondria and microsomes. The activity of this hydroxylase is not tightly regulated, and the resultant metabolite, 25-hydroxyvitamin D [25(OH)D], is the major circulating and storage form of vitamin D. Approximately 88% of 25(OH)D circulates bound to the vitamin D–binding protein, 0.03% is free, and the rest circulates bound to albumin. The half-life of 25(OH)D is ~2–3 weeks, with that of 25(OH)D$_2$ being shorter than that of 25(OH)D$_3$ due to a lower affinity of vitamin D-binding protein for the former. The half-life of 25(OH)D is also greatly shortened when vitamin D–binding protein levels are reduced, as can occur with increased urinary losses in the nephrotic syndrome.

The second hydroxylation, required for the formation of the mature hormone, occurs in the kidney (Fig. 402-5). The 25-hydroxyvitamin D-1α-hydroxylase is a tightly regulated cytochrome P450–like mixed-function oxidase expressed in the proximal convoluted tubule cells of the kidney. PTH and hypophosphatemia are the major inducers of this microsomal enzyme in the kidney, whereas calcium, FGF23, and the enzyme's product, 1,25(OH)$_2$D, repress it. The 25-hydroxyvitamin

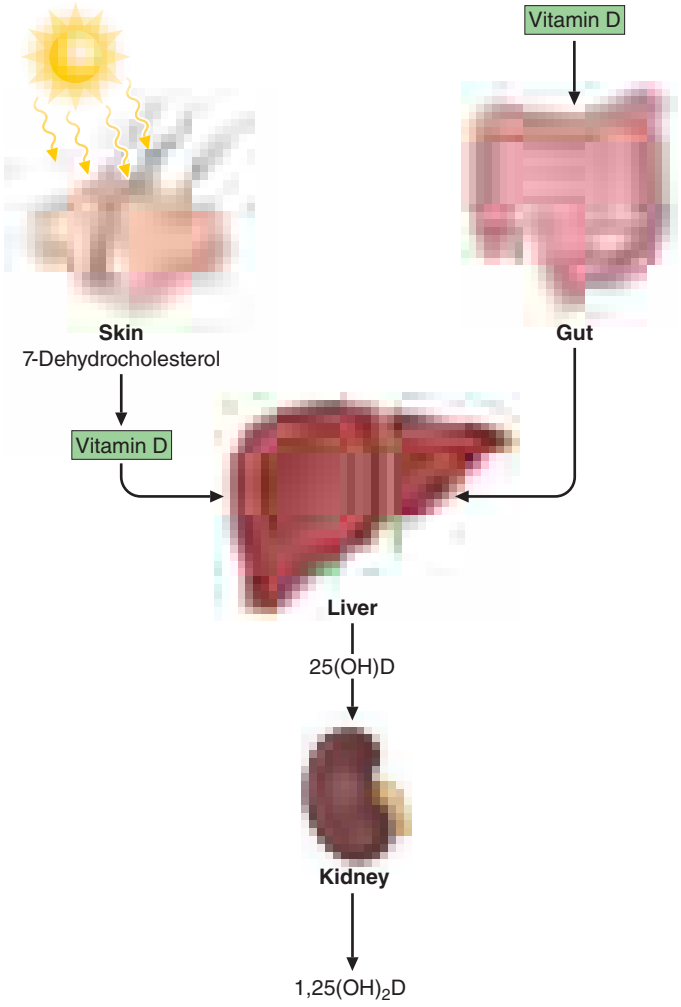


FIGURE 402-4 Vitamin D synthesis and activation. Vitamin D is synthesized in the skin in response to ultraviolet radiation and also is absorbed from the diet. It is then transported to the liver, where it undergoes 25-hydroxylation. This metabolite is the major circulating form of vitamin D. The final step in hormone activation, 1α-hydroxylation, occurs in the kidney.

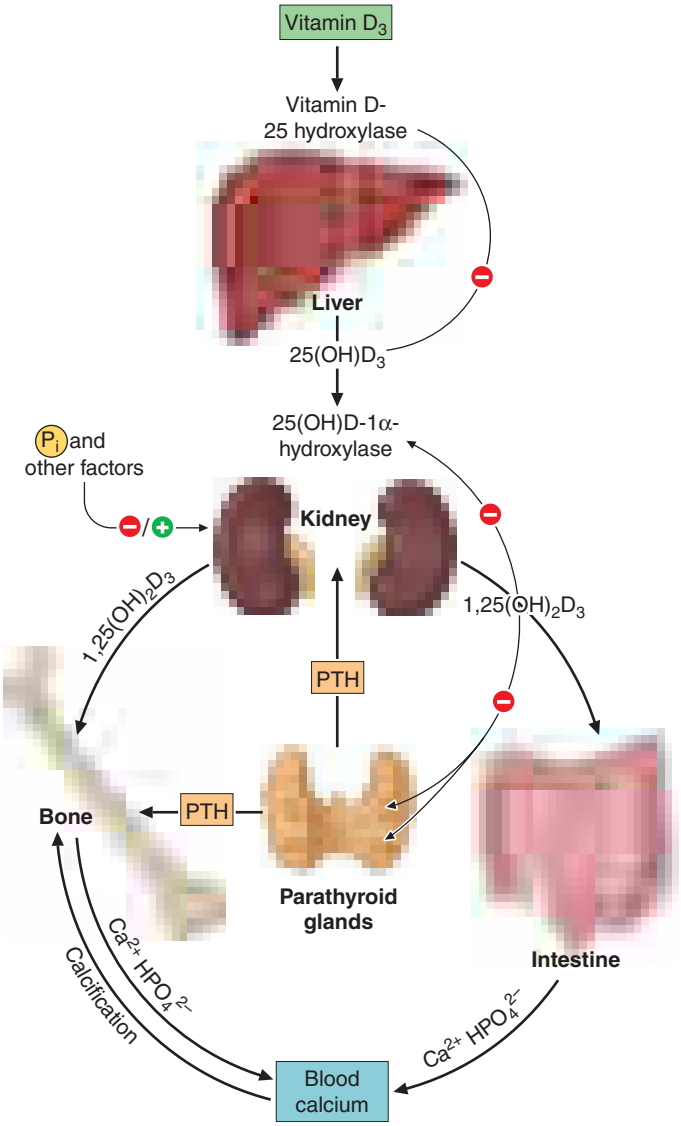


FIGURE 402-5 Schematic representation of the hormonal control loop for vitamin D metabolism and function. A reduction in the serum calcium below ~2.2 mmol/L (8.8 mg/dL) prompts a proportional increase in the secretion of parathyroid hormone (PTH) and so mobilizes additional calcium from the bone. PTH promotes the synthesis of 1,25(OH)$_2$D in the kidney, which in turn stimulates the mobilization of calcium from bone and intestine and regulates the synthesis of PTH by negative feedback.

D-1α-hydroxylase is also present in numerous other cell types, where it is not subject to hormonal regulation. It is expressed in epidermal keratinocytes, but keratinocyte production of $1,25(OH)_2D$ is not thought to contribute to circulating levels of this hormone. In addition to being present in the trophoblastic layer of the placenta, the 1α-hydroxylase is produced by macrophages associated with granulomas and lymphomas. In these latter pathologic states, the activity of the enzyme is induced by interferon γ and TNF-α but is not regulated by calcium or $1,25(OH)_2D$; therefore, hypercalcemia, associated with elevated levels of $1,25(OH)_2D$, may be observed. Treatment of sarcoidosis-associated hypercalcemia with glucocorticoids, ketoconazole, or chloroquine reduces $1,25(OH)_2D$ production and effectively lowers serum calcium. In contrast, chloroquine has not been shown to lower the elevated serum $1,25(OH)_2D$ levels in patients with lymphoma.

The major pathway for inactivation of vitamin D metabolites is an additional hydroxylation step by the vitamin D 24-hydroxylase, an enzyme that is expressed in most tissues. $1,25(OH)_2D$ is the major inducer of this enzyme; therefore, this hormone promotes its own inactivation, thereby limiting its biologic effects. FGF23 also induces this hydroxylase, thereby reducing circulating $1,25(OH)_2D$ levels by increasing its inactivation, as well as by impairing its synthesis. Mutations of the gene encoding this enzyme (CYP24 A1) can lead to infantile hypercalcemia and, in those less severely affected, long-standing hypercalciuria, nephrocalcinosis and nephrolithiasis can occur.

Polar metabolites of $1,25(OH)_2D$ are secreted into the bile and reabsorbed via the enterohepatic circulation. Impairment of this recirculation, which is seen with diseases of the terminal ileum, leads to accelerated losses of vitamin D metabolites.

ACTIONS OF $1,25(OH)_2D$

$1,25(OH)_2D$ mediates its biologic effects by binding to a member of the nuclear receptor superfamily, the vitamin D receptor (VDR). This receptor belongs to the subfamily that includes the thyroid hormone receptors, the retinoid receptors, and the peroxisome proliferator–activated receptors; however, in contrast to the other members of this subfamily, only one VDR isoform has been isolated. The VDR binds to target DNA sequences as a heterodimer with the retinoid X receptor, recruiting a series of coactivators that modify chromatin and approximate the VDR to the basal transcriptional apparatus, resulting in the induction of target gene expression. The mechanism of transcriptional repression by the VDR varies with different target genes but has been shown to involve either interference with the action of activating transcription factors or the recruitment of novel proteins to the VDR complex, resulting in transcriptional repression.

The affinity of the VDR for $1,25(OH)_2D$ is approximately three orders of magnitude higher than that for other vitamin D metabolites. In normal physiologic circumstances, these other metabolites are not thought to stimulate receptor-dependent actions. However, in states of vitamin D toxicity, the markedly elevated levels of 25(OH)D may lead to hypercalcemia by interacting directly with the VDR and by displacing $1,25(OH)_2D$ from vitamin D–binding protein, resulting in increased bioavailability of the active hormone.

The VDR is expressed in a wide range of cells and tissues. The molecular actions of $1,25(OH)_2D$ have been studied most extensively in tissues involved in the regulation of mineral ion homeostasis. This hormone is a major inducer of calbindin 9K, a calcium-binding protein expressed in the intestine, which is thought to play an important role in the active transport of calcium across the enterocyte. The two major calcium transporters expressed by intestinal epithelia, TRPV5 and TRPV6 (transient receptor potential vanilloid), are also vitamin D responsive. By inducing the expression of these and other genes in the small intestine, $1,25(OH)_2D$ increases the efficiency of intestinal calcium absorption, and it also has been shown to have several important actions in the skeleton. The VDR is expressed in osteoblasts and regulates the expression of several genes in this cell. These genes include the bone matrix proteins, osteocalcin and osteopontin, which are upregulated by $1,25(OH)_2D$, in addition to type I collagen, which is transcriptionally repressed by $1,25(OH)_2D$. Both $1,25(OH)_2D$ and PTH induce the expression of RANK ligand, which promotes osteoclast differentiation and increases osteoclast activity, by binding to RANK on osteoclast progenitors and mature osteoclasts. This is the mechanism by which $1,25(OH)_2D$ induces bone resorption. $1,25(OH)_2D$ regulates phosphate homeostasis, primarily by inducing the expression of FGF23 in osteocytes. However, the skeletal features associated with VDR-knockout mice (rickets, osteomalacia) are largely corrected by increasing calcium and phosphorus intake, underscoring the importance of vitamin D action in the gut.

The VDR is expressed in the parathyroid gland, and $1,25(OH)_2D$ has been shown to have antiproliferative effects on parathyroid cells and to suppress the transcription of the parathyroid hormone gene. These effects of $1,25(OH)_2D$ on the parathyroid gland are an important part of the rationale for current therapies directed at preventing and treating hyperparathyroidism associated with renal insufficiency.

The VDR is also expressed in tissues and organs that do not play a role in mineral ion homeostasis. Notable in this respect is the observation that $1,25(OH)_2D$ has an antiproliferative effect on several cell types, including keratinocytes, breast cancer cells, and prostate cancer cells. The effects of $1,25(OH)_2D$ and the VDR on keratinocytes are particularly intriguing, since the VDR is primarily a transcriptional repressor in these cells. Alopecia is seen in humans and mice with mutant VDRs but is not a feature of vitamin D deficiency; thus, the effects of the VDR on the hair follicle are ligand-independent.

VITAMIN D DEFICIENCY

The mounting concern about the relationship between solar exposure and the development of skin cancer has led to increased reliance on dietary sources of vitamin D. Although the prevalence of vitamin D deficiency varies, the third National Health and Nutrition Examination Survey (NHANES III) revealed that vitamin D deficiency is prevalent throughout the United States, the prevalence being >29% in obese children. The clinical syndrome of vitamin D deficiency can be a result of deficient production of vitamin D in the skin, lack of dietary intake, accelerated losses of vitamin D, impaired vitamin D activation, or resistance to the biologic effects of $1,25(OH)_2D$ **(Table 402-6)**. The elderly and nursing home residents are particularly at risk for vitamin D deficiency, since both the efficiency of vitamin D synthesis in the skin and the absorption of vitamin D from the intestine decline with age. The presence of terminal ileal disease also results in impaired enterohepatic circulation of vitamin D metabolites. While intestinal malabsorption of dietary fats and short bowel syndrome, including that associated with intestinal bypass surgery, lead to vitamin D deficiency, the cause of vitamin D deficiency in obese individuals is poorly understood. In addition to intestinal diseases, accelerated inactivation of vitamin D metabolites can be seen with drugs that induce hepatic cytochrome P450 mixed-function oxidases such as barbiturates, phenytoin, and rifampin. Impaired 25-hydroxylation, associated with severe liver disease or isoniazid, is an uncommon cause of vitamin D deficiency. A mutation in the gene responsible for 25-hydroxylation has been identified in a few kindreds. Impaired 1α-hydroxylation is prevalent in the population with profound renal dysfunction due to an increase in circulating FGF23 levels. Thus, therapeutic interventions should be considered in

TABLE 402-6 Causes of Impaired Vitamin D Action

Vitamin D deficiency	Impaired 1α-hydroxylation
Impaired cutaneous production	Hypoparathyroidism
Dietary absence	Ketoconazole
Malabsorption	1α-hydroxylase mutation
Accelerated loss of vitamin D	FGF23 excess
Increased metabolism (barbiturates, phenytoin, rifampin)	Oncogenic osteomalacia
Impaired enterohepatic circulation	X-linked hypophosphatemic rickets
Nephrotic syndrome	Renal Failure
Impaired 25-hydroxylation	Target Organ Resistance
Liver disease, isoniazid	Vitamin D receptor mutation
25-hydroxylase mutation	Phenytoin
	Other
	Obesity

patients whose creatinine clearance is <0.5 mL/s (30 mL/min). Mutations in the renal 1α-hydroxylase are the basis for the genetic disorder, pseudovitamin D–deficiency rickets. This autosomal recessive disorder presents with the syndrome of vitamin D deficiency in the first year of life. Patients present with growth retardation, rickets, and hypocalcemic seizures. Serum $1,25(OH)_2D$ levels are low despite normal 25(OH)D levels and elevated PTH levels. Treatment with vitamin D metabolites that do not require 1α-hydroxylation for activity results in disease remission, although lifelong therapy is required. A second autosomal recessive disorder, hereditary vitamin D–resistant rickets, a consequence of vitamin D receptor mutations, is a greater therapeutic challenge. These patients present in a similar fashion during the first year of life, but alopecia often accompanies the disorder, demonstrating a functional role of the VDR in the keratinocyte stem cell population required for hair follicle regeneration. Serum levels of $1,25(OH)_2D$ are dramatically elevated in these individuals both because of increased production due to stimulation of 1α-hydroxylase activity as a consequence of secondary hyperparathyroidism and because of impaired inactivation, since induction of the 24-hydroxylase by $1,25(OH)_2D$ requires an intact VDR. Since the receptor mutation results in hormone resistance, daily calcium and phosphorus infusions may be required to bypass the defect in intestinal mineral ion absorption.

Regardless of the cause, the clinical manifestations of vitamin D deficiency are largely a consequence of impaired intestinal calcium absorption. Mild to moderate vitamin D deficiency is asymptomatic, whereas long-standing vitamin D deficiency results in hypocalcemia accompanied by secondary hyperparathyroidism, impaired mineralization of the skeleton (osteopenia on X-ray or decreased bone mineral density), and proximal myopathy. Vitamin D deficiency also has been shown to be associated with an increase in overall mortality, including cardiovascular causes. In the absence of an intercurrent illness, the hypocalcemia associated with long-standing vitamin D deficiency rarely presents with acute symptoms of hypocalcemia such as numbness, tingling, and seizures. However, the concurrent development of hypomagnesemia, which impairs parathyroid function, or the administration of potent bisphosphonates, which impair bone resorption, can lead to acute symptomatic hypocalcemia in vitamin D–deficient individuals.

Rickets and Osteomalacia In children, before epiphyseal fusion, vitamin D deficiency results in growth retardation associated with an expansion of the growth plate known as *rickets*. Three layers of chondrocytes are present in the normal growth plate: the reserve zone, the proliferating zone, and the hypertrophic zone. Rickets associated with impaired vitamin D action is characterized by expansion of the hypertrophic chondrocyte layer. The proliferation and differentiation of the chondrocytes in the rachitic growth plate are normal, and the expansion of the growth plate is a consequence of impaired apoptosis of the late hypertrophic chondrocytes, an event that precedes replacement of these cells by osteoblasts during endochondral bone formation. Investigations in murine models demonstrate that hypophosphatemia, which in vitamin D deficiency is a consequence of secondary hyperparathyroidism, is a key etiologic factor in the development of the rachitic growth plate.

The hypocalcemia and hypophosphatemia that accompany vitamin D deficiency result in impaired mineralization of bone matrix proteins, a condition known as *osteomalacia*. Osteomalacia is also a feature of long-standing hypophosphatemia, which may result from renal phosphate wasting, or chronic use of etidronate or phosphate-binding antacids. This hypomineralized matrix is biomechanically inferior to normal bone; as a result, patients with osteomalacia are prone to bowing of weight-bearing extremities and skeletal fractures. Vitamin D and calcium supplementation have been shown to decrease the incidence of hip fracture among ambulatory nursing home residents in France, suggesting that undermineralization of bone contributes significantly to morbidity in the elderly. Proximal myopathy is a striking feature of severe vitamin D deficiency both in children and in adults. Rapid resolution of the myopathy is observed upon vitamin D treatment.

Though vitamin D deficiency is the most common cause of rickets and osteomalacia, many disorders lead to inadequate mineralization of the growth plate and bone. Calcium deficiency without vitamin D deficiency, the disorders of vitamin D metabolism previously discussed, and hypophosphatemia can all lead to inefficient mineralization. Even in the presence of normal calcium and phosphate levels, chronic acidosis and drugs such as bisphosphonates can lead to osteomalacia. The inorganic calcium/phosphate mineral phase of bone cannot form at low pH. Bisphosphonates bind to and prevent hydroxyapatite crystal growth. Since alkaline phosphatase is necessary for normal mineral deposition, probably because the enzyme can hydrolyze inhibitors of mineralization such as inorganic pyrophosphate, genetic inactivation of the alkaline phosphatase gene (hereditary hypophosphatasia) also can lead to osteomalacia in the setting of normal calcium and phosphate levels.

Diagnosis of Vitamin D Deficiency, Rickets, and Osteomalacia The most specific screening test for vitamin D deficiency in otherwise healthy individuals is a serum 25(OH)D level. Although the normal ranges vary, levels of 25(OH)D <37 nmol/L (<15 ng/mL) are associated with increasing PTH levels and lower bone density. The National Academy of Medicine has defined vitamin D sufficiency as a vitamin D level >50 nmol/L (>20 ng/mL), although higher levels may be required to optimize intestinal calcium absorption in the elderly and those with underlying disease states, including obesity. Vitamin D deficiency leads to impaired intestinal absorption of calcium, resulting in decreased serum total and ionized calcium values. This hypocalcemia results in secondary hyperparathyroidism, a homeostatic response that initially maintains serum calcium levels at the expense of the skeleton. Due to the PTH-induced increase in bone turnover, alkaline phosphatase levels are often increased. In addition to increasing bone resorption, PTH decreases urinary calcium excretion while promoting phosphaturia. This results in hypophosphatemia, which exacerbates the mineralization defect in the skeleton. With prolonged vitamin D deficiency resulting in osteomalacia, calcium stores in the skeleton become relatively inaccessible, since osteoclasts cannot resorb unmineralized osteoid, and frank hypocalcemia ensues. Since PTH is a major stimulus for the renal 25(OH)D 1α-hydroxylase, there is increased synthesis of the active hormone, $1,25(OH)_2D$. Paradoxically, levels of this hormone are often normal in severe vitamin D deficiency. Therefore, measurements of $1,25(OH)_2D$ are not accurate reflections of vitamin D stores and should not be used to diagnose vitamin D deficiency in patients with normal renal function.

Radiologic features of vitamin D deficiency in children include a widened, expanded growth plate that is characteristic of rickets. These findings not only are apparent in the long bones but also are present at the costochondral junction, where the expansion of the growth plate leads to swellings known as the "rachitic rosary." Impairment of intramembranous bone mineralization leads to delayed fusion of the calvarial sutures and a decrease in the radiopacity of cortical bone in the long bones. If vitamin D deficiency occurs after epiphyseal fusion, the main radiologic finding is a decrease in cortical thickness and relative radiolucency of the skeleton. A specific radiologic feature of osteomalacia, whether associated with phosphate wasting or vitamin D deficiency, is pseudofractures, or Looser's zones. These are radiolucent lines that occur where large arteries are in contact with the underlying skeletal elements; it is thought that the arterial pulsations lead to the radiolucencies. As a result, these pseudofractures are usually a few millimeters wide, are several centimeters long, and are seen particularly in the scapula, the pelvis, and the femoral neck.

TREATMENT

Vitamin D Deficiency

Based on the National Academy of Medicine 2010 report, the recommended daily intake of vitamin D is 600 IU from 1 to 70 years of age, and 800 IU for those over 70. Based on the observation that 800 IU of vitamin D, with calcium supplementation, decreases the risk of hip fractures in elderly women, this higher dose is thought to be an appropriate daily intake for prevention of vitamin D deficiency in adults. The safety margin for vitamin D is large, and vitamin D toxicity usually is observed only in patients taking doses in the

range of 40,000 IU daily. Treatment of vitamin D deficiency should be directed at the underlying disorder, if possible, and also should be tailored to the severity of the condition. Vitamin D should always be repleted in conjunction with calcium supplementation since most of the consequences of vitamin D deficiency are a result of impaired mineral ion homeostasis. In patients in whom 1α-hydroxylation is impaired, metabolites that do not require this activation step are the treatment of choice. They include $1,25(OH)_2D_3$ (calcitriol [Rocaltrol], 0.25–0.5 μg/d) and 1α-hydroxyvitamin D_2 (doxercalciferol [Hectorol], 2.5–5 μg/d). If the pathway required for activation of vitamin D is intact, severe vitamin D deficiency can be treated with pharmacologic repletion initially (50,000 IU weekly for 3–12 weeks), followed by maintenance therapy (800 IU daily). Pharmacologic doses may be required for maintenance therapy in patients who are taking medications such as barbiturates or phenytoin, that accelerate metabolism of, or cause resistance to $1,25(OH)_2D$. Polymorphisms in the 25 hydroxylase and the 24 hydroxylase genes can also lead to different responses to the normal recommended daily intake of vitamin D. Calcium supplementation should include 1.5–2 g/d of elemental calcium. Normocalcemia is usually observed within 1 week of the institution of therapy, although increases in PTH and alkaline phosphatase levels may persist for 3–6 months. The most efficacious methods to monitor treatment and resolution of vitamin D deficiency are serum and urinary calcium measurements. In patients who are vitamin D replete and are taking adequate calcium supplementation, the 24-h urinary calcium excretion should be in the range of 100–250 mg/24 h. Lower levels suggest problems with adherence to the treatment regimen or with absorption of calcium or vitamin D supplements. Levels >250 mg/24 h predispose to nephrolithiasis and should lead to a reduction in vitamin D dosage and/or calcium supplementation.

FURTHER READING

Bikle D et al: Vitamin D metabolites in captivity? Should we measure free or total 25(OH)D to assess vitamin D status? J Steroid Biochem Mol Biol 173:105, 2017.

Dallas SL et al: The osteocyte: An endocrine cell and more. Endocr Rev 34:658, 2013.

de Baaij JH et al: Magnesium in man: Implications for health and disease. Physiol Rev 95:1, 2015.

Felsenfeld AJ, Levine BS: Approach to treatment of hypophosphatemia. Am J Kidney Dis 60:655, 2012.

Feng X, Teitelbaum SL: Osteoclasts: New insights. Bone Res 1:11, 2013.

Heaney RP, Holick MF: Why the IOM recommendations for vitamin D are deficient. J Bone Miner Res 26:455, 2011.

Institute of Medicine (IOM): *Dietary Reference Intakes for Calcium and Vitamin D*. Washington, DC: National Academies Press, 2011.

Siddiqui JA, Partridge NC: Physiological bone remodeling: Systemic regulation and growth factor involvement. Physiology 31:233, 2016.

Zofkova I: Hypercalcemia. Pathophysiological aspects. Phys Res 65:1, 2016.

403 Disorders of the Parathyroid Gland and Calcium Homeostasis

John T. Potts, Jr., Harald W. Jüppner

The four parathyroid glands are located posterior to the thyroid gland. They produce parathyroid hormone (PTH), which is the primary regulator of calcium physiology. PTH acts directly on bone, where it induces calcium release, and on the kidney, where it enhances calcium reabsorption in the distal tubules, and in the proximal renal tubules the synthesis of 1,25-dihydroxyvitamin D ($1,25[OH]_2D$), a hormone that increases gastrointestinal calcium absorption. Serum PTH levels are tightly regulated by a negative feedback loop. Calcium, acting through the calcium-sensing receptor (CaSR), and vitamin D, acting through its nuclear receptor, reduce PTH release and synthesis. Additional evidence indicates that fibroblast growth factor 23 (FGF23), a phosphaturic hormone, can suppress PTH secretion. Understanding the hormonal pathways that regulate calcium levels and bone metabolism is essential for effective diagnosis and management of a wide array of hyper- and hypocalcemic disorders.

Hyperparathyroidism (HPT), characterized by excess production of PTH, is a common cause of hypercalcemia and is usually the result of autonomously functioning adenomas or hyperplasia. Surgery for this disorder is highly effective and has been shown to reverse some of the deleterious effects of long-standing PTH excess on bone density. Humoral hypercalcemia of malignancy is also common and is usually due to the overproduction of parathyroid hormone–related peptide (PTHrP) by cancer cells. The similarities in the biochemical characteristics of HPT and humoral hypercalcemia of malignancy, first noted by Albright in 1941, are now known to reflect the actions of PTH and PTHrP through the same G protein–coupled PTH/PTHrP receptor.

The genetic basis of multiple endocrine neoplasia (MEN) types 1 and 2, familial hypocalciuric hypercalcemia (FHH), different forms of pseudohypoparathyroidism (PHP), Jansen's syndrome, disorders of vitamin D synthesis and action, and the molecular events associated with parathyroid gland neoplasia have provided new insights into the regulation of calcium homeostasis. PTH and *possibly some* of its analogues are promising therapeutic agents for the treatment of postmenopausal or senile osteoporosis, and calcimimetic agents, which activate the CaSR, have provided new approaches for PTH suppression.

PARATHYROID HORMONE

PHYSIOLOGY

The primary function of PTH is to maintain the extracellular fluid (ECF) calcium concentration within a narrow normal range. The hormone acts directly on bone and kidney and indirectly on the intestine through its effects on synthesis of $1,25(OH)_2D$ to increase serum calcium concentrations; in turn, PTH production is closely regulated by the concentration of serum ionized calcium. This feedback system is the critical homeostatic mechanism for maintenance of ECF calcium. Any tendency toward hypocalcemia, as might be induced by calcium- or vitamin D-deficient diets, is counteracted by an increased secretion of PTH. This in turn (1) increases the rate of dissolution of bone mineral, thereby increasing the flow of calcium from bone into blood; (2) reduces the renal clearance of calcium, returning more of the calcium and phosphate filtered at the glomerulus into ECF; (3) increases the efficiency of calcium absorption in the intestine by stimulating the production of $1,25(OH)_2D$. Immediate control of blood calcium is due to PTH effects on bone and, to a lesser extent, on renal calcium clearance. Maintenance of steady-state calcium balance, on the other hand, probably results from the effects of $1,25(OH)_2D$ on calcium absorption **(Chap. 402)**. The renal actions of the hormone are exerted at multiple sites and include inhibition of phosphate transport (proximal tubule), augmentation of calcium reabsorption (distal tubule), and stimulation of the renal 25(OH)D-1α-hydroxylase. As much as 12 mmol (500 mg) calcium is transferred between the ECF and bone each day (a large amount in relation to the total ECF calcium pool), and PTH has a major effect on this transfer. The homeostatic role of the hormone can preserve calcium concentration in blood at the cost of bone demineralization.

PTH has multiple actions on bone, some direct and some indirect. PTH-mediated changes in bone calcium release can be seen within minutes. The chronic effects of PTH are to increase the number of bone cells, both osteoblasts and osteoclasts, and to increase the remodeling of bone; these effects are apparent within hours after the hormone is given and persist for hours after PTH is withdrawn. Continuous exposure to elevated PTH (as in HPT or long-term infusions in animals) leads to increased osteoclast-mediated bone resorption. However, the intermittent administration of PTH, elevating hormone levels for 1–2 hours each

day, leads to a net stimulation of bone formation rather than bone breakdown. Striking increases, especially in trabecular bone in the spine and hip, have been reported with the use of PTH in combination with estrogen. PTH(1–34) as monotherapy caused a highly significant reduction in fracture incidence in a worldwide placebo-controlled trial.

Osteoblasts (or stromal cell precursors), which have PTH/PTHrP receptors, are crucial to this bone-forming effect of PTH; osteoclasts, which mediate bone breakdown, lack such receptors. PTH-mediated stimulation of osteoclasts is indirect, acting in part, through cytokines released from osteoblasts to activate osteoclasts; in experimental studies of bone resorption in vitro, osteoblasts must be present for PTH to activate osteoclasts to resorb bone **(Chap. 402)**.

STRUCTURE

PTH is an 84-amino-acid single-chain peptide. The amino-terminal portion, PTH(1–34), is highly conserved and is critical for the biologic actions of the molecule. Modified synthetic fragments of the amino-terminal sequence as small as PTH(1–11) are sufficient to activate the PTH/PTHrP receptor (see below). The carboxyl-terminal regions of the full-length PTH(1–84) molecule also can bind to a separate binding protein/receptor (cPTH-R), but this receptor has been incompletely characterized. Fragments shortened at the amino-terminus possibly by binding to cPTH-R can reduce, directly or indirectly, some of the biologic actions of full-length PTH(1–84) and of PTH(1–34).

BIOSYNTHESIS, SECRETION, AND METABOLISM

Synthesis Parathyroid cells have multiple methods of adapting to increased needs for PTH production. Most rapid (within minutes) is secretion of preformed hormone in response to hypocalcemia. Second, within hours, PTH mRNA expression is induced by sustained hypocalcemia. Finally, protracted challenge leads within days to cellular replication to increase parathyroid gland mass.

PTH is initially synthesized as a larger molecule (preproparathyroid hormone, consisting of 115 amino acids). After a first cleavage step to remove the "pre" sequence of 25 amino acid residues, a second cleavage step removes the "pro" sequence of 6 amino acid residues before secretion of the mature peptide comprising 84 residues. Mutations in the preprotein region of the gene can cause hypoparathyroidism by interfering with hormone synthesis, transport, or secretion.

Transcriptional suppression of the PTH gene by calcium is nearly maximal at physiologic calcium concentrations. Hypocalcemia increases transcriptional activity within hours. $1,25(OH)_2D$ strongly suppresses PTH gene transcription. In patients with renal failure, IV administration of supraphysiologic levels of $1,25(OH)_2D$ or analogues of this active metabolite can dramatically suppress PTH overproduction, which is sometimes difficult to control due to severe secondary HPT. Regulation of proteolytic destruction of preformed hormone (posttranslational regulation of hormone production) is an important mechanism for mediating rapidly (within minutes) changes in hormone availability. High calcium increases and low calcium inhibits the proteolytic destruction of stored hormone.

Regulation of PTH Secretion PTH secretion increases steeply to a maximum value of about five times the basal rate of secretion as the calcium concentration falls from normal to the range of 1.9–2.0 mmol/L (7.6–8.0 mg/dL; measured as total calcium). However, the ionized fraction of blood calcium is the important determinant of hormone secretion. Severe intracellular magnesium deficiency impairs PTH secretion (see below).

ECF calcium controls PTH secretion by interaction with a CaSR, a G protein–coupled receptor (GPCR) for which Ca^{2+} ions act as the primary ligand (see below). This receptor is a member of a distinctive subgroup of the GPCR superfamily that mediates its actions through two related signaling G proteins, namely Gq and G11, is characterized by a large extracellular domain suitable for "clamping" the small-molecule ligand. Stimulation of the CaSR by high calcium levels suppresses PTH secretion. The CaSR is present in parathyroid glands and the calcitonin-secreting cells of the thyroid (C cells), as well as in multiple other sites, including brain and kidney. Genetic evidence has revealed a key biologic role for the CaSR in parathyroid gland responsiveness to calcium and in renal calcium clearance. Heterozygous loss-of-function mutations in CaSR cause the syndrome of FHH, in which the blood calcium abnormality resembles that observed in HPT but with hypocalciuria; two more recently defined variants of FHH, FHH2, and FHH3, are caused either by heterozygous mutations in G11, one of the signaling proteins down-stream of the CaSR, or by heterozygous mutations in *AP2A1*. Homozygous loss-of-function mutations in the CaSR are the cause of severe neonatal HPT, a disorder that can be lethal if not treated within the first days of life. On the other hand, heterozygous gain-of-function mutations cause a form of hypocalcemia resembling hypoparathyroidism (see below).

Metabolism The secreted form of PTH is indistinguishable by immunologic criteria and by molecular size from the 84-amino-acid peptide (PTH[1–84]) extracted from glands. However, much of the immunoreactive material found in the circulation is smaller than the extracted or secreted hormone. The principal circulating fragments of immunoreactive hormone lack a portion of the critical amino-terminal sequence required for biologic activity and, hence, are biologically inactive fragments (so-called middle and carboxyl-terminal fragments). Much of the proteolysis of hormone occurs in the liver and kidney. Peripheral metabolism of PTH does not appear to be regulated by physiologic states (high versus low calcium, etc.); hence, peripheral metabolism of hormone, although responsible for rapid clearance of secreted hormone, appears to be a high-capacity, metabolically invariant catabolic process.

The rate of clearance of the secreted 84-amino-acid peptide from blood is more rapid than the rate of clearance of the biologically inactive fragment(s) corresponding to the middle and carboxyl-terminal regions of PTH. Consequently, the interpretation of results obtained with earlier PTH radioimmunoassays was influenced by the nature of the peptide fragments detected by the antibodies.

Although the problems inherent in PTH measurements have been largely circumvented by use of double-antibody immunometric assays, it is now known that some of these assays detect, besides the intact molecule, large amino-terminally truncated forms of PTH, which are present in normal and uremic individuals in addition to PTH(1–84). The concentration of these fragments relative to that of intact PTH(1–84) is higher with induced hypercalcemia than in eucalcemic or hypocalcemic conditions and is higher in patients with impaired renal function. PTH(7–84) has been identified as a major component of these amino-terminally truncated fragments. Growing evidence suggests that the PTH(7–84) (and probably related amino-terminally truncated fragments) can act, through yet undefined mechanisms, as an inhibitor of PTH action and may be of clinical significance, particularly in patients with chronic kidney disease (CKD). In this group of patients, efforts to prevent secondary HPT by a variety of measures (vitamin D analogues, higher calcium intake, higher dialysate calcium, phosphate-lowering strategies, and calcimetic drugs) can lead to oversuppression of the parathyroid glands since some amino-terminally truncated PTH fragments, such as PTH(7–84), react in many immunometric PTH assays (now termed second-generation assays; see below under "Diagnosis"), thus overestimating the levels of biologically active, intact PTH. Such excessive parathyroid gland suppression in CKD can lead to adynamic bone disease (see below), which has been associated in children with further impaired growth and increased bone fracture rates in adults, and can furthermore lead to significant hypercalcemia. The measurement of PTH with newer third-generation immunometric assays, which use detection antibodies directed against extreme amino-terminal PTH epitopes and thus detect only full-length PTH(1–84), may provide some advantage to prevent bone disease in CKD.

PARATHYROID HORMONE–RELATED PROTEIN (PTHrP)

PTHrP is responsible for most instances of humoral hypercalcemia of malignancy **(Chap. 89)**, a syndrome that resembles primary HPT but without elevated PTH levels. Most cell types normally produce PTHrP, including brain, pancreas, heart, lung, mammary tissue, placenta,

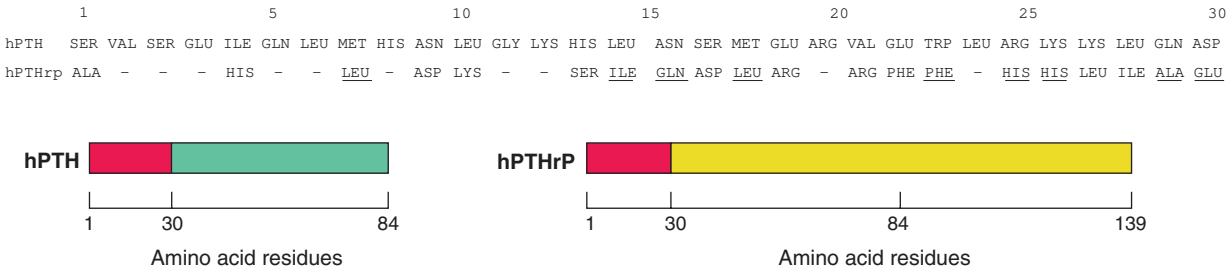


FIGURE 403-1 Schematic diagram to illustrate similarities and differences in structure of human parathyroid hormone (PTH) and human PTH-related peptide (PTHrP). Close structural (and functional) homology exists between the first 30 amino acids of hPTH and hPTHrP (red area). The PTHrP sequence may be ≥139 amino acid residues in length. PTH is only 84 residues long; after residue 30, there is little structural homology between the two. Dashed lines in the PTHrP sequence indicate identity; underlined residues, although different from those of PTH, still represent conservative changes (charge or polarity preserved). Ten amino acids are identical, and a total of 20 of 30 are homologues.

endothelial cells, and smooth muscle. In fetal animals, PTHrP directs transplacental calcium transfer, and high concentrations of PTHrP are produced in mammary tissue and secreted into milk, but the biologic significance of the very high concentrations of this hormone in breast milk is unknown. PTHrP also plays an essential role in endochondral bone formation and in branching morphogenesis of the breast, and possibly in uterine contraction and other biologic functions.

PTH and PTHrP, although products of different genes, exhibit considerable functional and structural homology **(Fig. 403-1)** and have evolved from a shared ancestral gene. The structure of the gene encoding human PTHrP, however, is more complex than that of PTH, containing multiple additional exons, which can undergo alternate splicing patterns during formation of the mature mRNA. Protein products of 139, 141, and 173 amino acids are produced, and other molecular forms may result from tissue-specific degradation at accessible internal cleavage sites. The biologic roles of these various molecular species and the nature of the circulating forms of PTHrP are unclear. In fact, it is uncertain whether PTHrP circulates at any significant level in adults. As a paracrine factor, PTHrP may be produced, act, and be destroyed locally within tissues. In adults, PTHrP appears to have little influence on calcium homeostasis, except in disease states, when large tumors, especially of the squamous cell type as well as renal cell carcinomas, lead to massive overproduction of the hormone and hypercalcemia.

PTH AND PTHrP HORMONE ACTION

Both PTH and PTHrP bind to and activate the PTH/PTHrP receptor. The PTH/PTHrP receptor (also known as the PTH-1 receptor, PTH1R) belongs to a subfamily of GPCRs that includes the receptors for calcitonin, glucagon, secretin, vasoactive intestinal peptide, and other peptides. Although both ligands activate the PTH1R, the two peptides induce distinct responses in the receptor, which explains how a single receptor without isoforms can serve two biologic roles. The extracellular regions of the receptor are involved in hormone binding, and the intracellular domains, after hormone activation, bind G protein subunits to transduce hormone signaling into cellular responses through the stimulation of second messenger formation. A second receptor that binds PTH, originally termed the *PTH-2 receptor* (PTH2R), is primarily expressed in brain, pancreas, and testis. Different mammalian PTH1Rs respond equivalently to PTH and PTHrP, at least when tested with traditional assays, whereas only the human PTH2R responds efficiently to PTH (but not to PTHrP). PTH2Rs from other species show little or no stimulation of second-messenger formation in response to PTH or PTHrP. The endogenous ligand of the PTH2R was shown to be a hypothalamic peptide referred to as tubular infundibular peptide of 39 residues, TIP39, that is distantly related to PTH and PTHrP. The PTH1R and the PTH2R can be traced backward in evolutionary time to fish; in fact, the zebrafish genome contains, in addition to the PTH1R and the PTH2R orthologs, a third receptor, the PTH3R, that is more closely related to the fish PTH1R than to the fish PTH2R. The evolutionary conservation of structure and function suggests important biologic roles for these receptors, even in fish, which lack discrete parathyroid glands but produce two molecules that are closely related to mammalian PTH.

Studies using the cloned PTH1R confirm that it can be coupled to more than one G protein and second-messenger pathway, apparently explaining the multiplicity of pathways stimulated by PTH. Activation of protein kinases (A and C) and calcium transport channels is associated with a variety of hormone-specific tissue responses. These responses include inhibition of phosphate and bicarbonate transport, stimulation of calcium transport, and activation of renal 1α-hydroxylase in the kidney. The responses in bone include effects on collagen synthesis, alkaline phosphatase, ornithine decarboxylase, citrate decarboxylase, and glucose-6-phosphate dehydrogenase activities, phospholipid synthesis, as well as calcium and phosphate transport. Ultimately, these biochemical events lead to an integrated hormonal response in bone turnover and calcium homeostasis. PTH also activates Na^+/Ca^{2+} exchangers at renal distal tubular sites and stimulates translocation of preformed calcium transport channels, moving them from the interior to the apical surface to increase tubular uptake of calcium. PTH-dependent stimulation of phosphate excretion (reducing reabsorption—the opposite effect from actions on calcium in the kidney) involves the down-regulation of two sodium-dependent phosphate co-transporters, NPT2a and NPT2c, and their expression at the apical membrane, thereby reducing phosphate reabsorption in the proximal renal tubules. Similar mechanisms may be involved in other renal tubular transporters that are influenced by PTH. Recent studies reaffirm the critical linkage of blood phosphate lowering to net calcium entry into blood by PTH action and emphasize the participation of bone cells other than osteoclasts in the rapid calcium elevating actions of PTH.

PTHrP exerts important developmental influences on fetal bone development and in adult physiology. A homozygous ablation of the gene encoding PTHrP (or disruption of the PTH1R gene) in mice causes a lethal phenotype in which animals are born with pronounced acceleration of chondrocyte maturation that resembles a lethal form of chondrodysplasia in humans that is caused by homozygous or compound heterozygous, inactivating PTH1R mutations **(Fig. 403-2)**. Heterozygous PTH1R mutations in humans furthermore can be a cause of delayed tooth eruption and mice that are heterozygous for ablation of the PTHrP gene display reduced mineral density consistent with osteoporosis. Experiments with these mouse models point to a hitherto unappreciated role of PTHrP as a paracrine/autocrine factor that modulates bone metabolism in adults as well as during bone development.

CALCITONIN

(See also Chap. 381) Calcitonin is a hypocalcemic peptide hormone that in several mammalian species acts as an indirect antagonist to the calcemic actions of PTH. Calcitonin seems to be of limited physiologic significance in humans, at least with regard to calcium homeostasis. It is of medical significance because of its role as a tumor marker in sporadic and hereditary cases of medullary carcinoma and its medical use as an adjunctive treatment in severe hypercalcemia and in Paget's disease of bone.

The hypocalcemic activity of calcitonin is accounted for primarily by inhibition of osteoclast-mediated bone resorption and secondarily by stimulation of renal calcium clearance. These effects are mediated by receptors on osteoclasts and renal tubular cells. Calcitonin exerts additional

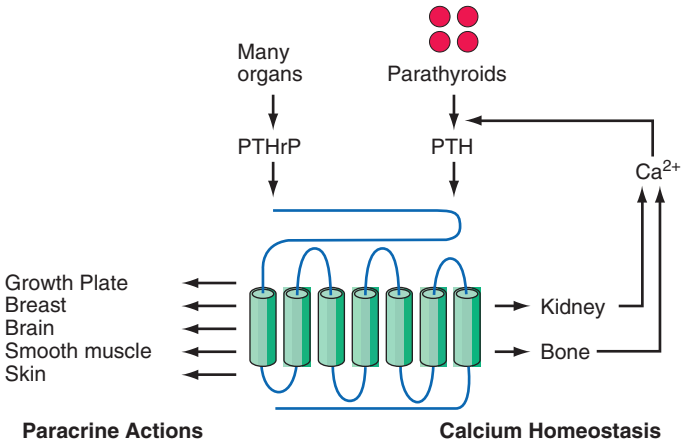


FIGURE 403-2 Dual role for the actions of the PTH/PTHrP receptor (PTH1R). Parathyroid hormone (PTH; endocrine-calcium homeostasis) and PTH-related peptide (PTHrP; paracrine–multiple tissue actions including growth plate cartilage in developing bone) use the single receptor for their disparate functions mediated by the amino-terminal 34 residues of either peptide. Other regions of both ligands interact with other receptors (not shown).

effects through receptors present in the brain, the gastrointestinal tract, and the immune system. The hormone, for example, exerts analgesic effects directly on cells in the hypothalamus and related structures, possibly by interacting with receptors for related peptide hormones such as calcitonin gene–related peptide (CGRP) or amylin. Both of these ligands have specific high-affinity receptors that share considerable structural similarity with the PTH1R and can also bind to and activate calcitonin receptors. The calcitonin receptor shares considerable structural similarity with the PTH1R.

The thyroid is the major source of the hormone, and the cells involved in calcitonin synthesis arise from neural crest tissue. During embryogenesis, these cells migrate into the ultimobranchial body, derived from the last branchial pouch. In submammalian vertebrates, the ultimobranchial body constitutes a discrete organ, anatomically separate from the thyroid gland; in mammals, the ultimobranchial gland fuses with and is incorporated into the thyroid gland.

The naturally occurring calcitonins consist of a peptide chain of 32 amino acids. There is considerable sequence variability among species. Calcitonin from salmon, which is used therapeutically, is 10–100 times more potent than mammalian forms in lowering serum calcium.

There are two calcitonin genes, α and β; the transcriptional control of these genes is complex. Two different mRNA molecules are transcribed from the α gene; one is translated into the precursor for calcitonin, and the other message is translated into an alternative product, CGRP. CGRP is synthesized wherever the calcitonin mRNA is expressed (e.g., in medullary carcinoma of the thyroid). The β, or CGRP-2, gene is transcribed into the mRNA for CGRP in the central nervous system (CNS); this gene does not produce calcitonin, however. CGRP has cardiovascular actions and may serve as a neurotransmitter or play a developmental role in the CNS.

The circulating level of calcitonin in humans is lower than that in many other species. In humans, even extreme variations in calcitonin production do not change calcium and phosphate metabolism; no definite effects are attributable to calcitonin deficiency (totally thyroidectomized patients receiving only replacement thyroxine) or excess (patients with medullary carcinoma of the thyroid, a calcitonin-secreting tumor) **(Chap. 381)**. Calcitonin has been a useful pharmacologic agent to suppress bone resorption in Paget's disease **(Chap. 405)** and osteoporosis **(Chap. 404)** and in the treatment of hypercalcemia of malignancy (see below). However, bisphosphates are usually more effective and the physiologic role, if any, of calcitonin in humans is uncertain. On the other hand, ablation of the calcitonin gene (combined because of the close proximity with ablation of the CGRP gene) in mice leads to reduced bone mineral density, suggesting that its biologic role in mammals is still not fully understood.

HYPERCALCEMIA

(See also Chap. 50) Hypercalcemia can be a manifestation of a serious illness such as malignancy or can be detected coincidentally by laboratory testing in a patient with no obvious illness. The number of patients recognized with asymptomatic hypercalcemia, usually HPT, increased in the late twentieth century.

Whenever hypercalcemia is confirmed, a definitive diagnosis must be established. Although HPT, a frequent cause of asymptomatic hypercalcemia, is a chronic disorder in which manifestations, if any, may be expressed only after months or years, hypercalcemia can also be the earliest manifestation of malignancy, the second most common cause of hypercalcemia in the adult. The causes of hypercalcemia are numerous **(Table 403-1)**, but HPT and cancer account for 90% of all cases.

Before undertaking a diagnostic workup, it is essential to be sure that true hypercalcemia, not a false-positive laboratory test, is present. A false-positive diagnosis of hypercalcemia is usually the result of inadvertent hemoconcentration during blood collection or elevation in serum proteins such as albumin. Hypercalcemia is a chronic problem, and it is cost-effective to obtain several serum calcium measurements; these tests need not be in the fasting state.

Clinical features are helpful in differential diagnosis. Hypercalcemia in an adult who is asymptomatic is usually due to primary HPT. In malignancy-associated hypercalcemia, the disease is usually not occult; rather, symptoms of malignancy bring the patient to the physician, and hypercalcemia is discovered during the evaluation. In such patients, the interval between detection of hypercalcemia and death, especially without vigorous treatment, is often <6 months. Accordingly, if an asymptomatic individual has had hypercalcemia or some manifestation of hypercalcemia such as kidney stones for >1 or 2 years, it is unlikely that malignancy is the cause. Nevertheless, differentiating primary HPT from occult malignancy can occasionally be difficult, and careful evaluation is required, particularly when the duration of the hypercalcemia is unknown. Hypercalcemia not due to HPT or malignancy can result from excessive vitamin D action, impaired metabolism of $1,25(OH)_2D$, high bone turnover from any of several causes, or from renal failure (Table 403-1). Dietary history and a history of ingestion

TABLE 403-1 Classification of Causes of Hypercalcemia

I. Parathyroid-Related
- A. Primary hyperparathyroidism
  - 1. Adenoma(s)
  - 2. Multiple endocrine neoplasia
  - 3. Carcinoma
- B. Lithium therapy
- C. Familial hypocalciuric hypercalcemia

II. Malignancy-Related
- A. Solid tumor with metastases (breast)
- B. Solid tumor with humoral mediation of hypercalcemia (lung, kidney)
- C. Hematologic malignancies (multiple myeloma, lymphoma, leukemia)

III. Vitamin D–Related
- A. Vitamin D intoxication
- B. ↑ $1,25(OH)_2D$; sarcoidosis and other granulomatous diseases
- C. ↑ $1,25(OH)_2D$; impaired $1,25(OH)_2D$ metabolism due to 24-hydroxylase deficiency and inactivating mutations in the sodium-dependent phosphate co-transporters

IV. Associated with High Bone Turnover
- A. Hyperthyroidism
- B. Immobilization
- C. Thiazides
- D. Vitamin A intoxication
- E. Fat necrosis

V. Associated with Renal Failure
- A. Severe secondary hyperparathyroidism
- B. Aluminum intoxication
- C. Milk-alkali syndrome

of vitamins or drugs are often helpful in diagnosing some of the less frequent causes. Immunometric PTH assays serve as the principal laboratory test in establishing the diagnosis.

Hypercalcemia from any cause can result in fatigue, depression, mental confusion, anorexia, nausea, vomiting, constipation, reversible renal tubular defects, increased urine output, a short QT interval in the electrocardiogram, and, in some patients, cardiac arrhythmias. There is a variable relation from one patient to the next between the severity of hypercalcemia and the symptoms. Generally, symptoms are more common at calcium levels >2.9–3.0 mmol/L (11.6–12.0 mg/dL), but some patients, even at this level, are asymptomatic. When the calcium level is >3.2 mmol/L (12.8 mg/dL), calcification in kidneys, skin, vessels, lungs, heart, and stomach occurs and renal insufficiency may develop, particularly if blood phosphate levels are normal or elevated due to impaired renal excretion. Severe hypercalcemia, usually defined as ≥3.7–4.5 mmol/L (14.8–18.0 mg/dL), can be a medical emergency; coma and cardiac arrest can occur.

Acute management of the hypercalcemia is usually successful. The type of treatment is based on the severity of the hypercalcemia and the nature of associated symptoms, as outlined below.

PRIMARY HYPERPARATHYROIDISM

Natural History and Incidence Primary HPT is a generalized disorder of calcium, phosphate, and bone metabolism due to an increased secretion of PTH. The elevation of circulating hormone usually leads to hypercalcemia and hypophosphatemia. There is great variation in the manifestations. Patients may present with multiple signs and symptoms, including recurrent nephrolithiasis, peptic ulcers, mental changes, and, less frequently, extensive bone resorption. However, with greater awareness of the disease and wider use of multiphasic screening tests, including measurements of blood calcium, the diagnosis is frequently made in patients who have no symptoms and minimal, if any, signs of the disease other than hypercalcemia and elevated levels of PTH. The manifestations may be subtle, and the disease may have a benign course for many years or a lifetime. This milder form of the disease is usually termed *asymptomatic HPT*. Rarely, HPT develops or worsens abruptly and causes severe complications such as marked dehydration and coma, so-called hypercalcemic parathyroid crisis.

The annual incidence of the disease is calculated to be as high as 0.2% in patients >60, with an estimated prevalence, including undiscovered asymptomatic patients, of ≥1%; some reports suggest the incidence may be declining. If confirmed, these changing estimates may reflect less frequent routine testing of serum calcium in recent years, earlier overestimates in incidence, or unknown factors. The disease has a peak incidence between the third and fifth decades but occurs in young children and in the elderly.

Etiology Parathyroid tumors are most often encountered as isolated adenomas without other endocrinopathy. They may also arise in hereditary syndromes such as MEN syndromes. As many as 10% of patients with HPT are found to have mutations in 1 of 11 genes (see below). Parathyroid tumors may also arise as secondary to underlying disease (excessive stimulation in secondary HPT, especially chronic renal failure), or after other forms of excessive stimulation such as lithium therapy. These etiologies are discussed below.

SOLITARY ADENOMAS A single abnormal gland is the cause in ~80% of patients; the abnormality in the gland is usually a benign neoplasm or adenoma and rarely a parathyroid carcinoma. Some surgeons and pathologists report that the enlargement of multiple glands is common; double adenomas are reported. In ~15% of patients, all glands are hyperfunctioning; *chief cell parathyroid hyperplasia* is usually hereditary and frequently associated with other endocrine abnormalities.

HEREDITARY SYNDROMES AND MULTIPLE PARATHYROID TUMORS Hereditary HPT can occur without other endocrine abnormalities but is usually part of a *multiple endocrine neoplasia* syndrome **(Chap. 381)**. MEN1 (Wermer's syndrome) consists of HPT and tumors of the pituitary and pancreas, often associated with gastric hypersecretion and peptic ulcer disease (Zollinger-Ellison syndrome). MEN2A is characterized by pheochromocytoma and medullary carcinoma of the thyroid, as well as HPT; MEN2B has additional associated features such as multiple neuromas but usually lacks HPT. Each of these MEN syndromes is transmitted in an apparent autosomal dominant manner, although, as noted below, the genetic basis of MEN1 involves biallelic loss of a tumor suppressor.

The *hyperparathyroidism jaw tumor* (HPT-JT) syndrome occurs in families with parathyroid tumors (sometimes carcinomas) in association with benign jaw tumors. This disorder is caused by mutations in *CDC73* (*HRPT2*) and mutations in this gene are also observed in parathyroid cancers. Some kindreds exhibit hereditary HPT without other endocrinopathies. This disorder is often termed *nonsyndromic familial isolated hyperparathyroidism* (FIHP). There is speculation that these families may be examples of variable expression of the other syndromes such as MEN 1, MEN 2, or the HPT-JT syndrome, but they may also have distinctive, still unidentified genetic causes. For example, different heterozygous GCM2 mutations co-segregate with the disease in several FIHP kindreds; some of these mutations enhanced activity of a GCM2-dependent reporter.

Pathology Adenomas are most often located in the inferior parathyroid glands, but in 6–10% of patients, parathyroid adenomas may be located in the thymus, the thyroid, the pericardium, or behind the esophagus. Adenomas are usually 0.5–5 g in size but may be as large as 10–20 g (normal glands weigh 25 mg on average). Chief cells are predominant in both hyperplasia and adenoma. With chief cell hyperplasia, the enlargement may be so asymmetric that some involved glands appear grossly normal. If generalized hyperplasia is present, however, histologic examination reveals a uniform pattern of chief cells and disappearance of fat even in the absence of an increase in gland weight. Thus, microscopic examination of biopsy specimens of several glands can be helpful to interpret findings at surgery.

Parathyroid carcinoma is often not aggressive. Long-term survival without recurrence is common if at initial surgery the entire gland is removed without rupture of the capsule. Recurrent parathyroid carcinoma is usually slow-growing with local spread in the neck, and surgical correction of recurrent disease may be feasible. Occasionally, however, parathyroid carcinoma is more aggressive, with distant metastases (lung, liver, and bone) found at the time of initial operation. It may be difficult to appreciate initially that a primary tumor is carcinoma; increased numbers of mitotic figures and increased fibrosis of the gland stroma may precede invasion. The diagnosis of carcinoma is often made in retrospect. HPT from a parathyroid carcinoma may be indistinguishable from other forms of primary HPT but is usually more severe clinically. A potential clue to the diagnosis is offered by the degree of calcium elevation. Calcium values of 3.5–3.7 mmol/L (14–15 mg/dL) are frequent with carcinoma and may alert the surgeon to remove the abnormal gland with care to avoid capsular rupture. Recent findings concerning the genetic basis of parathyroid carcinoma (distinct from that of benign adenomas) indicate the need, in these kindreds, for family screening (see below).

GENETIC DEFECTS ASSOCIATED WITH HPT

As in many other types of neoplasia, two fundamental types of genetic defects have been identified in parathyroid gland tumors: (1) overactivity of protooncogenes and (2) loss of function of tumor-suppressor genes. The former, by definition, can lead to uncontrolled cellular growth and function by activation (gain-of-function mutation) of a single allele of the responsible gene, whereas the latter requires loss of function of both allelic copies. Biallelic loss of function of a tumor-suppressor gene is usually characterized by a germ-line defect (all cells) and an additional somatic deletion/mutation in the tumor **(Fig. 403-3)**.

Mutations in the *MEN1* gene locus, encoding the protein MENIN, on chromosome 11q13 are responsible for causing MEN1; the normal allele of this gene fits the definition of a tumor-suppressor gene. Inheritance of one mutated allele in this hereditary syndrome, followed by loss of the other allele via somatic cell mutation, leads to monoclonal expansion and tumor development. Also, in ~15–20% of sporadic

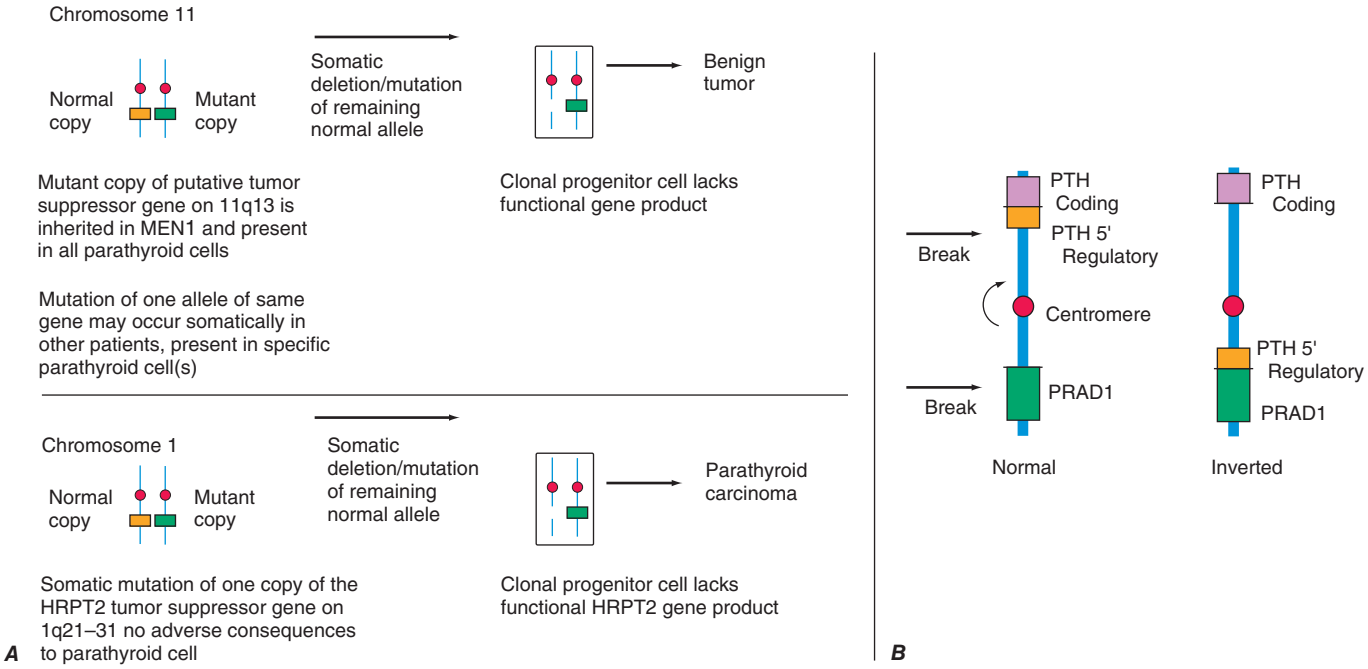


FIGURE 403-3 **A.** Schematic diagram indicating molecular events in tumor susceptibility. The patient with the hereditary abnormality (multiple endocrine neoplasia, or MEN) is envisioned as having one defective gene inherited from the affected parent on chromosome 11, but one copy of the normal gene is present from the other parent. In the monoclonal tumor (benign tumor), a somatic event, here partial chromosomal deletion, removes the remaining normal gene from a cell. In nonhereditary tumors, two successive somatic mutations must occur, a process that takes a longer time. By either pathway, the cell, deprived of growth-regulating influence from this gene, has unregulated growth and becomes a tumor. A different genetic locus also involving loss of a tumor-suppressor gene termed HRPT2 is involved in the pathogenesis of parathyroid carcinoma. *(From A Arnold: J Clin Endocrine Metab 77:1108, 1993. Copyright 1993, The Endocrine Society.)* **B.** Schematic illustration of the mechanism and consequences of gene rearrangement and overexpression of the PRAD 1 protooncogene (pericentromeric inversion of chromosome 11) in parathyroid adenomas. The excessive expression of PRAD1 (a cell cycle control protein, cyclin D1) by the highly active PTH gene promoter in the parathyroid cell contributes to excess cellular proliferation. *(From J Habener et al, in L DeGroot, JL Jameson [eds]: Endocrinology, 4th ed. Philadelphia, Saunders, 2001; with permission.)*

parathyroid adenomas, both alleles of the *MEN1* locus on chromosome 11 are somatically deleted, implying that the same defect responsible for MEN1 can also cause the sporadic disease **(Fig. 403-3*A*)**. Consistent with the Knudson hypothesis for two-step neoplasia in certain inherited cancer syndromes **(Chap. 67)**, the earlier onset of HPT in the hereditary syndromes reflects the need for only one mutational event to trigger the monoclonal outgrowth. In sporadic adenomas, typically occurring later in life, two different somatic events must occur before the *MEN1* gene is silenced.

Other presumptive anti-oncogenes involved in HPT include a still unidentified gene mapped to chromosome 1p seen in 40% of sporadic parathyroid adenomas and a gene mapped to chromosome Xp11 in patients with secondary HPT and renal failure, who progressed to "tertiary" HPT, now known to reflect monoclonal outgrowths within previously hyperplastic glands.

A more complex pattern, still incompletely resolved, arises with genetic defects and carcinoma of the parathyroids. This appears to be due to biallelic loss of a functioning copy of a gene, *HRPT2* (or CDC73), originally identified as the cause of the HPT-JT syndrome. Several inactivating mutations have been identified in *HRPT2* (located on chromosome 1q21-31), which encodes a 531-amino-acid protein called parafibromin. The responsible genetic mutations in *HRPT2* appear to be necessary, but not sufficient, for parathyroid cancer.

In general, the detection of additional genetic defects in these parathyroid tumor–related syndromes and the variations seen in phenotypic expression/penetrance indicate the multiplicity of the genetic factors responsible. Nonetheless, the ability to detect the presence of the major genetic contributors has greatly aided a more informed management of family members of patients identified in the hereditary syndromes such as MEN1, MEN2, and HPT-JT.

An important contribution from studies on the genetic origin of parathyroid carcinoma has been the realization that the mutations involve a different pathway than that involved with the benign gland enlargements. Unlike the pathogenesis of genetic alterations seen in colon cancer, where lesions evolve from benign adenomas to malignant disease by progressive genetic changes, the alterations commonly seen in most parathyroid cancers (*HRPT2* mutations) are infrequently seen in sporadic parathyroid adenomas.

Abnormalities at the *Rb* gene were the first to be noted in parathyroid cancer. The *Rb* gene, a tumor-suppressor gene located on chromosome 13q14, was initially associated with retinoblastoma but has since been implicated in other neoplasias, including parathyroid carcinoma. Early studies implicated allelic deletions of the *Rb* gene in many parathyroid carcinomas and decreased or absent expression of the Rb protein. However, because there are often large deletions in chromosome 13 that include many genes in addition to the *Rb* locus (with similar findings in some pituitary carcinomas), it remains possible that other tumor-suppressor genes on chromosome 13 may be playing a role in parathyroid carcinoma.

Study of the parathyroid cancers found in some patients with the HPT-JT syndrome has led to identification of a much larger role for mutations in the *HRPT2* gene in most parathyroid carcinomas, including those that arise sporadically, without apparent association with the HPT-JT syndrome. Mutations in the coding region have been identified in 75–80% of all parathyroid cancers analyzed, leading to the conclusion that, with addition of presumed mutations in the noncoding regions, this genetic defect may be seen in essentially all parathyroid carcinomas. Of special importance was the discovery that, in some sporadic parathyroid cancers, germ-line mutations have been found; this, in turn, has led to careful investigation of the families of these patients and a new clinical indication for genetic testing in this setting.

Hypercalcemia occurring in family members (who are also found to have the germ-line mutations) can lead to the finding, at parathyroid surgery, of premalignant parathyroid tumors.

Overall, it seems there are multiple factors in parathyroid cancer, in addition to the *HRPT2* and *Rb* gene, although the *HRPT2* gene mutation is the most invariant abnormality. *RET* encodes a tyrosine kinase type receptor; specific inherited germ-line mutations lead to a constitutive activation of the receptor, thereby explaining the autosomal dominant mode of transmission and the relatively early onset of neoplasia. In the MEN 2 syndrome, the *RET* protooncogene may be responsible for the earliest disorder detected, the polyclonal disorder (C cell hyperplasia, which then

is transformed into a clonal outgrowth—a medullary carcinoma with the participation of other, still uncharacterized genetic defects).

In some parathyroid adenomas, activation of a protooncogene has been identified **(Fig. 403-3*B*)**. A reciprocal translocation involving chromosome 11 has been identified that juxtaposes the *PTH* gene promoter upstream of a gene product termed PRAD1, encoding a cyclin D protein that plays a key role in normal cell division. This translocation plus other mechanisms that cause an equivalent overexpression of cyclin D1 are found in 20–40% of parathyroid adenomas.

Mouse models have confirmed the role of several of the major identified genetic defects in parathyroid disease and the MEN syndromes. Loss of the *MEN1* gene locus or overexpression of the PRAD1 protooncogene or the mutated *RET* protooncogene have been analyzed by genetic manipulation in mice, with the expected onset of parathyroid tumors or medullary carcinoma, respectively.

Signs and Symptoms Many patients with HPT are asymptomatic. Manifestations of HPT involve primarily the kidneys and the skeletal system. Kidney involvement, due either to deposition of calcium in the renal parenchyma or to recurrent nephrolithiasis, was present in 60–70% of patients prior to 1970. With earlier detection, renal complications occur in <20% of patients in many large series. Renal stones are usually composed of either calcium oxalate or calcium phosphate. In occasional patients, repeated episodes of nephrolithiasis or the formation of large calculi may lead to urinary tract obstruction, infection, and loss of renal function. Nephrocalcinosis may also cause decreased renal function and phosphate retention.

The distinctive bone manifestation of HPT is *osteitis fibrosa cystica*, which occurred in 10–25% of patients in series reported 50 years ago. Histologically, the pathognomonic features are an increase in the giant multinucleated osteoclasts in scalloped areas on the surface of the bone (Howship's lacunae) and a replacement of the normal cellular and marrow elements by fibrous tissue. X-ray changes include resorption of the phalangeal tufts and replacement of the usually sharp cortical outline of the bone in the digits by an irregular outline (subperiosteal resorption). In recent years, osteitis fibrosa cystica is very rare in primary HPT, probably due to the earlier detection of the disease.

Dual-energy x-ray absorptiometry (DXA) of the spine provides reproducible quantitative estimates (within a few percent) of spinal bone density. Similarly, bone density in the extremities can be quantified by densitometry of the hip or of the distal radius at a site chosen to be primarily cortical. CT is a very sensitive technique for estimating spinal bone density, but reproducibility of standard CT is no better than 5%. Newer CT techniques (spiral, "extreme" CT) are more reproducible but are currently available in a limited number of medical centers. Cortical bone density is reduced while cancellous bone density, especially in the spine, is relatively preserved. In symptomatic patients, dysfunctions of the CNS, peripheral nerve and muscle, gastrointestinal tract, and joints also occur. It has been reported that severe neuropsychiatric manifestations may be reversed by parathyroidectomy. When present in symptomatic patients, neuromuscular manifestations may include proximal muscle weakness, easy fatigability, and atrophy of muscles and may be so striking as to suggest a primary neuromuscular disorder. The distinguishing feature is the complete regression of neuromuscular disease after surgical correction of the HPT.

Gastrointestinal manifestations are sometimes subtle and include vague abdominal complaints and disorders of the stomach and pancreas. Again, cause and effect are unclear. In MEN 1 patients with HPT, duodenal ulcer may be the result of associated pancreatic tumors that secrete excessive quantities of gastrin (Zollinger-Ellison syndrome). Pancreatitis has been reported in association with HPT, but the incidence and the mechanism are not established.

Much attention has been paid in recent years to the manifestations of and optimum management strategies for asymptomatic HPT. This is now the most prevalent form of the disease. *Asymptomatic primary hyperparathyroidism* is defined as biochemically confirmed HPT (elevated or inappropriately normal PTH levels despite hypercalcemia) with the absence of signs and symptoms typically associated with more severe HPT such as features of renal or bone disease.

TABLE 403-2 Guidelines for Surgery in Asymptomatic Primary Hyperparathyroidism[a]

PARAMETER	GUIDELINE
Serum calcium (above normal)	>1 mg/dL
Renal	Creatinine clearance <60 mL/min 24-h urine for calcium >400 mg/d and increased stone risk by biochemical stone risk analysis Presence of nephrolithiasis or nephrocalcinosis by X-ray, ultrasound, or CT
Skeletal	BMD by DXA: T-score <-2.5 at lumbar spine, total hip, femoral neck, or distal 1/3 radius Vertebral fracture by X-ray, CT, MRI, or VFA
Age	<50

[a]JP Bilezikian et al: Guidelines for the management of asymptomatic primary hyperparathyroidism: Summary statement from the fourth international workshop. J Clin Endocrinol Metab 99(10):3561, 2014. Creatinine clearance calculated by Cockcroft-Gault equation or Modification of Diet in Renal Disease (MDRD) equation.

Four conferences on the topic have been held in the United States over the past two decades, with the most recent in 2013. The published proceedings include discussion of more subtle manifestations of disease, its natural history (without parathyroidectomy), and guidelines both for indications for surgery and medical monitoring in nonoperated patients.

Issues of concern include the potential for cardiovascular deterioration, the presence of subtle neuropsychiatric symptoms, and the longer-term status of skeletal integrity in patients not treated surgically. The current consensus is that medical monitoring rather than surgical correction of HPT may be justified in certain patients. The current recommendation is that patients who show mild disease, as defined by not meeting guidelines **(Table 403-2)**, can be safely followed under management guidelines **(Table 403-3)**. There is, however, growing uncertainty about subtle disease manifestations and whether surgery is therefore indicated in most patients. Among the issues is the evidence of eventual (>8 years) deterioration in bone mineral density after a decade of relative stability. There is concern that this late-onset deterioration in bone density in nonoperated patients could contribute significantly to the well-known age-dependent fracture risk (osteoporosis). Significant and sustained improvements in bone mineral density are seen after successful parathyroidectomy and some evidence for reduction in fractures.

Cardiovascular disease including left ventricular hypertrophy, cardiac functional defects, and endothelial dysfunction have been reported as reversible in European patients with more severe symptomatic disease after surgery, leading to numerous studies of these cardiovascular features in those with milder disease. There are reports of endothelial dysfunction in patients with mild asymptomatic HPT, but more observation is needed the expert panels concluded, especially whether there is reversibility with surgery.

A topic of considerable interest and some debate is assessment of neuropsychiatric status and health-related quality of life (QOL) status in hyperparathyroid patients both before surgery and in response to parathyroidectomy. Several observational studies suggest improvements in symptom score after surgery. Randomized studies of surgery

TABLE 403-3 Guidelines for Monitoring in Asymptomatic Primary Hyperparathyroidism

PARAMETER	GUIDELINE
Serum calcium	Annually
Renal	eGFR, annually; serum creatinine, annually. If renal stones suspected, 24-h biochemical stone profile, renal imaging by X-ray, ultrasound, or CT
Serum creatinine	Annually
Skeletal	Every 1-2 y (3 sites), X-ray or VFA (Vertebral Fracture Assessment) of spine if clinically indicated (e.g., height loss, back pain)

Source: JP Bilezikian et al: J Clin Endocrinol Metab 99(10):3561, 2014.

versus observation, however, have yielded inconclusive results, especially regarding benefits of surgery. Many studies report that HPT is associated with increased neuropsychiatric symptoms, but it is not possible at present to determine which patients might improve after surgery.

DIAGNOSIS

The diagnosis is typically made by detecting an elevated immunoreactive PTH level in a patient with asymptomatic hypercalcemia (see "Differential Diagnosis: Special Tests," below). Serum phosphate is usually low but may be normal, especially if renal failure has developed.

Several modifications in PTH assays have been introduced in efforts to improve their utility in light of information about metabolism of PTH (as discussed above). First-generation assays were based on displacement of radiolabeled PTH from antibodies that reacted with PTH (often also PTH fragments). Double-antibody or immunometric assays (one antibody that is usually directed against the carboxyl-terminal portion of intact PTH to capture the hormone and a second radio- or enzyme-labeled antibody that is usually directed against the amino-terminal portion of intact PTH) greatly improved the diagnostic discrimination of the tests by eliminating interference from circulating biologically inactive fragments, detected by the original first-generation assays. Double-antibody assays are now referred to as second-generation. Such PTH assays have in some centers and testing laboratories been replaced by third-generation assays after it was discovered that large PTH fragments, devoid of only the extreme amino-terminal portion of the PTH molecule, are also present in blood and are detected, incorrectly as intact PTH. These amino-terminally truncated PTH fragments were prevented from registering in the newer third-generation assays by use of a detection antibody directed against the extreme amino-terminal epitope. These assays may be useful for clinical research studies as in management of chronic renal disease, but the consensus is that either second- or third-generation assays are useful in the diagnosis of primary HPT and for the diagnosis of high-turnover bone disease in CKD.

Many tests based on renal responses to excess PTH (renal calcium and phosphate clearance; blood phosphate, chloride, magnesium; nephrogenous cyclic AMP) were used in earlier decades. These tests have low specificity for HPT and are therefore not cost-effective; they have been replaced by PTH immunometric assays combined with simultaneous blood calcium measurements **(Fig. 403-4)**.

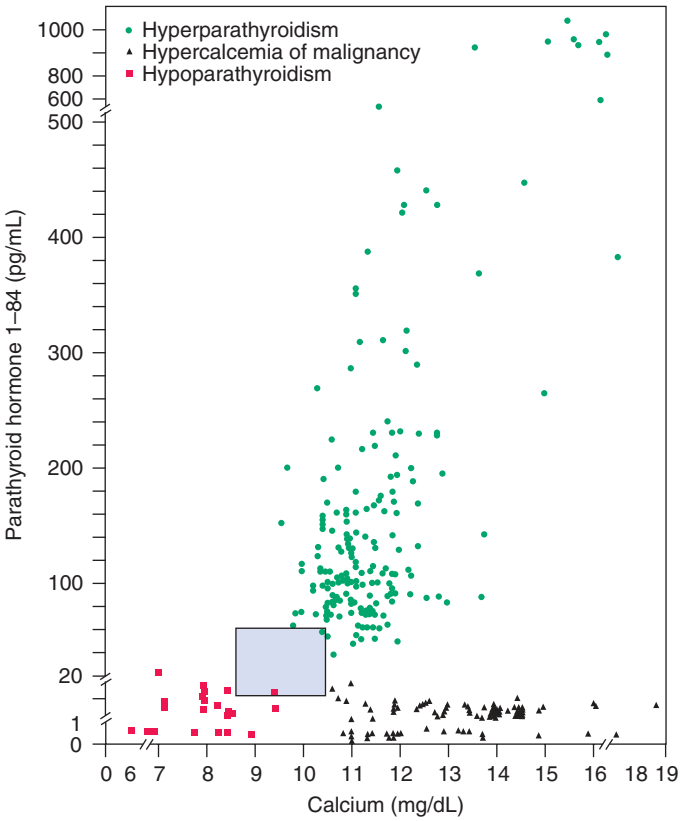


FIGURE 403-4 Levels of immunoreactive parathyroid hormone (PTH) detected in patients with primary hyperparathyroidism, hypercalcemia of malignancy, and hypoparathyroidism. Boxed area represents the upper and normal limits of blood calcium and/or immunoreactive PTH. *(From SR Nussbaum, JT Potts, Jr, in L DeGroot, JL Jameson [eds]: Endocrinology, 4th ed. Philadelphia, Saunders, 2001; with permission.)*

TREATMENT

Hyperparathyroidism

Surgical excision of the abnormal parathyroid tissue is the definitive therapy for this disease. As noted above, medical surveillance without operation for patients with mild, asymptomatic disease is, however, still preferred by some physicians and patients, particularly when the patients are more elderly. Evidence favoring surgery, if medically feasible, is growing because of concerns about skeletal, cardiovascular, and neuro-psychiatric disease, even in mild HPT.

Two surgical approaches are generally practiced. The conventional parathyroidectomy procedure was neck exploration with general anesthesia; this procedure is being replaced in many centers, whenever feasible, by an outpatient procedure with local anesthesia, termed *minimally invasive parathyroidectomy*. Parathyroid exploration is challenging and should be undertaken by an experienced surgeon. Certain features help in predicting the pathology (e.g., multiple abnormal glands in familial cases). However, some critical decisions regarding management can be made only during the operation.

With conventional surgery, one approach is still based on the view that typically only one gland (the adenoma) is abnormal. If an enlarged gland is found, a normal gland should be sought. In this view, if a biopsy of a normal-sized second gland confirms its histologic (and presumed functional) normality, no further exploration, biopsy, or excision is needed. At the other extreme is the minority viewpoint that all four glands be sought and that most of the total parathyroid tissue mass be removed. The concern with the former approach is that the recurrence rate of HPT may be high if a second abnormal gland is missed; the latter approach could involve unnecessary surgery and an unacceptable rate of hypoparathyroidism. When normal glands are found in association with one enlarged gland, excision of the single adenoma usually leads to cure or at least years free of symptoms. Long-term follow-up studies to establish true rates of recurrence are limited.

Recently, there has been growing experience with new surgical strategies that feature a minimally invasive approach guided by improved preoperative localization and intraoperative monitoring by PTH assays. Preoperative ^{99m}Tc sestamibi scans with single-photon emission CT (SPECT) are used to predict the location of an abnormal gland and intraoperative sampling of PTH before and at 5-minute intervals after removal of a suspected adenoma to confirm a rapid fall (>50%) to normal levels of PTH. In several centers, a combination of preoperative sestamibi imaging, cervical block anesthesia, minimal surgical incision, and intraoperative PTH measurements has allowed successful outpatient surgical management with a clear-cut cost benefit compared to general anesthesia and more extensive neck surgery. The use of these minimally invasive approaches requires clinical judgment to select patients unlikely to have multiple gland disease (e.g., MEN or secondary HPT). The growing acceptance of the technique and its relative ease for the patient has lowered the threshold for surgery.

Severe hypercalcemia may provide a preoperative clue to the presence of parathyroid carcinoma. In such cases, when neck exploration is undertaken, the tissue should be widely excised; care is taken to avoid rupture of the capsule to prevent local seeding of tumor cells.

Multiple-gland hyperplasia, as predicted in familial cases, poses more difficult questions of surgical management. Once a diagnosis

of hyperplasia is established, all the glands must be identified. Two schemes have been proposed for surgical management. One is to totally remove three glands with partial excision of the fourth gland; care is taken to leave a good blood supply for the remaining gland. Other surgeons advocate total parathyroidectomy with immediate transplantation of a portion of a removed, minced parathyroid gland into the muscles of the forearm, with the view that surgical excision is easier from the ectopic site in the arm if there is recurrent hyperfunction.

In a minority of cases, if no abnormal parathyroid glands are found in the neck, the issue of further exploration must be decided. There are documented cases of five or six parathyroid glands and of unusual locations for adenomas such as in the mediastinum.

When a second parathyroid exploration is indicated, the minimally invasive techniques for preoperative localization such as ultrasound, CT scan, and isotope scanning are combined with venous sampling and/or selective digital arteriography in one of the centers specializing in these procedures. Intraoperative monitoring of PTH levels by rapid PTH immunoassays may be useful in guiding the surgery. At one center, long-term cures have been achieved with selective embolization or injection of large amounts of contrast material into the end-arterial circulation feeding the parathyroid tumor.

A decline in serum calcium occurs within 24 h after successful surgery; usually blood calcium falls to low-normal values for 3–5 days until the remaining parathyroid tissue resumes full hormone secretion. Acute postoperative hypocalcemia is likely only if severe bone mineral deficits are present or if injury to all the normal parathyroid glands occurs during surgery. In general, there are few problems encountered in patients with uncomplicated disease such as a single adenoma (the clear majority), who do not have symptomatic bone disease nor a large deficit in bone mineral, who are vitamin D and magnesium sufficient, and who have good renal and gastrointestinal function. The extent of postoperative hypocalcemia varies with the surgical approach. If all glands are biopsied, hypocalcemia may be transiently symptomatic and more prolonged. Hypocalcemia is more likely to be symptomatic after second parathyroid explorations, particularly when normal parathyroid tissue was removed at the initial operation and when the manipulation and/or biopsy of the remaining normal glands are more extensive in the search for the missing adenoma.

Patients with HPT have efficient intestinal calcium absorption due to the increased levels of $1,25(OH)_2D$ stimulated by PTH excess. Once hypocalcemia signifies successful surgery, patients can be put on a high-calcium intake or be given oral calcium supplements. Despite mild hypocalcemia, most patients do not require parenteral therapy. If the serum calcium falls to <2 mmol/L (8 mg/dL), *and if the phosphate level rises simultaneously*, the possibility that surgery has caused hypoparathyroidism must be considered. With unexpected hypocalcemia, coexistent hypomagnesemia should be considered, as it interferes with PTH secretion and causes functional hypoparathyroidism **(Chap. 402)**.

Signs of hypocalcemia include symptoms such as muscle twitching, a general sense of anxiety, and positive Chvostek's and Trousseau's signs coupled with serum calcium consistently <2 mmol/L (8 mg/dL). Parenteral calcium replacement at a low level should be instituted when hypocalcemia is symptomatic. The rate and duration of IV therapy are determined by the severity of the symptoms and the response of the serum calcium to treatment. An infusion of 0.5–2 mg/kg per hour or 30–100 mL/h of a 1-mg/mL solution usually suffices to relieve symptoms. Usually, parenteral therapy is required for only a few days. If symptoms worsen or if parenteral calcium is needed for >2–3 days, therapy with a vitamin D analogue and/or oral calcium (2–4 g/d) should be started (see below). It is cost-effective to use calcitriol (doses of 0.5–1 μg/d) because of the rapidity of onset of effect and prompt cessation of action when stopped, in comparison to other forms of vitamin D. A rise in blood calcium after several months of vitamin D replacement may indicate restoration of parathyroid function to normal. It is also appropriate to monitor serum PTH serially to estimate gland function in such patients.

If magnesium deficiency was present, it can complicate the postoperative course since magnesium deficiency impairs the secretion of PTH. Hypomagnesemia should be corrected whenever detected. Magnesium replacement can be effective orally (e.g., $MgCl_2$, $MgOH_2$), but parenteral repletion is usual to ensure postoperative recovery, if magnesium deficiency is suspected due to low blood magnesium levels. Because the depressant effect of magnesium on central and peripheral nerve functions does not occur at levels <2 mmol/L (normal range 0.8–1.2 mmol/L), parenteral replacement can be given rapidly. A cumulative dose as great as 0.5–1 mmol/kg of body weight can be administered if severe hypomagnesemia is present; often, however, total doses of 20–40 mmol are sufficient.

MEDICAL MANAGEMENT

The guidelines for recommending surgical intervention, if feasible (Table 403-2), as well as for monitoring patients with asymptomatic HPT who elect not to undergo parathyroidectomy (Table 403-3), reflect the changes over time since the first conference on the topic in 1990. Medical monitoring rather than corrective surgery is still acceptable, but it is clear that surgical intervention is the more frequently recommended option for the reasons noted above. Tightened guidelines favoring surgery include lowering the recommended level of serum calcium elevation, more careful attention to skeletal integrity through reference to peak skeletal mass at baseline (T scores) rather than age-adjusted bone density (Z scores), as well as the presence of any fragility fracture. The other changes noted in the two guidelines (Tables 403-2 and 403-3) reflect accumulated experience and practical consideration, such as a difficulty in quantity of urine collections. Despite the usefulness of the guidelines, the importance of individual patient and physician judgment and preference are clear in all recommendations.

When surgery is not selected, or not medically feasible, there is interest in the potential value of specific medical therapies. There is no long-term experience regarding specific clinical outcomes such as fracture prevention, but it has been established that bisphosphonates increase bone mineral density significantly without changing serum calcium (as does estrogen, but the latter is not favored because of reported adverse effects in other organ systems). Calcimimetics that lower PTH secretion lower calcium but do not affect bone mass density (BMD).

OTHER PARATHYROID-RELATED CAUSES OF HYPERCALCEMIA

Lithium Therapy Lithium, used in the management of bipolar depression and other psychiatric disorders, causes hypercalcemia in ~10% of treated patients. The hypercalcemia is dependent on continued lithium treatment, remitting and recurring when lithium is stopped and restarted. The parathyroid adenomas reported in some hypercalcemic patients with lithium therapy may reflect the presence of an independently occurring parathyroid tumor; a permanent effect of lithium on parathyroid gland growth need not be implicated as most patients have complete reversal of hypercalcemia when lithium is stopped. However, long-standing stimulation of parathyroid cell replication by lithium may predispose to development of adenomas (as is documented in secondary HPT and renal failure).

At the levels achieved in blood in treated patients, lithium can be shown in vitro to shift the PTH secretion curve to the right in response to calcium; i.e., higher calcium levels are required to lower PTH secretion, probably acting at the calcium sensor (see below). This effect can cause elevated PTH levels and consequent hypercalcemia in otherwise normal individuals. Fortunately, there are usually alternative medications for the underlying psychiatric illness. Parathyroid surgery should not be recommended unless hypercalcemia and elevated PTH levels persist after lithium is discontinued.

GENETIC DISORDERS CAUSING HYPERPARATHYROID-LIKE SYNDROMES

Familial Hypocalciuric Hypercalcemia FHH (also called *familial benign hypercalcemia*) is inherited as an autosomal dominant trait. Affected individuals are discovered because of asymptomatic hypercalcemia. Most cases of FHH (FHH1) are caused by an inactivating mutation in a single allele of the CaSR (see below), leading to inappropriately normal or even increased secretion of PTH, whereas another hypercalcemic disorder, namely the exceedingly rare Jansen's disease, is caused by a constitutively active PTH/PTHrP receptor in target tissues. Neither FHH1 nor Jansen's disease, however, are growth disorders of the parathyroids. Other forms of FHH are caused either by heterozygous mutations in *GNA11* (encoding Gα11), one of the signaling proteins downstream of the CaSR (FHH2), or by mutations in *AP1A1* (FHH3).

The pathophysiology of FHH1 is now understood. The primary defect is abnormal sensing of the blood calcium by the parathyroid gland and renal tubule, causing inappropriate secretion of PTH and excessive reabsorption of calcium in the distal renal tubules. The CaSR is a member of the third family of GPCRs (type C or type III). The receptor responds to increased ECF calcium concentration by suppressing PTH secretion through second-messenger signaling involving the G proteins Gα11 and Gαq, thereby providing negative-feedback regulation of PTH secretion. Many different inactivating CaSR mutations have been identified in patients with FHH1. These mutations lower the capacity of the sensor to bind calcium, and the mutant receptors function as though blood calcium levels were low; excessive secretion of PTH occurs from an otherwise normal gland. Approximately two-thirds of patients with FHH have mutations within the protein-coding region of the CaSR gene. The remaining one-third of kindreds may have mutations in the promoter of the CaSR gene or are caused by mutations in other genes.

Even before elucidation of the pathophysiology of FHH, abundant clinical evidence served to separate the disorder from primary HPT; these clinical features are still useful in differential diagnosis. Patients with primary HPT have <99% renal calcium reabsorption, whereas most patients with FHH have >99% reabsorption. The hypercalcemia in FHH is often detectable in affected members of the kindreds in the first decade of life, whereas hypercalcemia rarely occurs in patients with primary HPT or the MEN syndromes who are aged <10 years. PTH may be elevated in the different forms of FHH, but the values are usually normal or lower for the same degree of calcium elevation than is observed in patients with primary HPT. Parathyroid surgery performed in a few patients with FHH before the nature of the syndrome was understood led to permanent hypoparathyroidism; nevertheless, hypocalciuria persisted, establishing that hypocalciuria is not PTH-dependent (now known to be due to the abnormal CaSR in the kidney).

Few clinical signs or symptoms are present in patients with FHH, while other endocrine abnormalities are not. Most patients are detected as a result of family screening after hypercalcemia is detected in a proband. In those patients inadvertently operated upon for primary HPT, the parathyroids appeared normal or moderately hyperplastic. Parathyroid surgery is not appropriate, nor, in view of the lack of symptoms, does medical treatment seem needed to lower the calcium. One striking exception to the rule against parathyroid surgery in this syndrome is the occurrence, usually in consanguineous marriages (due to the rarity of the gene mutation), of a homozygous or compound heterozygote state, resulting in severe impairment of CaSR function. In this condition, neonatal severe hypercalcemia, total parathyroidectomy is mandatory, but calcimetics have been used as a temporary measure. Rare but well-documented cases of acquired hypocalciuric hypercalcemia are reported due to antibodies against the CaSR. They appear to be a complication of an underlying autoimmune disorder and respond to therapies directed against the underlying disorder.

Jansen's Disease Activating mutations in the PTH/PTHrP receptor (PTH1R) have been identified as the cause of this rare autosomal dominant syndrome. Because the mutations lead to constitutive activation of receptor function, one abnormal copy of the mutant receptor is sufficient to cause the disease, thereby accounting for its dominant mode of transmission. The disorder leads to short-limbed dwarfism due to abnormal regulation of chondrocyte maturation in the growth plates of the bone that are formed through the endochondral process. In adult life, there are numerous abnormalities in bone, including multiple cystic resorptive areas resembling those seen in severe HPT. Hypercalcemia and hypophosphatemia with undetectable or low PTH levels are typically observed. The pathogenesis of the growth plate abnormalities in Jansen's disease has been confirmed by transgenic experiments in which targeted expression of the mutant PTH/PTHrP receptor to the proliferating chondrocyte layer of growth plate emulated several features of the human disorder. Some of these genetic mutations in the parathyroid gland or PTH target cells that affect Ca^{2+} metabolism are illustrated in **Figure 403-5**.

MALIGNANCY-RELATED HYPERCALCEMIA

Clinical Syndromes and Mechanisms of Hypercalcemia

Hypercalcemia due to malignancy is common (occurring in as many as 20% of cancer patients, especially with certain types of tumor such as lung carcinoma), often severe and difficult to manage, and, on rare occasions, difficult to distinguish from primary HPT. Although malignancy is often clinically obvious or readily detectable by medical history, hypercalcemia can occasionally be due to an occult tumor. Previously, hypercalcemia associated with malignancy was thought to be due to local invasion and destruction of bone by tumor cells; many cases are now known to result from the elaboration by the malignant cells of humoral mediators of hypercalcemia. PTHrP is the responsible humoral agent in most solid tumors that cause hypercalcemia.

The histologic character of the tumor is more important than the extent of skeletal metastases in predicting hypercalcemia. Small cell carcinoma (oat cell) and adenocarcinoma of the lung, though the most common lung tumors associated with skeletal metastases, rarely cause hypercalcemia. By contrast, many patients with squamous cell carcinoma of the lung develop hypercalcemia. Histologic studies of bone in patients with squamous cell or epidermoid carcinoma of the lung, in sites invaded by tumor as well as areas remote from tumor invasion, reveal increased bone resorption.

Two main mechanisms of hypercalcemia are operative in cancer hypercalcemia. Many solid tumors associated with hypercalcemia, particularly squamous cell and renal tumors, produce and secrete PTHrP that causes increased bone resorption and mediate the hypercalcemia through systemic actions on the skeleton. Alternatively, direct bone marrow invasion occurs with hematologic malignancies such as leukemia, lymphoma, and multiple myeloma. Lymphokines and cytokines (including PTHrP) produced by cells involved in the marrow response to the tumors promote resorption of bone through local destruction. Several hormones, hormone analogues, cytokines, and growth factors have been implicated as the result of clinical assays, in vitro tests, or chemical isolation. The etiologic factor produced by activated normal lymphocytes and by myeloma and lymphoma cells, originally termed *osteoclast activation factor*, now appears to represent the biologic action of several different cytokines, probably interleukin 1 and lymphotoxin or tumor necrosis factor (TNF). In some lymphomas, there is a third mechanism, caused by an increased blood level of $1,25(OH)_2D$, produced by the abnormal lymphocytes.

In the more common mechanism, usually termed *humoral hypercalcemia of malignancy*, solid tumors (cancers of the lung and kidney, in particular), in which bone metastases are absent, minimal, or not detectable clinically, secrete PTHrP measurable by immunoassay. Secretion by the tumors of the PTH-like factor, PTHrP, activates the PTH1R, resulting in a pathophysiology closely resembling HPT, but with normal or suppressed PTH levels. The clinical picture resembles primary HPT (hypophosphatemia accompanies hypercalcemia), and elimination or regression of the primary tumor leads to disappearance of the hypercalcemia.

As in HPT, patients with the humoral hypercalcemia of malignancy have elevated urinary nephrogenous cyclic AMP excretion, hypophosphatemia, and increased urinary phosphate clearance. However, in humoral hypercalcemia of malignancy, immunoreactive PTH is

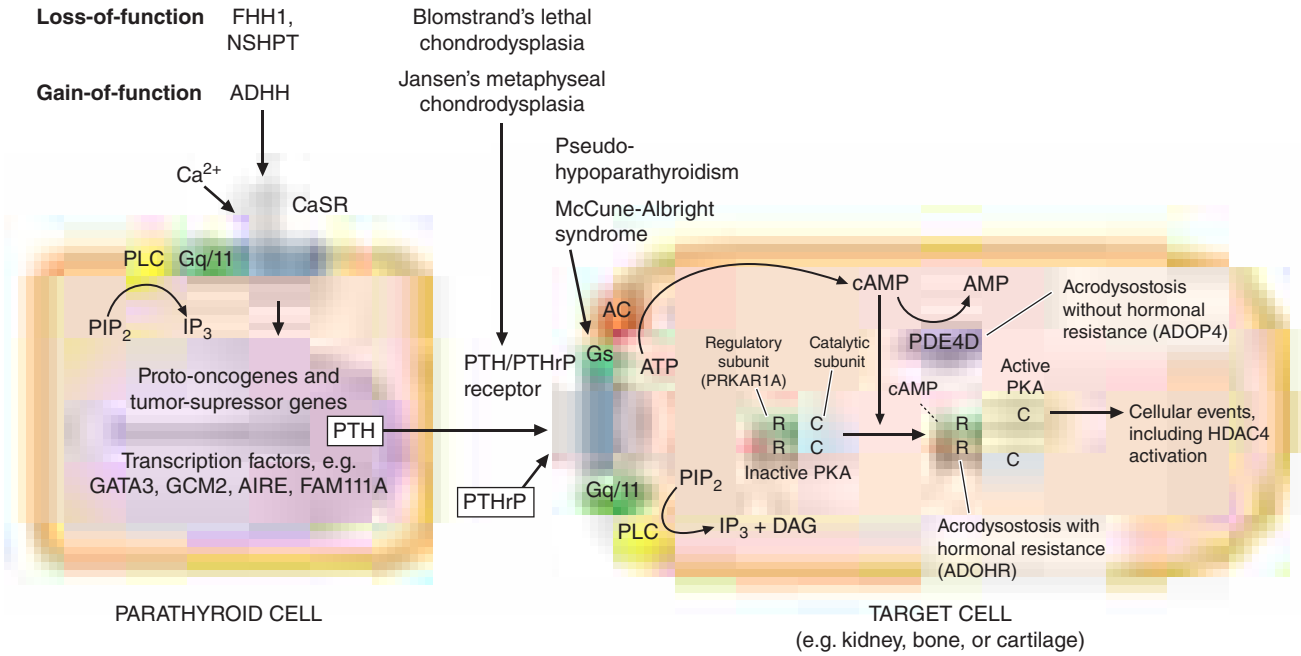


FIGURE 403-5 Illustration of some genetic mutations that alter calcium metabolism by effects on the parathyroid cell or target cells of parathyroid hormone (PTH) action. Alterations in PTH production by the parathyroid cell can be caused by changes in the response to extracellular fluid calcium (Ca^{2+}) that are detected by the calcium-sensing receptor (CaSR). Furthermore, PTH (or PTHrP) can show altered efficacy in target cells such as in proximal tubular cells, by altered function of its receptor (PTH/PTHrP receptor) or the signal transduction proteins, G proteins such as $G_s\alpha$ that is linked to adenylate cyclase (AC), the enzyme responsible for producing cyclic AMP (cAMP) (also illustrated are $G\alpha q/G\alpha 11$, which activate an alternate pathway of receptor signal transmission involving the generation of inositol triphosphate [IP_3] or diacylglycerol [DAG]). Heterozygous loss-of-function mutations in the CaSR cause familial hypocalciuric hypercalcemia (FHH) and homozygous mutations (both alleles mutated) and severe neonatal hyperparathyroidism (NSHPT); heterozygous gain-of-functions causes autosomal dominant hypercalciuric hypocalcemia (ADHH). Other defects in parathyroid cell function that occur at the level of gene regulation (oncogenes or tumor suppressor genes) or transcription factors are discussed in the text. Blomstrand's lethal chondrodysplasia is due to homozygous or compound heterozygous loss-of-function mutations in the PTH/PTHrP receptor, a neonatally lethal disorder, while pseudohypoparathyroidism involves inactivation at the level of the G proteins, specifically mutations that eliminate or reduce $G_s\alpha$ activity in the kidney (see text for details). Acrodysostosis can occur with (ADOHR; mutant regulatory subunit of PKA) or without hormonal resistance (ADOP4; mutant PDE4D). Jansen's metaphyseal chondrodysplasia and McCune-Albright syndrome represent gain-of-function mutations in the PTH/PTHrP receptor and $G_s\alpha$ protein, respectively.

undetectable or suppressed, making the differential diagnosis easier. Other features of the disorder differ from those of true HPT. Although the biologic actions of PTH and PTHrP are exerted through the same receptor, subtle differences in receptor activation by the two ligands must account for some of the discordance in pathophysiology, when an excess of one or the other peptide occurs. Other cytokines elaborated by the malignancy may contribute to the variations from HPT in these patients as well. Patients with humoral hypercalcemia of malignancy may have low to normal levels of $1,25(OH)_2D$ instead of elevated levels as in true HPT. In some patients with the humoral hypercalcemia of malignancy, osteoclastic resorption is unaccompanied by an osteoblastic or bone-forming response, implying inhibition of the normal coupling of bone formation and resorption.

Several different assays (single- or double-antibody, different epitopes) have been developed to detect PTHrP. Most data indicate that circulating PTHrP levels are undetectable (or low) in normal individuals except perhaps in pregnancy (high in human milk) and elevated in most cancer patients with the humoral syndrome. The etiologic mechanisms in cancer hypercalcemia may be multiple in the same patient. For example, in breast carcinoma (metastatic to bone) and in a distinctive type of T cell lymphoma/leukemia initiated by human T cell lymphotropic virus I, hypercalcemia is caused by direct local lysis of bone as well as by a humoral mechanism involving excess production of PTHrP. HPT has been reported to coexist with the humoral cancer syndrome and, rarely, ectopic HPT due to tumor elaboration of true PTH is reported.

Diagnostic Issues Levels of PTH measured by the double-antibody technique are undetectable or extremely low in tumor hypercalcemia, as would be expected with the mediation of the hypercalcemia by a factor other than PTH (the hypercalcemia suppresses the normal parathyroid glands). In a patient with minimal symptoms referred for hypercalcemia, low or undetectable PTH levels would focus attention on a possible occult malignancy (except for very rare cases of ectopic HPT).

Ordinarily, the diagnosis of cancer hypercalcemia is not difficult because tumor symptoms are prominent when hypercalcemia is detected. Indeed, hypercalcemia may be noted incidentally during the workup of a patient with known or suspected malignancy. Clinical suspicion that malignancy is the cause of the hypercalcemia is heightened when there are other signs or symptoms of a paraneoplastic process such as weight loss, fatigue, muscle weakness, or unexplained skin rash, or when symptoms specific for a particular tumor are present. Squamous cell tumors are most frequently associated with hypercalcemia, particularly tumors of the lung, kidney, head and neck, and urogenital tract. Radiologic examinations can focus on these areas when clinical evidence is unclear. Bone scans with technetium-labeled bisphosphonate are useful for detection of osteolytic metastases; the sensitivity is high, but specificity is low; results must be confirmed by conventional x-rays to be certain that areas of increased uptake are due to osteolytic metastases per se. Bone marrow biopsies are helpful in patients with anemia or abnormal peripheral blood smears.

TREATMENT

Malignancy-Related Hypercalcemia

Treatment of the hypercalcemia of malignancy is first directed to control of tumor; reduction of tumor mass usually corrects hypercalcemia. If a patient has severe hypercalcemia yet has a good chance for effective tumor therapy, treatment of the hypercalcemia should be vigorous while awaiting the results of definitive therapy (see general approach to hypercalcemic states below). If hypercalcemia occurs in the late stages of a tumor that is resistant to antitumor therapy, the treatment of the hypercalcemia should be judicious as high calcium levels can have a mild sedating effect. Standard therapies for hypercalcemia (discussed below) are applicable to patients with malignancy.

VITAMIN D–RELATED HYPERCALCEMIA

Vitamin D–mediated hypercalcemia can be due to excessive ingestion vitamin D analogs or abnormal metabolism of the vitamin. Abnormal metabolism of the vitamin is usually acquired in association with a widespread granulomatous disorder. Vitamin D metabolism is carefully regulated, particularly the activity of renal 1α-hydroxylase, the enzyme responsible for the production of $1,25(OH)_2D$ **(Chap. 402)**. The regulation of 1α-hydroxylase and the normal feedback suppression by $1,25(OH)_2D$ seem to work less well in infants than in adults and to operate poorly, if at all, in sites other than the renal tubule; these phenomena may explain the occurrence of hypercalcemia secondary to excessive $1,25(OH)_2D_3$ production in infants with Williams' syndrome (see below) and in adults with sarcoidosis or lymphoma.

Vitamin D Intoxication Chronic ingestion of 40–100 times the normal physiologic requirement of vitamin D (amounts >40,000–100,000 U/d) is usually required to produce significant hypercalcemia in otherwise healthy individuals. The stated upper limit of safe dietary intake is 2000 U/d (50 μg/d) in adults because of concerns about potential toxic effects of cumulative supraphysiologic doses. These recommendations are now regarded as too restrictive, since some estimates are that in elderly individuals in northern latitudes, 2000 U/d or more may be necessary to avoid vitamin D insufficiency.

Hypercalcemia in vitamin D intoxication is due to an excessive biologic action of the vitamin, perhaps the consequence of increased levels of 25(OH)D rather than merely increased levels of the active metabolite $1,25(OH)_2D$ (the latter may not be elevated in vitamin D intoxication). 25(OH)D has definite, if low, biologic activity in the intestine and bone. The production of 25(OH)D is less tightly regulated than is the production of $1,25(OH)_2D$. Hence concentrations of 25(OH)D are elevated several-fold in patients with excess vitamin D intake.

The diagnosis is substantiated by documenting elevated levels of 25(OH)D >100 ng/mL. Hypercalcemia is usually controlled by restriction of dietary calcium intake and appropriate attention to hydration. These measures, plus discontinuation of vitamin D, usually lead to resolution of hypercalcemia. However, vitamin D stores in fat may be substantial, and vitamin D intoxication may persist for weeks after vitamin D ingestion is terminated. Such patients are responsive to glucocorticoids, which in doses of 100 mg/d of hydrocortisone or its equivalent usually return serum calcium levels to normal over several days; severe intoxication may require intensive therapy.

Sarcoidosis and Other Granulomatous Diseases In patients with sarcoidosis and other granulomatous diseases, such as tuberculosis and fungal infections, excess $1,25(OH)_2D$ is synthesized in macrophages or other cells in the granulomas. Indeed, increased $1,25(OH)_2D$ levels have been reported in anephric patients with sarcoidosis and hypercalcemia. Macrophages obtained from granulomatous tissue convert 25(OH)D to $1,25(OH)_2D$ at an increased rate. There is a positive correlation in patients with sarcoidosis between 25(OH)D levels (reflecting vitamin D intake) and the circulating concentrations of $1,25(OH)_2D$, whereas normally there is no increase in $1,25(OH)_2D$ with increasing 25(OH)D levels due to multiple feedback controls on renal 1α-hydroxylase **(Chap. 402)**. The usual regulation of active metabolite production by calcium and phosphate or by PTH does not operate in these patients. Clearance of $1,25(OH)_2D$ from blood may be decreased in sarcoidosis as well. PTH levels are usually low and $1,25(OH)_2D$ levels are elevated, but primary HPT and sarcoidosis may coexist in some patients.

Management of the hypercalcemia can often be accomplished by avoiding excessive sunlight exposure and limiting vitamin D and calcium intake. Presumably, however, the abnormal sensitivity to vitamin D and abnormal regulation of $1,25(OH)_2D$ synthesis will persist as long as the disease is active. Alternatively, glucocorticoids in the equivalent of 100 mg/d of hydrocortisone or equivalent doses of glucorticoids may help control hypercalcemia. Glucocorticoids appear to act by blocking excessive production of $1,25(OH)_2D$, as well as the response to it in target organs.

Idiopathic Hypercalcemia of Infancy This rare disorder, usually referred to as *Williams' syndrome*, is an autosomal dominant disorder characterized by multiple congenital development defects, including supravalvular aortic stenosis, mental retardation, and an elfin facies, in association with hypercalcemia due to abnormal sensitivity to vitamin D. The hypercalcemia associated with the syndrome was first recognized in England after fortification of milk with vitamin D. The cardiac and developmental abnormalities were independently described, but the connections between these defects and hypercalcemia were not described until later. Levels of $1,25(OH)_2D$ can be elevated, ranging from 46 to 120 nmol/L (150–500 pg/mL). The mechanism of the abnormal sensitivity to vitamin D and of the increased circulating levels of $1,25(OH)_2D$ is still unclear. Studies suggest that genetic mutations involving microdeletions at the elastin locus and perhaps other genes on chromosome 7 may play a role in the pathogenesis. Other causes of hypercalcemia in infants and young children are 24-hydroxylase deficiency that impairs metabolism of $1,25(OH)_2D$, or mutations involving the sodium-dependent phosphate transporters (NPT2a or NPT2c).

HYPERCALCEMIA ASSOCIATED WITH HIGH BONE TURNOVER

Hyperthyroidism As many as 20% of hyperthyroid patients have high-normal or mildly elevated serum calcium concentrations; hypercalciuria is even more common. The hypercalcemia is due to increased bone turnover, with bone resorption exceeding bone formation. Severe calcium elevations are not typical, and the presence of such suggests a concomitant disease such as HPT. Usually, the diagnosis is obvious, but signs of hyperthyroidism may occasionally be occult, particularly in the elderly **(Chap. 375)**. Hypercalcemia is managed by treatment of the hyperthyroidism. Reports that thyroid-stimulating hormone (TSH) itself normally has a bone-protective effect suggest that suppressed TSH levels also play a role in hypercalcemia.

Immobilization Immobilization is a rare cause of hypercalcemia in adults in the absence of an associated disease but may cause hypercalcemia in children and adolescents, particularly after spinal cord injury and paraplegia or quadriplegia. With resumption of ambulation, the hypercalcemia in children usually returns to normal.

The mechanism appears to involve a disproportion between bone formation and bone resorption; the former decreased and the latter increased. Hypercalciuria and increased mobilization of skeletal calcium can develop in normal volunteers subjected to extensive bed rest, although hypercalcemia is unusual. Immobilization of an adult with a disease associated with high bone turnover, however, such as Paget's disease, may cause hypercalcemia.

Thiazides Administration of benzothiadiazines (thiazides) can cause hypercalcemia in patients with high rates of bone turnover. Traditionally, thiazides are associated with aggravation of hypercalcemia in primary HPT, but this effect can be seen in other high-bone-turnover states as well. The mechanism of thiazide action is complex. Chronic thiazide administration leads to reduction in urinary calcium; the hypocalciuric effect appears to reflect the enhancement of proximal tubular resorption of sodium and calcium in response to sodium depletion. Some of this renal effect is due to augmentation of PTH action and is more pronounced in individuals with intact PTH secretion. However, thiazides cause hypocalciuria in hypoparathyroid patients on high-dose vitamin D and oral calcium replacement if sodium intake is restricted. This finding is the rationale for the use of thiazides as an adjunct to therapy in hypoparathyroid patients, as discussed below. Thiazide administration to normal individuals causes a transient increase in blood calcium (usually within the high-normal range) that reverts to preexisting levels after a week or more of continued administration. If hormonal function and calcium and bone metabolism are normal, homeostatic controls are reset to counteract the calcium-elevating effect of the thiazides. In the presence of HPT or increased bone turnover from another cause, homeostatic mechanisms are ineffective. The abnormal effects of the thiazide on calcium metabolism disappear within days of cessation of the drug.

Vitamin A Intoxication Vitamin A intoxication is a rare cause of hypercalcemia and is most commonly a side effect of dietary faddism

(Chap. 326). Calcium levels can be elevated into the 3–3.5-mmol/L (12–14 mg/dL) range after the ingestion of 50,000–100,000 units of vitamin A daily (10–20 times the minimum daily requirement). Typical features of severe hypercalcemia include fatigue, anorexia, and, in some, severe muscle and bone pain. Excess vitamin A intake is presumed to increase bone resorption.

The diagnosis can be established by history and by measurement of vitamin A levels in serum. Occasionally, skeletal x-rays reveal periosteal calcifications, particularly in the hands. Withdrawal of the vitamin is usually associated with prompt disappearance of the hypercalcemia and reversal of the skeletal changes. As in vitamin D intoxication, administration of 100 mg/d hydrocortisone or its equivalent leads to a rapid return of the serum calcium to normal.

HYPERCALCEMIA ASSOCIATED WITH RENAL FAILURE

Severe Secondary HPT The pathogenesis of secondary HPT in CKD is incompletely understood. Resistance to the normal level of PTH is a major factor contributing to the development of hypocalcemia, which, in turn, is a stimulus to parathyroid gland enlargement. However, recent findings have indicated that an increase of FGF23 production by osteocytes (and possibly osteoblasts) in bone occurs well before an elevation in PTH is detected. FGF23 is a potent inhibitor of the renal 1-alpha hydroxylase and the FGF23-dependent reduction in $1,25(OH)_2$ vitamin D seems to be an important stimulus for the development of secondary HPT.

Secondary HPT occurs not only in patients with renal failure but also in those with osteomalacia due to multiple causes (Chap. 402), including deficiency of vitamin D action and PHP (deficient response to PTH downstream of PTHR1). For both disorders, hypocalcemia seems to be the common denominator in initiating the development of secondary HPT. Primary (1°) and secondary (2°) HPT can be distinguished conceptually by the autonomous growth of the parathyroid glands in primary HPT (presumably irreversible) and the adaptive response of the parathyroids in secondary HPT (typically reversible). In fact, reversal over weeks from an abnormal pattern of secretion, presumably accompanied by involution of parathyroid gland mass to normal, occurs in patients with osteomalacia who have been treated effectively with calcium and vitamin D. However, it is now recognized that a true clonal outgrowth (irreversible) can arise in long-standing, inadequately treated CKD (e.g., tertiary [3°] HPT; see below).

Patients with secondary HPT may develop bone pain, ectopic calcification, and pruritus. The bone disease seen in patients with secondary HPT and CKD is termed *renal osteodystrophy* and affects primarily bone turnover. However, osteomalacia is frequently encountered as well and may be related to the circulating levels of FGF23.

Two other skeletal disorders have been frequently associated in the past with CKD patients treated by long-term dialysis, who received aluminum-containing phosphate binders. Aluminum deposition in bone (see below) leads to an osteomalacia-like picture. The other entity is a low- turnover bone disease termed "aplastic" or "adynamic" bone disease; PTH levels are lower than typically observed in CKD patients with secondary HPT. It is believed that the condition is caused, at least in part, by excessive PTH suppression, which may be even greater than previously appreciated in light of evidence that some of the immunoreactive PTH detected by most commercially available PTH assays is not the full-length biologically active molecule (as discussed above) but may consist of amino-terminally truncated fragments that do not activate the PTH1R.

TREATMENT

Hypercalcemia in Secondary HPT

Medical therapy to reverse secondary HPT in CKD includes reduction of excessive blood phosphate by restriction of dietary phosphate, the use of nonabsorbable phosphate binders, and careful, selective addition of calcitriol (0.25–2 μg/d) or related analogues. Calcium carbonate became preferred over aluminum-containing antacids to prevent aluminum-induced bone disease. However, synthetic gels that also bind phosphate (such as sevelamer; Chap. 305) are now widely used, with the advantage of avoiding not only aluminum retention, but excess calcium loading, which may contribute to cardiovascular calcifications. Intravenous calcitriol (or related analogues), administered as several pulses each week, helps control secondary HPT. Aggressive but carefully administered medical therapy can often, but not always, reverse HPT and its symptoms and manifestations.

Occasional patients develop severe manifestations of secondary HPT, including hypercalcemia, pruritus, extraskeletal calcifications, and painful bones, despite aggressive medical efforts to suppress the HPT. PTH hypersecretion no longer responsive to medical therapy, a state of severe HPT in patients with CKD that requires surgery, has been referred to as *tertiary hyperparathyroidism*. Parathyroid surgery is necessary to control this condition. Based on genetic evidence from examination of tumor samples in these patients, the emergence of autonomous parathyroid function is due to a monoclonal outgrowth of one or more previously hyperplastic parathyroid glands. The adaptive response has become an independent contributor to disease; this finding seems to emphasize the importance of optimal medical management to reduce the proliferative response of the parathyroid cells that enables the irreversible genetic change.

Aluminum Intoxication Aluminum intoxication (and often hypercalcemia as a complication of medical treatment) in the past occurred in patients on chronic dialysis; manifestations included acute dementia and unresponsive and severe osteomalacia. Bone pain, multiple nonhealing fractures, particularly of the ribs and pelvis, and a proximal myopathy occur. Hypercalcemia develops when these patients are treated with vitamin D or calcitriol because of impaired skeletal responsiveness. Aluminum is present at the site of osteoid mineralization, osteoblastic activity is minimal, and calcium incorporation into the skeleton is impaired. The disorder is now rare because of the avoidance of aluminum-containing antacids or aluminum excess in the dialysis regimen (Chap. 408).

Milk-Alkali Syndrome The milk-alkali syndrome is due to excessive ingestion of calcium and absorbable antacids such as milk or calcium carbonate. It is much less frequent since proton-pump inhibitors and other treatments became available for peptic ulcer disease. For a time, the increased use of calcium carbonate in the management of secondary HPT led to reappearance of the syndrome. Several clinical presentations—acute, subacute, and chronic—have been described, all of which feature hypercalcemia, alkalosis, and renal failure. The chronic form of the disease, termed *Burnett's syndrome*, is associated with irreversible renal damage. The acute syndromes reverse if the excess calcium and absorbable alkali are stopped.

Individual susceptibility is important in the pathogenesis, as some patients are treated with calcium carbonate and alkali regimens without developing the syndrome. One variable is the fractional calcium absorption as a function of calcium intake. Some individuals absorb a high fraction of calcium, even with intakes ≥2 g of elemental calcium per day, instead of reducing calcium absorption with high intake, as occurs in most normal individuals. Resultant mild hypercalcemia after meals in such patients is postulated to contribute to the generation of alkalosis. Development of hypercalcemia causes increased sodium excretion and some depletion of total-body water. These phenomena and perhaps some suppression of endogenous PTH secretion due to mild hypercalcemia lead to increased bicarbonate resorption and to alkalosis in the face of continued calcium carbonate ingestion. Alkalosis per se selectively enhances calcium resorption in the distal nephron, thus aggravating the hypercalcemia. The cycle of mild hypercalcemia → bicarbonate retention → alkalosis → renal calcium retention → severe hypercalcemia perpetuates and aggravates hypercalcemia and alkalosis as long as calcium and absorbable alkali are ingested.

DIFFERENTIAL DIAGNOSIS: SPECIAL TESTS

Differential diagnosis of hypercalcemia is best achieved by using clinical criteria, but immunometric assays to measure PTH are especially

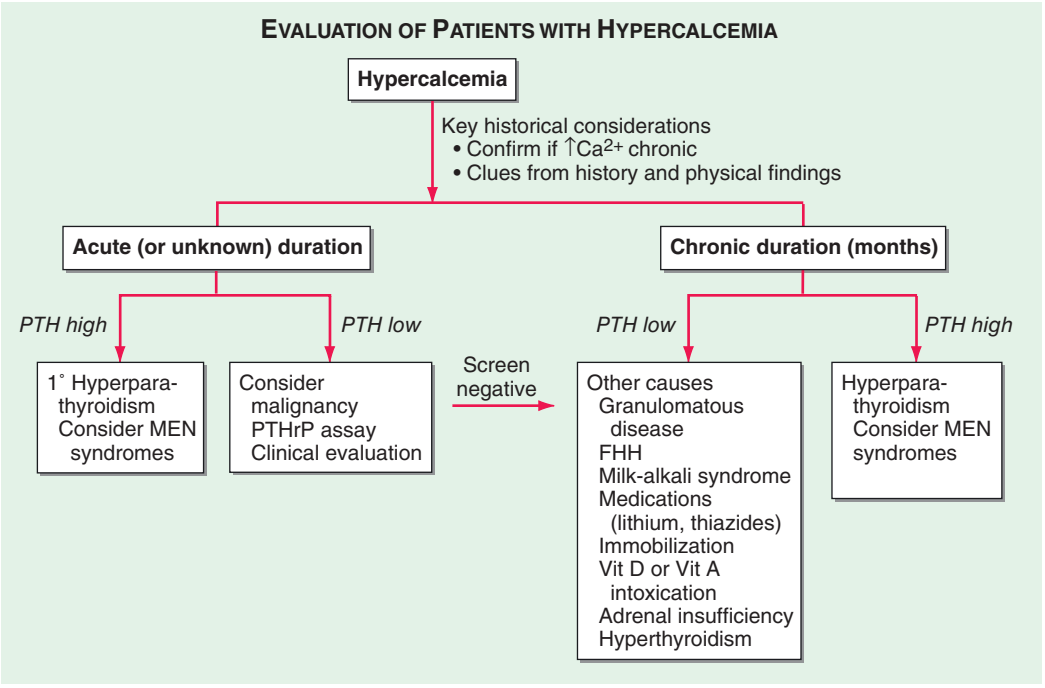


FIGURE 403-6 Algorithm for the evaluation of patients with hypercalcemia. See text for details. FHH, familial hypocalciuric hypercalcemia; MEN, multiple endocrine neoplasia; PTH, parathyroid hormone; PTHrP, parathyroid hormone–related peptide.

useful in distinguishing among major causes (Fig. 403-6). The clinical features that deserve emphasis are the presence or absence of symptoms or signs of disease and evidence of chronicity. If one discounts fatigue or depression, >90% of patients with primary HPT have *asymptomatic hypercalcemia*; symptoms of malignancy are usually present in cancer-associated hypercalcemia. Disorders other than HPT and malignancy cause <10% of cases of hypercalcemia, and some of the nonparathyroid causes are associated with clear-cut manifestations such as renal failure.

HPT is the likely diagnosis in patients with *chronic hypercalcemia*. If hypercalcemia has been manifest for >1 year, malignancy can usually be excluded as the cause. A striking feature of malignancy-associated hypercalcemia is the rapidity of the course, whereby signs and symptoms of the underlying malignancy are evident within months of the detection of hypercalcemia. Although clinical considerations are helpful in arriving at the correct diagnosis of the cause of hypercalcemia, appropriate laboratory testing is essential for definitive diagnosis. The immunoassay for PTH usually separates HPT from all other causes of hypercalcemia (exceptions are very rare reports of ectopic production of excess PTH by nonparathyroid tumors). Patients with HPT have elevated PTH levels despite hypercalcemia, whereas patients with malignancy and the other causes of hypercalcemia (except for disorders mediated by PTH such as lithium-induced hypercalcemia) have levels of hormone below normal or undetectable. Assays based on the double-antibody method for PTH exhibit very high sensitivity (especially if serum calcium is simultaneously evaluated) and specificity for the diagnosis of primary HPT (Fig. 403-4).

In summary, PTH values are elevated in >90% of parathyroid-related causes of hypercalcemia, undetectable or low in malignancy-related hypercalcemia, and undetectable or normal in vitamin D–related and high-bone-turnover causes of hypercalcemia. In view of the specificity of the PTH immunoassay and the high frequency of HPT in hypercalcemic patients, it is cost-effective to measure the PTH level in all hypercalcemic patients unless malignancy or a specific nonparathyroid disease is obvious. False-positive PTH assay results are rare. Immunoassays for PTHrP are helpful in diagnosing certain types of malignancy-associated hypercalcemia. Although FHH is parathyroid-related, the disease should be managed distinctively from HPT. Clinical features and the low urinary calcium excretion can help make the distinction. Because the incidence of malignancy and HPT both increase with age, they can coexist as two independent causes of hypercalcemia.

$1,25(OH)_2D$ levels are elevated in many (but not all) patients with primary HPT. In other disorders associated with hypercalcemia, concentrations of $1,25(OH)_2D$ are low or, at the most, normal. However, this test is of low specificity and is not cost-effective, as not all patients with HPT have elevated $1,25(OH)_2D$ levels and not all nonparathyroid hypercalcemic patients have suppressed $1,25(OH)_2D$. Measurement of $1,25(OH)_2D$ is, however, critically valuable in establishing the cause of hypercalcemia in sarcoidosis and certain lymphomas.

A useful general approach is outlined in Fig. 403-6. If the patient is *asymptomatic* and there is evidence of *chronicity* to the hypercalcemia, HPT is almost certainly the cause. If PTH levels (usually measured at least twice) are elevated, the clinical impression is confirmed and little additional evaluation is necessary. If there is only a short history or no data as to the duration of the hypercalcemia, *occult malignancy* must be considered; if the PTH levels are not elevated, then a thorough workup must be undertaken for malignancy, including chest x-ray, CT of chest and abdomen, and bone scan. Immunoassays for PTHrP may be especially useful in such situations. Attention should also be paid to clues for underlying hematologic disorders such as anemia, increased plasma globulin, and abnormal serum immunoelectrophoresis; bone scans can be negative in some patients with metastases such as in multiple myeloma. Finally, if a patient with chronic hypercalcemia is asymptomatic and malignancy therefore seems unlikely on clinical grounds, but PTH values are not elevated, it is useful to search for other chronic causes of hypercalcemia such as occult sarcoidosis. A careful history of dietary supplements and drug use may suggest intoxication with vitamin D or vitamin A or the use of thiazides.

TREATMENT

General Approach to Hypercalcemic States

The approach to medical treatment of hypercalcemia varies with its severity. Mild hypercalcemia, <3 mmol/L (12 mg/dL), can be managed by hydration. More severe hypercalcemia (levels of 3.2–3.7 mmol/L [13–15 mg/dL]) must be managed aggressively; above that level, hypercalcemia can be life-threatening and requires emergency measures (Table 403-4). By using a combination of approaches in severe hypercalcemia, the serum calcium concentration can be decreased by 0.7–2.2 mmol/L (3–9 mg/dL) within 24–48 h in most patients, enough to relieve acute symptoms, prevent death from

TABLE 403-4 Therapies for Severe Hypercalcemia

TREATMENT	ONSET OF ACTION	DURATION OF ACTION	ADVANTAGES	DISADVANTAGES
Most Useful Therapies				
Hydration with normal saline	Hours	During infusion	Rehydration invariably needed	Volume overload
Forced diuresis; normal saline plus loop diuretic	Hours	During treatment	Rapid action	Volume overload, cardiac decompensation, intensive monitoring, electrolyte disturbance, inconvenience
Pamidronate	1–2 days	10–14 days to weeks	High potency; intermediate onset of action	Fever in 20%, hypophosphatemia, hypocalcemia, hypomagnesemia, rarely jaw necrosis, atypical femoral fracture
Zolendronate	1–2 days	>3 weeks	Same as for pamidronate (lasts longer)	Same as pamidronate above
Denosumab	1-2 days	>3 weeks	Strongest antiresorptive	Occasional severe hypocalcemia, rarely jaw necrosis, skin infections, atypical femoral fracture
Special Use Therapies				
Calcitonin	Hours	1–2 days	Rapid onset of action; useful as adjunct in severe hypercalcemia	Rapid tachyphylaxis
Phosphate Oral	24 h	During use	Chronic management (with hypophosphatemia); low toxicity if P <4 mg/dL	Limited use except as adjuvant or chronic therapy
Glucocorticoids	Days	Days, weeks	Oral therapy, antitumor agent	Active only in certain malignancies, vitamin D excess and sarcoidosis; glucocorticoid side effects
Dialysis	Hours	During use and 24–48 h afterward	Useful in renal failure; onset of effect in hours; can immediately reverse life-threatening hypercalcemia	Complex procedure, reserved for extreme or special circumstances

hypercalcemic crisis, and permit diagnostic evaluation. Therapy can then be directed at the underlying disorder—the second priority.

Hypercalcemia develops because of excessive skeletal calcium release, increased intestinal calcium absorption, or inadequate renal calcium excretion. Understanding the particular pathogenesis helps guide therapy. For example, hypercalcemia in patients with malignancy is primarily due to excessive skeletal calcium release and is, therefore, minimally improved by restriction of dietary calcium. On the other hand, patients with vitamin D hypersensitivity or vitamin D intoxication have excessive intestinal calcium absorption, and restriction of dietary calcium is beneficial. Decreased renal function or ECF depletion decreases urinary calcium excretion. In such situations, rehydration may rapidly reduce or reverse the hypercalcemia, even though increased bone resorption persists. As outlined below, the more severe the hypercalcemia, the greater the number of combined therapies that should be used. Rapid acting (hours) approaches—rehydration, forced diuresis, and calcitonin—can be used with the most effective antiresorptive agents such as bisphosphonates (since severe hypercalcemia usually involves excessive bone resorption).

HYDRATION, INCREASED SALT INTAKE, MILD AND FORCED DIURESIS

The first principle of treatment is to restore normal hydration. Many hypercalcemic patients are dehydrated because of vomiting, inanition, and/or hypercalcemia-induced defects in urinary concentrating ability. The resultant drop in glomerular filtration rate is accompanied by an additional decrease in renal tubular sodium and calcium clearance. Restoring a normal ECF volume corrects these abnormalities and increases urine calcium excretion by 2.5–7.5 mmol/d (100–300 mg/d). Increasing urinary sodium excretion to 400–500 mmol/d increases urinary calcium excretion even further than simple rehydration. After rehydration has been achieved, saline can be administered or furosemide or ethacrynic acid can be given twice daily to depress the tubular reabsorptive mechanism for calcium (care must be taken to prevent dehydration). The combined use of these therapies can increase urinary calcium excretion to ≥12.5 mmol/d (500 mg/d) in most hypercalcemic patients. Since this is a substantial percentage of the exchangeable calcium pool, the serum calcium concentration usually falls 0.25–0.75 mmol/L (1–3 mg/dL) within 24 h. Precautions should be taken to prevent potassium and magnesium depletion; calcium-containing renal calculi are a potential complication.

Under life-threatening circumstances, the preceding approach can be pursued more aggressively, but the availability of effective agents to block bone resorption (such as bisphosphonates) has reduced the need for extreme diuresis regimens **(Table 403-5)**. Depletion of potassium and magnesium is inevitable unless replacements are given; pulmonary edema can be precipitated. The potential complications can be reduced by careful monitoring of central venous pressure and plasma or urine electrolytes; catheterization of the bladder may be necessary. Dialysis treatment may be needed when renal function is compromised.

BISPHOSPHONATES

The bisphosphonates are analogues of pyrophosphate, with high affinity for bone, especially in areas of increased bone turnover, where they are powerful inhibitors of bone resorption. These

TABLE 403-5 Functional Classification of Hypocalcemia (Excluding Neonatal Conditions)

PTH Absent	
Hereditary hypoparathyroidism	Hypomagnesemia
Acquired hypoparathyroidism	
PTH Ineffective	
Chronic kidney disease	Active vitamin D ineffective
Active vitamin D lacking	Intestinal malabsorption
↓ Dietary intake or sunlight	Vitamin D–dependent rickets type II
Defective metabolism:	
Anticonvulsant therapy	Pseudohypoparathyroidism
Vitamin D–dependent rickets type I	Mutant, less active PTH
PTH Overwhelmed	
Severe, acute hyperphosphatemia	Osteitis fibrosa after parathyroidectomy
Tumor lysis	
Acute kidney injury	
Rhabdomyolysis	

Abbreviation: PTH, parathyroid hormone.

bone-seeking compounds are stable in vivo because phosphatase enzymes cannot hydrolyze the central carbon-phosphorus-carbon bond. The bisphosphonates are concentrated in areas of high bone turnover and are taken up by and inhibit osteoclast action; the mechanism of action is complex. The bisphosphonate molecules that contain amino groups in the side chain structure (see below) interfere with prenylation of proteins and can lead to cellular apoptosis. The highly active nonamino group–containing bisphosphonates are also metabolized to cytotoxic products.

The initial bisphosphonate widely used in clinical practice, etidronate, was effective but had several disadvantages, including the capacity to inhibit bone formation as well as blocking resorption. Subsequently, a number of second- or third-generation compounds have become the mainstays of antiresorptive therapy for treatment of hypercalcemia and osteoporosis. The newer bisphosphonates have a highly favorable ratio of blocking resorption versus inhibiting bone formation; they inhibit osteoclast-mediated skeletal resorption yet do not cause mineralization defects at ordinary doses. Though the bisphosphonates have similar structures, the routes of administration, efficacy, toxicity, and side effects vary. The potency of the compounds for inhibition of bone resorption varies more than 10,000-fold, increasing in the order of etidronate, tiludronate, pamidronate, alendronate, risedronate, and zolendronate. The IV use of pamidronate and zolendronate is approved for the treatment of hypercalcemia; between 30 and 90 mg pamidronate, given as a single IV dose over a few hours, returns serum calcium to normal within 24–48 h with an effect that lasts for weeks in 80–100% of patients. Zolendronate given in doses of 4 or 8 mg/5-min infusion has a more rapid and more sustained effect than pamidronate in direct comparison.

These drugs are used extensively in cancer patients. Absolute survival improvements are noted with pamidronate and zolendronate in multiple myeloma, for example. However, though still rare, there are increasing reports of jaw necrosis, especially after dental surgery, mainly in cancer patients treated with multiple doses of the more potent bisphosphonates.

DENOSUMAB

Denosumab is the most recent antiresorptive therapy to be approved for the treatment of hypercalcemia. It is a monoclonal antibody that binds to receptor activator of nuclear factor-κB (RANKL) and prevents it from binding to the receptor RANK on osteoclast precursors and mature osteoclasts. The inhibition of differentiation, activation, and function of osteoclasts leads to a reduction in bone resorption. It has a profound suppressive effect on biochemical markers of bone resorption and is the most powerful antiresorptive agent currently available. Repeated doses of denosumab, 120 mg given subcutaneously, may be effective in patients with hypercalcemia of malignancy who have lost responsiveness to bisphosphonates.

OTHER THERAPIES

Calcitonin acts within a few hours of its administration, principally through receptors on osteoclasts, to block bone resorption. Calcitonin, after 24 h of use, is no longer effective in lowering calcium. Tachyphylaxis, a known phenomenon with this drug, seems to explain this effect, since the drug is initially often effective. Therefore, in life-threatening hypercalcemia, calcitonin can be used effectively within the first 24 h in combination with rehydration and saline diuresis while waiting for more sustained effects from a simultaneously administered bisphosphonate such as pamidronate. Usual doses of calcitonin are 2–8 U/kg of body weight IV, SC, or IM every 6–12 h. *Plicamycin* (formerly mithramycin), which inhibits bone resorption and *gallium nitrate*, which exerts a hypocalcemic action also by inhibiting bone resorption, is no longer used because of superior alternatives such as bisphosphonates.

Glucocorticoids have utility, especially in hypercalcemia complicating certain malignancies. They increase urinary calcium excretion and decrease intestinal calcium absorption when given in pharmacologic doses, but they also cause negative skeletal calcium balance. In normal individuals and in patients with primary HPT, glucocorticoids neither increase nor decrease the serum calcium concentration. In patients with hypercalcemia due to certain osteolytic malignancies, however, glucocorticoids may be effective as a result of antitumor effects. The malignancies in which hypercalcemia responds to glucocorticoids include multiple myeloma, leukemia, Hodgkin's disease, other lymphomas, and carcinoma of the breast, at least early in the course of the disease. Glucocorticoids are also effective in treating hypercalcemia due to vitamin D intoxication and sarcoidosis. Glucocorticoids are also useful in the rare form of hypercalcemia, now recognized in certain autoimmune disorders in which inactivating antibodies against the receptor imitate FHH. Elevated PTH and calcium levels are effectively lowered by the glucocorticoids. In all the preceding situations, the hypocalcemic effect develops over several days, and the usual glucocorticoid dosage is 40–100 mg prednisone (or its equivalent) daily in four divided doses. The side effects of chronic glucocorticoid therapy may be acceptable in some circumstances.

Dialysis is often the treatment of choice for severe hypercalcemia complicated by renal failure, which is difficult to manage medically. Peritoneal dialysis with calcium-free dialysis fluid can remove 5–12.5 mmol (200–500 mg) of calcium in 24–48 h and lower the serum calcium concentration by 0.7–2.2 mmol/L (3–9 mg/dL). Large quantities of phosphate are lost during dialysis, and serum inorganic phosphate concentration usually falls, potentially aggravating hypercalcemia. Therefore, the serum inorganic phosphate concentration should be measured after dialysis, and phosphate supplements should be added to the diet or to dialysis fluids if necessary.

Phosphate therapy, PO or IV, has a limited role in certain circumstances **(Chap. 402)**. Correcting hypophosphatemia lowers the serum calcium concentration by several mechanisms, including bone/calcium exchange. The usual oral treatment is 1–1.5 g phosphorus per day for several days, given in divided doses. It is generally believed, but not established, that toxicity does not occur if therapy is limited to restoring serum inorganic phosphate concentrations to normal.

Raising the serum inorganic phosphate concentration above normal decreases serum calcium levels, sometimes strikingly. Intravenous phosphate is one of the most dramatically effective treatments available for severe hypercalcemia but is toxic and even dangerous (fatal hypocalcemia). For these reasons, it is used rarely and only in severely hypercalcemic patients with cardiac or renal failure where dialysis, the preferable alternative, is not feasible or is unavailable.

SUMMARY

The various therapies for hypercalcemia are listed in Table 403-4. The choice depends on the underlying disease, the severity of the hypercalcemia, the serum inorganic phosphate level, and the renal, hepatic, and bone marrow function. Mild hypercalcemia (≤3 mmol/L [12 mg/dL)]) can usually be managed by hydration. Severe hypercalcemia (≥3.7 mmol/L [15 mg/dL]) requires rapid correction. IV pamidronate, or zolendronate, or subcutaneous denosumab should be administered. In addition, for the first 24–48 h, aggressive sodium-calcium diuresis with IV saline should be given and, following rehydration, large doses of furosemide or ethacrynic acid, but only if appropriate monitoring is available and cardiac and renal function are adequate. Intermediate degrees of hypercalcemia between 3 and 3.7 mmol/L (12 and 15 mg/dL) should be approached with vigorous hydration and then the most appropriate selection for the patient of the combinations used with severe hypercalcemia.

HYPOCALCEMIA

(See also Chap. 50)

PATHOPHYSIOLOGY OF HYPOCALCEMIA: CLASSIFICATION BASED ON MECHANISM

Chronic hypocalcemia is less common than hypercalcemia; causes include chronic renal failure, hereditary and acquired hypoparathyroidism, vitamin D deficiency, PHP, and hypomagnesemia.

Acute rather than chronic hypocalcemia is seen in critically ill patients or as a consequence of certain medications and often does not require specific treatment. Transient hypocalcemia is seen with severe sepsis, burns, acute kidney injury, and extensive transfusions with citrated blood. Although as many as one-half of patients in an intensive care setting are reported to have calcium concentrations of <2.1 mmol/L (8.5 mg/dL), most do not have a reduction in ionized calcium. Patients with severe sepsis may have a decrease in ionized calcium (true hypocalcemia), but in other severely ill individuals, hypoalbuminemia is the primary cause of the reduced total calcium concentration. Alkalosis increases calcium binding to proteins, and in this setting direct measurements of ionized calcium should be made.

Medications such as protamine, heparin, and glucagon may cause transient hypocalcemia. These forms of hypocalcemia are usually not associated with tetany and resolve with improvement in the overall medical condition. The hypocalcemia after repeated transfusions of citrated blood usually resolves quickly.

Patients with *acute pancreatitis* have hypocalcemia that persists during the acute inflammation and varies in degree with disease severity. The cause of hypocalcemia remains unclear. PTH values are reported to be low, normal, or elevated, and both resistance to PTH and impaired PTH secretion have been postulated. Occasionally, a chronic low total calcium and low ionized calcium concentration are detected in an elderly patient without obvious cause and with a paucity of symptoms; the pathogenesis is unclear.

Chronic hypocalcemia, however, is usually symptomatic and requires treatment. Neuromuscular and neurologic manifestations of chronic hypocalcemia include muscle spasms, carpopedal spasm, facial grimacing, and, in extreme cases, laryngeal spasm and convulsions. Respiratory arrest may occur. Increased intracranial pressure occurs in some patients with long-standing hypocalcemia, often in association with papilledema. Mental changes include irritability, depression, and psychosis. The QT interval on the electrocardiogram is prolonged, in contrast to its shortening with hypercalcemia. Arrhythmias occur, and digitalis effectiveness may be reduced. Intestinal cramps and chronic malabsorption may occur. Chvostek's or Trousseau's sign can be used to confirm latent tetany.

The classification of hypocalcemia shown in Table 403-5 is based on an organizationally useful premise that PTH is responsible for minute-to-minute regulation of plasma calcium concentration and, therefore, that the occurrence of hypocalcemia must mean a failure of the homeostatic action of PTH. Failure of the PTH response can occur if there is hereditary or acquired parathyroid gland failure, if a mutant PTH is secreted, or if PTH is ineffective in target organs, or if the action of the hormone is overwhelmed by the loss of calcium from the ECF at a rate faster than it can be replaced.

PTH ABSENT

Whether hereditary or acquired, hypoparathyroidism has a number of common components. The disease is rare with estimates from all causes to be ~25–35 patients/100,000 of the population (based on U.S. and Danish estimates). Symptoms of untreated hypocalcemia are shared by both types of hypoparathyroidism, although the onset of hereditary hypoparathyroidism can be more gradual and associated with other developmental defects. Basal ganglia calcification and extrapyramidal syndromes are more common and earlier in onset in hereditary hypoparathyroidism. Acquired hypoparathyroidism secondary to surgery in the neck is still more common than hereditary hypoparathyroidism, but the frequency of surgically induced parathyroid failure has diminished as a result of improved surgical techniques that spare the parathyroid glands and increased use of nonsurgical therapy for hyperthyroidism. PHP, an example of ineffective PTH action rather than a failure of parathyroid gland production, may share several features with hypoparathyroidism, including extraosseous calcification and extrapyramidal manifestations such as choreoathetotic movements and dystonia.

Papilledema and raised intracranial pressure may occur in both hereditary and acquired hypoparathyroidism, as do chronic changes in fingernails and hair and lenticular cataracts, the latter usually reversible with treatment of hypocalcemia. Certain skin manifestations, including alopecia and candidiasis, are characteristic of hereditary hypoparathyroidism associated with autoimmune polyglandular failure **(Chap. 381)**.

Hypocalcemia associated with hypomagnesemia is associated with both deficient PTH release and impaired responsiveness to the hormone. Patients with hypocalcemia secondary to hypomagnesemia have absent or low levels of circulating PTH, indicative of diminished hormone release despite a maximum physiologic stimulus by hypocalcemia. Plasma PTH levels return to normal with correction of the hypomagnesemia. Thus hypoparathyroidism with low levels of PTH in blood can be due to hereditary gland failure, acquired gland failure, or acute but reversible gland dysfunction (hypomagnesemia).

Genetic Abnormalities and Hereditary Hypoparathyroidism

Hereditary hypoparathyroidism can occur as an isolated entity without other endocrine or dermatologic manifestations. More typically it is syndromic, occurring in association with other abnormalities such as defective development of the thymus or failure of other endocrine organs such as the adrenal, thyroid, or ovary **(Chap. 381)**. Hereditary hypoparathyroidism is often manifest within the first decade but may appear later.

Genetic defects associated with hypoparathyroidism serve to illuminate the complexity of organ development, hormonal biosynthesis and secretion, and tissue-specific patterns of endocrine effector function When hypoparathyroidism is associated with other developmental or organ defects, treatment of the hypocalcemia can still be effective.

A form of hypoparathyroidism associated with defective development of both the thymus and the parathyroid glands is termed the *DiGeorge syndrome*, or the *velocardiofacial syndrome*. Congenital cardiovascular, facial, and other developmental defects are present, and patients may die in early childhood with severe infections, hypocalcemia and seizures, or cardiovascular complications. Patients can survive into adulthood, and milder, incomplete forms occur. Most cases are sporadic, but an autosomal dominant form involving microdeletions of chromosome 22q11.2 has been described. Smaller deletions in chromosome 22 are seen in incomplete forms of the DiGeorge syndrome, appearing in childhood or adolescence, that are manifest primarily by parathyroid gland failure. The chromosome 22 defect is now termed *DSG1;* more recently, a defect in chromosome 10p is also recognized—now called *DSG2*. The phenotypes seem similar. Studies on the chromosome 22 defect have pinpointed a transcription factor, TBX1. Deletions of the orthologous mouse gene show a phenotype similar to the human syndrome.

Another autosomal dominant developmental defect, featuring hypoparathyroidism, deafness, and renal dysplasia (HDR) has been studied at the genetic level. Cytogenic abnormalities in some, but not all kindred, point to translocation defects on chromosome 10, as in DiGeorge syndrome. However, the lack of immunodeficiency and heart defects distinguishes the two syndromes. Mouse models, as well as deletional analysis in some HDR patients, have identified the transcription factor GATA3, which is important in embryonic development and is expressed in developing kidney, ear structures, and the parathyroids.

Another pair of linked developmental disorders involving the parathyroids is recognized. *Kenney-Caffey syndrome type I* features hypoparathyroidism, short stature, osteosclerosis, and thick cortical bones. A defect seen in Middle Eastern patients, particularly in Saudi Arabia, termed *Sanjad-Sakati syndrome*, also exhibits growth failure and other dysmorphic features. This syndrome, which is clearly autosomal recessive, involves a gene on chromosome 1q42-q43. Both syndromes apparently involve a chaperone protein, called *TBCE*, relevant to tubulin function. Recently, a defect in FAM111A was identified as the cause of *Kenney-Caffey syndrome type 2*.

Hypoparathyroidism can occur in association with a complex hereditary autoimmune syndrome involving failure of the adrenals, the ovaries, the immune system, and the parathyroids in association with recurrent mucocutaneous candidiasis, alopecia, vitiligo, and pernicious anemia **(Chap. 381)**. The responsible gene on chromosome 21q22.3 has been identified. The protein product, which resembles a transcription

factor, has been termed the *autoimmune regulator*, or AIRE. A stop codon mutation occurs in many Finnish families with the disorder, commonly referred to as *polyglandular autoimmune type 1 deficiency*, while another AIRE mutation (Y85C) is typically observed in Jews of Iraqi and Iranian descent.

Hypoparathyroidism is seen in two disorders associated with mitochondrial dysfunction and myopathy, one termed the *Kearns-Sayre syndrome* (KSS), with ophthalmoplegia and pigmentary retinopathy, and the other termed the *MELAS syndrome*, *m*itochondrial *e*ncephalopathy, *l*actic *a*cidosis, and *s*troke-like episodes. Mutations or deletions in mitochondrial genes have been identified.

Several forms of hypoparathyroidism, each rare in frequency, are seen as isolated defects; the genetic mechanisms are varied. The inheritance includes autosomal dominant, autosomal recessive, and X-linked modes. Three separate autosomal defects involving the parathyroid gene have been recognized: one is dominant and the other two are recessive. The dominant form has a point mutation in the signal sequence, a critical region involved in intracellular transport of the hormone precursor. An Arg for Cys mutation interferes with processing of the precursor and is believed to trigger an apoptotic cellular response, hence acting as a dominant negative. The other two forms are recessive. A point mutation also blocks cleavage of the PTH precursor but requires both alleles to be mutated. Another involves a single-nucleotide base change that results in an exon splicing defect; the lost exon contains the promoter—hence, the gene is silenced. An autosomal recessive form of the disease was found to be caused by an Arg to Cys mutation at amino acid residue 25 of the secreted PTH. An X-linked recessive form of hypoparathyroidism has been described in males and the defect has been localized to chromosome Xq26-q27, perhaps involving the *SOX3* gene.

Abnormalities in the CaSR are detected in three distinctive hypocalcemic disorders. All are rare but more than 10 different gain-of-function mutations have been found in one form of hypocalcemia termed *autosomal dominant hypocalcemic hypercalciuria (ADHH)*. The receptor senses the ambient calcium level as excessive and suppresses PTH secretion, leading to hypocalcemia. The hypocalcemia is aggravated by constitutive receptor activity in the renal tubule causing excretion of inappropriate amounts of calcium. Recognition of the syndrome is important because efforts to treat the hypocalcemia with vitamin D analogues and increased oral calcium exacerbate the already excessive urinary calcium excretion (several grams or more per 24 h), leading to irreversible renal damage from stones and ectopic calcification.

Other causes of isolated hypoparathyroidism include homozygous, inactivating mutations in the parathyroid-specific transcription factor GCM2, which lead to an autosomal recessive form of the disease, or heterozygous point mutations in GCM2, which have a dominant negative effect on the wild-type protein and thus lead to an autosomal dominant form of hypoparathyroidism. Furthermore, heterozygous mutations in Gα11, one of the two signaling proteins downstream of the CaSR, have been identified as a cause of autosomal dominant hypoparathyroidism.

The *Bartter syndrome* is a group of disorders associated with disturbances in electrolyte and acid/base balance, sometimes with nephrocalcinosis and other features. Several types of ion channels or transporters are involved. Curiously, *Bartter syndrome type V* has the electrolyte and pH disturbances seen in the other syndromes but appears to be due to a gain-of-function in the CaSR. The defect may be more severe than in ADHH and explains the additional features seen beyond hypocalcemia and hypercalciuria.

As with autoimmune disorders that block the CaSR (discussed above under hypercalcemic conditions), there are autoantibodies that at least transiently activate the CaSR, leading to suppressed PTH secretion and hypocalcemia. This disorder may wax and wane.

Acquired Hypoparathyroidism *Acquired chronic hypoparathyroidism* is usually the result of inadvertent surgical removal of all the parathyroid glands; in some instances, not all the tissue is removed, but the remainder undergoes vascular supply compromise secondary to fibrotic changes in the neck after surgery. In the past, the most frequent cause of acquired hypoparathyroidism was surgery for hyperthyroidism. Hypoparathyroidism now usually occurs after surgery for hyperparathyroidism when the surgeon, facing the dilemma of removing too little tissue and thus not curing the HPT, removes too much. Parathyroid function may not be totally absent in all patients with postoperative hypoparathyroidism.

Rare causes of acquired chronic hypoparathyroidism include radiation-induced damage subsequent to radioiodine therapy of hyperthyroidism and glandular damage in patients with hemochromatosis or hemosiderosis after repeated blood transfusions. Infection may involve one or more of the parathyroids but usually does not cause hypoparathyroidism because all four glands are rarely involved.

Transient hypoparathyroidism is frequent following surgery for HPT. After a variable period of hypoparathyroidism, normal parathyroid function may return due to hyperplasia or recovery of remaining tissue. Occasionally, recovery occurs months after surgery.

TREATMENT

Acquired and Hereditary Hypoparathyroidism

Conventional treatment has involved replacement with vitamin D or $1,25(OH)_2D_3$ (calcitriol) combined with a high oral calcium intake. In most patients, blood calcium and phosphate levels are satisfactorily regulated, but some patients show resistance and a brittleness, with a tendency to alternate between hypocalcemia and hypercalcemia. For many patients, vitamin D in doses of 40,000–120,000 U/d (1–3 mg/d) combined with ≥1 g elemental calcium is satisfactory. The wide dosage range reflects the variation encountered from patient to patient; precise regulation of each patient is required. Compared to typical daily requirements in euparathyroid patients of 200 U/d (or in older patients as high as 800 U/d), the high dose of vitamin D (as much as 100-fold higher) reflects the reduced conversion of vitamin D to $1,25(OH)_2D$. Many physicians now use 0.5–1 μg of calcitriol in management of such patients, especially if they are difficult to control. Because of its storage in fat, when vitamin D is withdrawn, weeks are required for the disappearance of the biologic effects, compared with a few days for calcitriol, which has a rapid turnover.

Oral calcium and vitamin D restore the overall calcium-phosphate balance but do not reverse the lowered urinary calcium reabsorption typical of hypoparathyroidism. Therefore, care must be taken to avoid excessive urinary calcium excretion after vitamin D and calcium replacement therapy; otherwise, nephrocalcinosis and kidney stones can develop, and the risk of CKD is increased. Thiazide diuretics lower urine calcium by as much as 100 mg/d in hypoparathyroid patients on vitamin D, provided they are maintained on a low-sodium diet. Use of thiazides seems to be of benefit in mitigating hypercalciuria and easing the daily management of these patients.

Hypoparathyroidism is rare among endocrine disorders in not being treated with the missing hormone. Recent developments have changed that. Experimental use of PTH(1-34), the synthetic fragment used in treatment of osteoporosis showed promise. Subsequently, the full length molecule PTH(1-84) has been shown to be effective and is now FDA-approved for therapy of hypoparathyroidism. Published reports illustrate that its use substantially reduced the requirements for supplemental calcium and active vitamin D to maintain serum calcium. Recommendations offered by a recent conference on management of hypoparathyroidism suggest its use, particularly in patients with inadequate control of blood calcium, requirement for inconveniently/excessively high doses of calcium and active vitamin D replacement, and/or high urine calcium.

Hypomagnesemia Severe hypomagnesemia (<0.4 mmol/L; <0.8 meq/L) is associated with hypocalcemia **(Chap. 402)**. Restoration of the total-body magnesium deficit leads to rapid reversal of hypocalcemia. There are at least two causes of the hypocalcemia—impaired PTH secretion and reduced responsiveness to PTH. **For further discussion of causes and treatment of hypomagnesemia, see Chap. 402.**

The effects of magnesium on PTH secretion are similar to those of calcium; hypermagnesemia suppresses and hypomagnesemia stimulates PTH secretion. The effects of magnesium on PTH secretion are normally of little significance, however, because the calcium effects dominate. Greater change in magnesium than in calcium is needed to influence hormone secretion. Nonetheless, hypomagnesemia might be expected to increase hormone secretion. It is therefore surprising to find that severe hypomagnesemia is associated with blunted secretion of PTH. The explanation for the paradox is that severe, chronic hypomagnesemia leads to intracellular magnesium deficiency, which interferes with secretion and peripheral responses to PTH. The mechanism of the cellular abnormalities caused by hypomagnesemia is unknown, although effects on adenylate cyclase (for which magnesium is a cofactor) have been proposed.

PTH levels are undetectable or inappropriately low in severe hypomagnesemia despite the stimulus of severe hypocalcemia, and acute repletion of magnesium leads to a rapid increase in PTH level. Serum phosphate levels are often not elevated, in contrast to the situation with acquired or idiopathic hypoparathyroidism, probably because phosphate deficiency is often seen in hypomagnesemia **(Chap. 363)**.

Diminished peripheral responsiveness to PTH also occurs in some patients, as documented by subnormal response in urinary phosphorus and urinary cyclic AMP excretion after administration of exogenous PTH to patients who are hypocalcemic and hypomagnesemic. Both blunted PTH secretion and lack of renal response to administered PTH can occur in the same patient. When acute magnesium repletion is undertaken, the restoration of PTH levels to normal or supranormal may precede restoration of normal serum calcium by several days.

TREATMENT

Hypomagnesemia

Repletion of magnesium cures the condition. Repletion should be parenteral. Attention must be given to restoring the intracellular deficit, which may be considerable. After IV magnesium administration, serum magnesium may return transiently to the normal range, but unless replacement therapy is adequate, serum magnesium will again fall. If the cause of the hypomagnesemia is renal magnesium wasting, magnesium may have to be given long term to prevent recurrence **(Chap. 402)**.

PTH INEFFECTIVE

PTH is ineffective when the PTHR1–signaling protein complex is defective (as in the different forms of PHP, discussed below); when PTH action to promote calcium absorption from the diet via the synthesis of $1,25(OH)_2D$ is insufficient because of vitamin D deficiency or because vitamin D is ineffective (defects in vitamin D receptor or vitamin D synthesis); or in CKD in which the calcium-elevating action of PTH is impaired.

Typically, hypophosphatemia is more severe than hypocalcemia in vitamin D deficiency states because of the increased secretion of PTH, which, although only partly effective in elevating blood calcium, is readily capable of promoting urinary phosphate excretion.

PHP, on the other hand, has a pathophysiology that is different from the other disorders of ineffective PTH action. PHP resembles hypoparathyroidism (in which PTH synthesis is deficient) and is manifested by hypocalcemia and hyperphosphatemia, yet elevated PTH levels. The cause of the disorder is defective PTH-dependent activation of the stimulatory G protein complex or the downstream effector protein kinase A, resulting in failure of PTH to increase intracellular cAMP or to respond to elevated cAMP levels (see below).

Chronic Kidney Disease Improved medical management of CKD now allows many patients to survive for decades and hence time enough to develop features of renal osteodystrophy, which must be controlled to avoid additional morbidity. Impaired production of $1,25(OH)_2D$ is now thought to be the principal factor that causes calcium deficiency, secondary HPT, and bone disease; hyperphosphatemia typically occurs only in the later stages of the disease. Low levels of $1,25(OH)_2D$ due to increased FGF23 production in bone are critical in the development of hypocalcemia. The uremic state also causes impairment of intestinal absorption by mechanisms other than defects in vitamin D metabolism. Nonetheless, treatment with supraphysiologic amounts of vitamin D or calcitriol can correct the impaired calcium absorption. Since increased FGF23 levels are seen even in early stages of CKD, and have been reported to correlate with increased mortality and left ventricular hypertrophy, there is current interest in approaches to lower intestinal phosphate absorption early during the course of kidney disease and to thereby lower FGF23 levels. However, there is concern as to whether vitamin D supplementation increases the circulating FGF23 levels in CKD patients. Although vitamin D analogs improve survival in this patient population, it is notable that there are often dramatic elevations of FGF23.

Hyperphosphatemia in CKD lowers blood calcium levels by several mechanisms, including extraosseous deposition of calcium and phosphate, impairment of the bone-resorbing action of PTH, and reduction in $1,25(OH)_2D$ production due to elevated FGF23 and diminished renal tissue.

TREATMENT

Chronic Kidney Disease

Therapy of CKD **(Chap. 305)** involves appropriate management of patients prior to dialysis and adjustment of regimens once dialysis is initiated. Attention should be paid to restriction of phosphate in the diet; avoidance of aluminum-containing phosphate-binding antacids to prevent the problem of aluminum intoxication; provision of an adequate calcium intake by mouth, usually 1–2 g/d; and supplementation with 0.25–1 μg/d calcitriol or other activated forms of vitamin D. Each patient must be monitored closely. The aim of therapy is to restore normal calcium balance to prevent osteomalacia and severe secondary HPT (it is usually recommended to maintain PTH levels between 100 and 300 pg/mL) and, in light of evidence of genetic changes and monoclonal outgrowths of parathyroid glands in CKD patients, to prevent secondary from becoming autonomous HPT. Reduction of hyperphosphatemia and restoration of normal intestinal calcium absorption by calcitriol can improve blood calcium levels and reduce the manifestations of secondary HPT. Since adynamic bone disease can occur in association with low PTH levels, it is important to avoid excessive suppression of the parathyroid glands while recognizing the beneficial effects of controlling the secondary HPT. These patients should probably be closely monitored with PTH assays that detect only the full-length PTH(1–84) to ensure that biologically active PTH and not inactive, inhibitory PTH fragments are measured. Use of phosphate-binding agents such as sevelamer are approved only in end-stage renal disease (ESRD), but it may be necessary to initiate such treatment much earlier during the course of kidney disease to prevent the increase in FGF23 and its "off-target" effects.

Vitamin D Deficiency Due to Inadequate Diet and/or Sunlight Vitamin D deficiency due to inadequate intake of dairy products enriched with vitamin D, lack of vitamin supplementation, and reduced sunlight exposure in the elderly, particularly during winter in northern latitudes, is more common in the United States than previously recognized. Biopsies of bone in elderly patients with hip fracture (documenting osteomalacia) and abnormal levels of vitamin D metabolites, PTH, calcium, and phosphate indicate that vitamin D deficiency may occur in as many as 25% of elderly patients, particularly in northern latitudes in the United States. Concentrations of 25(OH)D are low or low-normal in these patients. Quantitative histomorphometric analysis of bone biopsy specimens from such individuals reveals widened osteoid seams consistent with osteomalacia **(Chap. 402)**. PTH hypersecretion compensates for the tendency for the blood calcium to fall but also increases renal phosphate excretion and thus causes osteomalacia.

TABLE 403-6 Classification of Pseudohypoparathyroidism (PHP) and Pseudopseudohypoparathyroidism (PPHP)

TYPE	HYPOCALCEMIA, HYPERPHOSPHATEMIA	RESPONSE OF URINARY cAMP TO PTH	SERUM PTH	$G_s\alpha$ SUBUNIT DEFICIENCY	AHO	RESISTANCE TO HORMONES OTHER THAN PTH
PHP1A	Yes	↓	↑	Yes	Yes	Yes
PPHP	No	Normal	Normal	Yes	Yes	No
PHP1B	Yes	↓	↑	No	No	Yes (in some patients)
PHP2	Yes	Normal	↑	No	No	No
Acrodysostosis with hormonal resistance	Yes	Normal (but ↓ phosphaturic response)	↑	No	Yes	Yes

Abbreviations: ↓, decreased;↑, increased; AHO, Albright's hereditary osteodystrophy; PTH, parathyroid hormone.

Treatment involves adequate replacement with vitamin D and calcium until the deficiencies are corrected. Severe hypocalcemia rarely occurs in moderately severe vitamin D deficiency of the elderly, but vitamin D deficiency must be considered in the differential diagnosis of mild hypocalcemia.

Mild hypocalcemia, secondary HPT, severe hypophosphatemia, and a variety of nutritional deficiencies occur with gastrointestinal diseases. Hepatocellular dysfunction can lead to reduction in 25(OH)D levels, as in portal or biliary cirrhosis of the liver, and malabsorption of vitamin D and its metabolites, including $1,25(OH)_2D$, may occur in a variety of bowel diseases, hereditary or acquired. Hypocalcemia itself can lead to steatorrhea, due to deficient production of pancreatic enzymes and bile salts. Depending on the disorder, vitamin D or its metabolites can be given parenterally, guaranteeing adequate blood levels of active metabolites.

Defective Vitamin D Metabolism • **ANTICONVULSANT THERAPY** Anticonvulsant therapy with any of several agents induces acquired vitamin D deficiency by increasing the conversion of vitamin D to inactive compounds and/or causing resistance to its action. The more marginal the vitamin D intake in the diet, the more likely that anticonvulsant therapy will lead to abnormal mineral and bone metabolism.

VITAMIN D–DEPENDENT RICKETS TYPE I Vitamin D–dependent rickets type I, previously termed *pseudo-vitamin D–resistant rickets*, differs from true vitamin D–resistant rickets (vitamin D–dependent rickets type II, see below) in that it is typically less severe and the biochemical and radiographic abnormalities can be reversed with appropriate doses of the vitamin's active metabolite, $1,25(OH)_2D$. Physiologic amounts of calcitriol cure the disease (Chap. 402). This finding fits with the pathophysiology of the disorder, which is autosomal recessive, and is now known to be caused by mutations in the gene encoding 25(OH)D-1α-hydroxylase. Both alleles are inactivated in affected patients and compound heterozygotes, harboring distinct mutations, are common.

Clinical features include hypocalcemia, often with tetany or convulsions, hypophosphatemia, secondary HPT, and osteomalacia, often associated with skeletal deformities and increased alkaline phosphatase. Treatment involves physiologic replacement doses of $1,25(OH)_2D$ (Chap. 402).

VITAMIN D–DEPENDENT RICKETS TYPE II Vitamin D–dependent rickets type II results from end-organ resistance to the active metabolite $1,25(OH)_2D_3$. The clinical features resemble those of the type I disorder and include hypocalcemia, hypophosphatemia, secondary HPT, and rickets but also partial or total alopecia. Plasma levels of $1,25(OH)_2D$ are elevated, in keeping with the refractoriness of the end organs. This disorder is caused by mutations in the gene encoding the vitamin D receptor; treatment is difficult and requires regular, usually nocturnal calcium infusions, which dramatically improve growth, but do not restore hair growth (Chap. 402).

Pseudohypoparathyroidism PHP refers to a group of distinct inherited disorders. Patients affected by PHP type Ia (PHP-Ia) are characterized by symptoms and signs of hypocalcemia in association with distinctive skeletal and developmental defects. The hypocalcemia is due to a deficient response to PTH, which is probably restricted to the proximal renal tubules. Hyperplasia of the parathyroids, a response to hormone-resistant hypocalcemia, causes elevation of PTH levels. Studies, both clinical and basic, have clarified some aspects of these disorders, including the variable clinical spectrum, the pathophysiology, the genetic defects, and their mode of inheritance.

A working classification of the various forms of PHP is given in Table 403-6. The classification scheme is based on the signs of ineffective PTH action (low calcium and high phosphate), low or normal urinary cAMP response to exogenous PTH, the presence or absence of *Albright's hereditary osteodystrophy* (AHO), and assays to measure the concentration of the $G_s\alpha$ subunit of the adenylate cyclase enzyme. Using these criteria, there are four types: PHP types Ia and Ib (PHP1A and PHP1B); pseudopseudohypoparathyroidism (PPHP) and PHP-II (PHP2).

PHP1A AND PHP1B Individuals with PHP1, the most common of the disorders, show a deficient urinary cyclic AMP response to administration of exogenous PTH. Patients with PHP1 are divided into type Ia (PHP1A) and type Ib (PHP1B). Patients with PHP1A show evidence for AHO and reduced amounts of $G_s\alpha$ protein/activity, as determined in readily accessible tissues such as erythrocytes, lymphocytes, and fibroblasts. Only some PHP1B patients show typically AHO features, but they have normal $G_s\alpha$ activity. PHP1C, sometimes listed as a third form of PHP1, is really a variant of PHP1A, although the mutant $G_s\alpha$ shows normal activity in certain in vitro assays.

Most patients who have PHP1A reveal characteristic features of AHO, which consist of short stature, early-onset obesity, round face, obesity, skeletal anomalies (brachydactyly), intellectual impairment, and/or heterotopic calcifications. Patients have low calcium and high phosphate levels, as with true hypoparathyroidism. PTH levels, however, are elevated, reflecting resistance to hormone action.

Amorphous deposits of calcium and phosphate are found in the basal ganglia in about one-half of patients. The defects in metacarpal and metatarsal bones are sometimes accompanied by short phalanges as well, possibly reflecting premature closing of the epiphyses. The typical findings are short fourth and fifth metacarpals and metatarsals. The defects are usually bilateral. Exostoses and radius curvus are frequent.

Inheritance and Genetic Defects Multiple defects at the *GNAS* locus have now been identified in PHP1A, PHP1B, and PPHP patients. This gene, which is located on chromosome 20q13.3, encodes the α-subunit of the stimulatory G-protein ($G_s\alpha$), among other products (see below). Mutations include abnormalities in splice junctions associated with deficient mRNA production, point mutations, insertions, and/or deletion that all result in a protein with defective function resulting in a 50% reduction of $G_s\alpha$ activity in erythrocytes or other cells.

Detailed analyses of disease transmission in affected kindreds have clarified many features of PHP1A, PPHP, and PHP1B (Fig. 403-7). The former two entities, often traced through multiple generations, have an inheritance pattern consistent with genetic imprinting. The phenomenon of gene imprinting, involving methylation of genetic loci, independent of any mutation, impairs transcription from either the maternal or the paternal allele (Chap. 456). The $G_s\alpha$ transcript is biallelically expressed in most tissues; expression from paternal allele is silenced through as-of-yet unknown mechanisms in some tissues including the proximal renal tubules and the thyroid; consequently, inheritance of a

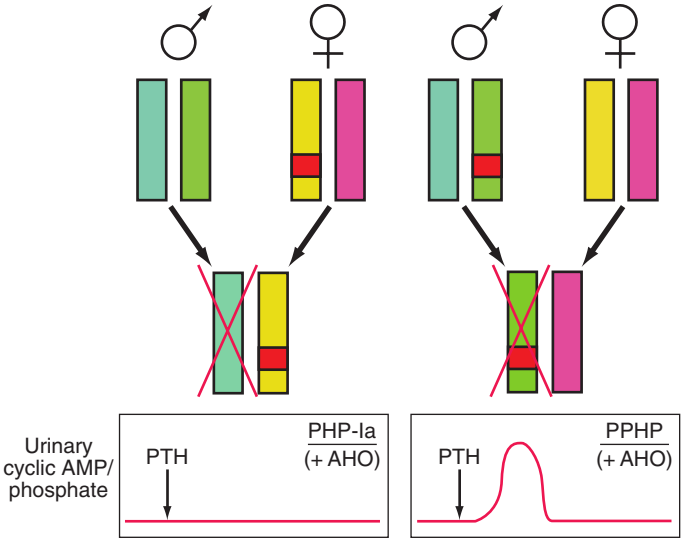


FIGURE 403-7 Paternal imprinting of renal parathyroid hormone (PTH) resistance (*GNAS* gene for $G_s\alpha$ subunit) in pseudohypoparathyroidism (PHP1A). An impaired excretion of urinary cyclic AMP and phosphate is observed in patients with PHP. In the renal cortex, there is selective silencing of the paternal $G_s\alpha$ expression. The disease becomes manifest only in patients who inherit the defective gene from an obligate female carrier (*left*). If the genetic defect is inherited from an obligate male gene carrier, there is no biochemical abnormality; administration of PTH causes an appropriate increase in the urinary cyclic AMP and phosphate concentration (pseudo-PHP [PPHP]; *right*). Both patterns of inheritance lead to Albright's hereditary osteodystrophy (AHO), perhaps because of haplotype insufficiency—i.e., both copies of $G_s\alpha$ must be active in the fetus for normal bone development.

defective paternal allele has no implications with regard to hormonal function. Thus, females affected by either PHP1A or PPHP will have offspring with PHP1A, if these children inherit the allele carrying the *GNAS* mutation; in contrast, if the mutant allele is inherited from a male affected by either disorder, the offspring will exhibit PPHP. Consistent with these data in humans, gene-ablation studies in mice have shown that inheritance of the mutant $G_s\alpha$ allele from the female causes much reduced $G_s\alpha$ protein in renal cortex, hypocalcemia, and resistance to PTH. Offspring inheriting the mutant allele from the male showed no evidence of PTH resistance or hypocalcemia.

Imprinting is tissue selective. Paternal $G_s\alpha$ expression is not silenced in most tissues. It seems likely, therefore, that the AHO phenotype recognized in PPHP as well as PHP1A reflects $G_s\alpha$ haploinsufficiency during embryonic or postnatal development.

The complex mechanisms that control the *GNAS* gene contribute to challenges involved in unraveling the pathogenesis of these disorders, especially that of PHP1B. Much intensive work with families in which multiple members are affected by PHP1B, as well as studies of the complex regulation of the *GNAS* gene locus, have now shown that autosomal dominant PHP1B is caused by microdeletions within or up-stream of the maternal *GNAS* locus, which are associated with a loss of DNA methylation at one or several loci of the maternal allele (Table 403-6). These abnormalities in methylation silence the expression of the gene. This leads in the proximal renal tubules—where $G_s\alpha$ appears to be expressed exclusively from the maternal allele—to PTH resistance.

PHP1B, lacking the AHO phenotype in most instances, shares with PHP1A the hypocalcemia and hyperphosphatemia caused by PTH resistance, and thus the blunted urinary cyclic AMP response to administered PTH, a standard test to assess the presence or absence of hormone resistance (Table 403-6). Furthermore, these endocrine abnormalities become apparent only if the disease-causing mutation is inherited maternally. Bone responsiveness may be excessive rather than blunted in PHP1B (and in PHP1A) patients, based on case reports that have emphasized an osteitis fibrosa–like pattern in several PHP1B patients.

PHP2 refers to patients with hypocalcemia and hyperphosphatemia, who have a normal urinary cyclic AMP, but an impaired urinary phosphaturic response to PTH. In a PHP2 variant, referred to as acrodysostosis with hormonal resistance (ADOHR), patients have a defect in the regulatory subunit of PKA (PRKAR1A) that mediates the response to PTH distal to cAMP production. Acrodysostosis without hormonal resistance is caused by mutations in the cAMP-selective phosphodiesterase 4 (ADOP4). It remains unclear why the PTH-resistance in some patients, labeled as PHP2 without bony abnormalities, resolves upon treatment with vitamin D supplements.

The diagnosis of these hormone-resistant states can usually be made without difficulty when there is a positive family history for features of AHO, in association with the signs and symptoms of hypocalcemia. In both categories—PHP1A and PHP1B—serum PTH levels are elevated, particularly when patients are hypocalcemic. However, patients with PHP1B or PHP2 without acrodysostosis present only with hypocalcemia and high PTH levels, as evidence for hormone resistance. In PHP1A and PHP1B, the response of urinary cyclic AMP to the administration of exogenous PTH is blunted. The diagnosis of PHP2, in the absence of acrodysostosis, is more complex and vitamin D deficiency must be excluded before such a diagnosis can be entertained.

TREATMENT

Pseudohypoparathyroidism

Treatment of PHP is similar to that of hypoparathyroidism, except that calcium and activated vitamin D doses are usually higher. Patients with PHP show no PTH resistance in the distal tubules—hence, urinary calcium clearance is typically reduced and they are not at risk of developing nephrocalcinosis as patients with true hypoparathyroidism, unless overtreatment occurs, for example, after the completion of pubertal development and skeletal mutation, when calcium and $1,25(OH)_2$ treatment should be reduced. Variability in response makes it necessary to establish the optimal regimen for each patient, based on maintaining appropriate blood calcium level and urinary calcium excretion, and keeping the PTH level within or slightly above the normal range.

PTH OVERWHELMED

Occasionally, loss of calcium from the ECF is so severe that PTH cannot compensate. Such situations include acute pancreatitis and severe, acute hyperphosphatemia, often in association with renal failure, conditions in which there is rapid efflux of calcium from the ECF. Severe hypocalcemia can occur quickly; PTH rises in response to hypocalcemia but does not return blood calcium to normal.

Severe, Acute Hyperphosphatemia Severe hyperphosphatemia is associated with extensive tissue damage or cell destruction (Chap. 402). The combination of increased release of phosphate from muscle and impaired ability to excrete phosphorus because of renal failure causes moderate to severe hyperphosphatemia, the latter causing calcium loss from the blood and mild to moderate hypocalcemia. Hypocalcemia is usually reversed with tissue repair and restoration of renal function as phosphorus and creatinine values return to normal. There may even be a mild hypercalcemic period in the oliguric phase of renal function recovery. This sequence, severe hypocalcemia followed by mild hypercalcemia, reflects widespread deposition of calcium in muscle and subsequent redistribution of some of the calcium to the ECF after phosphate levels return to normal.

Other causes of hyperphosphatemia include hypothermia, massive hepatic failure, and hematologic malignancies, either because of high cell turnover of malignancy or because of cell destruction by chemotherapy.

TREATMENT

Severe, Acute Hyperphosphatemia

Treatment is directed toward lowering of blood phosphate by the administration of phosphate-binding antacids or dialysis, often needed for the management of CKD. Although calcium replacement

may be necessary if hypocalcemia is severe and symptomatic, calcium administration during the hyperphosphatemic period tends to increase extraosseous calcium deposition and aggravate tissue damage. The levels of $1,25(OH)_2D$ may be low during the hyperphosphatemic phase and return to normal during the oliguric phase of recovery.

Osteitis Fibrosa after Parathyroidectomy Severe hypocalcemia after parathyroid surgery is rare now that osteitis fibrosa cystica is an infrequent manifestation of HPT. When osteitis fibrosa cystica is severe, however, bone mineral deficits can be large. After parathyroidectomy, hypocalcemia can persist for days if calcium replacement is inadequate. Treatment may require parenteral administration of calcium; addition of calcitriol and oral calcium supplementation is sometimes needed for weeks to a month or two until bone defects are filled (which, of course, is of therapeutic benefit in the skeleton), making it possible to discontinue parenteral calcium and/or reduce the amount.

DIFFERENTIAL DIAGNOSIS OF HYPOCALCEMIA

Care must be taken to ensure that true hypocalcemia is present; in addition, acute transient hypocalcemia can be a manifestation of a variety of severe, acute illnesses, as discussed above. *Chronic hypocalcemia*, however, can usually be ascribed to a few disorders associated with absent or ineffective PTH. Important clinical criteria include the duration of the illness, signs or symptoms of associated disorders, and the presence of features that suggest a hereditary abnormality. A nutritional history can be helpful in recognizing a low intake of vitamin D and calcium in the elderly, and a history of excessive alcohol intake may suggest magnesium deficiency.

Hypoparathyroidism and PHP are typically lifelong illnesses, usually (but not always) appearing by adolescence; hence, a recent onset of hypocalcemia in an adult is more likely due to nutritional deficiencies, renal failure, or intestinal disorders that result in deficient or ineffective vitamin D. Neck surgery, even long past, however, can be associated with a delayed onset of postoperative hypoparathyroidism. A history of seizure disorder raises the issue of anticonvulsive medication. Developmental defects may point to the diagnosis of PHP. Rickets and a variety of neuromuscular syndromes and deformities may indicate ineffective vitamin D action, either due to defects in vitamin D metabolism or to vitamin D deficiency.

A pattern of *low calcium with high phosphorus* in the absence of renal failure or massive tissue destruction almost invariably means hypoparathyroidism or PHP. A *low calcium and low phosphorus* pattern points to absent or ineffective vitamin D, thereby impairing the action of PTH on calcium metabolism (but not phosphate clearance). The relative ineffectiveness of PTH in calcium homeostasis in vitamin D deficiency, anticonvulsant therapy, gastrointestinal disorders, and hereditary defects in vitamin D metabolism leads to secondary HPT as a compensation. The excess PTH on renal tubule phosphate transport accounts for renal phosphate wasting and hypophosphatemia.

Exceptions to these patterns may occur. Most forms of hypomagnesemia are due to long-standing nutritional deficiency as seen in chronic alcoholics. Despite the fact that the hypocalcemia is principally due to an acute absence of PTH, phosphate levels are usually low, rather than elevated, as in hypoparathyroidism. Chronic renal failure is often associated with hypocalcemia and hyperphosphatemia, despite secondary HPT.

Diagnosis is usually established by application of the PTH immunoassay, tests for vitamin D metabolites, and measurements of the urinary cyclic AMP response to exogenous PTH. In hereditary and acquired hypoparathyroidism and in severe hypomagnesemia, PTH is either undetectable or inappropriately in the normal range (Fig. 403-4). This finding in a hypocalcemic patient is supportive of hypoparathyroidism, as distinct from ineffective PTH action, in which even mild hypocalcemia is associated with elevated PTH levels. Hence a failure to detect elevated PTH levels establishes the diagnosis of hypoparathyroidism; elevated levels suggest the presence of secondary HPT, as found in many of the situations in which the hormone is ineffective due to associated abnormalities in vitamin D action. Assays for 25(OH)D can be helpful. Low or low-normal 25(OH)D indicates vitamin D deficiency due to lack of sunlight, inadequate vitamin D intake, or intestinal malabsorption. Recognition that mild hypocalcemia, rickets, and hypophosphatemia are due to anticonvulsant therapy is made by history.

FURTHER READING

Bastepe M, Jüppner H: Pseudohypoparathyroidism, Albright's hereditary osteodystrophy, and progressive osseous heteroplasia: Disorders caused by inactivating GNAS mutations, in *Endocrinology*, 7th ed, JL Jameson, LJ DeGroot (eds), 2016, pp. 1147–1159.

Bilezikian J et al: Guidelines for the Management of Asymptomatic Primary Hyperparathyroidism: Summary Statement from the Fourth International Workshop. J Clin Endocrinol Metab 99:3561, 2014.

Brandi M et al: Management of Hypoparathyroidism: Summary Statement and Guidelines. J Clin Endocrinol Metab 101:2273, 2016.

Gardella T et al: Parathyroid hormone and the parathyroid hormone receptor type 1 in the regulation of calcium and phosphate homeostasis and bone metabolism, in *Endocrinology*, 7th ed, JL Jameson, LJ DeGroot (eds), 2016, pp. 969–990.

Thakker R et al: Regulation of calcium homeostasis and genetic disorders that affect calcium metabolism, in *Endocrinology*, 7th ed, JL Jameson, LJ DeGroot (eds), 2016, pp. 1063–1089.

404 Osteoporosis

Robert Lindsay, Felicia Cosman

Osteoporosis, a condition characterized by decreased bone strength, is prevalent among postmenopausal women but also occurs in both women and men as a function of age and with underlying conditions or major risk factors associated with bone demineralization. Its chief clinical manifestations are vertebral and hip fractures, although fractures can occur at almost any skeletal site. Osteoporosis affects >10 million individuals in the United States, but only a small proportion are diagnosed and treated.

DEFINITION

Osteoporosis is defined as a reduction in the strength of bone that leads to an increased risk of fractures. Loss of bone tissue is associated with deterioration in skeletal microarchitecture. The World Health Organization (WHO) operationally defines osteoporosis as a bone density that falls 2.5 standard deviations (SD) below the mean for young healthy adults of the same sex and race—also referred to as a *T-score* of –2.5. Postmenopausal women at the lower end of the young normal range (a T-score <–1.0) are defined as having low bone density and are also at increased risk of osteoporosis. Although risk is lower in this group, >50% of fractures among postmenopausal women, including hip fractures, occur in this group with low bone density. As a consequence, clinical assessment has evolved to include absolute risk of fracture, which incorporates bone mineral density (BMD) with age, gender, and other clinical risk factors to calculate 10-year fracture risk.

Osteoporosis-related fractures are adulthood fractures of any bone that occur in the setting of trauma less than or equal to a fall from standing height, with the exceptions of fingers, toes, face and skull.

EPIDEMIOLOGY

In the United States, as many as 8 million women and 2 million men have osteoporosis (BMD T-score <–2.5 at lumbar spine, total hip or femoral neck), and >40 million individuals have bone mass levels that put them at increased risk of developing osteoporosis (e.g. BMD T-score <–1.0). Osteoporosis occurs more frequently with increasing age, as bone tissue is lost progressively. In women, the loss of ovarian function at menopause (typically around age 50) precipitates rapid bone loss so

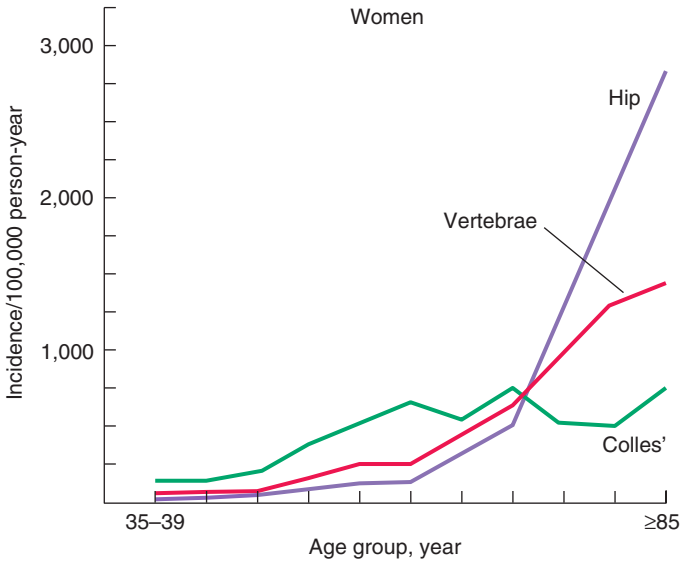


FIGURE 404-1 Epidemiology of vertebral, hip, and Colles' fractures with age. *(Adapted from C Cooper, LJ Melton III: Trends Endocrinol Metab 3:224, 1992; with permission.)*

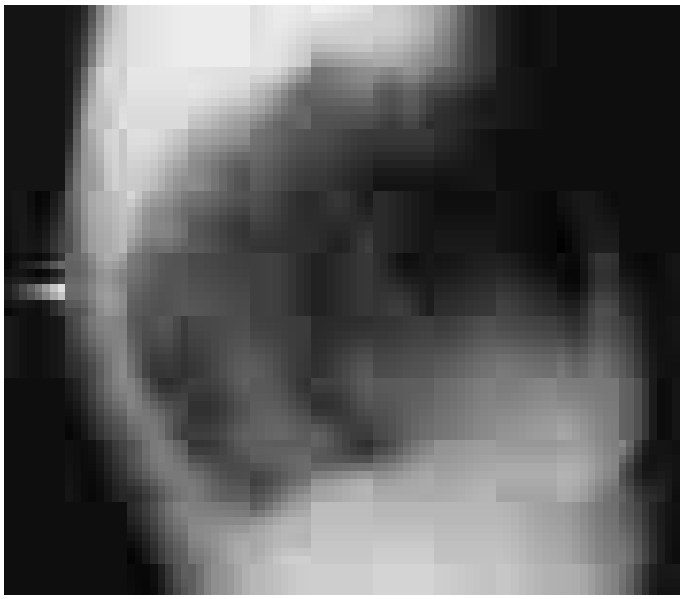


FIGURE 404-2 Lateral spine x-ray showing severe osteopenia and a severe wedge-type deformity (severe anterior compression).

that most women meet the diagnostic criterion for osteoporosis by age 70–80. As the population ages, the number of individuals with osteoporosis and fractures will also increase, despite a recognized reduction in age specific risk. It is estimated that currently about 2 million fractures occur each year in the United States as a consequence of osteoporosis. Many of the fractures defined as related to osteoporosis occur in individuals with low bone mass. Within that population, segregation of those at high risk of facture for treatment has become an important issue in clinical management.

The epidemiology of fractures follows the trend for loss of bone density, with most fractures, especially those of the hip and vertebrae, showing exponential increases with advancing age. **(Fig. 404-1)**. Lifetime osteoporotic fracture risk for a woman who reaches the age of 50 is about 50% and corresponding risk for a 50-year old man is about 20%.

About 300,000 hip fractures occur each year in the United States, almost all requiring hospital admission and emergency surgical intervention. The lifetime probability that a 50-year-old white individual will have a hip fracture is 14% for women and 5% for men; the risk for African-Americans is about half of those rates, and the risk for Asians and nonblack Hispanics appears similar to that for Caucasians. Hip fractures are associated with a high incidence of deep-vein thrombosis and pulmonary embolism and a mortality rate between 5 and 20% during the year after surgery, with higher mortality rates among males and African Americans. There is also significant morbidity after hip fracture, with about 30% of survivors requiring long-term care (at least temporarily), and many never regaining the independence that they had prior to the fracture.

There are about 500,000 symptomatic vertebral fractures per year in the United States, but probably >1,000,000 vertebral fractures occur yearly that are not recognized clinically at the time of the event. Many of these initially "silent" vertebral fractures are identified incidentally during radiography for other purposes **(Fig. 404-2)**. Even when asymptomatic, these vertebral fractures are a major sign of skeletal fragility and carry the same predictive value for subsequent fracture. Vertebral fractures rarely require hospitalization, but are associated with long-term morbidity and an increase in mortality rates, primarily related to pulmonary disease. Multiple vertebral fractures lead to height loss (often of several inches), kyphosis, and secondary pain and discomfort related to altered biomechanics of the back. Thoracic fractures can be associated with restrictive lung disease, whereas lumbar fractures are associated with abdominal symptoms that include distention, early satiety, and constipation.

Approximately 400,000 wrist fractures occur in the United States each year. Fractures of other bones (including about 150,000 pelvic fractures and >100,000 proximal humerus fractures) also occur with osteoporosis. Although some fractures result from major trauma, the threshold for fracture is reduced in osteoporotic bone **(Fig. 404-3)**. In addition to reduced bone density with advancing age, there are a number of risk factors for fracture; the common ones are summarized in **Table 404-1**. Prior fractures, a family history of osteoporosis-related fractures (particularly hip fractures), low body weight, cigarette consumption, and excessive alcohol consumption are all independent predictors of fracture. Chronic diseases with inflammatory components that increase skeletal remodeling, such as rheumatoid arthritis, increase the risk of osteoporosis, as do diseases associated with malabsorption. Chronic diseases that increase the risk of falling or frailty, including dementia, Parkinson's disease, and multiple sclerosis, also increase fracture risk (Table 404-1).

In the United States and Europe, osteoporosis-related fractures are more common among women than men, presumably due to a lower peak bone mass as well as postmenopausal bone loss in women. However, this gender difference in bone density and age-related increase in hip fractures is not as apparent in some other cultures, possibly due to genetics, physical activity level, or diet.

Fractures are themselves risk factors for future fractures (Table 404-1). Vertebral fractures increase the risk of other vertebral fractures as well as fractures of the peripheral skeleton such as the hip and wrist. Wrist fractures also increase the risk of vertebral and hip fractures. Among individuals aged >50, any fracture except those of the fingers, toes, face, and skull should be considered as potentially related to osteoporosis regardless of the specific circumstances of the fracture. Osteoporotic bone is more likely to fracture than is normal bone at any level of trauma, and a fracture in a person aged >50 should trigger evaluation

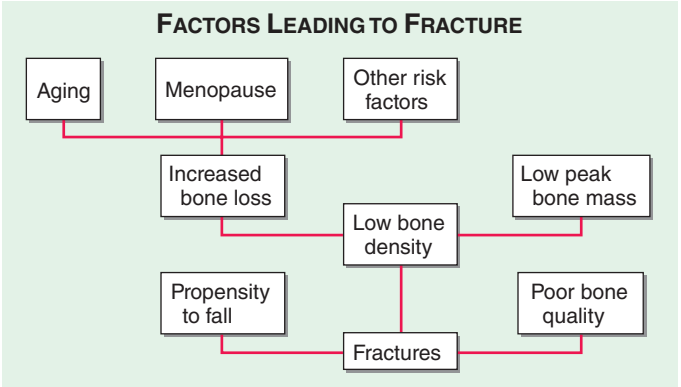


FIGURE 404-3 Factors leading to osteoporotic fractures.

TABLE 404-1 Risk Factors for Osteoporosis Fracture

NONMODIFIABLE	POTENTIALLY MODIFIABLE
Personal history of fracture as an adult	Current cigarette smoking
History of fracture in first-degree relative	Estrogen deficiency
Female gender	Early menopause (<45 years) or bilateral ovariectomy
Advanced age	Prolonged premenstrual amenorrhea (>1 year)
White race	Poor nutrition especially low calcium and vitamin D intake
Dementia	Alcoholism
	Impaired eyesight despite adequate correction
	Recurrent falls
	Inadequate physical activity
	Poor health/frailty

for osteoporosis. This often does not occur since postfracture care is fragmented. Recent attempts to coordinate care with one individual assuming the responsibility for guiding patients through the system and ensuring their evaluation and treatment for osteoporosis may improve care, but is more difficult to do in the open medical care systems in the United States. In countries with single payor systems, that approach does seem to be effective, as is also the case in closed health care systems in the United States.

The risk for future fracture after a first fracture is not linear. Highest risk occurs within the first 2 years after the first fracture. A recent large Medicare Database study indicated that almost 20% of women will have a second fracture within 2 years after the first. Risk diminishes to less than half of those rates in the subsequent 3 years and declines to baseline thereafter for most fracture types, though risk after a vertebral or hip fracture may persist a bit longer.

PATHOPHYSIOLOGY

BONE REMODELING

Osteoporosis results from bone loss due to age-related changes in bone remodeling as well as extrinsic and intrinsic factors that exaggerate this process. These changes may be superimposed on a low peak bone mass. Consequently, understanding the bone remodeling process is fundamental to understanding the pathophysiology of osteoporosis (Chap. 402). During growth, the skeleton increases in size by linear growth and by apposition of new bone tissue on the outer surfaces of the cortex (Fig. 404-4). The latter process is called *modeling*, a process that also allows the long bones to adapt in shape to the stresses placed on them. Increased sex hormone production at puberty is required for skeletal maturation, which reaches maximum mass and density in early adulthood. The sexual dimorphism in skeletal size becomes obvious after puberty, although true bone density remains similar between the sexes. Nutrition and lifestyle also play an important role in growth, though genetic factors primarily determine peak skeletal mass and density.

Numerous genes control skeletal growth, peak bone mass, and body size, as well as skeletal structure and density. Heritability estimates of 50–80% for bone density and size have been derived on the basis of twin studies. Though peak bone mass is often lower among individuals with a family history of osteoporosis, association studies of candidate genes (vitamin D receptors; type I collagen, estrogen receptors [ER], and interleukin 6 [IL-6]; and insulin-like growth factor I [IGF-I]) and bone mass, bone turnover, and fracture prevalence have been inconsistent. Linkage studies suggest that a genetic locus on chromosome 11 is associated with high bone mass. Families with high bone mass and without much apparent age-related bone loss have been shown to have a point mutation in LRP5, a low-density lipoprotein receptor–related protein. The role of this gene in the general population is not clear, although a nonfunctional mutation results in osteoporosis-pseudoglioma syndrome, and LRP5 signaling appears to be important in controlling bone formation. Genome-wide scans for low bone mass suggest multiple genes are involved, many of which are also implicated also in control of body size.

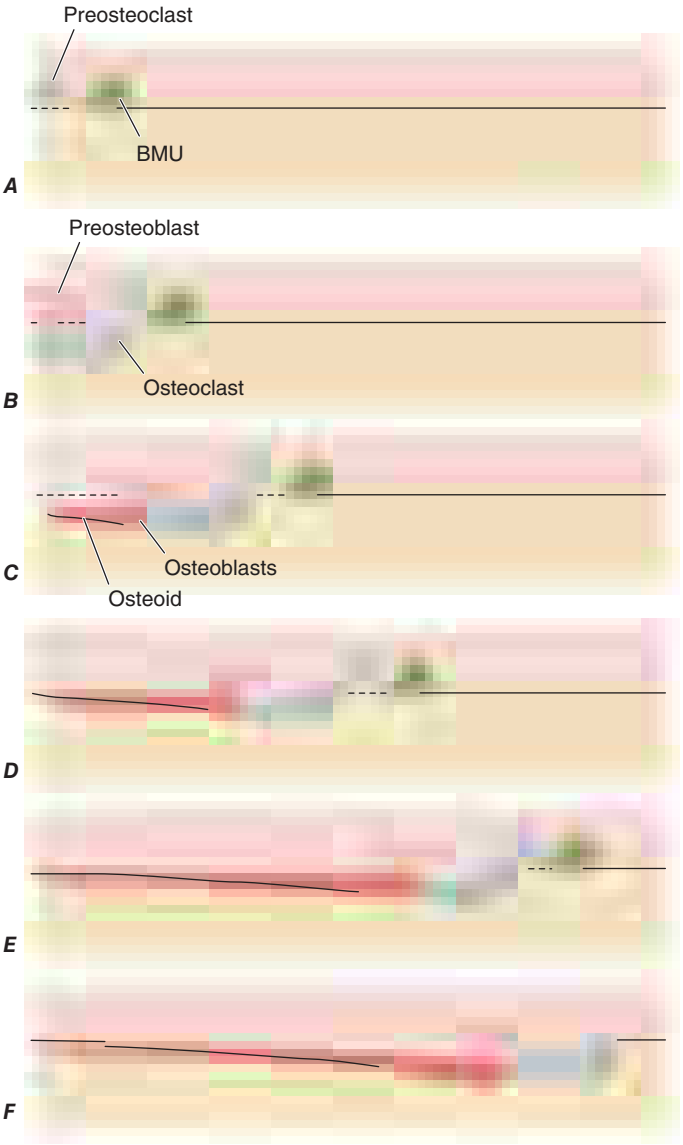


FIGURE 404-4 Mechanism of bone remodeling. The basic molecular unit (BMU) moves along the trabecular surface at a rate of about 10 μm/d. The figure depicts remodeling over ~120 days. ***A.*** Origination of BMU-lining cells contracts to expose collagen and attract preosteoclasts. ***B.*** Osteoclasts fuse into multinucleated cells that resorb a cavity. Mononuclear cells continue resorption, and preosteoblasts are stimulated to proliferate. ***C.*** Osteoblasts align at bottom of cavity and start forming osteoid (black). ***D.*** Osteoblasts continue formation and mineralization. Previous osteoid starts to mineralize (*horizontal lines*). ***E.*** Osteoblasts begin to flatten. ***F.*** Osteoblasts turn into lining cells; bone remodeling at initial surface (left of drawing) is now complete, but BMU is still advancing (to the right). *(Adapted from SM Ott, in JP Bilezikian et al [eds]: Principles of Bone Biology, vol. 18. San Diego, Academic Press, 1996, pp 231–241.)*

In adults, bone remodeling, not modeling, is the principal metabolic skeletal process. Bone remodeling has two primary functions: (1) to repair microdamage within the skeleton to maintain skeletal strength and ensure the relative youth of the skeleton and (2) to supply calcium when needed from the skeleton to maintain serum calcium. Remodeling may be activated by microdamage to bone as a result of excessive or accumulated stress. Acute demands for calcium involve osteoclast-mediated resorption as well as calcium transport by osteocytes. Chronic demands for calcium can result in secondary hyperparathyroidism, increased bone remodeling, and overall loss of bone tissue. Bone remodeling occurs through well coordinated activity of osteocytes, osteoblasts, and osteoclasts. Osteocytes are the terminal-differentiated cells derived from osteoblasts after incorporation into newly formed bone tissue. Osteoblasts derive from mesenchymal cell lineage and osteoclasts from monocyte/macrophage lineage.

Bone remodeling also is regulated by multiple hormones, including estrogens (in both genders), androgens, vitamin D, and parathyroid hormone (PTH), as well as locally produced growth factors, such as IGF-I, transforming growth factor β (TGF-β), PTH–related peptide (PTHrP), interleukins (ILs), prostaglandins, and members of the tumor necrosis factor (TNF) superfamily. These factors primarily modulate the rate at which new remodeling sites are activated, a process that results initially in bone resorption by osteoclasts, followed by a period of repair during which new bone tissue is synthesized by osteoblasts **(Chap. 402)**. The cytokine responsible for communication between the osteoblasts, other marrow cells, and osteoclasts is RANK ligand (RANKL) (receptor activator of nuclear factor-kappa-B [NFκB]; RANKL). RANKL, a member of the TNF family, is secreted by osteocytes, osteoblasts, and certain cells of the immune system. The osteoclast receptor for this protein is referred to as *RANK*. Activation of RANK by RANKL is a final common path in osteoclast development and activation. A humoral decoy for RANKL, also secreted by osteoblasts, is referred to as *osteoprotegerin* **(Fig. 404-5)**. Modulation of osteoclast recruitment and activity appears to be related to the interplay among these three factors. Additional influences include nutrition (particularly calcium intake) and physical activity level. RANKL production is in part regulated by the canonical Wnt signaling pathway. Wnt activation through mechanical loading, or by hormonal or cytokine factors, stimulates bone formation by increasing formation and activity of osteoblasts and decreases RANKL secretion, which inhibits production and activity of osteoclasts. Sclerostin, also an osteocyte protein, is a major inhibitor of Wnt activation and bone formation. Both the RANKL and Wnt pathways have become major targets for pharmacologic treatment of osteoporosis (see below).

In young adults, resorbed bone is replaced by an equal amount of new bone tissue. Thus, the mass of the skeleton remains constant after peak bone mass is achieved by the age of about 20. After age 30–45, however, the resorption and formation processes become imbalanced, and resorption exceeds formation. This imbalance may begin at different ages and varies at different skeletal sites; it becomes exaggerated in women after menopause. Excessive bone loss can be due to an increase in osteoclastic activity and/or a decrease in osteoblastic activity. In addition, an increase in remodeling activation frequency, and thus the number of remodeling sites, can magnify the small imbalance seen at each remodeling unit. Increased recruitment of bone remodeling sites produces a reversible reduction in bone tissue but also can result in permanent loss of tissue and disrupted skeletal architecture. In trabecular bone, if the osteoclasts penetrate trabeculae, they leave no template for new bone formation to occur, and, consequently, rapid bone loss ensues and cancellous connectivity becomes impaired. A higher number of remodeling sites increases the likelihood of this event. In cortical bone, increased activation of remodeling creates more porous bone. The effect of this increased porosity on cortical bone strength may be modest if the overall diameter of the bone is not changed. However, decreased apposition of new bone on the periosteal surface coupled with increased endocortical resorption of bone decreases the biomechanical strength of long bones. Even a slight exaggeration in normal bone loss increases the risk of osteoporosis-related fractures because of the architectural changes that occur, and osteoporosis is largely a disease of disordered skeletal architecture, although currently the only clinical tool generally available (dual-energy x-ray absorptiometry [DXA]) measures mass (an estimate of the mineral in bone) not architecture. Several tools are becoming available that may give more insight into the architecture of the skeleton (including Trabecular Bone Score, a noninvasive addition to DXA).

CALCIUM NUTRITION

Peak bone mass may be impaired by inadequate calcium intake during growth among other nutritional factors (calories, protein, and other minerals), leading to increased risk of osteoporosis later in life. During the adult phase of life, insufficient calcium intake contributes to secondary hyperparathyroidism and an increase in the rate of bone remodeling to assist in maintaining normal serum calcium levels. PTH stimulates the hydroxylation of vitamin D in the kidney, leading to increased levels of 1,25-dihydroxyvitamin D [1,25$(OH)_2$D] and enhanced gastrointestinal calcium absorption. PTH also reduces renal calcium loss. Although these are all appropriate compensatory homeostatic responses for adjusting calcium economy, the long-term effects

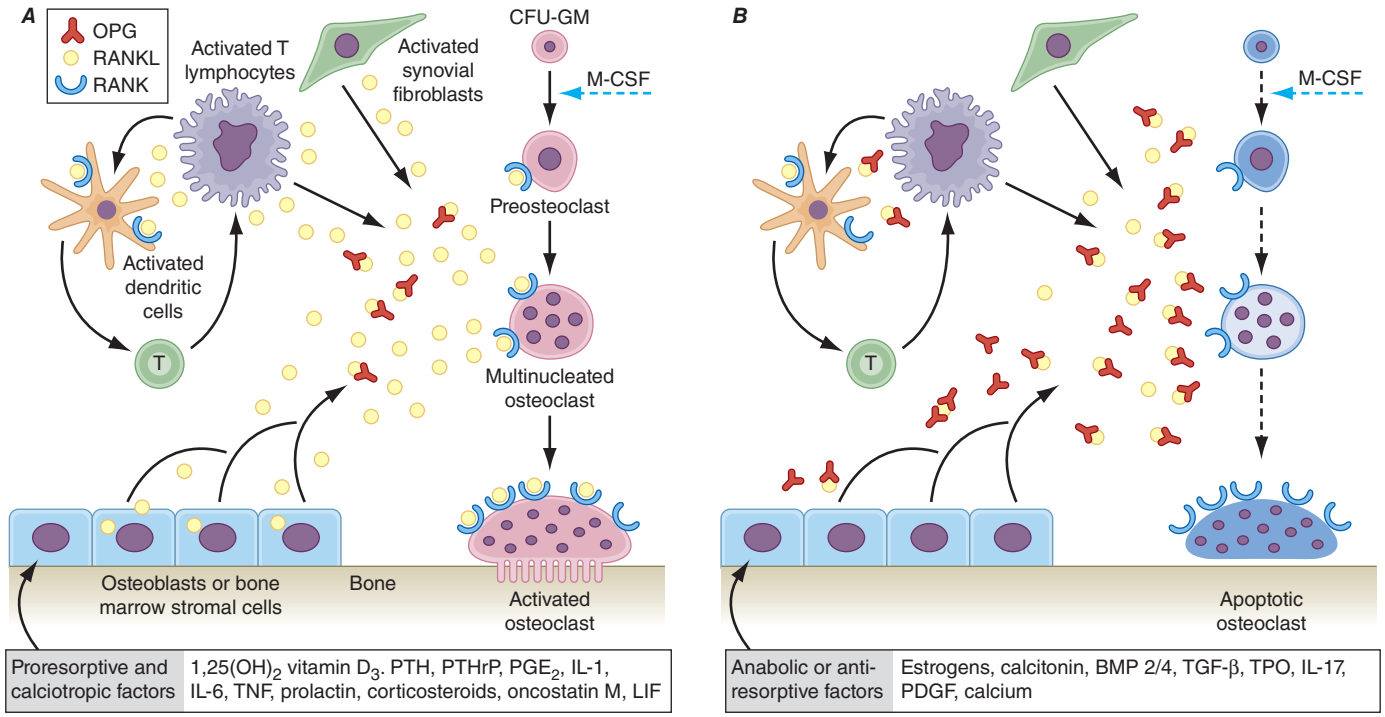


FIGURE 404-5 Hormonal control of bone resorption. ***A.*** Proresorptive and calciotropic factors. ***B.*** Anabolic and antiosteoclastic factors. RANKL expression is induced in osteoblasts, activated T cells, synovial fibroblasts, and bone marrow stromal cells. It binds to membrane-bound receptor RANK to promote osteoclast differentiation, activation, and survival. Conversely, osteoprotegerin (OPG) expression is induced by factors that block bone catabolism and promote anabolic effects. OPG binds and neutralizes RANKL, leading to a block in osteoclastogenesis and decreased survival of preexisting osteoclasts. CFU-GM, colony-forming units, granulocyte macrophage; M-CSF, macrophage colony-stimulating factor; RANKL, receptor activator of nuclear factor NFκB; PTH, parathyroid hormone; PGE_2, prostaglandin E_2; TNF, tumor necrosis factor; LIF, leukemia inhibitory factor; TPO, thrombospondin; PDGF, platelet-derived growth factor; OPG-L, osteoprotegerin-ligand; IL, interleukin; TGF-β, transforming growth factor β. *(From WJ Boyle et al: Nature 423: 337, 2003.)*

are detrimental to the skeleton because the increased remodeling rates and the ongoing imbalance between resorption and formation at remodeling sites combine to accelerate loss of bone tissue.

Total daily calcium intakes <400 mg are detrimental to the skeleton, and intakes in the range of 600–800 mg, which is about the average intake among adults in the United States, are also probably suboptimal. The recommended daily required intake of 1000–1200 mg for adults accommodates population heterogeneity in controlling calcium balance **(Chap. 325)**. Such intakes should preferentially come from dietary sources and supplements used only when dietary intakes fall short, and cannot be modified easily. The supplement should be enough to bring total intake to about 1200 mg/d. Recent studies have suggested that there may be differences in safety based on calcium source; high intakes primarily from supplement sources appear to result in a greater risk of renal stones, and perhaps cardiovascular calcifications (although the literature is inconsistent and controversial).

VITAMIN D

(See also Chap. 402) Severe vitamin D deficiency causes rickets in children and osteomalacia in adults. However, vitamin D insufficiency may be more prevalent than previously thought, particularly among individuals at increased risk such as the elderly; those living in northern latitudes; and individuals with poor nutrition, obesity, malabsorption, or chronic liver or renal disease. Dark-skinned individuals are also at high risk of vitamin D deficiency. Although there is considerable controversy about overall optimal health targets for serum 25-hydroxyvitamin D (25[OH]D), there is evidence that for optimal skeletal health, serum 25(OH)D should be >75 nmol/L (30 ng/mL). To achieve this level for most adults requires an intake of at least 800–1000 units/d, or higher in individuals with risk factors (as above).

Vitamin D insufficiency leads to compensatory secondary hyperparathyroidism and is an important risk factor for osteoporosis and fractures. Some studies have shown that >50% of inpatients on a general medical service exhibit biochemical features of vitamin D deficiency, including increased levels of PTH and alkaline phosphatase and lower levels of ionized calcium. In women living in northern latitudes, vitamin D levels decline during the winter months. This is associated with seasonal bone loss, reflecting increased bone turnover. Even among healthy ambulatory individuals, mild vitamin D deficiency is increasing in prevalence. In part this is due to decreased exposure to sunlight coupled with increased use of potent sunscreens. Treatment with vitamin D can return levels to normal [>75 μmol/L (30 ng/mL)] and prevent the associated increase in bone remodeling, bone loss, and fractures. Reduced falls and fracture rates also have been documented among individuals in northern latitudes who have greater vitamin D intake and have higher 25(OH)D levels (though one study suggested an increased fall risk with higher 25OHD levels). Although vitamin D levels might affect risk and/or severity of other diseases, including cancers (colorectal, prostate, and breast), autoimmune diseases, multiple sclerosis, cardiovascular disease and diabetes, most controlled clinical trials have not confirmed these effects.

ESTROGEN STATUS

Estrogen deficiency causes bone loss by two distinct but interrelated mechanisms: (1) activation of new bone remodeling sites and (2) exaggeration of the imbalance between bone formation and resorption. The change in activation frequency causes a transient bone loss until a new steady state between resorption and formation is achieved. The remodeling imbalance, however, results in a permanent decrement in mass. In addition, the very presence of more remodeling sites in the skeleton increases the probability that trabeculae will be penetrated, eliminating the template on which new bone can be formed and accelerating the loss of bony tissue.

The most common estrogen-deficient state is the cessation of ovarian function at the time of menopause, which occurs on average at age 51 **(Chap. 388)**. Thus, with current life expectancy, an average woman will spend about 30 years without an ovarian supply of estrogen. Breast cancer treatment with aromatase inhibitors is an increasingly common cause of estrogen deficiency. The mechanism by which estrogen deficiency causes bone loss is summarized in Fig. 404-5. Marrow cells (macrophages, monocytes, osteoclast precursors, mast cells) as well as bone cells (osteoblasts, osteocytes, osteoclasts) express ERs α and β. Loss of estrogen increases production of RANKL and reduces production of osteoprotegerin, increasing osteoclast formation and recruitment. Estrogen also may play a role in determining the life span of bone cells by controlling the rate of apoptosis. Thus, in situations of estrogen deprivation, the life span of osteoblasts may be decreased, whereas the longevity and activity of osteoclasts are increased. The rate and duration of bone loss after menopause are heterogeneous and unpredictable. Once surfaces are lost in cancellous bone, the rate of bone loss declines. In cortical bone, loss is slower but may continue for a longer time period.

Since remodeling is initiated at the surface of bone, it follows that trabecular bone—which has a considerably larger surface area (80% of the total) than cortical bone—will be affected preferentially by estrogen deficiency. Fractures occur earliest at sites where trabecular bone contributes most to bone strength; consequently, vertebral fractures are the most common early skeletal consequence of estrogen deficiency.

In males, estrogen may an important role in regulation of bone remodeling. In an experiment in which males were rendered estrogen and androgen deficient, restoring estrogen supply reduced remodeling rate more than restoring androgen.

PHYSICAL ACTIVITY

Inactivity, such as prolonged bed rest or paralysis, results in significant bone loss. Concordantly, athletes have higher bone mass than nonathletes. These changes in skeletal mass are most marked when the stimulus begins during growth and before the age of puberty. Adults are less capable than children of increasing bone mass after restoration of physical activity. Epidemiologic data support the beneficial effects on the skeleton of chronic high levels of physical activity. Fracture risk is lower in rural communities and in countries where physical activity is maintained into old age. However, when exercise is initiated during adult life, the effects of moderate exercise on the skeleton are modest, with a bone mass increase of 1–2% in short-term studies of <2 years' duration. It is argued that more active individuals are less likely to fall and are more capable of protecting themselves upon falling, thereby reducing fracture risk. Continuing physical activity into the later years appears to slow cognitive decline, another major reason for including exercise programs for the aging population.

CHRONIC DISEASES

Various genetic and acquired diseases are associated with an increase in the risk of osteoporosis **(Table 404-2)**. Mechanisms that contribute to bone loss are unique for each disease and typically result from multiple factors, including nutrition, reduced physical activity levels, and factors that affect rates of bone remodeling. In most, but not all circumstances, the primary diagnosis is made before osteoporosis presents clinically. Both Type I and Type II diabetes mellitus are associated with an increased fracture risk, with increased risk at higher bone density than in the non-diabetic population. This may be due to differences in the chemical composition of bone tissue that is more brittle than normal, a predilection for conversion of precursors to adipose cells rather than osteoblasts, and the sequelae of diabetes that increase the risk of falls and injury.

MEDICATIONS

A large number of medications used in clinical practice have potentially detrimental effects on the skeleton **(Table 404-3)**. *Glucocorticoids* are the most common cause of medication-induced osteoporosis. It is often not possible to determine the extent to which osteoporosis is related to glucocorticoid or to other factors, as the effects of medication are superimposed on the effects of the primary disease, which in itself may be associated with bone loss (e.g., rheumatoid arthritis). Excessive doses of thyroid hormone can accelerate bone remodeling and result in bone loss.

Other medications have less detrimental effects on the skeleton than pharmacologic doses of glucocorticoids. *Anticonvulsants* are thought to

TABLE 404-2 Diseases Associated with an Increased Risk of Generalized Osteoporosis in Adults

Hypogonadal states	**Hematologic disorders/malignancy**
Turner's syndrome	Multiple myeloma
Klinefelter's syndrome	Lymphoma and leukemia
Anorexia nervosa	Malignancy-associated parathyroid hormone (PTHrP) production
Hypothalamic amenorrhea	Mastocytosis
Hyperprolactinemia	Hemophilia
Other primary or secondary hypogonadal states	Thalassemia
Endocrine disorders	**Selected inherited disorders**
Cushing's syndrome	Osteogenesis imperfecta
Hyperparathyroidism	Marfan's syndrome
Thyrotoxicosis	Hemochromatosis
Diabetes mellitus (both type 1 and 2	Hypophosphatasia
Acromegaly	Glycogen storage diseases
Adrenal insufficiency	Homocystinuria
Nutritional and gastrointestinal disorders	Ehlers-Danlos syndrome
Malnutrition	Porphyria
Parenteral nutrition	Menkes' syndrome
Malabsorption syndromes	Epidermolysis bullosa
Gastrectomy	**Other disorders**
Severe liver disease, especially biliary cirrhosis	Immobilization
Pernicious anemia	Chronic obstructive pulmonary disease
Rheumatologic disorders	Pregnancy and lactation
Rheumatoid arthritis	Scoliosis
Ankylosing spondylitis	Multiple sclerosis
	Sarcoidosis
	Amyloidosis

increase the risk of osteoporosis, although many affected individuals have concomitant insufficiency of 1,25(OH)2D, as some anticonvulsants induce the cytochrome P450 system and vitamin D metabolism. Patients undergoing transplantation are at high risk for rapid bone loss and fracture not only from glucocorticoids but also from treatment with other *immunosuppressants* such as cyclosporine and tacrolimus (FK506). In addition these, patients often have underlying metabolic abnormalities such as hepatic or renal failure that predispose to bone loss. Recently the use of proton pump inhibitors has been shown in observational studies to be associated with a higher risk of fracture. Given their widespread and frequent longterm use, the skeletal effect is important from a public health perspective and when reviewing risk for fracture in individuals.

Aromatase inhibitors, which potently block the aromatase enzyme that converts androgens and other adrenal precursors to estrogen, reduce circulating postmenopausal estrogen supply dramatically. These agents, which are used in various stages for breast cancer treatment, also have been shown to have a detrimental effect on bone density and risk of fracture. Androgen deprivation therapies, used to treat men with prostate cancer, also result in rapid loss of bone and increased fracture risk. Various diabetes medications, including but not limited to thiazoledinediones, and antidepressants, including the Selective Serotonin Reuptake Inhibitor (SSRIs) increase risk of osteoporosis and fracture. It is difficult in some cases to separate the risk accrued by the underlying disease from that attributable the medication. Thus, both depression and diabetes are risk factors for fracture by themselves.

TABLE 404-3 Drugs Associated with an Increased Risk of Generalized Osteoporosis in Adults

Glucocorticoids	Excessive thyroxine
Cyclosporine	Aluminum
Cytotoxic drugs	Gonadotropin-releasing hormone agonists
Anticonvulsants	Heparin
Aromatase inhibitors	Lithium
SSRIs	Protein Pump Inhibitors
	Thiazolidinediones
	Androgen Deprivation Therapies

SMOKING

Smoking produces detrimental effects on bone mass mediated directly by toxic effects on osteoblasts or indirectly by modifying estrogen metabolism. On average, cigarette smokers reach menopause 1–2 years earlier than the general population. Cigarette smoking also produces secondary effects that can modulate skeletal status, including intercurrent respiratory and other illnesses, frailty, decreased exercise, poor nutrition, and the need for additional medications (e.g., glucocorticoids for lung disease).

DIAGNOSIS

MEASUREMENT OF BONE MASS

Several noninvasive techniques are available for estimating skeletal mass or BMD. They include DXA, single-energy x-ray absorptiometry (SXA), quantitative CT, and ultrasound (US). DXA is a highly accurate x-ray technique that has become the standard for measuring bone density. Though it can be used for measurement in any skeletal site, clinical determinations usually are made of the lumbar spine and hip. DXA also can be used to measure BMD of the wrist and body composition. In the DXA technique, two x-ray energies are used to estimate the area of mineralized tissue, and the mineral content is divided by bone area, which partially corrects for body and bone size. However, this correction is only partial since DXA is a two-dimensional scanning technique and cannot estimate the depth or posteroanterior length of the bone. Thus, small slim people tend to have lower than average BMD, a feature that is important in interpreting BMD measurements. Bone spurs, which are common in osteoarthritis, tend to falsely increase bone density mostly of the spine and are a particular problem in measuring spine BMD in older individuals. Because DXA measurement devices are provided by several different manufacturers, the output varies in absolute terms. Consequently, it has become standard practice to relate the results to "normal" values by using T-scores (a T-score of 1 equals 1 SD), which compare individual results to those in a young population that is matched for race and sex, with the average value given a score of zero and the range being +2.5 to –2.5 (i.e., 2.5 SDs above or below the mean). Z-scores (also SDs) compare individual results to those of an age and gender-matched reference population. Thus, a 60-year-old woman with a Z-score of –1 (1 SD below mean for age) has a T-score of –2.5 (2.5 SD below mean for a young control group) **(Fig. 404-6)**. A T-score <–2.5 in the lumbar spine, femoral neck, or total hip has been defined as osteoporosis.

As noted above, since >50% of fractures occur in individuals with low bone mass (i.e., a T-score between –1.0 and –2.5), rather than osteoporosis, attempts are ongoing to redefine the disease as a fracture risk rather than a specific BMD. To that end the absolute fracture risk assessment tool FRAX (Fracture Risk Assessment) often accompanies the report of bone density. FRAX estimates include age, gender, height, weight, fracture history, hip fracture in a parent, steroid use,

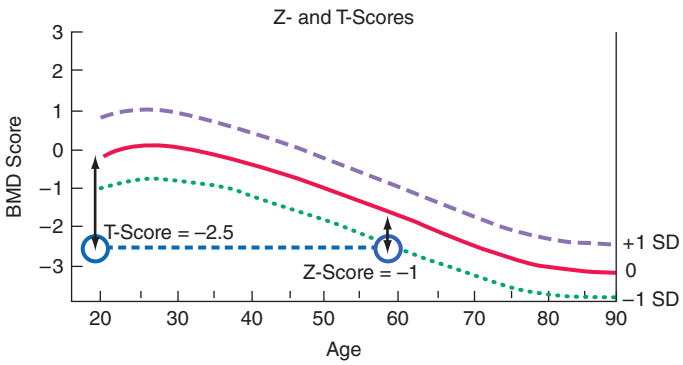


FIGURE 404-6 Relationship between Z-scores and T-scores in a 60-year-old woman. BMD, bone mineral density; SD, standard deviation.

rheumatoid arthritis, other secondary causes as well as bone density of the femoral neck. The program then calculates the estimated risk over a ten-year time frame for major osteoporosis-related fractures (clinical spine, hip, wrist and proximal humerus) as well as hip fracture.

CT can also be used to measure the spine and the hip, but is rarely used clinically, in part because the radiation exposure and cost are both much higher than with DXA. High resolution peripheral quantitative computed tomography (QCT) can be used to measure bone in the forearm or tibia, and is a research tool that provides information on skeletal architecture non-invasively. Magnetic resonance imaging can also be used to obtain some architectural information on the forearm and perhaps the hip, but again is primarily a research tool at present.

Ultrasound can be used to measure bone mass by calculating the attenuation of the signal as it passes through bone or the speed with which it traverses the bone. Although the ultrasound technique was purported to assesses properties of bone other than mass (e.g., quality), this has not been confirmed. Because of its relatively low cost and mobility, ultrasound bone density measurement is amenable for use as a screening procedure in stores or health fairs.

All these techniques for measuring BMD have been approved by the U.S. Food and Drug Administration (FDA) on the basis of their capacity to predict fracture risk. The hip is the preferred site of measurement in most individuals, since it predicts the risk of hip fracture, the most important consequence of osteoporosis, better than any other bone density measurement site. When hip measurements are performed by DXA, the spine can be measured at the same time. In younger individuals such as perimenopausal or early postmenopausal women, spine measurements may be the most sensitive indicator of bone loss. When the spine or hip is not measureable due to severe degenerative spine disease or scoliosis or prior spine or hip surgery, BMD of the wrist is often measured.

INDICATIONS FOR BONE MASS MEASUREMENT

Clinical guidelines have been developed for the use of bone densitometry in clinical practice **(Table 404-4)**. The National Osteoporosis Foundation (NOF) guidelines recommend bone mass measurements in postmenopausal women, assuming they have one or more risk factors for osteoporosis in addition to age, sex, and estrogen deficiency. The guidelines further recommend that bone mass measurement be considered in *all* women by age 65, a position ratified by the U.S. Preventive Health Services Task Force. In males the use of bone density determination is not recommended until the age of 70 years in the absence of multiple risk factors or the occurrence of an osteoporosis-related fracture.

Risk factors (age, prior fracture, family history of hip fracture, low body weight, cigarette consumption, excessive alcohol use, steroid use, and rheumatoid arthritis) can be combined with BMD to assess the 10-year fracture probabilities. Fracture risk probability calculators are available as part of the report from many DXA machines and also available online (*https://www.sheffield.ac.uk/FRAX/*) **(Fig. 404-7)**. In the United States it has been determined to be cost effective to treat if the 10-year fracture risk from FRAX is ≥20%, and/or the 10-year risk of hip fracture is ≥3%. FRAX is an imperfect tool, as it does not include any assessment of fall risk, and secondary causes are excluded when BMD is entered. More importantly, it does not distinguish the contribution toward of future fracture probability from an acute recent fracture versus the much lesser importance of the more remote fracture. Moreover, there is no mandate for vertebral fracture diagnosis and no additional fracture probability estimated for patients who have had multiple fractures. Nonetheless it is useful as an educational tool for patients, particularly for those who are excessively worried about BMD levels despite relative youth and health.

TABLE 404-4 Indications for BMD Testing

- Women aged ≥65 and men aged ≥70; regardless of clinical risk factors
- Younger postmenopausal women, women in the menopausal transition, and men aged from 50 to 69 with clinical risk factors for fracture
- Adults who have a fracture at or after age 50
- Adults with a condition (e.g., rheumatoid arthritis) or taking a medication (e.g., glucocorticoids at a daily dose >5 mg prednisone or equivalent for >3 months associated with low bone mass or bone loss

VERTEBRAL IMAGING

DXA equipment can also be used to obtain lateral images of the thoracic and lumbar spine, a technique called vertebral fracture assessment (VFA). While not as definitive as a radiograph it is an excellent screening tool for both women and men based on age and BMD even in the absence of any specific symptoms since the majority of vertebral fractures are asymptomatic for a long time. Furthermore, the VFA can be used to evaluate height loss or back pain that suggest the presence of an undiagnosed vertebral fracture.

Because vertebral fractures are often asymptomatic when they first occur, the diagnosis of vertebral fracture is rarely made at the time. Since vertebral fractures, whether symptomatic or asymptomatic, are associated with the same clinical sequelae, it is critical that patients with these fractures are identified. Vertebral fracture prevalence in the US based on the National Health and Nutrition Evaluation Studies (NHANES) population appears to be about 10% in the 1970s and 20% in the 1980s, when the strictest criteria for diagnosis are utilized. The NOF and other organizations have recommended that women by the age of 65 and men by the age of 70 undergo vertebral imaging if a T-Score is ≤–1.5 at the spine, hip, or femoral neck. For women by the age of 70 and men by the age of 80 if a T-Score is <–1.0. For younger individuals, vertebral imaging is recommended for those with an osteoporosis related fracture, height loss, or glucocorticoid use. (See **Table 404-5**.)

APPROACH TO THE PATIENT

Osteoporosis

The perimenopausal transition is a good opportunity to initiate a discussion about risk factors for osteoporosis and consideration of indications for a BMD test. A careful history and physical examination should be performed to identify risk factors for osteoporosis. A low Z-score increases the suspicion of a secondary cause for bone loss unless the low value can be explained by body size. Height loss >2.5–3.8 cm (>1–1.5 in.) is an indication for VFA by DXA or radiography to rule out asymptomatic vertebral fractures, as is the presence of significant kyphosis or back pain, particularly if it began after menopause. For patients who present with fractures, it is important to ensure that the fractures are not caused by an underlying malignancy. Usually this is clear on routine radiography, but on occasion, CT, MRI, or radionuclide scans may be necessary. In this regard it is important not to dismiss fractures simply because they happened on significant trauma. Persons with osteoporosis fracture more readily with any level of injury, a concept that needs continual emphasis.

ROUTINE LABORATORY EVALUATION

There is no established algorithm for the evaluation of women who present with osteoporosis. A general evaluation that includes complete blood count, serum and 24-h urine calcium, and renal and hepatic function tests is useful for identifying selected secondary causes of low bone mass, particularly for women with fractures or unexpectedly low Z-scores. An elevated serum calcium level suggests hyperparathyroidism or malignancy, whereas a reduced serum calcium level may reflect malnutrition or a malabsorption disease such as celiac disease. In the presence of hypercalcemia, a serum PTH level differentiates between hyperparathyroidism (PTH↑) and malignancy (PTH↓), and a high PTHrP level can help document the presence of humoral hypercalcemia of malignancy **(Chap. 403)**. A low urine calcium (<50 mg/24 h) suggests malnutrition, or malabsorption; a high urine calcium (>300 mg/24 h) during normal calcium intake (excluding calcium supplements for at least a week before the urine collection) is indicative of hypercalciuria. Hypercalciuria occurs primarily in three situations: (1) a renal calcium leak, which is more common in males with osteoporosis; (2) absorptive hypercalciuria, which can be idiopathic or associated with increased $1,25(OH)_2D$ in granulomatous disease; or (3) hematologic malignancies or conditions associated with excessive bone turnover such as Paget's disease, hyperparathyroidism, and hyperthyroidism. Renal hypercalciuria is treated with thiazide diuretics, which lower urine

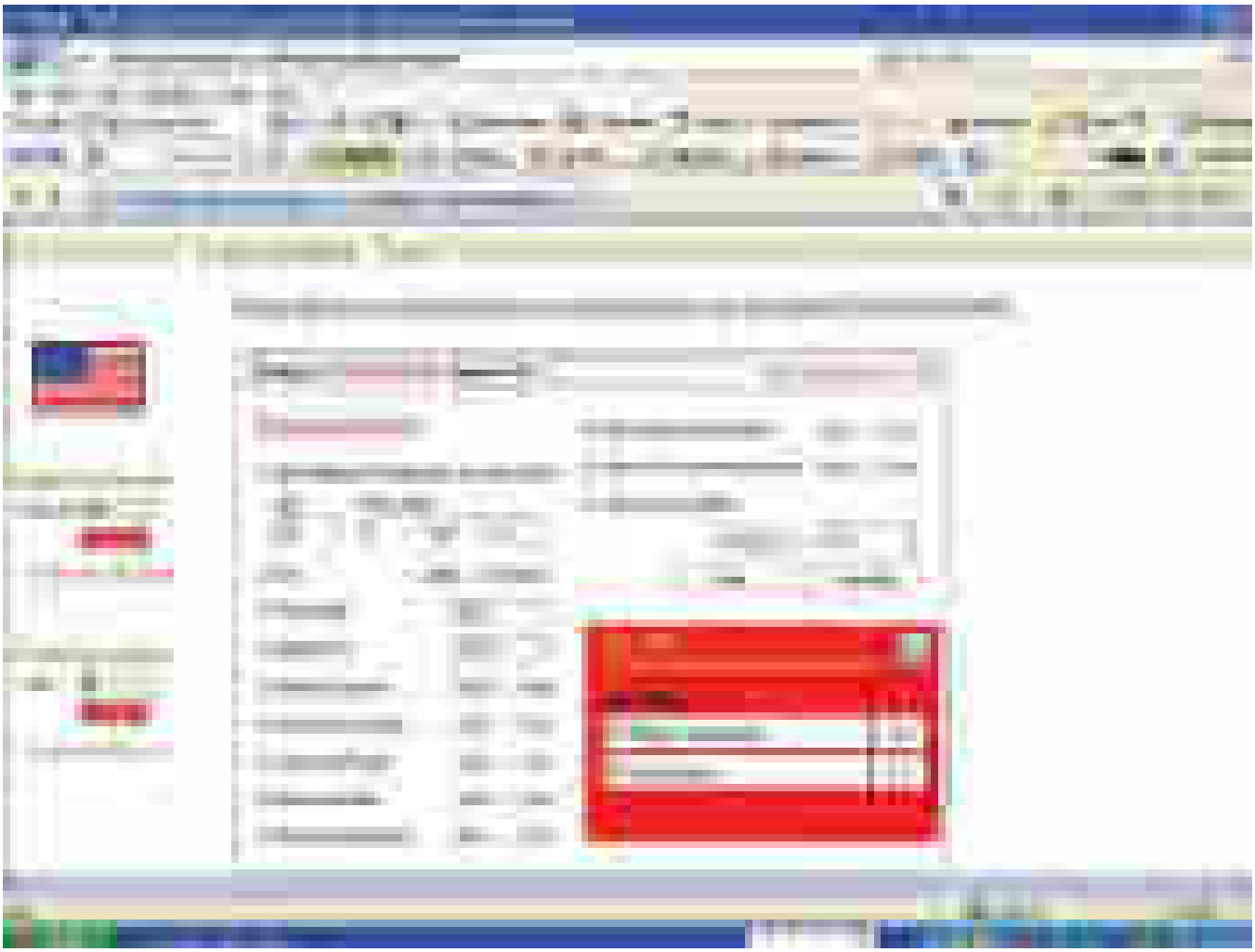


FIGURE 404-7 FRAX calculation tool. When the answers to the indicated questions are filled in, the calculator can be used to assess the 10-year probability of fracture. The calculator (available online at *http://www.shef.ac.uk/FRAX/tool.jsp?locationValue=9*) also can risk adjust for various ethnic groups.

calcium and help improve calcium economy. In this setting, thiazides alone can improve bone mass and possibly reduce risk of fracture. They might also reduce renal stone risk.

Individuals who have osteoporosis-related fractures or bone density in the osteoporotic range should have a measurement of serum 25(OH)D level, since the intake of vitamin D required to achieve a target level >30 ng/mL is highly variable. Hyperthyroidism should be evaluated by measuring thyroid-stimulating hormone (TSH).

When there is clinical suspicion of Cushing's syndrome, urinary free cortisol levels or a fasting serum cortisol should be measured after overnight dexamethasone. When bowel disease, malabsorption, or malnutrition is suspected, serum albumin, cholesterol, and a complete blood count should be checked. Asymptomatic malabsorption may be heralded by anemia (macrocytic—vitamin B_{12} or folate deficiency; microcytic—iron deficiency) or low serum cholesterol or urinary calcium levels. If these or other features suggest malabsorption, further evaluation is required. Asymptomatic celiac disease with selective malabsorption is being found with increasing frequency; the diagnosis can be made by testing for transglutaminase IgA antibodies, but may require confirmation by endoscopic biopsy. A trial of a gluten-free diet can also be confirmatory **(Chap. 318)**. When osteoporosis is found associated with symptoms of rash, multiple allergies, diarrhea, or flushing, mastocytosis should be considered and excluded by using 24-h urine histamine collection or serum tryptase.

Myeloma can masquerade as generalized osteoporosis, although it more commonly presents with bone pain and characteristic "punched-out" lesions on radiography. Serum and urine electrophoresis and/or evaluation for serum free light chains light chains in urine are required to exclude this diagnosis. More commonly a monoclonal gammopathy (MGUS) is found and the patient subsequently monitored to ensure that this is not an incipient myeloma. MGUS itself may be associated with an increased risk of osteoporosis. A bone marrow biopsy may be required to rule out myeloma (in patients with equivocal electrophoretic results) and also can be used to exclude mastocytosis, leukemia, and other marrow infiltrative disorders such as Gaucher's disease.

TABLE 404-5 Indications for Vertebral Testing

Consider vertebral imaging tests for the following individuals[a]

- All women aged ≥70 and all men aged ≥80 if BMD T-score at the spine, total hip, or femoral neck is <1.0
- Women aged from 65 to 69 and men aged from 70 to 79 if BMD T-score at the spine, total hip, or femoral neck is <1.5
- Postmenopausal women and men aged ≥50 with specific risk factors:
  - Low-trauma fracture during adulthood (aged ≥50)
  - Historical height loss of ≥1.5 in. (4 cm)[b]
  - Prospective height loss of ≥0.8 in. (2 cm)[c]
  - Recent or ongoing long-term glucocorticoid treatment

[a]If bone density testing is not available, vertebral imaging may be considered based on age alone. [b]Current height compared to peak height during childhood. [c]Cumulative height loss measured during interval medical assessment.

BONE BIOPSY

Tetracycline labeling of the skeleton allows determination of the rate of remodeling as well as evaluation for other metabolic bone

TABLE 404-6 Biochemical Markers of Bone Metabolism in Clinical Use

Bone formation
Serum bone-specific alkaline phosphatase
Serum osteocalcin
Serum propeptide of type I procollagen
Bone resorption
Urine and serum cross-linked N-telopeptide
Urine and serum cross-linked C-telopeptide

diseases. The current use of BMD tests, in combination with hormonal evaluation and biochemical markers of bone remodeling, has largely replaced the clinical use of bone biopsy, although it remains an important tool in the diagnosis of chronic kidney disease mineral bone disease (CKDMBD), in evaluating mechanism of action of osteoporosis pharmacologies and in clinical research.

BIOCHEMICAL MARKERS

Several biochemical tests are available that provide an index of the overall rate of bone remodeling **(Table 404-6)**. Biochemical markers usually are characterized as those related primarily to *bone formation* or *bone resorption*. These tests measure the overall state of bone remodeling at a single point in time. Clinical use of these tests has been hampered by biologic variability (in part related to circadian rhythm) as well as analytic variability, although the latter is improving.

For the most part, remodeling markers do not predict rates of bone loss well enough in individuals to make accurate assessment of potential future changes in bone density. However, they do provide adjunct information that assists in both evaluation of the patient and in assessment of treatment response. Markers of bone resorption may help in the prediction of fracture risk, independently of bone density, particularly in older individuals. In women ≥65 years, when bone density results are greater than the usual treatment thresholds noted above, a high level of bone resorption should prompt consideration of treatment. The primary use of biochemical markers is for monitoring the response to treatment. With the introduction of antiresorptive therapeutic agents, bone remodeling declines rapidly, with the fall in resorption occurring earlier than the fall in formation. Inhibition of bone resorption is maximal within 3 months or so. Thus, measurement of bone resorption (CTX is the preferred marker) before initiating therapy and 3–6 months after starting therapy provides an earlier estimate of patient response than does bone densitometry. A decline in resorptive markers can be ascertained after treatment with bisphosphonates, denosumab or estrogen; this effect is less marked after treatment with weaker agents such as raloxifene or calcitonin. Bone turnover markers are also useful in monitoring the effects of 1-34hPTH, or teriparatide, which rapidly increases bone formation (P1NP is the most sensitive but osteocalcin is also a very good formation marker) and later bone resorption. The recent suggestion of "drug holidays" (see below) has opened another use for biochemical markers, allowing evaluation of the off-effect of drugs such as bisphosphonates.

TREATMENT

Osteoporosis

MANAGEMENT OF PATIENTS WITH FRACTURES

Treatment of a patient with osteoporosis frequently involves management of acute fractures as well as treatment of the underlying disease. Hip fractures almost always require surgical repair if the patient is to become ambulatory again. Depending on the location and severity of the fracture, condition of the neighboring joint, and general status of the patient, procedures may include open reduction and internal fixation with pins and plates, hemiarthroplasties, and total arthroplasties. These surgical procedures are followed by intense rehabilitation in an attempt to return patients to their prefracture functional level. Long bone fractures often require either external or internal fixation. Other fractures (e.g., vertebral, rib, and pelvic fractures) usually are managed with supportive care, requiring no specific orthopedic treatment.

Only ~25–30% of vertebral compression fractures present with sudden-onset back pain. For acutely symptomatic fractures, treatment with analgesics is required, including nonsteroidal anti-inflammatory agents and/or acetaminophen, sometimes with the addition of a narcotic agent (codeine or oxycodone). (A few small, randomized clinical trials suggest that calcitonin may reduce pain related to acute vertebral compression fracture). A technique that involves percutaneous injection of artificial cement (polymethylmethacrylate) into the vertebral body (vertebroplasty or kyphoplasty), may offer significant pain relief in some patients, however controlled trials of these procedures have provided some doubt of their efficacy. Furthermore, risks include acute extravasation of cement outside of the vertebral body with neurologic impairment and possibly an increased risk of vertebral fracture in adjacent vertebrae due to increased rigidity of the treated vertebral body. Short periods of bed rest may be helpful for pain management, but in general, early mobilization is recommended as it helps prevent further bone loss associated with immobilization. Occasionally, use of a soft elastic-style brace may facilitate earlier mobilization. Muscle spasms often occur with acute compression fractures and can be treated with muscle relaxants and heat treatments. Severe pain usually resolves within 6–10 weeks. More chronic severe pain might suggest the possibility of multiple myeloma or other underlying conditions.

Vertebral fractures cause height loss because of the loss of vertebral body height during compression of the vertebral body. These fractures can produce kyphotic posture, particularly when wedge shaped or just loss of thoracic height. Chronic pain following vertebral fracture is probably not bony in origin; instead, it is related to abnormal strain on muscles, ligaments, and tendons and to secondary facet-joint arthritis associated with alterations in thoracic and/or abdominal shape. Chronic pain may also be the result of ribs sitting right on top of the iliac crest bones, particularly in patients who have had multiple vertebral compression fractures. Chronic pain is difficult to treat effectively and may require analgesics, sometimes including narcotic analgesics. Frequent intermittent rest in a supine or semireclining position is often required to allow the soft tissues, which are under tension, to relax. Back and core-strengthening exercises may be beneficial. Heat treatments help relax muscles and reduce the muscular component of discomfort. Various physical modalities, such as ultrasound and transcutaneous nerve stimulation, may be beneficial in some patients. Pain also occurs in the neck region, not as a result of compression fractures (which almost never occur in the cervical spine as a result of osteoporosis) but because of chronic strain associated with trying to elevate the head in a person with a significant thoracic kyphosis.

Multiple vertebral fractures often are associated with psychological symptoms; this is not always appreciated. The changes in body configuration and back pain can lead to marked loss of self-image and a secondary depression. Altered balance, precipitated by the kyphosis and the anterior movement of the body's center of gravity, leads to a fear of falling, a consequent tendency to remain indoors, and the onset of social isolation. These symptoms sometimes can be alleviated by family support and/or psychotherapy. Medication may be necessary when depressive features are present.

Multiple studies show that patients presenting with fractures after age 50 years (even fractures traditionally linked to osteoporosis) are largely not screened or treated for osteoporosis. Estimates suggest that fewer than 25% of fracture patients receive follow-up care. Recently several studies have demonstrated the effectiveness of a relatively simple and inexpensive program that reduces the risk of subsequent fractures. In the Kaiser system it is estimated that a 20% decline in hip fracture occurrence was seen with the introduction of a fracture liaison service. This involves a health care professional

(usually a nurse or physician's assistant) whose job is to educate patients, and coordinate evaluation and osteoporosis treatment as patients move from an emergency room, acute care hospital, rehabilitation hospital and/or orthopedic practice to outpatient management. If the Kaiser experience can be repeated, not only would there be significant savings of health care dollars, but also a dramatic drop in hip fracture incidence and a marked improvement in morbidity and mortality among the aging population.

MANAGEMENT OF THE UNDERLYING DISEASE

Risk Factor Reduction After risk assessment patients should be thoroughly educated to reduce the impact of modifiable risk factors associated with bone loss and falling. Medications should be reviewed to ensure that all are necessary and taken at the lowest required dose. Glucocorticoid medication, if present, should be evaluated to determine that it is truly indicated and is being given in doses that are as low as possible. For those on thyroid hormone replacement, TSH testing should be performed to determine that an excessive dose is not being used, as iatrogenic thyrotoxicosis can be associated with increased bone loss. In patients who smoke, efforts should be made to facilitate smoking cessation. Reducing risk factors for falling also include alcohol abuse treatment and a review of the medical regimen for any drugs that might be associated with orthostatic hypotension and/or sedation, including hypnotics and anxiolytics. If nocturia occurs, the frequency should be reduced, if possible (e.g., by decreasing or modifying diuretic use), as arising in the middle of sleep is a common precipitant of a fall. Patients should be instructed about environmental safety with regard to eliminating exposed wires, curtain strings, slippery rugs, and mobile tables. Avoiding stocking feet on wood floors, checking carpet condition (particularly on stairs), and providing good light in paths to bathrooms and outside the home are important preventive measures. Treatment for impaired vision is recommended, particularly a problem with depth perception, which is specifically associated with increased falling risk. Elderly patients with neurologic impairment (e.g., stroke, Parkinson's disease, Alzheimer's disease) are particularly at risk of falling and require specialized supervision and care.

Nutritional Recommendations • **Calcium** A large body of data indicates that optimal calcium intake reduces bone loss and suppresses bone turnover. Recommended intakes from an Institute of Medicine report are shown in **Table 404-7**. The National Health and Nutritional Evaluation Studies (NHANES) have consistently documented that average calcium intakes fall considerably short of these recommendations. The preferred source of calcium is diet, but many patients require calcium supplementation to bring intake to about 1200 mg/d. Best sources of calcium include dairy products (milk, yogurt, and cheese), nondaily milks (almond, rice, soy), and fortified foods such as certain cereals, waffles, snacks, juices, and crackers. Some of these fortified foods contain as much calcium per serving as milk. Various vegetables and fruits, such as kale, broccoli, and dried figs contain reasonably high calcium content, though some of it may not be fully bioavailable. Calcium intake calculators are available at *NOF.org* or *NYSOPEP.org* and will give a rough idea of total calcium intake.

TABLE 404-7 Adequate Calcium Intake

LIFE STAGE GROUP	ESTIMATED ADEQUATE DAILY CALCIUM INTAKE, mg/d
Young children (1–3 years)	500
Older children (4–8 years)	800
Adolescents and young adults (9–18 years)	1300
Men and women (19–50 years)	1000
Men and women (51 and older)	1200

Note: Pregnancy and lactation needs are the same as for nonpregnant women (e.g., 1300 mg/d for adolescents/young adults and 1000 mg/d for ≥19 years).

Source: Adapted from the Standing Committee on the Scientific Evaluation of Dietary Reference Intakes. Food and Nutrition Board. Institute of Medicine. Washington, DC, 1997, National Academy Press.

TABLE 404-8 Elemental Calcium Content of Various Oral Calcium Preparations

CALCIUM PREPARATION	ELEMENTAL CALCIUM CONTENT
Calcium citrate	60 mg/300 mg
Calcium lactate	80 mg/600 mg
Calcium gluconate	40 mg/500 mg
Calcium carbonate	400 mg/g
Calcium carbonate +5 μg vitamin D_2 (OsCal 250)	250 mg/tablet
Calcium carbonate (Tums 500)	500 mg/tablet

Source: Adapted from SM Krane, MF Holick, Chap. 355, in *Harrison's Principles of Internal Medicine*, 14th ed. New York, McGraw-Hill, 1998.

If calcium supplements are required, they should be taken in doses sufficient to bring total intake to the required level (1200 mg/d). Doses of supplements should be ≤600 mg per single dose, as the calcium absorption fraction decreases at higher doses. Calcium supplements should be calculated on the basis of the elemental calcium content of the supplement, not the weight of the calcium salt **(Table 404-8)**. Calcium supplements containing carbonate are best taken with food since they require acid for solubility. Calcium citrate supplements can be taken at any time. To confirm bioavailability, calcium supplements can be placed in distilled vinegar. They should dissolve within 30 min.

Several controlled clinical trials of calcium, mostly with accompanying vitamin D, have confirmed reductions in clinical fractures, including fractures of the hip (~20–30% risk reduction), particularly in elderly individuals who are more likely to be dietarily deficient. All recent studies of pharmacologic agents have been conducted in the context of calcium replacement (± vitamin D). Thus, it is standard practice to ensure an adequate calcium and vitamin D intake in patients with osteoporosis whether they are receiving additional pharmacologic therapy or not. A systematic review confirmed a greater BMD response to antiresorptive therapy when calcium intake was adequate.

Although side effects from supplemental calcium are minimal (eructation and constipation mostly with carbonate salts), individuals with a history of kidney stones should have a 24-h urine calcium determination before starting increased calcium to avoid exacerbating hypercalciuria. A recent analysis of published data has suggested that high intakes of calcium particularly from supplements are associated with an increase in the risk of heart disease. This is an evolving story with data both confirming and refuting the finding. Since high calcium intakes also increase the risk of renal stones and confer no extra benefit to the skeleton, the recommendation that total intakes should be between 1000 and 1500 mg/d seems reasonable.

Vitamin D Diet alone rarely contains sufficient vitamin D to maintain target circulating levels [serum 25(OH)D consistently >75 μmol/L (30 ng/mL) Vitamin D is synthesized from a precursor in the skin under the influence of heat and ultraviolet light **(Chap. 402)** but production is blocked by sunscreen and sun avoidance. Therefore, large segments of the population do not obtain sufficient vitamin D from either skin production or dietary sources. Since vitamin D supplementation at doses that would achieve these serum levels is safe and inexpensive, the National Academy of Medicine (formerly, Institute of Medicine, IOM) recommends daily intakes of 200 IU for adults <50 years of age, 400 IU for those 50–70 years, and 600 IU for those >70 years (based on obtaining a serum level of 20 ng/mL, lower than the level recommended by most other guidelines). Multivitamin tablets usually contain 400 IU, and many calcium supplements also contain vitamin D. Some data suggest that higher doses (≥1000 IU) may be required in the elderly and chronically ill. The IOM report suggests that it is safe to take up to 4000 IU/d. For those with osteoporosis or those at risk of osteoporosis 1000–2000IU/day can usually maintain serum 25(OH)D above 30 ng/mL.

Other Nutrients Other nutrients such as salt, high animal protein intakes, and caffeine may have modest effects on calcium excretion or absorption. Adequate vitamin K status is required for optimal carboxylation of osteocalcin. States in which vitamin K nutrition or metabolism is impaired, such as with long-term warfarin therapy, have been associated with reduced bone mass. Research concerning cola intake is controversial but suggests a possible link to reduced bone mass through factors that are independent of caffeine.

Magnesium is abundant in foods, and magnesium deficiency is quite rare in the absence of a serious chronic disease. Magnesium supplementation may be warranted in patients with inflammatory bowel disease, celiac disease, chemotherapy, severe diarrhea, malnutrition, or alcoholism. Dietary phytoestrogens, which are derived primarily from soy products and legumes (e.g., garbanzo beans [chickpeas] and lentils), exert some estrogenic activity but are insufficiently potent to justify their use in place of a pharmacologic agent in the treatment of osteoporosis.

Patients with hip fractures are often frail and relatively malnourished. Some data suggest an improved outcome in such patients when they are provided calorie and protein supplementation. Excessive protein intake can increase renal calcium excretion, but this can be corrected by an adequate calcium intake.

Exercise Exercise in young individuals increases the likelihood that they will attain the maximal genetically determined peak bone mass. Meta-analyses of studies performed in postmenopausal women indicate that weight-bearing exercise helps prevent bone loss but does not appear to result in substantial gain of bone mass. This beneficial effect wanes if exercise is discontinued. Most of the studies are short term, and a more substantial effect on bone mass is likely if exercise is continued over a long period. Exercise also has beneficial effects on neuromuscular function, and it improves coordination, balance, and strength, thereby reducing the risk of falling. A walking program is a practical way to start. Other activities such as dancing, racquet sports, cross-country skiing, and use of gym equipment, are also recommended, depending on the patient's personal preference and general condition. Even women who cannot walk benefit from swimming or water exercises, not so much for the effects on bone, which are quite minimal, but because of effects on muscle. Exercise habits should be consistent, optimally at least three times a week. For most patients we suggest participation in exercise regimes that the patient enjoys, in order to improve adherence. We also emphasize the importance of making exercise a social activity, again to improve adherence.

PHARMACOLOGIC TREATMENT OF OSTEOPOROSIS

Patients presenting with typical osteoporosis related fractures (certainly hip and spine), in the setting of a BMD in the low bone mass or osteoporosis range should be treated with pharmacologic agents. Most guidelines also suggest that patients be considered for treatment when BMD T-Score is ≤–2.5, a level consistent with the diagnosis of osteoporosis. Treatment also should also be considered in postmenopausal women with fracture or multiple risk factors even if BMD is not in the osteoporosis range. Treatment thresholds depend on cost-effectiveness analyses but in the United States are >20% for 10-year major fracture probability and >3% 10-year hip fracture probability. It must be emphasized, however, that as with other diseases, risk assessment is an inexact science when applied to individual patients. Fractures are chance occurrences that can happen to anyone and do! Patients often accept risks that are higher than the physician might like out of concern for the (usually considerably lower) risks of adverse events of drugs.

Pharmacologic therapies for osteoporosis are either antiresorptive or anabolic. The antiresorptive agents include medications that have broad effects such as hormone/estrogen therapy and selective estrogen receptor modulators (SERMS) as well as those agents that are specific for the treatment of osteoporosis (bisphosphonates, denosumab, and calcitonin). The only currently approved anabolic agent is teriparatide, but two additional anabolic agents are currently under FDA review for treatment of osteoporosis (abaloparatide and romosozumab).

Antiresorptive Agents • **Estrogens** A large body of clinical trial data indicates that various types of estrogens (conjugated equine estrogens, estradiol, estrone, esterified estrogens, ethinyl estradiol, and mestranol) reduce bone turnover, prevent bone loss, and induce small increases in bone mass of the spine, hip, and total body. The effects of estrogen are seen in women with natural or surgical menopause and in late postmenopausal women with or without established osteoporosis. Estrogens are efficacious when administered orally or transdermally. For both oral and transdermal routes of administration, combined estrogen/progestin preparations are now available in many countries, obviating the problem of taking two tablets or using a patch and oral progestin.

For oral estrogens, the standard recommended doses have been 0.3 mg/d for esterified estrogens, 0.625 mg/d for conjugated equine estrogens, and 5 µg/d for ethinyl estradiol. For transdermal estrogen, the commonly used dose supplies 50 µg estradiol per day, but a lower dose may be appropriate for some individuals. Dose-response data for conjugated equine estrogens indicate that lower doses (0.3 and 0.45 mg/d) are effective. Doses even lower have also been shown to slow bone loss.

Fracture Data Epidemiologic databases indicate that women who take estrogen replacement have a 50% reduction, on average, of osteoporosis related fractures, including hip fractures. The beneficial effect of estrogen is greatest among those who start replacement early and continue the treatment; the benefit declines after discontinuation to the extent that there is no residual protective effect against fracture by 10 years after discontinuation. The first clinical trial evaluating fractures as secondary outcomes, the Heart and Estrogen-progestin Replacement Study (HERS) trial, showed no effect of hormone therapy on hip or other clinical fractures in women with established coronary artery disease. These data made the results of the Women's Health Initiative (WHI) exceedingly important **(Chap. 388)**. The estrogen-progestin arm of the WHI in >16,000 postmenopausal healthy women indicated that hormone therapy reduces the risk of hip and clinical spine fracture by 34% and that of all clinical fractures by 24%.

A few smaller clinical trials have evaluated spine fracture occurrence as an outcome with estrogen therapy. They have consistently shown that estrogen treatment reduces the incidence of vertebral compression fracture.

The WHI has provided a vast amount of data on the multisystemic effects of hormone therapy. Although earlier observational studies suggested that estrogen replacement might reduce heart disease, the WHI showed that combined estrogen-progestin treatment increased risk of fatal and nonfatal myocardial infarction by ~29%, confirming data from the HERS study. Other important relative risks included a 40% increase in stroke, a 100% increase in venous thromboembolic disease, and a 26% increase in risk of breast cancer. Subsequent analyses have confirmed the increased risk of stroke and in a substudy showed a twofold increase in dementia. Benefits other than the fracture reductions noted above included a 37% reduction in the risk of colon cancer. These relative risks have to be interpreted in light of absolute risk **(Fig. 404-8)**. For example, out of 10,000 women treated with estrogen-progestin for 1 year, there will be 8 excess heart attacks, 8 excess breast cancers, 18 excess venous thromboembolic events, 5 fewer hip fractures, 44 fewer clinical fractures, and 6 fewer colorectal cancers. These numbers must be multiplied by years of hormone treatment. There was no effect of hormone treatment on the risk of uterine cancer or total mortality.

It is important to note that these WHI findings apply specifically to hormone treatment in the form of conjugated equine estrogen plus medroxyprogesterone acetate. The relative benefits and risks of unopposed estrogen in women who had hysterectomies vary somewhat. They still show benefits against fracture occurrence and increased risk of venous thrombosis and stroke, similar in magnitude to the risks for combined hormone therapy. In contrast, though,

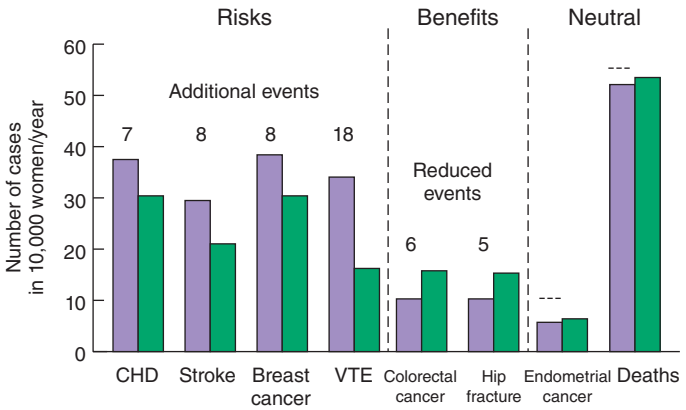


FIGURE 404-8 Effects of hormone therapy on event rates: green, placebo; purple, estrogen and progestin. CHD, coronary heart disease; VTE, venous thromboembolic events. *(Adapted from Women's Health Initiative. WHI HRT Update. Available at http://www.nhlbi.nih.gov/health/women/upd2002.htm.)*

the estrogen-only arm of WHI indicated no increased risk of heart attack or breast cancer. The data suggest that at least some of the detrimental effects of combined therapy are related to the progestin component. In addition, there is the possibility, suggested by primate data that the risk accrues mainly to women who have some years of estrogen deficiency before initiating treatment. Nonetheless there is marked reluctance among women for ET/HT and the US preventive services task force has specifically suggested that ET/HT not be used for disease prevention.

Mode of Action Two subtypes of ERs, α and β, have been identified in bone and other tissues. Cells of monocyte lineage express both ERα and ERβ, as do osteoblasts. Estrogen-mediated effects vary with the receptor type. Using ER knockout mouse models, elimination of ERα produces a modest reduction in bone mass, whereas mutation of ERβ has less of an effect on bone. A male patient with a homozygous mutation of ERα had markedly decreased bone density as well as abnormalities in epiphyseal closure, confirming the important role of ERα in bone biology. The mechanism of estrogen action in bone is an area of active investigation (Fig. 404-5). Although data are conflicting, estrogens may inhibit osteoclasts directly. However, the majority of estrogen (and androgen) effects on bone resorption are mediated indirectly through paracrine factors produced by osteoblasts. These actions include (1) increasing OPG production by osteoblasts (2) Increasing IGF-I and TGF-β and (3) suppressing IL-1 (α and β), IL-6, TNF-α, and osteocalcin synthesis. The indirect estrogen actions primarily decrease bone resorption.

Progestins In women with a uterus, daily progestin or cyclical progestins at least 12 days per month are prescribed in combination with estrogens to reduce the risk of uterine cancer. Medroxyprogesterone acetate and norethindrone acetate blunt the high-density lipoprotein response to estrogen, but micronized progesterone does not. Neither medroxyprogesterone acetate nor micronized progesterone appears to have an independent effect on bone; at lower doses of estrogen, norethindrone acetate may have an additive benefit. On breast tissue, progestins may account for the increase the risk of breast cancer with combination treatment.

SERMs Two SERMs are used currently in postmenopausal women: raloxifene, which is FDA-approved for the prevention and treatment of osteoporosis as well as the prevention of breast cancer, and tamoxifen, which is approved for the prevention and treatment of breast cancer. A third SERM, bazodoxiphene, is marketed in combination with conjugated estrogen for treatment of menopausal symptoms and prevention of bone loss. Bazedoxifene protects the uterus and breast from effects of estrogen and makes the use of progestin unnecessary.

Tamoxifen reduces bone turnover and bone loss in postmenopausal women compared with placebo groups. These findings support the concept that tamoxifen acts as an estrogenic agent in bone. There are limited data on the effect of tamoxifen on fracture risk, but the Breast Cancer Prevention study indicated a possible reduction in clinical vertebral, hip, and Colles' fractures. Tamoxifen is not FDA approved for prevention or treatment of osteoporosis. The major benefit of tamoxifen is on breast cancer occurrence. The breast cancer prevention trial indicated that tamoxifen administration over 4–5 years reduced the incidence of new invasive and noninvasive breast cancer by ~45% in women at increased risk of breast cancer. The incidence of ER-positive breast cancers was reduced by 65%. Tamoxifen increases the risk of uterine cancer in postmenopausal women, limiting its use for breast cancer prevention in women at low or moderate risk.

Raloxifene (60 mg/d) has effects on bone turnover and bone mass that are very similar to those of tamoxifen, indicating that this agent is also estrogenic on the skeleton. The effect of raloxifene on bone density (+1.4–2.8% versus placebo in the spine, hip, and total body) is somewhat less than that seen with standard doses of estrogens. Raloxifene reduces the occurrence of vertebral fracture by 30–50%, depending on the population; however, there are no data confirming that raloxifene can reduce the risk of nonvertebral fractures >8 years of observation.

Raloxifene, like tamoxifen and estrogen, has effects in other organ systems. The most beneficial effect appears to be a reduction in invasive breast cancer (mainly decreased ER-positive) occurrence of ~65% in women who take raloxifene compared to placebo. In a head-to-head study raloxifene was as effective as tamoxifen in preventing breast cancer in high-risk women and raloxifene is FDA approved for this indication. In a further study raloxifene had no effect on heart disease in women with increased risk for this outcome. In contrast to tamoxifen, raloxifene is not associated with an increase in the risk of uterine cancer or benign uterine disease. Raloxifene increases the occurrence of hot flashes but reduces serum total and low-density lipoprotein cholesterol, lipoprotein(a), and fibrinogen. Raloxifene with positive effects on breast cancer and vertebral fractures has become a useful agent for the treatment of the younger asymptomatic postmenopausal woman. In some women, a recurrence of menopausal symptoms may occur. Usually this is evanescent but occasionally is sufficiently impactful on daily life and sleep, that the drug must be withdrawn. Raloxifene increases the risk of deep vein thrombosis and may increase the risk of death from stroke among older women. Consequently it is not usually recommended for women aged >70 years.

Mode of Action of SERMs All SERMs bind to the ER, but each agent produces a unique receptor-drug conformation. As a result, specific coactivator or co-repressor proteins are bound to the receptor **(Chap. 370)**, resulting in differential effects on gene transcription that vary depending on other transcription factors present in the cell. Another aspect of selectivity is the affinity of each SERM for the different ERα and ERβ subtypes, which are expressed differentially in various tissues. These tissue-selective effects of SERMs offer the possibility of tailoring estrogen therapy to best meet the needs and risk factor profile of an individual patient.

Bisphosphonates Alendronate, risedronate, ibandronate, and zoledronic acid are approved for the prevention and treatment of postmenopausal osteoporosis. Alendronate, risedronate, and zoledronic acid are also approved for the treatment of steroid-induced osteoporosis, and risedronate and zoledronic acid are approved for prevention of steroid-induced osteoporosis. Alendronate, risedronate, and zoledronic acid are also approved for treatment of osteoporosis in men.

Alendronate decreases bone turnover and increases bone mass in the spine by up to 8% versus placebo and by 6% versus placebo in the hip. Multiple trials have evaluated its effect on fracture occurrence. The Fracture Intervention Trial provided evidence in >2000 women with prevalent vertebral fractures that daily alendronate treatment (5 mg/d for 2 years and 10 mg/d for 9 months afterward) reduces

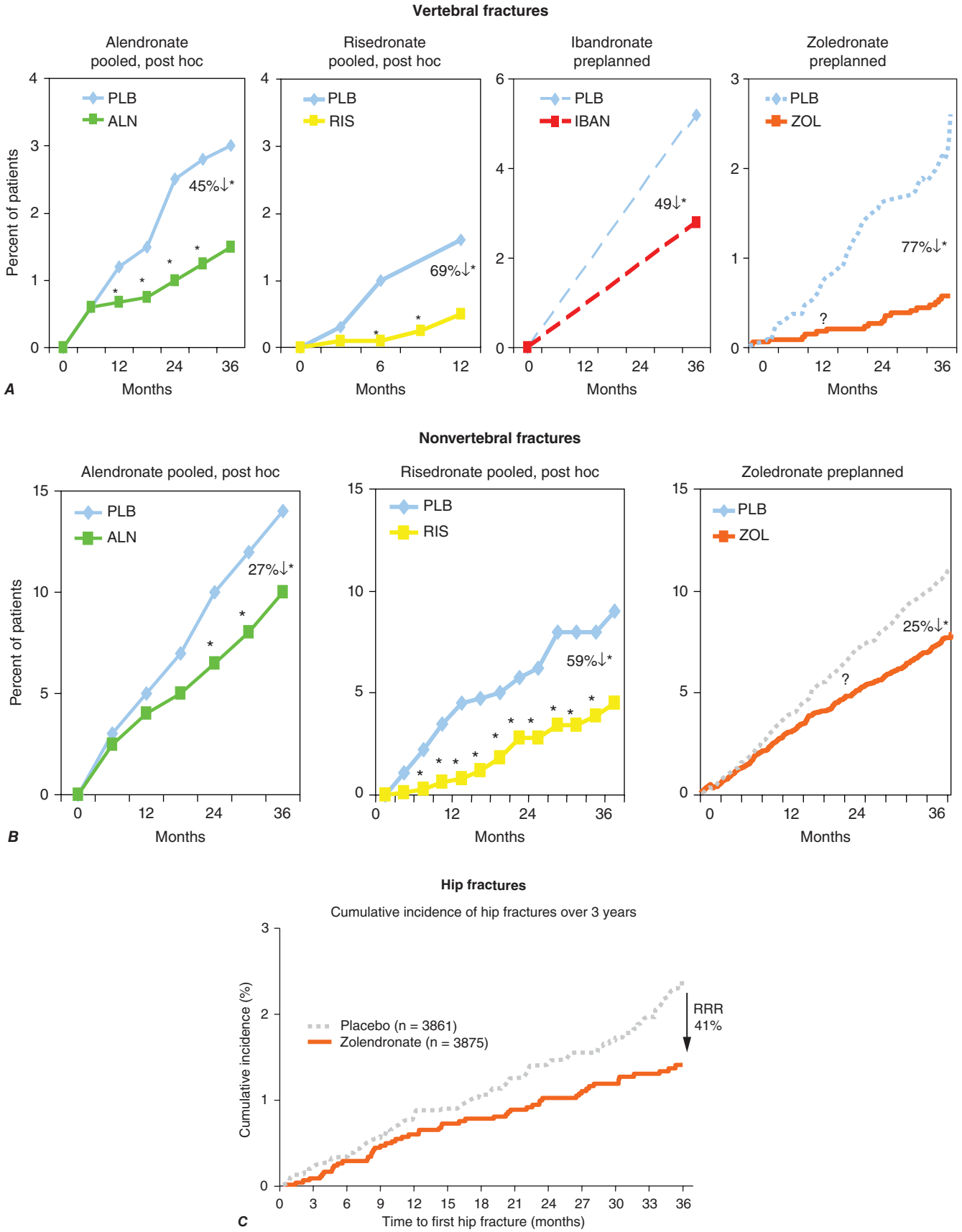


FIGURE 404-9 Effects of various bisphosphonates on clinical vertebral fractures *A.* nonvertebral fractures ***B.,*** and hip fractures ***C.*** Plb, placebo; RRR, relative risk reduction. (*After DM Black et al: J Clin Endocrinol Metab 85:4118, 2000; C Roux et al: Curr Med Res Opin 4:433, 2004; CH Chesnut et al: J Bone Miner Res 19: 1241, 2004; DM Black et al: N Engl J Med 356:1809, 2007; JT Harrington et al: Calcif Tissue Int 74:129, 2003.*)

vertebral fracture risk by about 50%, multiple vertebral fractures by up to 90%, and hip fractures by up to 50%. Several subsequent trials have confirmed these findings **(Figs. 404-9 and 404-10)**. For example, in a study of >1900 women with low bone mass treated with alendronate (10 mg/d) versus placebo, the incidence of all nonvertebral fractures was reduced by ~47% after only 1 year. In the United States the 10mg dose is approved for treatment of osteoporosis and 5 mg/d for prevention.

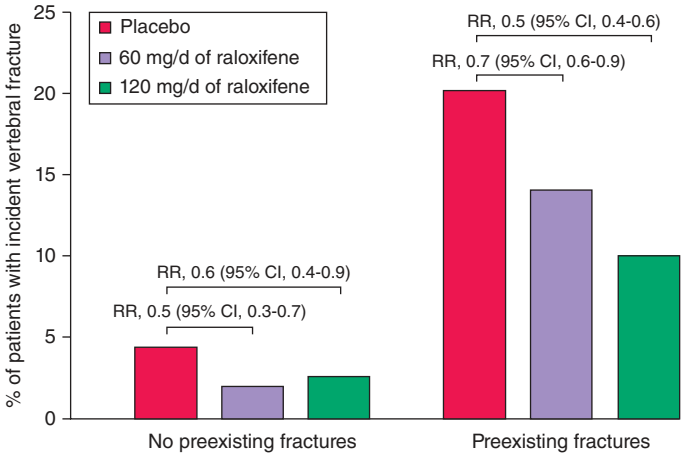


FIGURE 404-10 Effects of two doses of raloxifene on incident vertebral fractures in the MORE trial. *(After B Ettinger et al: JAMA:282:637, 1999.)*

Trials comparing once-weekly alendronate, 70 mg, with daily 10-mg dosing have shown equivalence with regard to bone mass and bone turnover responses. Consequently, once-weekly therapy generally is preferred because of the low incidence of gastrointestinal side effects and ease of administration. Alendronate should be taken with a full glass of water before breakfast after an overnight fast, as bisphosphonates are poorly absorbed. Because of the potential for esophageal irritation, alendronate is contraindicated in patients who have stricture or inadequate emptying of the esophagus. It is recommended that patients remain upright (standing or sitting)t for at least 30 min after taking the medication to avoid esophageal irritation, and that food and fluids (other than water) be avoided for the same duration. In clinical trials, overall gastrointestinal symptomatology was no different with alendronate than with placebo, but all oral bisphosphonates have been associated with esophageal irritation and inflammation.

Risedronate also reduces bone turnover and increases bone mass. Controlled clinical trials have demonstrated 40–50% reduction in vertebral fracture risk over 3 years, accompanied by a 40% reduction in clinical non-spine fractures. The only clinical trial specifically designed to evaluate hip fracture outcome (HIP) indicated that risedronate reduced hip fracture risk in women in their seventies with confirmed osteoporosis by 40%. In contrast, risedronate was not effective at reducing hip fracture occurrence in older women (80+ years) without proven osteoporosis. Studies have shown that 35 mg of risedronate administered once weekly is therapeutically equivalent to 5 mg/d. Patients should take risedronate with a full glass of plain water to facilitate delivery to the stomach and should not lie down for 30 min after taking the drug. (There is also a preparation of risedronate that can be taken with food; it is the only bisphosphonate that has this dosing flexibility.) The incidence of gastrointestinal side effects in trials with risedronate was similar to that of placebo.

Ibandronate is the third amino-bisphosphonate approved in the United States. Ibandronate (2.5 mg/d) has been shown in clinical trials to reduce vertebral fracture risk by ~40% but with no overall effect on non-vertebral fractures. In a post hoc analysis of subjects with a femoral neck T-score of ≤–3, ibandronate reduced the risk of nonvertebral fractures by ~60%. In clinical trials, ibandronate doses of 150 mg/month PO or 3 mg every 3 months IV had greater effects on turnover and bone mass than did 2.5 mg/d. Patients should take oral ibandronate in the same way as other bisphosphonates, but with 1 h elapsing before other food or drink (other than plain water).

Zoledronic acid is a potent bisphosphonate with a unique administration regimen (5 mg by 15 min IV infusion annually). Zoledronic acid data confirm that it is highly effective in fracture risk reduction. In a study of >7000 women followed for 3 years, zoledronic acid 5 mg IV annually) reduced the risk of vertebral fractures by 70%, nonvertebral fractures by 25%, and hip fractures by 40%. These results were associated with less height loss and disability. In the treated population, there was an increased risk of almost 25% of an acute phase reaction in patients with no prior bisphosphonate exposure (fever, myalgias, headache, malaise), but effects were short-lived (2–3 days). Detailed evaluation of all bisphosphonates failed to confirm a risk of atrial fibrillation. Zoledronic acid has also been studied in a placebo controlled trial of women and men within 3 months of an acute hip fracture. The risk of recurrent fracture was reduced by 35% and there was a 28% reduction in mortality.

Common Bisphosphonate Adverse Events All bisphosphonates have been associated with some musculoskeletal and joint pains of unclear etiology, which are occasionally severe. There is potential for renal toxicity and bisphosphonates are contraindicated in those with estimated GFR <30–35 mL/min. Hypocalcemia can occur.

Recently there has been concern about two potential side effects associated with bisphosphonate use. The first is osteonecrosis of the jaw (ONJ). ONJ usually follows a dental procedure in which bone is exposed (dental extractions and implants). It is presumed that the exposed bone becomes infected and dies. ONJ is more common among cancer victims receiving high doses of bisphosphonates for skeletal metastases. It is rare among persons with osteoporosis on usual doses of bisphosphonates. Oral antibiotic rinses and oral systemic antibiotics may be useful to prevent this rare adverse event if risk is perceived to be particularly high. The second is called atypical femoral fracture. These are unusual fractures that occur in the subtrochanteric femoral region or across the femoral shaft distal to the lesser trochanter. They are often preceded by pain in the lateral thigh or groin, that can be present for weeks, months or even years before the fracture. The fractures occur on trivial trauma, are horizontal with a medial beak and are non-comminuted. A committee put together by the American Society for Bone and Mineral Research described the major and minor criteria for these fractures, which appear to be related to duration of bisphosphonate therapy. The overall risk appears quite low, especially when compared to the number of hip fractures saved by these therapies, but they often require surgical fixation and are difficult to heal. Some evidence suggests that if the fractures are found early, when there is evidence of periosteal stress reaction or stress fracture, prior to the occurrence of overt fracture, that teriparatide can help heal the fracture and preclude the need for surgical repair. We routinely inform patients initiating bisphosphonates that if they develop thigh or groin pain they should inform us. Routine x-rays will sometimes pick up cortical thickening or even a stress fracture, but more commonly MRI or Technetium bone scan is required. The presence of an abnormality requires at minimum a period of modified weight bearing and may need prophylactic rodding of the femur. It is important to realize that these may be bilateral (about 50% of the time) and when an abnormality is found the other femur should be checked. It is unknown whether patients who have these atypical femur fractures can ever receive antiresorptive therapies again in the future, but it seems prudent to avoid their use for the majority of these individuals.

Mode of Action Bisphosphonates are structurally related to pyrophosphates, compounds that are incorporated into bone matrix. Bisphosphonates specifically impair osteoclast function and reduce osteoclast number, in part by inducing apoptosis. Recent evidence suggests that the nitrogen-containing bisphosphonates also inhibit protein prenylation, one of the end products in the mevalonic acid pathway, by inhibiting the enzyme farnesyl pyrophosphate synthase. This effect disrupts intracellular protein trafficking and ultimately may lead to apoptosis. Some bisphosphonates have very long retention in the skeleton and may exert long-term effects. The consequences of this, if any, are unknown.

Calcitonin Calcitonin is a polypeptide hormone produced by the thyroid gland **(Chap. 403)**. Its physiologic role is unclear as no skeletal disease has been described in association with calcitonin deficiency or excess. Calcitonin preparations are approved by the FDA for Paget's disease, hypercalcemia, and osteoporosis in women >5 years past menopause.

Injectable calcitonin produces small increments in bone mass of the lumbar spine. However, difficulty of administration and frequent reactions, including nausea and facial flushing, make general use limited. A nasal spray containing calcitonin (200 IU/d) is available for treatment of osteoporosis in postmenopausal women. One study suggests that nasal calcitonin produces small increments in bone mass and a small reduction in new vertebral fractures in calcitonin-treated patients (at one dose) versus those on calcium alone. There has been no proven effectiveness against nonvertebral fractures. Calcitonin is not indicated for prevention of osteoporosis and is not sufficiently potent to prevent bone loss in early postmenopausal women. Calcitonin might have an analgesic effect on bone pain, both in the subcutaneous and possibly in the nasal form. Recently concerns have been raised about an increase in the incidence of cancer associated with calcitonin use. Initially, the cancer noted was of the prostate, but an analysis of all data suggested a more general increase in cancer risk. In Europe the EMA have removed the osteoporosis indication, and an FDA Advisory Committee has voted for a similar change in the United States.

Mode of Action Calcitonin suppresses osteoclast activity by direct action on the osteoclast calcitonin receptor. Osteoclasts exposed to calcitonin cannot maintain their active ruffled border, which normally maintains close contact with underlying bone.

Denosumab A novel agent that given twice yearly by subcutaneous administration in a randomized controlled trial in postmenopausal women with osteoporosis has been shown to increase BMD in the spine, hip, and forearm and reduce vertebral, hip, and nonvertebral fractures over a 3-year period by 70, 40, and 20%, respectively (Fig. 404-11). Other clinical trials indicate ability to increase bone mass in postmenopausal women with low bone mass (above osteoporosis range) and in postmenopausal women with breast cancer treated with aromatase inhibitor therapies. In the oncology literature, denosumab reduces the risk of fractures in women on aromatase inhibitors and also reduces the risk of breast cancer recurrence significantly. In a study of men with prostate cancer treated with androgen deprivation therapy denosumab increased bone mass and reduced vertebral fracture occurrence. Denosumab was approved by the FDA in 2010 for the treatment of postmenopausal women who have a high risk for osteoporotic fractures, including those with a history of fracture or multiple risk factors for fracture, and those who have failed or are intolerant to other osteoporosis therapy. Denosumab is also approved for the treatment of osteoporosis in men at high risk for fracture, and women with breast cancer on aromatase inhibitors and men with prostate cancer on androgen deprivation treatment. A very long-term observational extension of the pivotal trial in postmenopausal women has provided evidence that BMD continues to increase in both the spine and hip with 3–10 years of denosumab treatment, with fracture rates that are at least as low as those seen during the active placebo controlled trial.

Denosumab may increase the risk of ONJ and atypical femur fractures similarly to bisphosphonates. Estimated incidence is 5/10,000 patient years for ONJ and 1/10,000 patient years for atypical femur fractures. Denosumab can cause hypersensitivity reactions, hypocalcemia and skin reactions including dermatitis, rash, and eczema. Early concerns about an imbalance in infections with denosumab have largely been allayed.

When denosumab is discontinued, there is a rebound increase in bone turnover and an apparent acceleration of bone loss. This likely reflects the maturation of osteoclast precursors that have accumulated in marrow when the drug was administered and can become mature bone resorbing cells once the drug is withdrawn. The consequences of this rebound increase in remodeling associated bone loss is a rapid increase in the risk of fracture, particularly vertebral fracture, and a specific increase in the occurrence of multiple vertebral fractures. In patients who need to stop denosumab or in patients in whom BMD and fracture risk reduction goals have been met, temporary use of bisphosphonate treatment can prevent the rebound increase in remodeling and rapid bone loss.

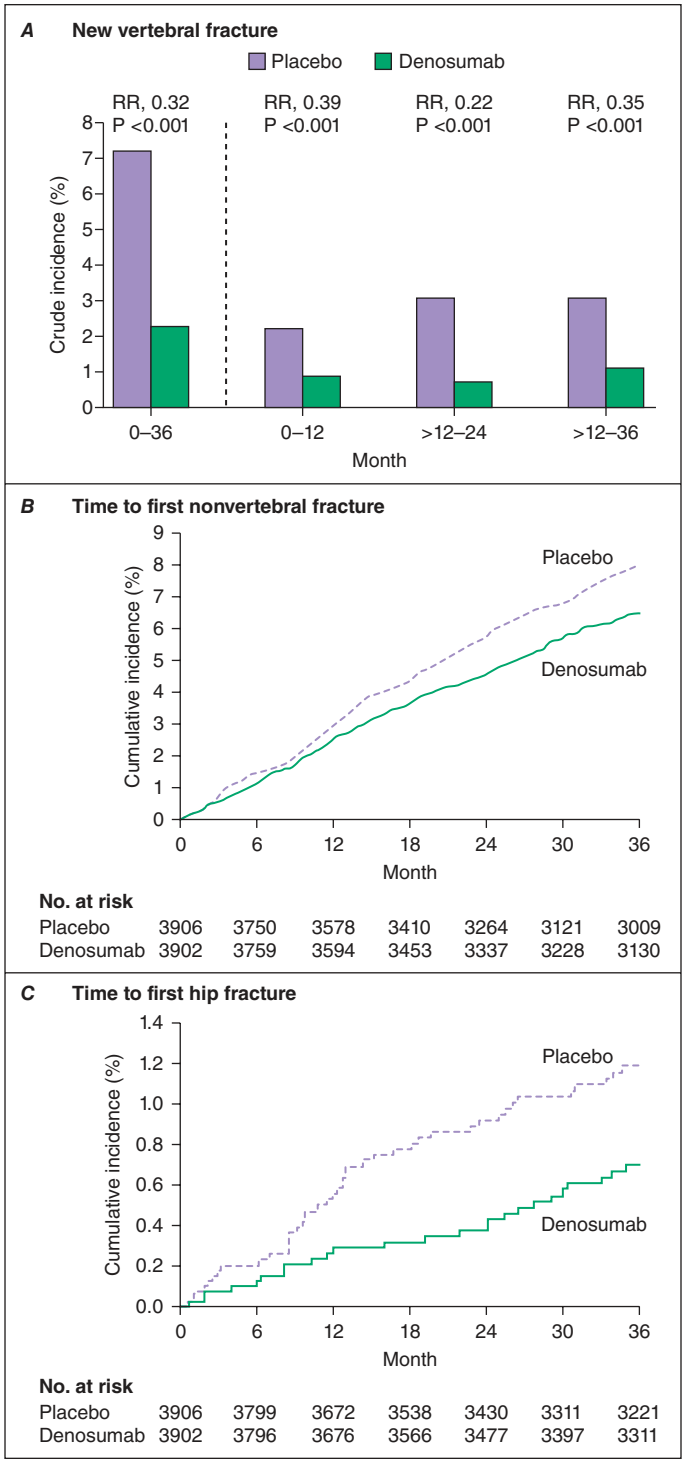


FIGURE 404-11 Effects of denosumab on new vertebral fractures *A.* and times to nonvertebral and hip fracture ***B.*** and ***C.*** *(After SR Cummings et al: N Engl J Med:361:756, 2009.)*

Mode of Action Denosumab is a fully human monoclonal antibody to RANKL, the final common effector of osteoclast formation, activity, and survival. Denosumab binds to RANKL, inhibiting its ability to initiate formation of mature osteoclasts from osteoclast precursors and to bring mature osteoclasts to the bone surface and initiate bone resorption. Denosumab also plays a role in reducing the survival of the osteoclast. Through these actions on the osteoclast, denosumab induces potent antiresorptive action, as assessed biochemically and histomorphometrically.

Anabolic Agents • **Parathyroid Hormone** Endogenous PTH is an 84-amino-acid peptide that is largely responsible for calcium homeostasis (Chap. 403). Although chronic elevation of PTH, as

occurs in hyperparathyroidism, is associated with bone loss (particularly cortical bone), PTH also can exert anabolic effects on bone. Consistent with this, some observational studies have indicated that mild endogenous hyperparathyroidism is associated with maintenance of trabecular bone mass, but loss of cortical bone. On the basis of these findings, early small-scale observational studies showed that PTH analogues could augment trabecular BMD. Subsequent controlled clinical trials have confirmed that PTH can increase bone mass and reduce fracture occurrence. The first randomized controlled trial in postmenopausal women showed that (1–34)PTH (teriparatide), when superimposed on ongoing estrogen therapy, produced substantial increments in bone mass (13% over a 3-year period compared with estrogen alone) and reduced the risk of vertebral compression deformity. In the pivotal study (median, 19 months' duration), 20 μgPTH(1–34) daily by subcutaneous injection (with no additional therapy) reduced vertebral fractures by 65% and nonvertebral fractures by 40–50% **(Fig. 404-12)**. Teriparatide produces rapid and robust increases in bone formation and then bone remodeling overall, resulting in substantial increases in bone mass and improvements in microarchitecture, including cancellous connectivity and cortical width. The BMD effects, particularly in the hip, are lower when patients switch from bisphosphonates to teriparatide, possibly in proportion to the potency of the antiresorptive agent. The hip BMD effect is particularly impaired when patients switch from denosumab to teriparatide. In patients on denosumab who need teriparatide treatment, there may be a role for combination therapy. In previously untreated women, teriparatide is best administered as monotherapy and followed by a potent antiresorptive agent such as denosumab or a bisphosphonate.

In women with painful acute osteoporotic vertebral fractures, teriparatide reduced subsequent vertebral fractures by about 50% compared with risedronate. There was no difference in nonvertebral fracture outcome between the two medications. A recent study comparing teriparatide with risedronate in patients with prevalent vertebral fractures showed significant benefit for teriparatide against vertebral fractures and nearly significant benefit for teriparatide against nonvertebral fractures.

Side effects of teriparatide are generally mild and can include muscle pain, weakness, dizziness, headache, and nausea. Rodents given prolonged treatment with PTH in high doses (3 to 60 times the human dose) developed osteogenic sarcomas after ~18 months of treatment. Rare cases of osteosarcoma have been described in patients treated with teriparatide consistent with the background incidence of osteosarcoma adults. Long-term surveillance studies of a high proportion of patients diagnosed with osteosarcoma as adults in both the United States and Scandinavia reveal no prior exposure to teriparatide in any of the cases.

PTH use may be limited by its mode of administration; alternative modes of delivery are being investigated. Because of the rodent osteosarcoma data, and the maximum duration of teriparatide in the pivotal trial of 2 years, the FDA has limited teriparatide treatment to 2 years.

Mode of Action Exogenously administered PTH appears to have direct actions on osteoblast activity, with biochemical and histomorphometric evidence of de novo bone formation early in response to PTH, before activation of bone resorption. Subsequently, PTH activates bone remodeling but still appears to favor bone formation over bone resorption. PTH given by daily injection stimulates osteoblast recruitment and activity through activation of Wnt signaling. Unlike all other treatments, PTH produces a true increase in bone tissue and an apparent restoration of bone microarchitecture **(Fig. 404-13)**.

Abaloparatide Abaloparatide is a synthetic analogue of human PTH-related peptide (PTHrP), which has significant homology to PTH and also binds the PTH Type 1 Receptor. Abaloparatide and teriparatide exert different binding affinities to the two different receptor conformations, R^0 and RG. Compared to TPTD,

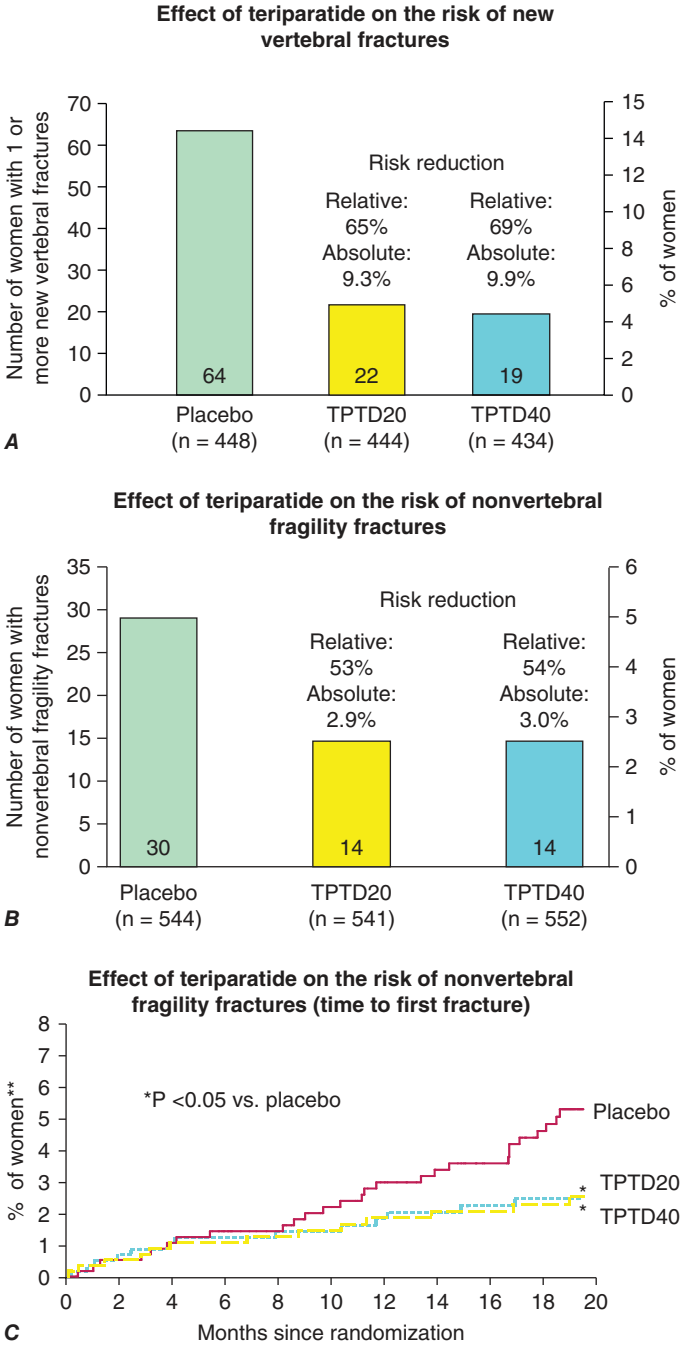


FIGURE 404-12 Effects of teriparatide on new vertebral fractures *A.* and nonvertebral fragility fractures ***B.*** and ***C.*** *(After RM Neer et al: N Engl J Med 344:1434, 2001.)*

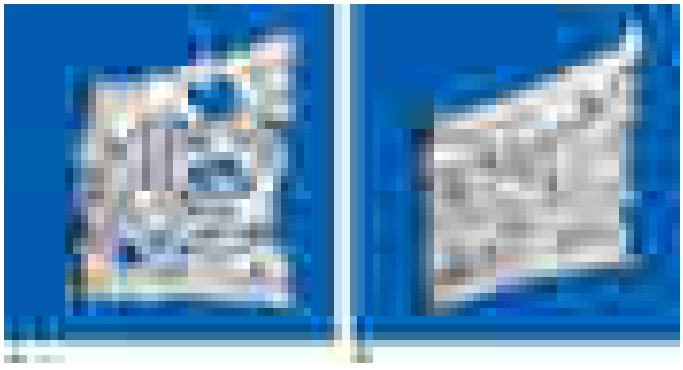


FIGURE 404-13 Effect of parathyroid hormone (PTH) treatment on bone microarchitecture. Paired biopsy specimens from a 64-year-old woman before ***A.*** and after ***B.*** treatment with PTH. *(From DW Dempster et al: J Bone Miner Res 16:1846, 2001.)*

abaloparatide binds with similar high affinity to the RG conformation, but with much lesser affinity to the R^0 conformation. These differences appear to result in a similar bone formation stimulus but lesser bone resorption stimulus and abaloparatide was specifically chosen for development among a large number of PTH and PTHrP analogues for what appeared to be an optimized anabolic profile.

In the Phase 3 Abaloparatide Comparator Trial in Vertebral Endpoints (ACTIVE) study, 2463 postmenopausal women with osteoporosis were randomized to blinded daily subcutaneous abaloparatide vs placebo or open label TPTD. At 18 months, spine BMD increase was similar with abalo and TPTD (11.2% abaloparatide and 10.5% Teriparatide); in the total hip, BMD increments were slightly larger with abaloparatide (4.2% vs 3.3%). New vertebral fracture incidence was reduced by 86% with abaloparatide and 80% with teriparatide compared with placebo (both $p<0.001$). Nonvertebral fractures were reduced by 43% with abaloparatide ($p=0.05$) and 28% with TPTD (NS, $p=0.22$). The ACTIVE study is in an ongoing extension where 92% of eligible participants from the abaloparatide and placebo arms transitioned to open label alendronate for a total treatment period of 24 months of alendronate.

Romosozumab Romosozumab is a humanized antibody that blocks the osteocyte production of sclerostin, resulting in an increase in bone formation and decline in bone resorption. In the pivotal trial (FRAME), 7180 postmenopausal women with osteoporosis were randomized to receive blinded monthly subcutaneous romosozumab (210 mg) or placebo for 1 year followed by transition to open-label subcutaneous denosumab (60 mg) every 6 months for an additional year. BMD increased over 13% in the spine and almost 7% in the hip in one year with romosozumab. At 1 year, the incidence of new vertebral fractures in the romosozumab group was significantly reduced by 73% compared with placebo. Clinical fracture risk (nonvertebral fractures and clinical vertebral fractures combined) was significantly reduced by 36%. Nonvertebral fractures were also reduced but the difference just missed statistical significance perhaps due to geographical differences; in the high enrolling Latin American region, there was no significant reduction in nonvertebral fractures, probably due to a very low background incidence in that region. In the rest of the world, nonvertebral fractures were significantly reduced by >40%. During the second year of the FRAME study, both groups transitioned to denosumab. Over 24 months, women who had received romosozumab during the first 12 months and then denosumab had 75% fewer new vertebral fractures than those who had received placebo for a year followed by denosumab. There were also nearly significant trends toward reduced clinical and nonvertebral fractures in the romosozumab/denosumab group. Compared with baseline, BMD increased by 17.6% in the spine and 8.8% in the total hip in the romosozumab/denosumab group. Safety and tolerability of the two drugs was similar with a slightly higher incidence of injection site reactions in the denosumab group. The FRAME study is in an ongoing extension where all participants received continued denosumab for an additional year. A parallel trial of very high risk patients, all of whom have prevalent vertebral fractures is also ongoing which compares romosozumab to alendronate for 1 year, followed by transition to or continuation of alendronate for 2 additional years.

OTHER UNAPPROVED PHARMACOLOGIC AGENTS

Odanacatib, a cathepsin K inhibitor, inhibits the osteoclast collagenase enzyme, preventing bone resorption but not affecting osteoclast viability. This agent was in late stage drug development. In a very large controlled clinical trial (~17,000 postmenopausal women with osteoporosis), bone mass increased substantially in spine and hip and reduced vertebral, hip and all nonvertebral fractures. Unfortunately, the medication was associated with a significantly increased risk of stroke and the development of this agent was aborted in September 2016.

Testosterone has been used to treat osteoporosis associated with low testosterone levels in men. There are data that indicate that testosterone can increase bone density, but no fracture endpoints. Since there are many other effects of testosterone, especially in older men (including prostate hypertrophy), decisions to use it for treatment of osteoporosis have to take the multisystemic effects into account.

Sodium fluoride was tested in two large parallel clinical trials in the late 1980s. Although BMD increased substantially, the increase was in part due to fluoride incorporation in the hydroxyapatite crystal. Fracture risk was not reduced and in fact was increased in nonvertebral sites. Therefore, fluoride is no longer considered a viable option for osteoporosis treatment.

Strontium ranelate has never been approved for osteoporosis in the United States but is approved in Europe and other exUS countries. It increases bone mass throughout the skeleton, but some of the increase is related to strontium incorporation into hydroxyapatite. In clinical trials, the drug reduced the risk of vertebral fractures by 37% and that of nonvertebral fractures by 14%. It appears to be modestly antiresorptive while at the same time not causing as much of a decrease in bone formation (measured biochemically). In 2014, the use of strontium was restricted because of an increased risk of cardiovascular disease and severe skin reactions. Small increased risks of venous thrombosis also occur.

Several small studies of growth hormone (GH), alone or in combination with other agents, have not shown consistent or substantial positive effects on skeletal mass.

NONPHARMACOLOGIC APPROACHES

Protective pads worn around the outer thigh, which cover the trochanteric region of the hip, can prevent hip fractures in elderly residents in nursing homes. The use of hip protectors is limited largely by issues of compliance and comfort, but new devices are being developed that may circumvent these problems and provide adjunctive treatments.

Kyphoplasty and *vertebroplasty* are also useful non-pharmacologic approaches for the treatment of painful vertebral fractures. However, no long-term data are available.

TREATMENT MONITORING

There are currently no well-accepted guidelines for monitoring treatment of osteoporosis. Because most osteoporosis treatments produce small or moderate bone mass increments on average, it is reasonable to consider BMD as a monitoring tool. Changes must exceed ~4% in the spine and 6% in the hip to be considered significant in any individual. The hip is the preferred site due to larger surface area and greater reproducibility. Medication-induced increments may require several years to produce changes of this magnitude (if they do at all). Consequently, it can be argued that BMD should be repeated at intervals >2 years. Only significant BMD reductions should prompt a change in medical regimen, as it is expected that many individuals will not show responses greater than the detection limits of the current measurement techniques.

Biochemical markers of bone turnover can help in treatment monitoring, but little hard evidence currently supports this concept; it remains unclear which endpoint is most useful. If bone turnover markers are used, a determination should be made before therapy is started and repeated ≥4 months after therapy is initiated. In general, a change in bone turnover markers must be 30–40% lower than the baseline to be significant because of the biologic and technical variability in these tests. A positive change in biochemical markers and/or bone density can be useful to help patients adhere to treatment regimens. Because markers change more rapidly than bone density they are often early signs of treatment effect. Currently collagen C-telopeptide measured on a fasting serum sample in the morning is the preferred marker of bone resorption, and osteocalcin or the propeptide of type 1 collagen (P1NP) for formation.

GLUCOCORTICOID-INDUCED OSTEOPOROSIS

Osteoporotic fractures are a well-characterized consequence of the hypercortisolism associated with Cushing's syndrome. However, the therapeutic use of glucocorticoids is by far the most common form of

glucocorticoid-induced osteoporosis (GCIO). Glucocorticoids are used widely in the treatment of a variety of disorders, including chronic lung disorders, rheumatoid arthritis, and other connective tissue diseases, inflammatory bowel disease, and after transplantation. Osteoporosis and related fractures are serious side effects of chronic glucocorticoid therapy. Because the effects of glucocorticoids on the skeleton are often superimposed on the consequences of aging and menopause, it is not surprising that women and the elderly are most frequently affected. The skeletal response to steroids is remarkably heterogeneous, however, and even young, growing individuals treated with glucocorticoids can present with fractures.

The risk of fractures depends on the dose and duration of glucocorticoid therapy, although recent data suggest that there may be no completely safe dose. Bone loss is more rapid during the early months of treatment, and trabecular bone is affected more severely than cortical bone. As a result, fractures have been shown to increase within 3 months of steroid treatment. There is an increase in fracture risk in both the axial skeleton and the appendicular skeleton, including risk of hip fracture. Bone loss can occur with any route of steroid administration, including high-dose inhaled glucocorticoids and intraarticular injections. Alternate-day delivery does not appear to ameliorate the skeletal effects of glucocorticoids.

PATHOPHYSIOLOGY

Glucocorticoids increase bone loss by multiple mechanisms, including (1) inhibition of osteoblast function and an increase in osteoblast apoptosis, resulting in impaired synthesis of new bone; (2) stimulation of bone resorption, probably as a secondary effect; (3) impairment of the absorption of calcium across the intestine, probably by a vitamin D–independent effect; (4) increase of urinary calcium loss and perhaps induction of some degree of secondary hyperparathyroidism; (5) reduction of adrenal androgens and suppression of ovarian and testicular secretion of estrogens and androgens; and (6) induction of glucocorticoid myopathy, which may exacerbate effects on skeletal and calcium homeostasis as well as increase the risk of falls.

EVALUATION OF THE PATIENT

Because of the prevalence of glucocorticoid-induced bone loss, it is important to evaluate the status of the skeleton in all patients starting or already receiving long-term glucocorticoid therapy. Modifiable risk factors should be identified, including those for falls. Examination should include testing of height and muscle strength. Laboratory evaluation should include an assessment of 24-h urinary calcium. All patients on long-term (>3 months) glucocorticoids should have measurement of bone mass at both the spine and the hip using DXA. If only one skeletal site can be measured, it is best to assess the spine in individuals <60 years and the hip in those >60 years.

PREVENTION

Bone loss caused by glucocorticoids can be prevented, and the risk of fractures significantly reduced. Strategies must include using the lowest dose of glucocorticoid for disease management. Topical and inhaled routes of administration are preferred, where appropriate. Risk factor reduction is important, including smoking cessation, limitation of alcohol consumption, and participation in weight-bearing exercise, when appropriate. All patients should receive an adequate calcium and vitamin D intake from the diet or from supplements.

TREATMENT

Glucocorticoid-Induced Osteoporosis

Several bisphosphonates (alendronate, risedronate, and zoledronic acid) have been demonstrated in large clinical trials to reduce the risk of fractures in patients being treated with glucocorticoids and are FDA-approved for the treatment of GCIO. Teriparatide is also approved for treatment of glucocorticoid induced osteoporosis. In a trial comparing teriparatide to alendronate, BMD increases were much greater and vertebral fracture risk reduction far more substantial with teriparatide compared to alendronate. A study of denosumab has just been completed and indicates greater efficacy of denosumab compared with risedronate for treatment of GCIO. The American College of Rheumatology has just released new guidelines for the management of GCIO (in 2016).

FURTHER READING

Amin S et al: Trends in fracture incidence: A population-based study over 20 years. J Bone Miner Res 29:581, 2014.

Black DM, Rosen CJ: Postmenopausal osteoporosis. N Engl J Med 374:2096, 2016.

Cosman F et al: Treatment sequence matters: Anabolic and antiresorptive therapy for osteoporosis. J Bone Miner Res 32:198, 2017.

Cosman F et al: Clinician's guide to prevention and treatment of osteoporosis. Osteoporos Int 25:2359, 2014.

Cosman F et al: Spine fracture prevalence in a nationally representative sample of US women and men aged >/=40 years: results from the National Health and Nutrition Examination Survey (NHANES) 2013–2014. Osteoporos Int 28:2319, 2017.

Panday K et al: Medication-induced osteoporosis: Screening and treatment strategies. Ther Adv Musculoskel Dis 6:185, 2014.

Roux C, Briot K: Imminent fracture risk. Osteoporos Int. 28:1765, 2017.

Sapir-Koren R, Livshits G: Osteocyte control of bone remodeling: Is sclerostin a key molecular coordinator of the balanced bone resorption-formation cycles? Osteoporos Int 25:2685, 2014.

Wright NC et al: The recent prevalence of osteoporosis and low bone mass in the United States based on bone mineral density at the femoral neck or lumbar spine. J Bone Miner Res 29:2520, 2014.

405 Paget's Disease and Other Dysplasias of Bone

Murray J. Favus, Tamara J. Vokes

PAGET'S DISEASE OF BONE

Paget's disease is a localized bone-remodeling disorder that affects widespread, noncontiguous areas of the skeleton. The pathologic process is initiated by overactive osteoclastic bone resorption followed by a compensatory increase in osteoblastic new bone formation, resulting in a structurally disorganized mosaic of woven and lamellar bone. Pagetic bone is expanded, less compact, and more vascular; thus, it is more susceptible to deformities and fractures. Although most patients are asymptomatic, symptoms resulting directly from bony involvement (bone pain, secondary arthritis, fractures) or secondarily from the expansion of bone causing compression of surrounding neural tissue are not uncommon.

Epidemiology There is a marked geographic variation in the frequency of Paget's disease, with high prevalence in Western Europe (Great Britain, France, and Germany, but not Switzerland or Scandinavia) and among those who have immigrated to Australia, New Zealand, South Africa, and North and South America. The disease is rare in native populations of the Americas, Africa, Asia, and the Middle East; when it does occur, the affected subjects usually have evidence of European ancestry, supporting the migration theory. For unclear reasons, the prevalence and severity of Paget's disease are decreasing, and the age of diagnosis is increasing.

The prevalence is greater in males and increases with age. Autopsy series reveal Paget's disease in about 3% of those over age 40. Prevalence of positive skeletal radiographs in patients aged >55 years is 2.5% for men and 1.6% for women. Elevated alkaline phosphatase (ALP) levels in asymptomatic patients have an age-adjusted incidence of 12.7 and 7 per 100,000 person-years in men and women, respectively.

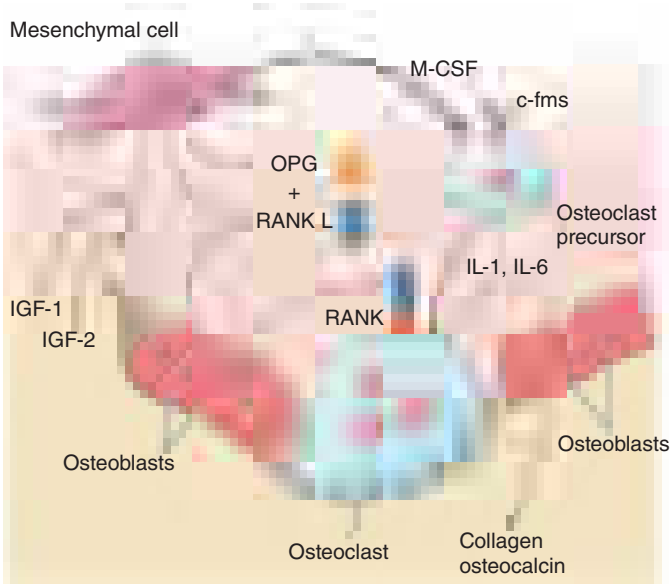


FIGURE 405-1 Diagram illustrating factors that promote differentiation and function of osteoclasts and osteoblasts and the role of the RANK pathway. Stromal bone marrow (mesenchymal) cells and differentiated osteoblasts produce multiple growth factors and cytokines, including macrophage colony-stimulating factor (M-CSF), to modulate osteoclastogenesis. RANKL (receptor activator of nuclear factor-κB ligand) is produced by osteoblast progenitors and mature osteoblasts and can bind to a soluble decoy receptor known as osteoprotegerin (OPG) to inhibit RANKL action. Alternatively, a cell-cell interaction between osteoblast and osteoclast progenitors allows RANKL to bind to its membrane-bound receptor, RANK, thereby stimulating osteoclast differentiation and function. RANK binds intracellular proteins called tumor necrosis factor receptor–associated factors (TRAFs) that mediate receptor signaling through transcription factors such as NF-κB. M-CSF binds to its receptor, c-fms, which is the cellular homologue of the *fms* oncogene. See text for the potential role of these pathways in disorders of osteoclast function such as Paget's disease and osteopetrosis. IL, interleukin; IGF, insulin-like growth factor.

Etiology The etiology of Paget's disease of bone remains unknown, but evidence supports both genetic and viral etiologies. A positive family history is found in 15–25% of patients and, when present, raises the prevalence of the disease seven- to tenfold among first-degree relatives.

A clear genetic basis has been established for several rare familial bone disorders that clinically and radiographically resemble Paget's disease but have more severe presentation and earlier onset. A homozygous deletion of the *TNFRSF11B* gene, which encodes osteoprotegrin **(Fig. 405-1)**, causes *juvenile Paget's disease*, also known as *familial idiopathic hyperphosphatasia*, a disorder characterized by uncontrolled osteoclastic differentiation and resorption. Familial patterns of disease in several large kindred are consistent with an autosomal dominant pattern of inheritance with variable penetrance. *Familial expansile osteolysis, expansile skeletal hyperphosphatasia, and early-onset Paget's disease* are associated with mutations in *TNFRSF11A* gene, which encodes RANK (receptor activator of nuclear factor-κB), a member of the tumor necrosis factor superfamily critical for osteoclast differentiation (Fig. 405-1). Finally, mutations in the gene for valosin-containing protein cause a rare syndrome with autosomal dominant inheritance and variable penetrance known as *inclusion body myopathy with Paget's disease and frontotemporal dementia* (*IBMPFD*). The role of genetic factors is less clear in the more common form of late-onset Paget's disease. The most common mutations identified in familial and sporadic cases of Paget's disease have been in the *SQSTM1* gene (sequestasome-1 or p62 protein) in the C-terminal ubiquitin-binding domain. The other candidate genes include: CSF1 (1p13), which encodes macrophage colony stimulating factor (M-CSF), a cytokine that is required for osteoclast differentiation; RIN3 (14q32), which encodes a guanine exchange factor called Rab and Ras interactor 3; OPTN (10p13), which is involved in regulating NFκB, TNFRSF11A (18q21) which encodes Receptor Activator of Nuclear Factor-κB (RANK), a receptor that is essential for osteoclast differentiation; and TM7SF4, which encodes Dendritic Cell-Specific Transmembrane Protein (DC-STAMP), a molecule that is essential for fusion of osteoclast. The phenotypic variability in patients with *SQSTM1* mutations suggests that additional factors, such as other genetic influences or viral infection, may influence clinical expression of the disease.

Several lines of evidence suggest that a viral infection may contribute to the clinical manifestations of Paget's disease, including (1) the presence of cytoplasmic and nuclear inclusions resembling paramyxoviruses (measles and respiratory syncytial virus) in pagetic osteoclasts and (2) viral mRNA in precursor and mature osteoclasts. The viral etiology is further supported by conversion of osteoclast precursors to pagetic-like osteoclasts by vectors containing the measles virus nucleocapsid or matrix genes. The decline in the incidence of Paget's disease coincides with the widespread vaccination against measles, also consistent with the potential role of virus in the development of the disease. However, the viral etiology has been questioned by the inability to culture a live virus from pagetic bone and by failure to clone the full-length viral genes from material obtained from patients with Paget's disease.

Pathophysiology The principal abnormality in Paget's disease is the increased number and activity of osteoclasts. Pagetic osteoclasts are large, increased 10- to 100-fold in number, and have a greater number of nuclei (as many as 100 compared to 3–5 nuclei in the normal osteoclast). The overactive osteoclasts may create a sevenfold increase in resorptive surfaces and an erosion rate of 9 μg/d (normal is 1 μg/d). Several causes for the increased number and activity of pagetic osteoclasts have been identified: (1) osteoclastic precursors are hypersensitive to $1,25(OH)_2D_3$; (2) osteoclasts are hyperresponsive to RANK ligand (RANKL), the osteoclast stimulatory factor that mediates the effects of most osteotropic factors on osteoclast formation; (3) marrow stromal cells from pagetic lesions have increased RANKL expression; (4) osteoclast precursor recruitment is increased by interleukin (IL) 6, which is increased in the blood of patients with active Paget's disease and is overexpressed in pagetic osteoclasts; (5) expression of the protooncogene c-*fos*, which increases osteoclastic activity, is increased; and (6) the antiapoptotic oncogene *Bcl*-2 in pagetic bone is overexpressed. Numerous osteoblasts are recruited to active resorption sites and produce large amounts of new bone matrix. As a result, bone turnover is high, and bone mass is normal or increased, not reduced, unless there is concomitant deficiency of calcium and/or vitamin D.

The characteristic feature of Paget's disease is increased bone resorption accompanied by accelerated bone formation. An initial osteolytic phase involves prominent bone resorption and marked hypervascularization. Radiographically, this manifests as an advancing lytic wedge, or "blade of grass" lesion. The second phase is a period of very active bone formation and resorption that replaces normal lamellar bone with haphazard (woven) bone. Fibrous connective tissue may replace normal bone marrow. In the final sclerotic phase, bone resorption declines progressively and leads to a hard, dense, less vascular pagetic or mosaic bone, which represents the so-called burned-out phase of Paget's disease. All three phases may be present at the same time at different skeletal sites.

Clinical Manifestations Diagnosis is often made in asymptomatic patients because they have elevated ALP levels on routine blood chemistry testing or an abnormality on a skeletal radiograph obtained for another indication. The skeletal sites most commonly involved are the pelvis, vertebral bodies, skull, femur, and tibia. Familial cases with an early presentation often have numerous active sites of skeletal involvement.

The most common presenting symptom is pain, which may result from increased bony vascularity, expanding lytic lesions, fractures, bowing, or other deformities. Bowing of the femur or tibia causes gait abnormalities and abnormal mechanical stresses with secondary osteoarthritis of the hip or knee joints. Long bone bowing also causes extremity pain by stretching the muscles attached to the bone softened by the pagetic process. Back pain results from enlarged pagetic vertebrae, vertebral compression fractures, spinal stenosis, degenerative

changes of the joints, and altered body mechanics with kyphosis and forward tilt of the upper back. Rarely, spinal cord compression may result from bone enlargement or from the vascular steal syndrome. Skull involvement may cause headaches, symmetric or asymmetric enlargement of the parietal or frontal bones (frontal bossing), and increased head size. Cranial expansion may narrow cranial foramens and cause neurologic complications including hearing loss from cochlear nerve damage from temporal bone involvement, cranial nerve palsies, and softening of the base of the skull (*platybasia*) with the risk of brainstem compression. Pagetic involvement of the facial bones may cause facial deformity; loss of teeth and other dental conditions; and, rarely, airway compression.

Fractures are serious complications of Paget's disease and usually occur in long bones at areas of active or advancing lytic lesions. Common fracture sites are the femoral shaft and subtrochanteric regions. Neoplasms arising from pagetic bone are rare (<0.5%). The incidence of sarcoma appears to be decreasing, possibly because of earlier, more effective treatment with potent antiresorptive agents. The majority of tumors are osteosarcomas, which usually present with new pain in a long-standing pagetic lesion. Osteoclast-rich benign giant cell tumors may arise in areas adjacent to pagetic bone, and they respond to glucocorticoid therapy.

Cardiovascular complications may occur in patients with involvement of large (15–35%) portions of the skeleton and a high degree of disease activity (ALP four times above normal). The extensive arteriovenous shunting and marked increases in blood flow through the vascular pagetic bone lead to a high-output state and cardiac enlargement. However, high-output heart failure is relatively rare and usually develops in patients with concomitant cardiac pathology. In addition, calcific aortic stenosis and diffuse vascular calcifications have been associated with Paget's disease.

Diagnosis The diagnosis may be suggested on clinical examination by the presence of an enlarged skull with frontal bossing, bowing of an extremity, or short stature with simian posturing. An extremity with an area of warmth and tenderness to palpation may suggest an underlying pagetic lesion. Other findings include bony deformity of the pelvis, skull, spine, and extremities; arthritic involvement of the joints adjacent to lesions; and leg-length discrepancy resulting from deformities of the long bones.

Paget's disease is usually diagnosed from radiologic and biochemical abnormalities. Radiographic findings typical of Paget's disease include enlargement or expansion of an entire bone or area of a long bone, cortical thickening, coarsening of trabecular markings, and typical lytic and sclerotic changes. Skull radiographs **(Fig. 405-2)** reveal regions of "cotton wool," or osteoporosis circumscripta, thickening of diploic areas, and enlargement and sclerosis of a portion or all of one or more skull bones. Vertebral cortical thickening of the superior and inferior end plates creates a "picture frame" vertebra. Diffuse radiodense enlargement of a vertebra is referred to as "ivory vertebra." Pelvic radiographs may demonstrate disruption or fusion of the sacroiliac joints; porotic and radiodense lesions of the ilium with whorls of coarse trabeculation; thickened and sclerotic iliopectineal line (brim sign); and softening with protrusio acetabuli, with axial migration of the hips and functional flexion contracture. Radiographs of long bones reveal bowing deformity and typical pagetic changes of cortical thickening and expansion and areas of lucency and sclerosis **(Fig. 405-3)**. Radionuclide ^{99m}Tc bone scans are less specific but are more sensitive than standard radiographs for identifying sites of active skeletal lesions. Although computed tomography (CT) and magnetic resonance imaging (MRI) studies are not necessary in most cases, CT may be useful for the assessment of possible fracture, and MRI is necessary to assess the possibility of sarcoma, giant cell tumor, or metastatic disease in pagetic bone. Definitive diagnosis of malignancy often requires bone biopsy.

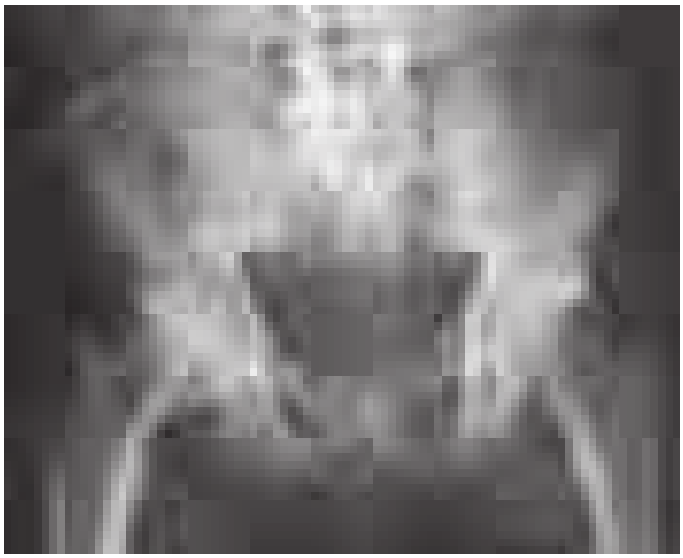


FIGURE 405-3 Radiograph of a 73-year-old man with Paget's disease of the right proximal femur. Note the coarsening of the trabecular pattern with marked cortical thickening and narrowing of the joint space consistent with osteoarthritis secondary to pagetic deformity of the right femur.

Biochemical evaluation is useful in the diagnosis and management of Paget's disease. The marked increase in bone turnover can be monitored using biochemical markers of bone formation and resorption. The parallel rise in markers of bone formation and resorption confirms the coupling of bone formation and resorption in Paget's disease. The degree of bone marker elevation reflects the extent and severity of the disease. For most patients, serum total ALP remains the test of choice

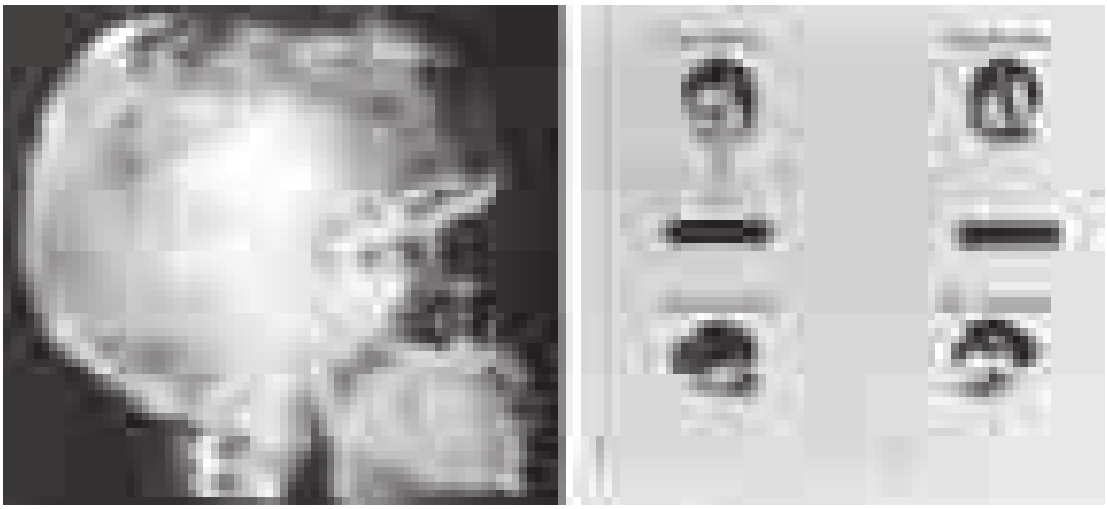


FIGURE 405-2 A 48-year-old woman with Paget's disease of the skull. ***Left.*** Lateral radiograph showing areas of both bone resorption and sclerosis. ***Right.*** ^{99m}Tc HDP bone scan with anterior, posterior, and lateral views of the skull showing diffuse isotope uptake by the frontal, parietal, occipital, and petrous bones.

both for diagnosis and assessing response to therapy. Occasionally, a symptomatic patient with evidence of progression at a single site may have a normal total ALP level but increased bone-specific ALP. For unclear reasons, serum osteocalcin, another marker of bone formation, is not always elevated and is not recommended for use in diagnosis or management of Paget's disease. In contrast, bone formation marker P1NP does reflect the activity of the disease and can be used instead of total ALP. Bone resorption markers (serum or urine N-telopeptide or C-telopeptide measured in the blood or urine) are also elevated in active Paget's disease and decrease more rapidly in response to therapy than does ALP.

Serum calcium and phosphate levels are normal in Paget's disease. Immobilization of a patient with active Paget's disease may rarely cause hypercalcemia and hypercalciuria and increase the risk for nephrolithiasis. However, the discovery of hypercalcemia, even in the presence of immobilization, should prompt a search for another cause of hypercalcemia. In contrast, hypocalcemia or mild secondary hyperparathyroidism may develop in Paget's patients with very active bone formation and insufficient calcium and vitamin D intake, particularly during bisphosphonate therapy when bone resorption is rapidly suppressed and active bone formation continues. Therefore, adequate calcium and vitamin D intake should be instituted prior to administration of bisphosphonates.

TREATMENT

Paget's Disease of Bone

The development of effective and potent pharmacologic agents (Table 405-1) has changed the treatment philosophy from treating only symptomatic patients to treating asymptomatic patients who are at risk for complications. According to the Endocrine Society Clinical Practice Guidelines published in 2014, pharmacologic therapy is indicated for most patients with active Paget's disease who are at risk of complications. Treatment may be initiated to control symptoms caused by metabolically active Paget's disease such as bone pain, fracture, headache, pain from pagetic radiculopathy or arthropathy, or neurologic complications; to decrease local blood flow and minimize operative blood loss in patients who need surgery at an active pagetic site; to reduce hypercalciuria that may occur during immobilization; and to decrease the risk of complications when disease activity is high (elevated ALP) and when the site of involvement involves weight-bearing bones, areas adjacent to major joints, vertebral bodies, and the skull. Whether or not early therapy prevents late complications remains to be determined. A randomized study of over 1200 patients from the United Kingdom showed no difference in bone pain, fracture rates, quality of life, and hearing loss between patients who received pharmacologic therapy to control symptoms (bone pain) and those receiving bisphosphonates to normalize serum ALP. However, the conclusions of that study are debatable since the most potent agent (zoledronic acid, the current drug of choice) was not used/available. It seems likely that the restoration of normal bone architecture following suppression of pagetic activity will prevent further deformities and complications.

Agents approved for treatment of Paget's disease suppress the very high rates of bone resorption and secondarily decrease the high rates of bone formation (Table 405-1). As a result of decreasing bone turnover, pagetic structural patterns, including areas of poorly mineralized woven bone, are replaced by more normal cancellous or lamellar bone. Reduced bone turnover can be documented by a decline in serum formation markers (ALP and P1NP) and urine or serum resorption markers (N-telopeptide, C-telopeptide).

Bisphosphonates are the mainstay of pharmacologic therapy of Paget's disease. Among them, zoledronic acid is currently recommended as the first choice, particularly for those who have severe disease or need rapid normalization of bone turnover (neurologic symptoms, severe bone pain due to a lytic lesion, risk of an impending fracture, or pretreatment prior to elective surgery in an area of active disease). Zoledronic acid normalized bone turnover faster and in a high proportion of patients (over 90%) than oral bisphosphonates with the therapeutic effect persisting for months or even years. It is given at a dose of 5 mg as an intravenous infusion over 20 min although slower rates of infusion are recommended for elderly or those with mild impairment of renal function. More significant renal impairment (GFR <35 mL/min) is a contraindication for use of zoledronic acid due to higher risk of further deterioration of renal function. About 20–25% of patients experience a flu-like syndrome after the first infusion, which can be partly ameliorated by pretreatment with acetaminophen or nonsteroidal anti-inflammatory drugs (NSAIDs). Oral bisphosphonates, alendronate and risedronate, can be used in subjects who have mild disease or some degree of renal impairment. Oral bisphosphonates should be taken first thing in the morning on an empty stomach, followed by maintenance of upright posture with no food, drink, or other medications for 30–60 min. The first clinically useful agent, etidronate, is no longer used due to its low potency and higher risk of inducing osteomalacia. The efficacy of different agents, based on their ability to normalize or decrease ALP levels, is summarized in Table 405-1, although the response rates are not comparable because they are obtained from different studies.

The subcutaneous injectable form of salmon calcitonin is approved for the treatment of Paget's disease but is rarely used due to its low potency and should be reserved for patients who either do not tolerate bisphosphonates or have a contraindication to their use. For patients with contraindication to bisphosphonates, another alternative is denosumab, an antibody to RANKL, which has been reported to result in reduction in ALP. However, it has not been approved for this indication and has less complete and less durable effect than bisphosphonates.

TABLE 405-1 Pharmacologic Agents Approved for Treatment of Paget's Disease

NAME	DOSE AND MODE OF DELIVERY	NORMALIZATION OF ALKALINE PHOSPHATASE (ALP)
Zoledronic acid	5 mg IV over 15 min	90% of patients at 6 mo
Pamidronate	30 mg IV/d over 4 h on 3 days	~50% of patients
Risedronate	30 mg PO/d for 2 mo	73% of patients
Alendronate	40 mg PO/d for 6 mo	63% of patients
Tiludronate	800 mg PO daily for 3 mo	35% of patients
Etidronate	200–400 mg PO/d × 6 mo	15% of patients
Calcitonin (Miacalcin)	100 U SC daily for 6–18 mo (may reduce to 50 U 3 × per wk)	(Reduction of ALP by up to 50%)

SCLEROSING BONE DISORDERS

OSTEOPETROSIS

Osteopetrosis refers to a group of disorders caused by severe impairment of osteoclast-mediated bone resorption. Other terms that are often used include marble bone disease, which captures the solid x-ray appearance of the involved skeleton, and Albers-Schonberg disease, which refers to the milder, adult form of osteopetrosis also known as autosomal dominant osteopetrosis type II. The major types of osteopetrosis include malignant (severe, infantile, autosomal recessive) osteopetrosis and benign (adult, autosomal dominant) osteopetrosis types I and II. A rare autosomal recessive intermediate form has a more benign prognosis. Autosomal recessive carbonic anhydrase (CA) II deficiency produces osteopetrosis of intermediate severity associated with renal tubular acidosis and cerebral calcification.

Etiology and Genetics Naturally occurring and gene-knockout animal models with phenotypes similar to those of the human disorders have been used to explore the genetic basis of osteopetrosis. The primary defect in osteopetrosis is the loss of osteoclastic bone resorption and preservation of normal osteoblastic bone

formation. Osteoprotegerin (OPG) is a soluble decoy receptor that binds osteoblast-derived RANK ligand, which mediates osteoclast differentiation and activation (Fig. 405-1). Transgenic mice that overexpress OPG develop osteopetrosis, presumably by blocking RANK ligand. Mice deficient in RANK lack osteoclasts and develop severe osteopetrosis.

Recessive mutations of CA II prevent osteoclasts from generating an acid environment in the clear zone between its ruffled border and the adjacent mineral surface. Absence of CA II, therefore, impairs osteoclastic bone resorption. Other forms of human disease have less clear genetic defects. About one-half of the patients with malignant infantile osteopetrosis have a mutation in the *TCIRG1* gene encoding the osteoclast-specific subunit of the vacuolar proton pump, which mediates the acidification of the interface between bone mineral and the osteoclast ruffled border. Mutations in the *CICN7* chloride channel gene cause autosomal dominant osteopetrosis type II.

Clinical Presentation The incidence of autosomal recessive severe (malignant) osteopetrosis ranges from 1 in 200,000 to 1 in 500,000 live births. As bone and cartilage fail to undergo modeling, paralysis of one or more cranial nerves may occur due to narrowing of the cranial foramens. Failure of skeletal modeling also results in inadequate marrow space, leading to extramedullary hematopoiesis with hypersplenism and pancytopenia. Hypocalcemia due to lack of osteoclastic bone resorption may occur in infants and young children. The untreated infantile disease is fatal, often before age 5.

Adult (benign) osteopetrosis is an autosomal dominant disease that is usually diagnosed by the discovery of typical skeletal changes in young adults who undergo radiologic evaluation of a fracture. The prevalence is 1 in 100,000 to 1 in 500,000 adults. The course is not always benign, because fractures may be accompanied by loss of vision, deafness, psychomotor delay, mandibular osteomyelitis, and other complications usually associated with the juvenile form. In some kindred, nonpenetrance results in skip generations, while in other families, severely affected children are born into families with benign disease. The milder form of the disease does not usually require treatment.

Radiography Typically, there are generalized symmetric increases in bone mass with thickening of both cortical and trabecular bone. Diaphyses and metaphyses are broadened, and alternating sclerotic and lucent bands may be seen in the iliac crests, at the ends of long bones, and in vertebral bodies. The cranium is usually thickened, particularly at the base of the skull, and the paranasal and mastoid sinuses are underpneumatized.

Laboratory Findings The only significant laboratory findings are elevated serum levels of osteoclast-derived tartrate-resistant acid phosphatase (TRAP) and the brain isoenzyme of creatine kinase. Serum calcium may be low in severe disease, and parathyroid hormone and 1,25-dihydroxyvitamin D levels may be elevated in response to hypocalcemia.

TREATMENT

Osteopetrosis

Allogeneic HLA-identical bone marrow transplantation has been successful in some children. Following transplantation, the marrow contains progenitor cells and normally functioning osteoclasts. A cure is most likely when children are transplanted before age 4. Marrow transplantation from nonidentical HLA-matched donors has a much higher failure rate. Limited studies in small numbers of patients have suggested variable benefits following treatment with interferon γ-1β, 1,25-dihydroxyvitamin D (which stimulates osteoclasts directly), methylprednisolone, and a low-calcium/high-phosphate diet.

Surgical intervention is indicated to decompress optic or auditory nerve compression. Orthopedic management is required for the surgical treatment of fractures and their complications including malunion and postfracture deformity.

PYKNODYSOSTOSIS

This is an autosomal recessive form of osteosclerosis that is believed to have affected the French impressionist painter Henri de Toulouse-Lautrec. The molecular basis involves mutations in the gene that encodes cathepsin K, a lysosomal metalloproteinase highly expressed in osteoclasts and important for bone-matrix degradation. Osteoclasts are present but do not function normally. Pyknodysostosis is a form of short-limb dwarfism that presents with frequent fractures but usually a normal life span. Clinical features include short stature; kyphoscoliosis and deformities of the chest; high arched palate; proptosis; blue sclerae; dysmorphic features including small face and chin, frontooccipital prominence, pointed beaked nose, large cranium, and obtuse mandibular angle; and small, square hands with hypoplastic nails. Radiographs demonstrate a generalized increase in bone density, but in contrast to osteopetrosis, the long bones are normally shaped. Separated cranial sutures, including the persistent patency of the anterior fontanel, are characteristic of the disorder. There may also be hypoplasia of the sinuses, mandible, distal clavicles, and terminal phalanges. Persistence of deciduous teeth and sclerosis of the calvarium and base of the skull are also common. Histologic evaluation shows normal cortical bone architecture with decreased osteoblastic and osteoclastic activities. Serum chemistries are normal, and unlike osteopetrosis, there is no anemia. There is no known treatment for this condition, and there are no reports of attempted bone marrow transplant.

PROGRESSIVE DIAPHYSEAL DYSPLASIA

Also known as *Camurati-Engelmann disease*, progressive diaphyseal dysplasia is an autosomal dominant disorder that is characterized radiographically by diaphyseal hyperostosis and a symmetric thickening and increased diameter of the endosteal and periosteal surfaces of the diaphyses of the long bones, particularly the femur and tibia, and, less often, the fibula, radius, and ulna. The genetic defect responsible for the disease has been localized to the area of chromosome 19q13.2 encoding tumor growth factor (TGF)-β1. The mutation promotes activation of TGF-β1. The clinical severity is variable. The most common presenting symptoms are pain and tenderness of the involved areas, fatigue, muscle wasting, and gait disturbance. The weakness may be mistaken for muscular dystrophy. Characteristic body habitus includes thin limbs with little muscle mass yet prominent and palpable bones and, when the skull is involved, large head with prominent forehead and proptosis. Patients may also display signs of cranial nerve palsies, hydrocephalus, central hypogonadism, and Raynaud's phenomenon. Radiographically, patchy progressive endosteal and periosteal new bone formation is observed along the diaphyses of the long bones. Bone scintigraphy shows increased radiotracer uptake in involved areas.

Treatment with low-dose glucocorticoids relieves bone pain and may reverse the abnormal bone formation. Intermittent bisphosphonate therapy has produced clinical improvement in a limited number of patients.

HYPEROSTOSIS CORTICALIS GENERALISATA

This is also known as *van Buchem's disease*; it is an autosomal recessive disorder characterized by endosteal hyperostosis in which osteosclerosis involves the skull, mandible, clavicles, and ribs. The major manifestations are due to narrowed cranial foramens with neural compressions that may result in optic atrophy, facial paralysis, and deafness. Adults may have an enlarged mandible. Serum ALP levels may be elevated, which reflect the uncoupled bone remodeling with high osteoblastic formation rates and low osteoclastic resorption. As a result, there is increased accumulation of normal bone. Endosteal hyperostosis with syndactyly, known as *sclerosteosis*, is a more severe form. The genetic defects for both sclerosteosis and van Buchem's disease have been associated with mutations in the *SOST* gene.

MELORHEOSTOSIS

Melorheostosis (Greek, "flowing hyperostosis") may occur sporadically or follow a pattern consistent with an autosomal recessive disorder. The major manifestation is progressive linear hyperostosis in one or more bones of one limb, usually a lower extremity. The name comes

from the radiographic appearance of the involved bone, which resembles melted wax that has dripped down a candle. Symptoms appear during childhood as pain or stiffness in the area of sclerotic bone. There may be associated ectopic soft tissue masses, composed of cartilage or osseous tissue, and skin changes overlying the involved bone, consisting of scleroderma-like areas and hypertrichosis. The disease does not progress in adults, but pain and stiffness may persist. Laboratory tests are unremarkable. No specific etiology is known. There is no specific treatment. Surgical interventions to correct contractures are often unsuccessful.

OSTEOPOIKILOSIS

The literal translation of osteopoikilosis is "spotted bones"; it is a benign autosomal dominant condition in which numerous small, variably shaped (usually round or oval) foci of bony sclerosis are seen in the epiphyses and adjacent metaphyses. The lesions may involve any bone except the skull, ribs, and vertebrae. They may be misidentified as metastatic lesions. The main differentiating points are that bony lesions of osteopoikilosis are stable over time and do not accumulate radionucleotide on bone scanning. In some kindred, osteopoikilosis is associated with connective tissue nevi known as *dermatofibrosis lenticularis disseminata*, also known as *Buschke-Ollendorff syndrome*. Histologic inspection reveals thickened but otherwise normal trabeculae and islands of normal cortical bone. No treatment is indicated.

HEPATITIS C–ASSOCIATED OSTEOSCLEROSIS

Hepatitis C–associated osteosclerosis (HCAO) is a rare acquired diffuse osteosclerosis in adults with prior hepatitis C infection. After a latent period of several years, patients develop diffuse appendicular bone pain and a generalized increase in bone mass with elevated serum ALP. Bone biopsy and histomorphometry reveal increased rates of bone formation, decreased bone resorption with a marked decrease in osteoclasts, and dense lamellar bone. One patient had increased serum OPG levels, and bone biopsy showed large numbers of osteoblasts positive for OPG and reduced osteoclast number. Empirical therapy includes pain control, and there may be beneficial response to bisphosphonate. Long-term antiviral therapy may reverse the bone disease.

DISORDERS ASSOCIATED WITH DEFECTIVE MINERALIZATION

HYPOPHOSPHATASIA

This is a rare inherited disorder that presents as rickets in infants and children or osteomalacia in adults with paradoxically low serum levels of ALP. The frequency of the severe neonatal and infantile forms is about 1 in 100,000 live births in Canada, where the disease is most common because of its high prevalence among Mennonites and Hutterites. It is rare in African Americans. The severity of the disease is remarkably variable, ranging from intrauterine death associated with profound skeletal hypomineralization at one extreme to premature tooth loss as the only manifestation in some adults. Severe cases are inherited in an autosomal recessive manner, but the genetic patterns are less clear for the milder forms. The disease is caused by a deficiency of tissue nonspecific (bone/liver/kidney) ALP (TNSALP), which, although ubiquitous, results only in bone abnormalities. Protein levels and functions of the other ALP isozymes (germ cell, intestinal, placental) are normal. Defective ALP permits accumulation of its major naturally occurring substrates including phosphoethanolamine (PEA), inorganic pyrophosphate (PPi), and pyridoxal 5′-phosphate (PLP). The accumulation of PPi interferes with mineralization through its action as a potent inhibitor of hydroxyapatite crystal growth.

Perinatal hypophosphatasia becomes manifest during pregnancy and is often complicated by polyhydramnios and intrauterine death. The infantile form becomes clinically apparent before the age of 6 months with failure to thrive, rachitic deformities, functional craniosynostosis despite widely open fontanels (which are actually hypomineralized areas of the calvarium), raised intracranial pressure, and flail chest with predisposition to pneumonia. Hypercalcemia and hypercalciuria are common. This form has a mortality rate of about 50%. Prognosis seems to improve for the children who survive infancy. Childhood hypophosphatasia has variable clinical presentation. Premature loss of deciduous teeth (before age 5) is the hallmark of the disease. Rickets causes delayed walking with waddling gait, short stature, and dolichocephalic skull with frontal bossing. The disease often improves during puberty but may recur in adult life. Adult hypophosphatasia presents during middle age with painful, poorly healing metatarsal stress fractures or thigh pain due to femoral pseudofractures. It is important to recognize hypophosphatasia in adults because treatment with bisphosphonates can result in increased rather than decreased bone fragility.

Laboratory investigation reveals low ALP levels and normal or elevated levels of serum calcium and phosphorus despite clinical and radiologic evidence of rickets or osteomalacia. Serum parathyroid hormone, 25-hydroxyvitamin D, and 1,25-dihydroxyvitamin D levels are normal. The elevation of PLP is specific for the disease and may even be present in asymptomatic parents of severely affected children. Because vitamin B_6 increases PLP levels, vitamin B_6 supplements should be discontinued 1 week before testing. Clinical testing is available to detect loss-of-function mutation(s) within the *ALPL* gene that encodes TNSALP.

In contrast to other forms of rickets and osteomalacia, calcium and vitamin D supplementation should be avoided because they may aggravate hypercalcemia and hypercalciuria. A low-calcium diet, glucocorticoids, and calcitonin have been used in a small number of patients with variable responses. Because fracture healing is poor, placement of intramedullary rods is best for acute fracture repair and for prophylactic prevention of fractures. In 2015, asfotase alfa, a tissue nonspecific ALP has been approved as enzyme replacement therapy perinatal/infantile- and juvenile-onset forms.

AXIAL OSTEOMALACIA

This is a rare disorder characterized by defective skeletal mineralization despite normal serum calcium and phosphate levels. Clinically, the disorder presents in middle-aged or elderly men with chronic axial skeletal discomfort. Cervical spine pain may also be present. Radiographic findings are mainly osteosclerosis due to coarsened trabecular patterns typical of osteomalacia. Spine, pelvis, and ribs are most commonly affected. Histologic changes show defective mineralization and flat, inactive osteoblasts. The primary defect appears to be an acquired defect in osteoblast function. The course is benign, and there is no established treatment. Calcium and vitamin D therapies are not effective.

FIBROGENESIS IMPERFECTA OSSIUM

This is a rare condition of unknown etiology. It presents in both sexes; in middle age or later; and with progressive, intractable skeletal pain and fractures; worsening immobilization; and a debilitating course. Radiographic evaluation reveals generalized osteomalacia, osteopenia, and occasional pseudofractures. Histologic features include a tangled pattern of collagen fibrils with abundant osteoblasts and osteoclasts. There is no effective treatment. Spontaneous remission has been reported in a small number of patients. Calcium and vitamin D have not been beneficial.

FIBROUS DYSPLASIA AND MCCUNE-ALBRIGHT SYNDROME

Fibrous dysplasia is a sporadic disorder characterized by the presence of one (monostotic) or more (polyostotic) expanding fibrous skeletal lesions composed of bone-forming mesenchyme. The association of the polyostotic form with café au lait spots and hyperfunction of an endocrine system such as pseudoprecocious puberty of ovarian origin is known as *McCune-Albright syndrome* (MAS). A spectrum of the phenotypes is caused by activating mutations in the *GNAS1* gene, which encodes the α subunit of the stimulatory G protein ($G_s\alpha$). As the postzygotic mutations occur at different stages of early development, the extent and type of tissue affected are variable and explain the mosaic pattern of skin and bone changes. GTP binding activates the $G_s\alpha$ regulatory protein and mutations in regions of $G_s\alpha$ that selectively inhibit GTPase activity, which results in constitutive stimulation of the cyclic

AMP–protein kinase A signal transduction pathway. Such mutations of the $G_s\alpha$ protein–coupled receptor may cause autonomous function in bone (parathyroid hormone receptor); skin (melanocyte-stimulating hormone receptor); and various endocrine glands including ovary (follicle-stimulating hormone receptor), thyroid (thyroid-stimulating hormone receptor), adrenal (adrenocorticotropic hormone receptor), and pituitary (growth hormone–releasing hormone receptor). The skeletal lesions are composed largely of mesenchymal cells that do not differentiate into osteoblasts, resulting in the formation of imperfect bone. In some areas of bone, fibroblast-like cells develop features of osteoblasts in that they produce extracellular matrix that organizes into woven bone. Calcification may occur in some areas. In other areas, cells have features of chondrocytes and produce cartilage-like extracellular matrix.

Clinical Presentation Fibrous dysplasia occurs with equal frequency in both sexes, whereas MAS with precocious puberty is more common (10:1) in girls. The monostotic form is the most common and is usually diagnosed in patients between 20 and 30 years of age without associated skin lesions. The polyostotic form typically manifests in children <10 years old and may progress with age. Early-onset disease is generally more severe. Lesions may become quiescent in puberty and progress during pregnancy or with estrogen therapy. In polyostotic fibrous dysplasia, the lesions most commonly involve the maxilla and other craniofacial bones, ribs, and metaphyseal or diaphyseal portions of the proximal femur or tibia. Expanding bone lesions may cause pain, deformity, fractures, and nerve entrapment. Sarcomatous degeneration involving the facial bones or femur is infrequent (<1%). The risk of malignant transformation is increased by radiation, which has proven to be ineffective treatment. In rare patients with widespread lesions, renal phosphate wasting and hypophosphatemia may cause rickets or osteomalacia. Hypophosphatemia may be due to production of a phosphaturic factor by the abnormal fibrous tissue.

MAS patients may have café au lait spots, which are flat, hyperpigmented skin lesions that have rough borders ("coast of Maine") in contrast to the café au lait lesions of neurofibromatosis that have smooth borders ("coast of California"). The most common endocrinopathy is isosexual pseudoprecocious puberty in girls. Other less common endocrine disorders include thyrotoxicosis, Cushing's syndrome, acromegaly, hyperparathyroidism, hyperprolactinemia, and pseudoprecocious puberty in boys.

Radiographic Findings In long bones, the fibrous dysplastic lesions are typically well-defined, radiolucent areas with thin cortices and a ground-glass appearance. Lesions may be lobulated with trabeculated areas of radiolucency **(Fig. 405-4)**. Involvement of facial bones usually presents as radiodense lesions, which may create a leonine appearance (leontiasis osea). Expansile cranial lesions may narrow foramens and cause optic lesions, reduce hearing, and create other manifestations of cranial nerve compression.

Laboratory Results Serum ALP is occasionally elevated but calcium, parathyroid hormone, 25-hydroxyvitamin D, and 1,25-dihydroxy-vitamin D levels are normal. Patients with extensive polyostotic lesions may have hypophosphatemia, hyperphosphaturia, and osteomalacia. The hypophosphatemia and phosphaturia are directly related to the levels of fibroblast growth factor 23 (FGF23). Biochemical markers of bone turnover may be elevated.

TREATMENT

Fibrous Dysplasia and MAS

Spontaneous healing of the lesions does not occur, and there is no established effective treatment. Improvement in bone pain and partial or complete resolution of radiographic lesions have been reported after IV bisphosphonate therapy. Surgical stabilization is used to prevent pathologic fracture or destruction of a major joint space and to relieve nerve root or cranial nerve compression or sinus obstruction.

FIGURE 405-4 Radiograph of a 16-year-old male with fibrous dysplasia of the right proximal femur. Note the multiple cystic lesions, including the large lucent lesion in the proximal midshaft with scalloping of the interior surface. The femoral neck contains two lucent cystic lesions.

OTHER DYSPLASIAS OF BONE AND CARTILAGE

PACHYDERMOPERIOSTOSIS

Pachydermoperiostosis, or hypertrophic osteoarthropathy (primary or idiopathic), is an autosomal dominant disorder characterized by periosteal new bone formation that involves the distal extremities. The lesions present as clubbing of the digits and hyperhidrosis and thickening of the skin, primarily of the face and forehead. The changes usually appear during adolescence, progress over the next decade, and then become quiescent. During the active phase, progressive enlargement of the hands and feet produces a paw-like appearance, which may be mistaken for acromegaly. Arthralgias, pseudogout, and limited mobility may also occur. The disorder must be differentiated from secondary hypertrophic osteopathy that develops during the course of serious pulmonary disorders. The two conditions can be differentiated by standard radiography of the digits in which secondary pachydermoperiostosis has exuberant periosteal new bone formation and a smooth and undulating surface. In contrast, primary hypertrophic osteopathy has an irregular periosteal surface.

There are no diagnostic blood or urine tests. Synovial fluid does not have an inflammatory profile. There is no specific therapy, although a limited experience with colchicine suggests some benefit in controlling the arthralgias.

OSTEOCHONDRODYSPLASIAS

These include several hundred heritable disorders of connective tissue. These primary abnormalities of cartilage manifest as disturbances in cartilage and bone growth. Selected growth-plate chondrodysplasias are described here. **For discussion of chondrodysplasias, see Chap. 406.**

Achondrodysplasia This is a relatively common form of short-limb dwarfism that occurs in 1 in 15,000 to 1 in 40,000 live births. The disease is caused by a mutation of the fibroblast growth factor receptor 3 (*FGFR3*) gene that results in a gain-of-function state. Most cases are sporadic mutations. However, when the disorder appears in families, the inheritance pattern is consistent with an autosomal dominant disorder. The primary defect is abnormal chondrocyte proliferation at the growth plate that causes development of short, but proportionately thick, long bones. Other regions of the long bones may be relatively

unaffected. The disorder is manifest by the presence of short limbs (particularly the proximal portions), normal trunk, large head, saddle nose, and an exaggerated lumbar lordosis. Severe spinal deformity may lead to cord compression. The homozygous disorder is more serious than the sporadic form and may cause neonatal death. Pseudoachondroplasia clinically resembles achondrodysplasia but has no skull abnormalities.

Enchondromatosis This is also called *dyschondroplasia* or *Ollier's disease*; it is also a disorder of the growth plate in which the primary cartilage is not resorbed. Cartilage ossification proceeds normally, but it is not resorbed normally, leading to cartilage accumulation. The changes are most marked at the ends of long bones, where the highest growth rates occur. Chondrosarcoma develops infrequently. The association of enchondromatosis and cavernous hemangiomas of the skin and soft tissues is known as *Maffucci's syndrome*. Both Ollier's disease and Maffucci's syndrome are associated with various malignancies, including granulosa cell tumor of the ovary and cerebral glioma.

Multiple Exostoses This is also called *diaphyseal aclasis* or *osteochondromatosis*; it is a genetic disorder that follows an autosomal dominant pattern of inheritance. In this condition, areas of growth plates become displaced, presumably by growing through a defect in the perichondrium. The lesion begins with vascular invasion of the growth-plate cartilage, resulting in a characteristic radiographic finding of a mass that is in direct communication with the marrow cavity of the parent bone. The underlying cortex is resorbed. The disease is caused by inactivating mutations of the *EXT1* and *EXT2* genes, whose products normally regulate processing of chondrocyte cytoskeletal proteins. The products of the *EXT* gene likely function as tumor suppressors, with the loss-of-function mutation resulting in abnormal proliferation of growth-plate cartilage. Solitary or multiple lesions are located in the metaphyses of long bones. Although usually asymptomatic, the lesions may interfere with joint or tendon function or compress peripheral nerves. The lesions stop growing when growth ceases but may recur during pregnancy. There is a small risk for malignant transformation into chondrosarcoma.

EXTRASKELETAL (ECTOPIC) CALCIFICATION AND OSSIFICATION

Deposition of calcium phosphate crystals (*calcification*) or formation of true bone (*ossification*) in nonosseous soft tissue may occur by one of three mechanisms: (1) metastatic calcification due to a supranormal calcium × phosphate concentration product in extracellular fluid; (2) dystrophic calcification due to mineral deposition into metabolically impaired or dead tissue despite normal serum levels of calcium and phosphate; and (3) ectopic ossification, or true bone formation. Disorders that may cause extraskeletal calcification or ossification are listed in Table 405-2.

TABLE 405-2 Diseases and Conditions Associated with Ectopic Calcification and Ossification

Metastatic calcification	Dystrophic calcification
Hypercalcemic states	Inflammatory disorders
Primary hyperparathyroidism	Scleroderma
Sarcoidosis	Dermatomyositis
Vitamin D intoxication	Systemic lupus erythematosus
Milk-alkali syndrome	Trauma-induced
Renal failure	Ectopic ossification
Hyperphosphatemia	Myositis ossificans
Tumoral calcinosis	Postsurgery
Secondary hyperparathyroidism	Burns
Pseudohypoparathyroidism	Neurologic injury
Renal failure	Other trauma
Hemodialysis	Fibrodysplasia ossificans progressiva
Cell lysis following chemotherapy	
Therapy with vitamin D and phosphate	

METASTATIC CALCIFICATION

Soft tissue calcification may complicate diseases associated with significant hypercalcemia, hyperphosphatemia, or both. In addition, vitamin D and phosphate treatments or calcium administration in the presence of mild hyperphosphatemia, such as during hemodialysis, may induce ectopic calcification. Calcium phosphate precipitation may complicate any disorder when the serum calcium × phosphate concentration product is >75. The initial calcium phosphate deposition is in the form of small, poorly organized crystals, which subsequently organize into hydroxyapatite crystals. Calcifications that occur in hypercalcemic states with normal or low phosphate have a predilection for kidney, lungs, and gastric mucosa. Hyperphosphatemia with normal or low serum calcium may promote soft tissue calcification with predilection for the kidney and arteries. The disturbances of calcium and phosphate in renal failure and hemodialysis are common causes of soft tissue (metastatic) calcification.

TUMORAL CALCINOSIS

This is a rare genetic disorder characterized by masses of metastatic calcifications in soft tissues around major joints, most often shoulders, hips, and ankles. Tumoral calcinosis differs from other disorders in that the periarticular masses contain hydroxyapatite crystals or amorphous calcium phosphate complexes, while in fibrodysplasia ossificans progressiva (below), true bone is formed in soft tissues. About one-third of tumoral calcinosis cases are familial, with both autosomal recessive and autosomal dominant modes of inheritance reported. The disease is also associated with a variably expressed abnormality of dentition marked by short bulbous roots, pulp calcification, and radicular dentin deposited in swirls. The primary defect responsible for the metastatic calcification appears to be hyperphosphatemia resulting from the increased capacity of the renal tubule to reabsorb filtered phosphate. Spontaneous soft tissue calcification is related to the elevated serum phosphate, which, along with normal serum calcium, exceeds the concentration product of 75.

The disease usually presents in childhood and continues throughout the patient's life. The calcific masses are typically painless and grow at variable rates, sometimes becoming large and bulky. The masses are often located near major joints but remain extracapsular. Joint range of motion is not usually restricted unless the tumors are very large. Complications include compression of neural structures and ulceration of the overlying skin with drainage of chalky fluid and risk of secondary infection. Small deposits not detected by standard radiographs may be detected by ^{99m}Tc bone scanning. The most common laboratory findings are hyperphosphatemia and elevated serum 1,25-dihydroxyvitamin D levels. Serum calcium, parathyroid hormone, and ALP levels are usually normal. Renal function is also usually normal. Urine calcium and phosphate excretions are low, and calcium and phosphate balances are positive.

An acquired form of the disease may occur with other causes of hyperphosphatemia, such as secondary hyperparathyroidism associated with hemodialysis, hypoparathyroidism, pseudohypoparathyroidism, and massive cell lysis following chemotherapy for leukemia. Tissue trauma from joint movement may contribute to the periarticular calcifications. Metastatic calcifications are also seen in conditions associated with hypercalcemia, such as in sarcoidosis, vitamin D intoxication, milk-alkali syndrome, and primary hyperparathyroidism. In these conditions, however, mineral deposits are more likely to occur in proton-transporting organs such as kidney, lungs, and gastric mucosa in which an alkaline milieu is generated by the proton pumps.

TREATMENT

Tumoral Calcinosis

Therapeutic successes have been achieved with surgical removal of subcutaneous calcified masses, which tend not to recur if all calcification is removed from the site. Reduction of serum phosphate by chronic phosphorus restriction may be accomplished using low dietary phosphorus intake alone or in combination with oral

phosphate binders. The addition of the phosphaturic agent acetazolamide may be useful. Limited experience using the phosphaturic action of calcitonin deserves further testing.

DYSTROPHIC CALCIFICATION

Posttraumatic calcification may occur with normal serum calcium and phosphate levels and normal ion-solubility product. The deposited mineral is either in the form of amorphous calcium phosphate or hydroxyapatite crystals. Soft tissue calcification complicating connective tissue disorders such as scleroderma, dermatomyositis, and systemic lupus erythematosus may involve localized areas of the skin or deeper subcutaneous tissue and is referred to as *calcinosis circumscripta*. Mineral deposition at sites of deeper tissue injury including periarticular sites is called *calcinosis universalis*.

ECTOPIC OSSIFICATION

True extraskeletal bone formation that begins in areas of fasciitis following surgery, trauma, burns, or neurologic injury is referred to as *myositis ossificans*. The bone formed is organized as lamellar or trabecular, with normal osteoblasts and osteoclasts conducting active remodeling. Well-developed haversian systems and marrow elements may be present. A second cause of ectopic bone formation occurs in an inherited disorder, *fibrodysplasia ossificans progressiva*.

FIBRODYSPLASIA OSSIFICANS PROGRESSIVA

This is also called *myositis ossificans progressiva*; it is a rare autosomal dominant disorder characterized by congenital deformities of the hands and feet and episodic soft tissue swellings that ossify. Ectopic bone formation occurs in fascia, tendons, ligaments, and connective tissue within voluntary muscles. Tender, rubbery induration, sometimes precipitated by trauma, develops in the soft tissue and gradually calcifies. Eventually, heterotopic bone forms at these sites of soft tissue trauma. Morbidity results from heterotopic bone interfering with normal movement and function of muscle and other soft tissues. Mortality is usually related to restrictive lung disease caused by an inability of the chest to expand. Laboratory tests are unremarkable.

There is no effective medical therapy. Bisphosphonates, glucocorticoids, and a low-calcium diet have largely been ineffective in halting progression of the ossification. Surgical removal of ectopic bone is not recommended, because the trauma of surgery may precipitate formation of new areas of heterotopic bone. Dental complications including frozen jaw may occur following injection of local anesthetics.

FURTHER READING

Cundy T et al: Durability of response to zoledronate treatment and competing mortality in Paget's disease of bone. Bone Miner Res 32:753, 2017.

Langston AL et al: Randomized trial of intensive bisphosphonate treatment versus symptomatic management in Paget's disease of bone. J Bone Miner Res 25:20, 2010.

Ralston SH et al: Letter to the Editor: The Endocrine Society Clinical Practice Guidelines on Paget's Disease: Many recommendations are not evidence based. J Clin Endocrinol Metab 100:L45, 2015.

Reid IR et al: Treatment of Paget's disease of bone with denosumab: Case report and literature review. Calcif Tissue Int 99:322, 2016.

Robinson C et al: Fibrous dysplasia/McCune-Albright syndrome: Clinical and translational perspectives. Curr Osteoporos Rep 14:178, 2016.

Singer FR: Paget's disease of the bone—Genetic and environmental factors. Nat Rev Endocrinol 11:662, 2015.

Singer FR et al: Paget's disease of bone: An endocrine society clinical practice guideline. J Clin Endocrinol Metab 99:4408, 2014.

Whyte MP: Sclerosing bone disorders, in *Primer on the Metabolic Bone Diseases and Disorders of Mineral Metabolism*, 8th ed, CJ Rosen (ed). Wiley-Blackwell, 2013, pp 769–785.

Whyte MP: Hypophosphatasia—Aetiology, nosology, pathogenesis, diagnosis and treatment. Nat Rev Endocrinol 12:233, 2016.

406 Heritable Disorders of Connective Tissue

Darwin J. Prockop, John F. Bateman

CLASSIFICATION OF CONNECTIVE TISSUE DISORDERS

Some of the most common conditions that are transmitted genetically in families are disorders that produce clinically obvious changes in the skeleton, skin, or other relatively acellular tissues that have been loosely defined as connective tissues. Because of their heritability, some of the disorders were recognized as potentially traceable to mutated genes soon after the principles of genetics were introduced into medicine by Garrod and others. About half a century later, McKusick emphasized the specificity of many of the diseases for selective connective tissues and suggested that they were probably caused by mutations in genes coding for the major proteins found in those tissues. In the last several decades, many of the disorders have been linked to mutations in several hundred different genes expressed in connective tissues. However, classifying the disorders on the basis of either their clinical presentations or the mutations causing them is continuing to present a challenge for both the clinician and the molecular biologist.

The information on the disorders has continued to develop on two levels. The initial clinical classifications suggested by McKusick and many others had to be refined as more patients were examined. For example, some patients had skin changes similar to those commonly seen in Ehlers-Danlos syndrome (EDS), but this feature was over-shadowed by other features such as extreme hypotonia or sudden rupture of large blood vessels. To account for the full spectrum of presentations in patients and families, many of the disorders have been re-classified several times and each has been divided into a series of sub-types. The task was daunting. For example, a recent effort to classify all the heritable disorders that alter the skeleton defined 436 distinctive conditions that were divided into 42 major groups.

The identification of mutations causing the diseases has developed on a parallel track. The first genes cloned for connective tissues were the two genes coding for Type I collagen, the most abundant protein in bones, skin, tendons and several other tissues. Some of the first assays in patients with osteogenesis imperfecta (OI) revealed mutations in Type I collagen genes. And biochemical data developed primarily with cultures of skin fibroblasts from the patients demonstrated that the mutations dramatically altered the synthesis or structure of collagen fibers. The results stimulated efforts to identify additional mutations in genes coding for structural proteins. Genes for collagens provided an attractive paradigm to search for mutations, since a series of different types of collagens were found in different connective tissues and the collagen genes were readily isolated by their unique signature sequences. Also, the collagen genes were vulnerable to a large number of different mutations because of unusual structural requirements of the protein. The search for mutations in collagen genes proved fruitful in that mutations were found in most patients with OI, in many patients with hyper-extensible skin, in some patients with dwarfism, and in patients with other disorders, including some such as the Alport syndrome (AS) that were not initially classified as disorders of connective tissue. Also mutations in collagen genes were found in subset of patients with a diagnosis of osteoarthritis (OA) and a subset of patients with the diagnosis of osteoporosis. However, the search for mutations quickly expanded to hundreds of other genes that included genes for other structural proteins, for the post-translational processing of the structural proteins, for growth factors and their receptors, and other genes whose functions are still not fully understood.

In many instances, the mutations helped to define the clinical subtype of the disorder. In some, however, it did not. Some patients with the

same clinical presentations were found to have mutations in different genes. Also, some patients with different manifestations were found to have mutations in same genes. In addition, it was difficult to establish whether a change in the structure of a gene caused the phenotypic changes in the patients and was not simply a neutral polymorphism. Therefore, there has been a continuing debate as to whether the disorders should be classified by their clinical presentations or by the genes at fault. As an illustration of the problems, mutations in 324 genes have been found associated with the 436 defined disorders of the skeleton. The latest nosology for the disorders remains a "hybrid between a list of clinically defined disorders, waiting for molecular clarification, and an annotated database documenting the phenotypic spectrum produced by mutations in a given gene." A simpler system of classification proved feasible for one rare heritable disorder of skin, epidermatolysis bullosa. The disorder was first defined by the presence of friction-induced blister. It was then divided into subtypes that were defined by the ultrastructural layers of the skin that cleaved and blistered. Most patients in each subtype were subsequently shown to have mutations in genes expressed in the corresponding layer of skin. Even with these patients, however, the strength of the genotype-phenotype correlation varies and mutations have not yet been found in every patient.

The best pathway through this maze of information is probably to begin by matching the signs and symptoms in a patient with the presentations that define each clinical classification. A major focus should be on the most common disorders, recognizing that the signs and symptoms may vary among different individuals and family members with the same diagnosis. Then, attempt to reach a decision, in consultation with the patient, parents and probably a specialist, as to whether a DNA analysis for the probable mutation is indicated. Among the considerations are the cost, the rigor with which the clinical classification has been linked to mutated genes, the reassurance the diagnosis can bring to patients and their families, the use of the information for prenatal diagnosis, and the possibility that mutation-specific therapies may be developed in the future. For patients with the most severe forms, it is probably best to consult a specialist in the disease to determine a program for therapy. Patient help groups have formed for many of the diseases and are an important source of information.

Patients with the most common forms of the disorders have mutations in a limited number of genes. This chapter will focus primarily on these. Also, it will provide a brief summary of biosynthesis and structure of connective tissues that may help guide the physician from the nature of the mutations to their clinical presentations.

COMPOSITION OF CONNECTIVE TISSUES

Connective tissues such as skin, bone, cartilage, ligaments, and tendons are the critical structural frameworks of the body. They consist of a complex interacting extracellular matrix network of collagens, proteoglycans, and a large number of non-collagenous glycoproteins and proteins. While these precise combinations of up to ~500 potential extracellular matrix building blocks provide tissue-specific function, there are many overarching similarities in composition such as the role of composite collagen fibrils in providing strength and form, elastin fibrils and proteoglycans and other interacting proteins, and glycoproteins that fine-tune function **(Table 406-1)**. The most abundant components of many connective tissues are three similar fibrillar collagens (Types I, II, and III). They have a similar tensile strength that is comparable to that of steel wires. The three fibrillar collagens are distributed in a tissue-specific manner: Type I collagen accounts for most of the protein of dermis, ligaments, tendons, and demineralized bone; Type I and Type III are the most abundant proteins of large blood vessels; and Type II is the most abundant protein of cartilage.

BIOSYNTHESIS AND TURNOVER OF CONNECTIVE TISSUES

Connective tissues are among the most stable components in living organisms, but they are not inert. During embryonic development,

TABLE 406-1 Constituents of Connective Tissues

CONNECTIVE TISSUE	MAJOR CONSTITUENTS	APPROXIMATE AMOUNTS, % DRY WT	CHARACTERISTICS OR FUNCTIONS
Dermis, ligaments, tendons	Type I collagen	80	Large bundles of fibrils
	Type III collagen	5–15	Thin fibrils
	Type IV collagen, laminins, and nidogen	<5	Form basal laminae under epithelium
	Types V, VI, and VII collagens	<5	V modifies Type I fibrils; VI forms beaded micro-fibrils; VII forms anchoring fibrils for epidermis
	Fibrillin aggregates/elastin	<5	Provide elasticity
	Fibronectin	<5	Associated with collagen fibers and cell surfaces
	Proteoglycans[a]/hyaluronan	<0.5	Provide resiliency
Bone (demineralized)	Type I collagen	90	Complex fibril network
	Type VI collagen	1–2	Beaded micro-fibrils
	Proteoglycans[a]/hyaluronan	1	Function unclear
	Osteonectin, osteopontin, osteocalcin, α2-glycoprotein, and sialoproteins	1–5	May regulate mineralization
Aorta	Type I collagen	20–40	Fibril network
	Type III collagen	20–40	Thin fibrils
	Fibrillin aggregates/elastin	20–40	Provide elasticity
	Type IV collagen, laminins, and nidogen	<5	Form basal lamina under endothelial cells
	Types V and VI collagens	<2	V modifies Type I fibrils; VI forms beaded micro-fibrils
	Proteoglycans[a]/hyaluronan	<3	Provide resiliency
Cartilage	Type II collagen	40–50	Arcades of thin fibrils
	Type IX collagen	5–10	Links Type II fibrils and other components
	Type VI collagen	<1	Beaded micro-fibrils, largely pericellular
	Type X collagen	5–10	Forms pericellular network in hypertrophic growth plate cartilage
	Type XI collagen	<10	Incorporated into some Type II fibrils
	Proteoglycans[a]/hyaluronan	15–50	Provide resiliency
	Small leucine-rich repeat proteins (SLRPs; >6 kinds)	<5	Multiple functions in assembly and function of the tissue

[a]Over 30 proteoglycans have been identified. They differ in the structures of their core proteins and their contents of glycosaminoglycan side chains of chondroitin-4-sulfate, chondroitin-6-sulfate, dermatan sulfate, and keratin sulfate. Basal lamina contains a proteoglycan with a side chain of heparan sulfate that resembles heparin.

connective tissue membranes appear as early as the four-cell blastocyst to provide a structural scaffold for the developing embryo. With the development of blood vessels and skeleton, there is a rapid increase in the synthesis, degradation, and resynthesis of connective tissues. The turnover continues at a slower, but still rapid pace throughout postnatal development and then spikes during the growth spurt of puberty. During adulthood, the metabolic turnover of most connective tissues is slow, but it continues at a moderate pace in bone. With age, malnutrition, physical inactivity, and low gravitational stress, the rate of degradation of most connective tissues, especially in bone and skin, begins to exceed the rate of synthesis and the tissues shrink. In starvation, a large fraction of the collagen in skin and other connective tissues is degraded and provides amino acids for gluconeogenesis **(Chap. 327)**. In both OA and rheumatoid arthritis, there is extensive degradation of articular cartilage collagen. Glucocorticoids weaken most tissues by decreasing collagen synthesis. In many pathologic states, however, collagen is deposited in excess. With most injuries to tissues, inflammatory and immune responses stimulate the deposition of collagen fibrils in the form of fibrotic scars. In humans, as distinct from many other species, the deposition of the fibrils is largely irreversible and prevents regeneration of normal tissues in diseases such as hepatic cirrhosis, pulmonary fibrosis, atherosclerosis, and nephrosclerosis.

Structure and Biosynthesis of Fibrillar Collagens The tensile strength of collagen fibers derives primarily from the self-assembly of protein monomers into large fibril structures in a process that resembles crystallization. The self-assembly requires monomers of highly uniform and relatively rigid structure. It also requires a complex series of posttranslational processing steps that maintain the solubility of the monomers until they are transported to the appropriate extracellular sites for fibril assembly. Because of the stringent requirements for correct self-assembly, it is not surprising that mutations in genes for fibrillar collagens cause many of the diseases of connective tissues.

The monomers of the three fibrillar collagens are formed from three polypeptide chains, called *α chains*, that are wrapped around each other into a rope-like triple-helical conformation. The triple helix is a unique structure among proteins, and it provides rigidity to the molecule. It also orients the side chains of amino acids in an "inside out" manner relative to most other proteins so that the charged and hydrophobic residues on the surface can direct self-assembly of the monomers into fibrils. The triple-helical conformation of the monomer is generated because each of the α chains has a repetitive amino acid sequence in which glycine (Gly) appears as every third amino acid. Each α chain contains about 1000 amino acids. Therefore, the sequence of each α chain can be designated as $(\text{-Gly-X-Y-})_n$, where X and Y represent amino acids other than glycine and n is >338. The presence of glycine, the smallest amino acid, in every third position in the sequence is critical because this residue must fit into a sterically restricted space in the middle of the helix where the three chains come together. The requirement for a glycine residue at every third position explains the severe effects of mutations that convert any of the glycine residues to an amino acid with a bulkier side chain (see below). Many of the X- and Y-position amino acids are proline and hydroxyproline, which, because of their ring structures, provide additional rigidity to the triple helix. Other X- and Y-positions are occupied by charged or hydrophobic amino acids that precisely direct lateral and longitudinal assembly of the monomers into highly ordered fibrils. Mutations that substitute amino acids in some X- and Y-positions can, in rare instances, also produce genetic diseases.

The fibers formed by the three fibrillar collagens differ in thickness and length, but they have a similar fine structure. As viewed by electron microscopy, they all have a characteristic pattern of cross-striations that are about one-quarter the length of the monomers and reflect the precise packing into fibrils. The three fibrillar collagens, however, differ in sequences found in the X- and Y-positions of the α chains and therefore in some of their physical properties. Type I collagen is composed of two identical α1(I) chains and a third α2(I) chain that differs slightly in its amino acid sequence. Type II collagen is composed of three identical α(II) chains. Type III collagen is also composed of three distinctive but identical α1(III) chains.

To deliver a monomer of the correct structure to the appropriate site of fibril assembly, the biosynthesis of fibrillar collagens involves a large number of unique processing steps **(Fig. 406-1)**. The monomer,

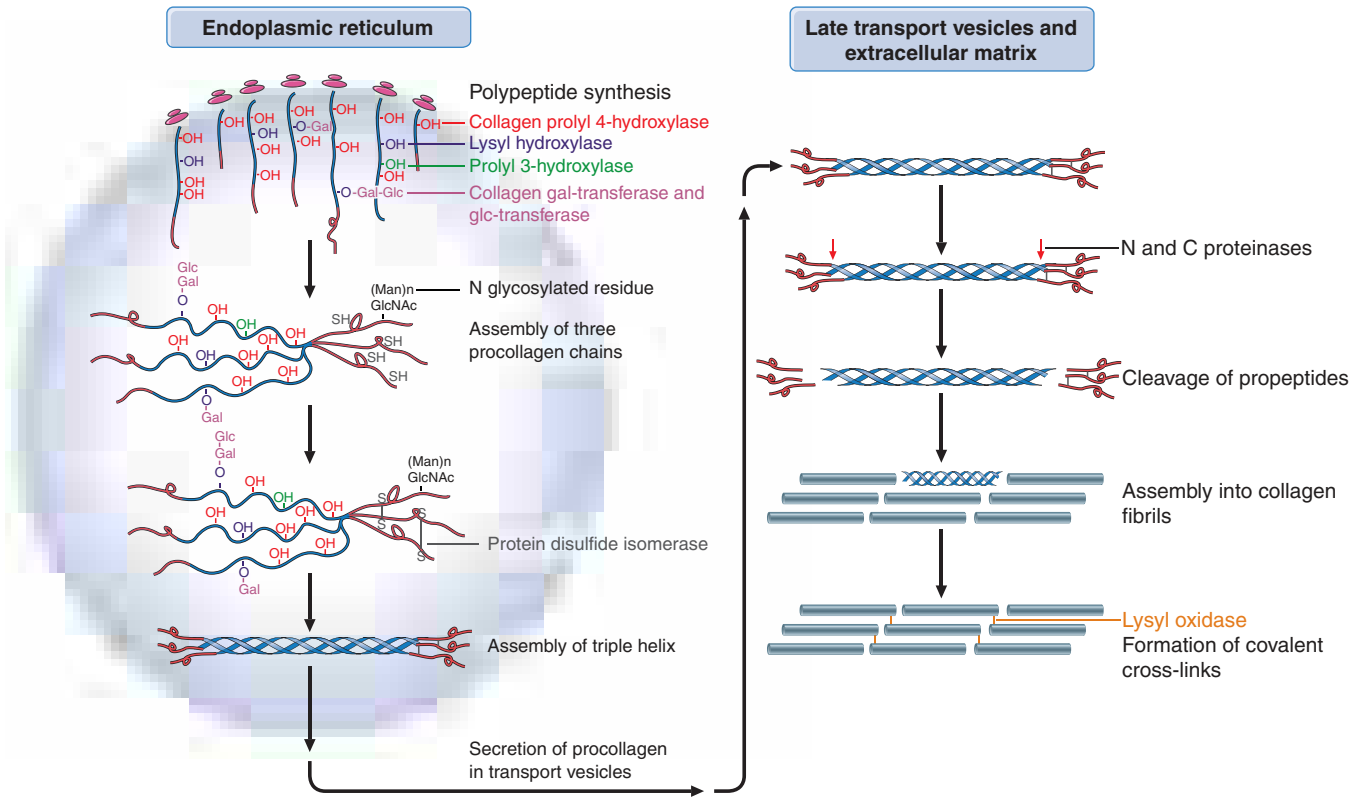


FIGURE 406-1 Schematic summary of biosynthesis of fibrillar collagens. *(Modified and reproduced with permission from J Myllyharju, KI Kivirikko: Trends in Genetics 20:33, 2004.)*

first synthesized as a soluble precursor called *procollagen*, contains an additional globular domain at each end. As the pre-proα chains of procollagen are synthesized on ribosomes, the free N-terminal ends move into the cisternae of the rough endoplasmic reticulum (ER). Signal peptides at the N-termini are cleaved, and additional posttranslational reactions begin. Proline residues in the Y-position of the repeating -Gly-X-Y- sequences are converted to hydroxyproline by the enzyme prolyl hydroxylase. The hydroxylation of prolyl residues is essential for the three α chains of the monomer to fold into a triple helix at body temperature. The enzyme requires ascorbic acid as one of its essential cofactors, an observation that explains why wounds fail to heal in scurvy **(Chap. 326)**. In scurvy, some of the underhydroxylated and unfolded protein accumulates in the cisternae of the rough ER and is degraded. Lysine residues in the Y-position are also hydroxylated to hydroxylysine by a separate lysyl hydroxylase. Many of the hydroxylysine residues are glycosylated with galactose or with galactose and glucose. A large mannose-rich oligosaccharide is assembled on the C-terminal propeptide of each chain. The proα chains are assembled by interactions among these C-terminal propeptides that control the selection of the appropriate partner chains to form hetero- or homotrimers and provide the correct chain registration required for subsequent formation of the collagen triple helix. After the C-terminal propeptides assemble the three proα chains, a nucleus of triple helix is formed near the C-terminus, and the helical conformation is propagated toward the N-terminus in a zipper-like manner that resembles crystallization. The folding into the triple helix is spontaneous in solution, but as discussed below, identification of rare mutations causing OI demonstrated that the folding *in cellulo* is assisted by a number of ancillary proteins which also prevent collagen fibril formation with the ER. The fully folded protein is then transported to the Golgi via a specific COPII vesicle process. After further modifications in the Golgi stack, the procollagen is secreted. After secretion, procollagen is processed to collagen by cleavage of the N-propeptides and C-propeptides by two specific proteinases. The release of the propeptides decreases the solubility of the protein about 1000-fold. The entropic energy that is released drives the self-assembly of the collagen into fibrils. Self-assembled collagen fibers have considerable tensile strength, but their strength is increased further by cross-linking reactions that form covalent bonds between α chains in one molecule and α chains in adjacent molecules. The resulting fibers, comprised of hundreds or thousands of triple-helical monomers, have some of the properties of a crystal but have innate imperfections that make them highly flexible.

Although the assembly of collagen monomers into fibers is largely a spontaneous reaction, the process in tissues is modulated by the presence of less abundant collagens (Type V with Type I, and Type XI with Type II) and by other components such as a series of small leucine-rich proteins (SLRPs). Some of the less abundant components alter the rate of fibril assembly, whereas others change the morphology of the fibers or their interactions with cells and other molecules. The presence of these other components is one explanation for why, in some tissues, the fibers are further assembled into large tendons; in others, into sheets; and in still others into complex structures such as the hexagonal array of fibers that provide both the strength and transparency of the cornea.

Collagen fibers are resistant to most proteases, but during degradation of connective tissues, they are cleaved by specific matrix metalloproteinases (collagenases) that cause partial unfolding of the triple helices into gelatin-like structures that are further degraded by less specific proteinases.

OTHER COLLAGENS AND RELATED MOLECULES

The unique properties of the triple helix are used to define a family of at least 28 collagens that contain repetitive -Gly-X-Y- sequences and form triple helices of varying length and complexity. The proteins are heterogeneous both in structure and function, and many are the sites of mutations causing genetic diseases. For example, the Type IV collagen found in basement membranes is composed of three α chains synthesized from any of six different genes. Mutations in any of the six genes can cause AS.

Fibrillin Aggregates and Elastin In addition to tensile strength, many tissues such as the lung, large blood vessels, and ligaments require elasticity. The elasticity was originally ascribed to an amorphous rubber-like protein named elastin. Subsequent analyses, largely sparked by discoveries of mutations causing the Marfan syndrome (MFS), demonstrated that the elasticity resided in thin fibrils composed primarily of large glycoproteins named fibrillins. The fibrillins contain large numbers of epidermal growth factor–like domains interspersed with characteristic cysteine-rich domains that are also found in latent transforming growth factor β (TGF-β) binding proteins. The fibrillins assemble into long beadlike strands that also contain numerous other components including small and variable amounts of elastin, bone morphogenic proteins (BMPs), and microfibril-associated glycoproteins (MAGPs). The principles whereby the fibrils provide elasticity to tissue and their biosynthetic assembly are still under investigation. As well as contributing to extracellular matrix structure, the fibrillins play a major role in TGF-β signaling.

Proteoglycans The resiliency to compression of connective tissues such as cartilage or the aorta is largely explained by the presence of proteoglycans. Proteoglycans are composed of a core protein to which are attached a large series of negatively charged polymers of disaccharides (largely chondroitin sulfates). At least 30 proteoglycans have been identified. They vary in their binding to collagens and other components of matrix, but specific functions have not been assigned to most. The major proteoglycan of cartilage, called aggrecan, has a core protein of 2000 amino acids that is decorated with about 100 side chains of chondroitin sulfate and keratin sulfate. The core protein, in turn, binds to long chains of the polymeric disaccharide hyaluronan to form proteoglycan aggregates, one of the largest soluble macromolecular structures in nature. Because of its highly negative charge and extended structure, the proteoglycan aggregate binds large amounts of water and small ions to distend the three-dimensional arcade of collagen fibers found in the same tissues. It thereby makes the cartilage resilient to pressure.

SPECIFIC DISORDERS

OSTEOGENESIS IMPERFECTA

The central feature of OI is a severe decrease in bone mass that makes bones brittle. The disorder is frequently associated with blue sclerae, dental abnormalities (dentinogenesis imperfecta), progressive hearing loss, and a positive family history. Most patients have mutations in one of the two genes coding for Type I collagen.

Classification OI was originally classified into two subtypes of congenita and tarda depending on the age of onset of the symptoms. Sillence suggested a series of sub-types based on clinical, radiological findings, and mode of inheritance. As with the other disorders discussed here, the description of rare recessive forms of OI and discovery of mutations in new genes has opened a debate as to whether the disorders should be classified by the clinical phenotypes or by the genes at fault. For the moment, the classification based on the clinical presentations seems the most useful **(Table 406-2)**.

Type I is the mildest subtype and can produce either mild or no apparent deformities of the skeleton. Most patients have distinctly blue sclerae. Type II produces bone so brittle that it is lethal *in utero* or shortly after birth; it can be subclassified into Types II A, B, and C, depending on radiologic findings. Of the nonlethal forms, Type III is progressively deforming with moderate to severe bone deformity and Type IV (common variable OI with normal sclerae) has mild to moderate bone fragility.

The classifications of patients by types of OI do not consistently predict the clinical course of the disease. Some patients appear normal at birth and become progressively worse; others have multiple fractures in infancy and childhood, improve after puberty, and fracture more frequently later in life. Women are particularly prone to fracture during pregnancy and after menopause. A few women from families with mild variants of OI do not develop fractures until after menopause, and their disease may be difficult to distinguish from postmenopausal osteoporosis.

TABLE 406-2 Classification of Osteogenesis Imperfecta (OI)

PHENOTYPE	TYPE	TYPICAL FEATURES	INHERITANCE	GENE/PROTEIN DEFECT	PROTEIN DEFECT
Non-deforming form	OI type 1	Mild to moderate bone fragility, normal or near normal stature, in most, blue sclerae, normal dentition in most hearing loss in ~50%.	AD	***COL1A1***	Collagen I haploinsufficiency
				COL1A2	
Perinatally lethal form	OI type 2	Extreme bone fragility, short stature, long bone bowing	AD	***COL1A1***	Collagen I structural mutations
				COL1A2	
			AR	*CRTAP*	Collagen post-translational modification and folding machinery
				LEPRE1/P3H1	
				PPIB/CYPB	
Progressively deforming OI	OI type 3	Moderate to severe bone deformity, blue sclerae at birth, hearing loss and abnormal dentition common.	AD	***COL1A1***	Collagen I structural mutations
				COL1A2	
			AR	*CRTAP*	Collagen post-translational modification and folding machinery
				LEPRE1/P3H1	
				PPIB/CYPB	
				FKBP10/FKBP65	
				PLOD2/ LH2	
				SERPINH1/HSP47	
				CREB3L1/OASIS	Protein folding/endoplasmic reticulum stress sensor
				SEC24D	COPII vesicle component collagen secretion
				BMP1	Proteolytic removal of procollagen N-propeptide
				WNT1	Wnt cell signaling pathway
				SERPINF1/PEDF	Signaling and collagen binding protein, important for mineralization
				TMEM38B/TRIC-B	Endoplasmic reticulum cation channel. Calcium homoestasis/collagen synthesis
				SP7/OSX	Transcription factor, bone formation
Common Variable OI with normal sclerae	OI type 4	Mild to moderate, bone fragility, normal sclerae, variable dentition, hearing loss in <10%.	AD	***COL1A1***	Collagen I structural mutations
				COL1A2	
				WNT1	Wnt cell signaling pathway
			AR	*CRTAP*	Collagen post-translational modification and folding machinery
				FKBP10/FKBP65	
				PPIB/CYPB	
				SERPINF1/PEDF	Signaling and collagen binding protein, important for mineralization
				SPARC	Collagen binding/extracellular matrix assembly
				SP7/OSX	Transcription factor, bone formation
OI with calcification of the interosseous membranes	OI type 5	Calcification of the interosseous membranes in forearm and legs and/or hypertrophic callus. Variable bone deformity, normal sclerae and dentition.	AD	*IFITM5*	Transcription factor, bone formation
Bruck syndrome type 1	BRKS1	Contractures with pterygia, fractures in infancy or early childhood, postnatal short stature, severe limb deformity, and progressive scoliosis	AR	*FKBP10*/FKBP65	Collagen folding machinery
Bruck syndrome type 2	BRKS2	As for Bruck syndrome type 1	AR	*PLOD2*/LH2	Collagen post-translational modification of lysine

Note: Predominant OI gene mutations (>90%) are in *COL1A1* and *COL1A2* (in bold typeface).

Abbreviations: AD, autosomal dominant; AR, autosomal recessive.

Incidence Type I OI has a frequency of about 1 in 15,000–20,000 births. Type II OI has a reported incidence of about 1 in 60,000. Only a limited number of patients with the severe forms of OI have been reported, and the combined incidence of the severe forms that are recognizable at birth (Types II, III, and IV) may be much higher than 1 in 60,000.

Skeletal Effects In Type I OI, the fragility of bones may be severe enough to limit physical activity or be so mild that individuals are unaware of any disability. Radiographs of the skull in patients with mild disease may show a mottled appearance because of small islands of irregular ossification. In Type II OI, ossification of many bones is frequently incomplete. Continuously beaded or broken ribs and crumpled long bones (accordina femora) may be present. For reasons that are not apparent, the long bones may be either thick or thin. In Types III and IV, multiple fractures from minor physical stress can produce severe deformities. Kyphoscoliosis can impair respiration, cause cor pulmonale, and predispose to pulmonary infections. The appearance of "popcorn-like" deposits of mineral in x-rays of the ends of long bones is an ominous sign. Progressive neurologic symptoms may result from basilar compression and communicating hydrocephalus. Type V OI is

recognized by the presence of dislocated radial heads and hyperplastic callus formation.

In all forms of OI, bone mineral density is decreased. However, the degree of osteopenia may be difficult to evaluate because recurrent fractures limit exercise and thereby diminish bone mass. Surprisingly, fractures appear to heal normally.

Ocular Features The sclerae can be normal, gray, slightly bluish, or bright blue. Blue sclerae, however, are an inherited trait in some families who do not have increased bone fragility.

Dentinogenesis The teeth may be normal, moderately discolored, or grossly abnormal. The enamel generally appears normal, but the teeth may have a characteristic amber, yellowish brown, or translucent bluish gray color because of a deficiency of dentin that is rich in Type I collagen. The deciduous teeth are usually smaller than normal, whereas permanent teeth are frequently bell-shaped and restricted at the base. In some patients, the teeth readily fracture and need to be extracted. Similar tooth defects, however, can be inherited without any evidence of OI.

Hearing Loss Hearing loss usually begins during the second decade of life and occurs in >50% of individuals aged >30. The loss can be conductive, sensorineural, or mixed, and it varies in severity. The middle ear usually exhibits maldevelopment, deficient ossification, persistence of cartilage in areas that are normally ossified, and abnormal calcium deposits.

Other Features Changes in other connective tissues can include thin skin that scars extensively, joint laxity with permanent dislocations indistinguishable from those of EDS, and occasionally, cardiovascular manifestations such as aortic regurgitation, floppy mitral valves, mitral incompetence, and fragility of large blood vessels. For unknown reasons, some patients develop bouts of a hypermetabolic state with elevated serum thyroxine levels, hyperthermia, and excessive sweating.

Molecular Defects Of the ~1720 unique gene mutations now described in OI, >86% are heterozygous mutations in either *COL1A1* and *COL1A2*, the genes coding for proα1 or proα2 chain of Type I procollagen (Table 406-2).

Most patients with Type I OI and blue sclerae have mutations that reduce the synthesis of proα1 chains to about one half. Mutations that reduce the synthesis of proα2 chains produce slightly more severe phenotypes and skin defects similar to EDS.

In contrast to the null mutations found in Type I OI, most of the severe variants (Types II, III, and IV) are caused by mutations that produce structurally abnormal proα chains that have compromised assembly or abnormal folding of the triple helix. As with collagen mutations in other connective tissue diseases, these structural mutations generally fall into two functional categories. The relatively rare mutations in the C-propeptide domain can prevent or seriously impair initial assembly of the procollagen trimers. These misfolded chains are retained in the ER and targeted for degradation by the ER-associated proteasomal pathway. Because these mutations induce an ER-stress response, the unfolded protein response (UPR) may have many downstream effects on cells. A possible role for ER stress in the pathology of OI is highlighted by the recent finding that a mutation in *CREB3L1*, the gene for OASIS, a member of the ATF6 family of ER stress sensors, produces severe recessive OI. ER stress is a new concept in the pathophysiology of connective tissue disease and has been best characterized for chondrodysplasias (CDs) (see below).

The most common Type I collagen mutations, however, are single base substitutions that introduce an amino acid with a bulky side chain for one of the glycine residues that appear as every third amino acid in the triple helix. In effect, any of the 338 glycine residues in the helical domain of either the proα1 or proα2 chain of Type I procollagen is a potential site for a disease-producing mutation. These mutations compromise the structural integrity of the triple helix, causing disruption to helix folding, retention of the mutant trimers in the ER, and increased posttranslational hydroxylation and glycosylation of lysines. Collagen-containing helix mutations can form insoluble aggregates in the ER that are degraded by the autophagosome-endosome system, rather than the proteasomes. The precise nature of any resulting ER stress response remains to be characterized.

A similar sequence of events occurs with less common mutations that produce partial gene deletions, partial gene duplications, and splicing mutations. In addition to their intracellular effects, the structurally abnormal mutant-containing collagen that is secreted by the cell can also have important extracellular affects. For example, the presence of one abnormal proα chain in a procollagen molecule can interfere with cleavage of the N-propeptide from the protein. The persistence of the N-propeptide on a fraction of the molecules interferes with the self-assembly of normal collagen so that thin and irregular collagen fibrils are formed. Furthermore, if structurally abnormal collagens are incorporated into fibrils, they may have a destabilizing effect and be selectively degraded, or they may alter the interactions of collagen with other connective tissue components, disturbing architecture and stability.

Several generalizations can be made about mutations in Type I collagen genes. One is that unrelated patients rarely have the same mutation in the same gene. Glycine substitutions in the N-terminal region of the triple helix tend to produce milder phenotypes, apparently because they have less effect on the zipper-like propagation of the triple-helical conformation from the C terminus. Rare substitutions of charged amino acids (Asp, Arg) or a branched amino acid (Val) in X- or Y-positions produce lethal phenotypes, apparently because they are located at sites for lateral assembly of the monomers or binding of other components of the matrix.

The search for mutations causing the less common and autosomal recessive forms of OI identified mutations in genes for a series of proteins that are essential for the timely folding of the procollagen monomer: cartilage-associated protein (*CRTAP*), prolyl-3-hydroxylase (*LEPRE1*/P3H1), cyclophilin B (*PPIB*), collagen chaperone-like protein HSP47 (*SERPINH1*), and the procollagen chaperone protein FKBP65 (*FKBP10*). Recently, mutations have been characterized the collagen vesicular secretion pathway (*SEC24D*) and in additional downstream components of the collagen fibrillogenesis pathway; *BMP1*, the gene coding for a metalloproteinase that cleaves the C-propeptide of Type I procollagen; *SPARC* (osteonectin), a collagen binding protein involved in extracellular matrix assembly; and *PLOD2* (LH2, lysyl oxidase 2) which is involved in establishing collagen cross-links. In addition to these mutations that affect the collagen assembly pathway, mutations have been characterized in genes involved in the regulation of bone formation and mineralization such as *SP7* (osterix), *IFITM5*, *WNT1*, and *TMEM38B* (Table 406-2).

Inheritance and Mosaicism in Germ-Line Cells and in Somatic Cells Type I OI is inherited as an autosomal dominant trait. However, some patients with Type I OI appear to represent sporadic new mutations or a diagnosis that was missed in earlier generations. Most lethal OI is the result of sporadic mutations that occur in the germ line in one of the parents. Because of the possibility for germ-line mosaicism for newly generated mutations, there is about a 7% probability that a second child could inherit a severe variant of OI.

Diagnosis OI is usually diagnosed on the basis of clinical criteria. The presence of fractures together with blue sclerae, dentinogenesis imperfecta, or family history of the disease is usually sufficient to make the diagnosis. Other causes of pathologic fractures must be excluded, including battered child syndrome, nutritional deficiencies, malignancies, and other inherited disorders such as CDs and hypophosphatasia that can have overlapping presentations. The absence of superficial bruises can be helpful in distinguishing OI from battered child syndrome. X-rays usually reveal a decrease in bone density that can be verified by photon or x-ray absorptiometry. Bone microscopy can be helpful in the diagnosis. The diagnosis, like other genetic disorders, is now routinely conducted using targeted candidate gene sequencing but whole genome sequencing is becoming increasingly common.

TREATMENT

Osteogenesis Imperfecta

Therapy should be directed toward decreasing the incidence of bone fractures, bone pain, and the restrictions on mobility. Physical therapy and occupational therapy are important. Diet should include adequate intake of calcium and vitamin D adjusted for the diminished weight of most patients. Orthopedic procedures are frequently required for deformities of long bone and scoliosis. Some surgeons recommend inserting rods into long bones. Drugs that have been developed for the therapy of osteoporosis are beneficial for some patients, but definitive data are difficult to obtain because of the small number of OI patients available for study. Bisphosphonates that inhibit osteoclasts are regarded as a mainstay of care in many medical centers for children with moderate to severe OI. Limited data are available on a series other therapies that are currently being tested: a monoclonal antibody that targets the receptor activator of nuclear factor-κB ligand (RANKL) to inhibit osteoclasts; inhibitors of cathepsin K to prevent digestion of proteins in the bone matrix; growth hormone to stimulate osteoblasts; a parathyroid analogue which stimulates osteoblasts; monoclonal antibodies to sclerostin that limit bone mass by inhibiting Wnt/β-catenin signaling in osteoblasts; and inhibitors to TGF-β that modulate osteoclast/osteoblast activity during bone remodeling. Cell therapy with infusion of mesenchymal stem/stromal cells improved a small cohort of children, but the improvements persisted for only a few months and the procedure has not been widely adopted.

EHLERS-DANLOS SYNDROME

EDS is characterized by hyper-extensible skin and hypermobile joints but the category includes rare patients with other distinctive features. Mutations in different types of collagen are found in many patients but other genes are at fault in rare forms. Contrary to initial expectations, no patients have been found with mutations in the gene for elastin in EDS.

Classification Several types of EDS have been defined, based on the extent to which the skin, joints, and other tissues are involved, mode of inheritance, and by molecular and biochemical analysis (Table 406-3). Classical EDS includes a severe form of the disease (Type I) and a milder form (Type II), both characterized by joint hypermobility and skin that is velvety in texture, hyper-extensible, and easily scarred. In hypermobile EDS (Type III), joint hypermobility is more prominent than skin changes. In vascular-type EDS (Type IV), the skin changes are more prominent than joint changes, and the patients are predisposed to sudden death from rupture of large blood vessels or other hollow organs. EDS V is similar to EDS II but is inherited as an X-linked trait. The ocular-scoliotic type of EDS (Type VI) is characterized by scoliosis, ocular fragility, and a cone-shaped deformity of the cornea (keratoconus). The arthrochalasic type of EDS (Type VII A and B) is characterized by marked joint hypermobility that is difficult to distinguish from EDS III except by the specific molecular defects in the processing of Type I procollagen to collagen. The periodontotic-type EDS (Type VIII) is distinguished by prominent periodontal changes. EDS IX, X, and XI were defined on the basis of preliminary biochemical and clinical data. EDS due to tenascin X deficiency has not been

TABLE 406-3 Different Forms of Ehlers-Danlos Syndrome

TYPE	TYPICAL FEATURES	INHERITANCE	GENE DEFECT	PROTEIN DEFECT
Classic (EDS I—severe and EDS II—mild)	Skin hyperextensibility and fragility, joint hypermobility, tissue fragility manifested by widened atrophic scarring	AD	*COL5A1* *COL5A2*	Collagen V
		AD	*COL1A1*	Proα1 (I) and proα2 (I) chains of procollagen I
		AD, AR	*COL1A2*	
Hypermobile (EDS III)	Joint hypermobility, moderate skin involvement, absence of tissue fragility	AD	*TNXB*	Tenascin X
Vascular (EDS IV)	Markedly reduced life span due to spontaneous rupture of internal organs such as arteries and intestines; skin is thin, translucent, and fragile, with extensive bruising; hypermobile minor joints; characteristic facial appearance	AD	*COL3A1*	Collagen III
Ocular-scoliotic EDS VI (EDS VIA and EDS VIB)	Features of classic EDS as well as severe muscular hypotonia after birth, progressive kyphoscoliosis, a Marfanoid habitus, osteopenia, occasionally rupture of the eye globe and great arteries	AR	*PLOD1*	Deficiency of procollagen-lysine 5-dioxygenase activity (EDS VIA)
			Unknown for EDS VIB	Unknown for EDS VIB
Arthrochalasic EDS VII (EDS VIIA and EDS VIIB)	Congenital bilateral hip dislocation, hypermobile joints, moderate skin involvement, osteopenia	AD	*COL1A1* *COL1A2*	Mutations that prevent cleavage of the N propeptides
Dermatosparactic EDS VII C	Redundant and fragile skin, prominent hernias, joint laxity, dysmorphic features	AR	*ADAMTS2*	Deficiency of procollagen I N-terminal proteinase
Periodontotic EDS VIII	Absorptive periodontosis with premature loss of permanent teeth, fragility of the skin, skin lesions	AD	C1R C1S	Components of the complement pathway
EDS due to tenascin X deficiency	Similar to EDS II	AR	*TNXB*	Tenascin X
EDS, cardiac valvular form	Similar to EDS II	AR	*COL1A2*	Type I collagen deficiency
EDS, progeroid form	Similar to EDS I-III with hair loss, hypotonia and aged appearance	AR	*B4GALT7*	Deficiency of galactosyltransferase 7 (defective synthesis of dermatan sulfate proteoglycans)
EDS, musculocontractural form	Hyperextensible and thin skin, hypermobility and contractures of hands and feet, kyphoscoliosis	AR	*CHST14* *DSE*	Dermatan 4-*O*-sulfotransferase 1 (*CHST14*) and DS epimerase 1 (*DSE*) leading to defective synthesis of dermatan sulfate proteoglycans

Abbreviations: AD, autosomal dominant; AR, autosomal recessive.

assigned a type; it is an autosomal recessive form of the syndrome similar to EDS II. The cardiac valvular form of EDS has similar features to EDS II, but also involves severe changes to the aorta. The progeroid form of EDS displays features of both EDS and progeria. Because of overlapping signs and symptoms, many patients and families with some of the features of EDS cannot be assigned to any of the defined types.

Incidence The overall incidence of EDS is about 1 in 5000 births, with a higher rate for blacks. Classical and hypermobile types of EDS are the most common. Patients with milder forms frequently do not seek medical attention.

Skin Skin changes vary from thin and velvety to skin that is either dramatically hyperextensible ("rubber person" syndrome) or easily torn or scarred. Patients with classical EDS develop characteristic "cigarette-paper" scars. In vascular-type EDS, extensive scars and hyperpigmentation develop over bony prominences, and the skin may be so thin that subcutaneous blood vessels are visible. In the periodontotic type of EDS, the skin is more fragile than hyperextensible, and it heals with atrophic, pigmented scars. Easy bruisability occurs in several types of EDS.

Ligament and Joint Changes Laxity and hypermobility of joints vary from mild to unreducible dislocations of hips and other large joints. In mild forms, patients learn to avoid dislocations by limiting physical activity. In more severe forms, surgical repair may be required. Some patients have progressive difficulty with age.

Other Features Mitral valve prolapse and hernias occur, particularly with Type I. Pes planus and mild to moderate scoliosis are common. Extreme joint laxity and repeated dislocations may lead to degenerative arthritis. In the ocular-scoliotic type of EDS, the eye may rupture with minimal trauma, and kyphoscoliosis can cause respiratory impairment. Also, sclerae may be blue.

Molecular Defects Subsets of patients with different types of EDS have mutations in the structural genes for collagens (Table 406-3). These include mutations in the *COL1A1* gene in a few patients with moderately severe classical EDS (Type I); mutations in *COL1A2* in rare patients with an aortic valvular form of EDS; mutations in two of the three genes (*COL5A1* and *COL5A2*) for Type V collagen, a minor collagen found in association with Type I collagen, in about half the patients with classical EDS (Types I and II); mutations in the *COL3A1* gene for Type III collagen that is abundant in the aorta in patients with the frequently lethal vascular EDS (Type IV).

Some of the Type I collagen-related mutations alter processing of the protein or genes for the processing enzymes. Arthrochalasic EDS (Type VII) is caused by mutations in the amino acid sequence that make Type I procollagen resistant to cleavage by procollagen N-proteinase or by mutations that decrease the activity of the enzyme. The persistence of the N-propeptide causes the formation of collagen fibrils that are thin and irregular. Some of the patients have fragile bones and therefore a phenotype that overlaps with OI. The ocular-scoliotic type of EDS (Type VI) is caused by homozygous or compound heterozygous mutations in the *PLOD 1* gene, which encodes procollagen-lysine 5-dioxygenase (lysyl hydroxylase 1), an enzyme required for formation of stable cross-links in collagen fibers.

Some patients with the hypermobile EDS (Type III) and a few with mild EDS (Type II) have mutations in the *TNXB* gene, which encodes tenascin X, another minor component of connective tissue that appears to regulate the assembly of collagen fibers. Mutations in proteoglycans have been found in a few patients. The progeroid and musculocontractural forms of EDS result from mutations in the key enzymes of proteoglycan biosynthesis: *B4GALT7*, the gene for β-1, 4-galactosyltransferase 7, *CHST14*, dermatan 4-*O*-sulfotransferase 1 and *DSE*, dermatan sulfate epimerase.

Diagnosis The diagnosis is based on clinical criteria and increasingly on DNA sequencing. Correlations between genotype and phenotype can be challenging, but gene or biochemical tests are particularly useful for the diagnosis of vascular-Type IV EDS with its dire prognosis.

As with other heritable diseases of connective tissue, there is a large degree of variability among members of the same family carrying the same mutation. Some patients have increased fractures and are difficult to distinguish from OI. A few families with heritable aortic aneurysms have mutations in the gene for Type III collagen without any evidence of EDS or OI.

TREATMENT

Ehlers-Danlos Syndrome

Patients with mild forms require little special therapy. They, or their families, frequently learn how to reset dislocated joints. In severe forms, surgical repair and tightening of joint ligaments require careful evaluation of individual patients, as the ligaments frequently do not hold sutures. Patients with easy bruisability should be evaluated for bleeding disorders. Patients with Type IV EDS and members of their families should be evaluated at regular intervals for early detection of aneurysms, but surgical repair may be difficult because of friable tissues. Also, women with Type IV EDS should be counseled about the increased risk of uterine rupture, bleeding, and other complications of pregnancy.

CHONDRODYSPLASIAS

(See also Chap. 405) CDs, also referred to as skeletal dysplasias, are heritable skeletal disorders that are characterized by dwarfism and abnormal body proportions. The category also includes some individuals with normal stature and body proportions who have features such as ocular changes or cleft palate, which are common in more severe CDs. Many patients develop degenerative joint changes; and mild CD in adults may be difficult to differentiate from primary generalized OA. An undefined number of patients have mutations in either the most abundant collagen in cartilage (Type II) or the less abundant collagens (Types X or XI). Other patients have mutations in genes that code for other components of cartilage or for proteins required for the embryonic development of cartilage, including a common mutation in a gene for a fibroblast growth factor receptor.

Classification Over 200 distinct types and subtypes have been defined based on criteria such as "bringing death" (thanatophoric), causing "twisted" bones (diastrophic), affecting metaphyses (metaphyseal), affecting epiphyses (epiphyseal), affecting spine (spondylo-), and producing histologic changes such as an apparent increase in the fibrous material in the epiphyses (fibrochondrogenesis). Also, a number of eponyms are based on the first or most comprehensive case reports. Severe forms of the diseases produce dwarfism with gross distortions of most cartilaginous structures and of other structures including the eye. Mild forms are more difficult to classify. Among the features are cataracts, degeneration of the vitreous, and retinal detachment, high forehead, hypoplastic facies, cleft palate, short extremities and gross distortions of the epiphyses, metaphyses, and joint surfaces. Patients with Stickler syndrome (hereditary arthro-ophthalmopathy) have been classified into three types based on a combination of the ocular phenotype and mutated genes.

Incidence The overall incidence of all forms of CD ranges from 1 per 2500 to 1 per 4000 births. Data on the frequency of individual CDs are incomplete, but the incidence of the Stickler syndrome is 1 in 10,000. Therefore, the disease is probably among the more common heritable disorders of connective tissue.

Molecular Defects Mutations in the *COL2A1* gene for the Type II collagen of cartilage are found in a fraction of patients with both mild and severe CDs. For example, a mutation in the gene substituting a cysteine residue for an arginine was found in three unrelated families with spondyloepiphyseal dysplasia (SED) and precocious generalized OA. Mutations in the gene were also found in some lethal CDs characterized by gross deformities of bones and cartilage, such as those found in SED congenita, spondyloepimetaphyseal dysplasia congenita, hypochondrogenesis/achondrogenesis Type II,

and Kniest syndrome. The highest incidence of *COL2A1* mutations, however, occurs in patients with the distinctive features of the Stickler syndrome, which is characterized by skeletal changes, orofacial abnormalities, and auditory abnormalities. Most of the mutations in *COL2A1* are premature stop codons that produce haploinsufficiency. In addition, some of the patients with the Stickler syndrome or a closely related syndrome have mutations in two genes specific for Type XI collagen, which is an unusual heterotrimer formed from α chains encoded by the gene for Type II collagen (*COL2A1*) and two distinctive genes for Type XI collagen (*COL11A1* and *COL11A2*). Mutations in the *COL11A1* gene are also found in patients with Marshall syndrome, which is similar to classic Stickler syndrome, but with more severe hearing loss and dysmorphic features, such as a flat or retracted midface with a flat nasal bridge, short nose, anteverted nostrils, long philtrum, and large-appearing eyes.

CDs are also caused by mutations in the less abundant collagens found in cartilage. For example, patients with Schmid metaphyseal CD have mutations in the gene for Type X collagen, a short, network-forming collagen found in the hypertrophic zone of endochondral cartilage. The syndrome is characterized by short stature, *coxa vara*, flaring metaphyses, and waddling gait. As with other collagen genes, the most common mutations are of two types: Nonsense mutations that lead to haploinsufficiency and structural mutations that compromise collagen assembly. In Type X collagen all the structural mutations detected occur in the C-terminal NC1 domain that coordinates the formation of the trimers. This NC1 domain is functionally equivalent to the C-propeptide of the fibrillar collagens. These mutations disturb the structure of the NC1 domain, leading to misfolding and initiation of cellular ER stress via the UPR. While the UPR evolved to allow cells to adjust their ER folding capacity to differing protein folding loads, it is deployed by cells when mutant misfolded proteins accumulate in the ER. Activation of the UPR attenuates protein translation and activates mutant protein degradation pathways such as ER-associated degradation. If these strategies do not sufficiently reduce the stress response, cell death may occur. In Schmid metaphyseal CD, mutant misfolded Type X collagen induces the UPR, resulting in downstream consequences that contribute to the pathophysiology. This general mechanism may also contribute to pathology in other chondrodyplasias (and in other connective tissues disorders) where gene mutations lead to protein structural abnormalities.

Some patients have mutations in genes for proteins that interact with collagens. Patients with pseudoachondroplasia or autosomal dominant multiple epiphyseal dysplasia have mutations in the gene for the cartilage oligomeric matrix protein (*COMP*), a protein that interacts with both collagens and proteoglycans in cartilage. However, some families with multiple epiphyseal dysplasia have a defect in one of the three genes for Type IX collagen (*COL9A1*, *COL9A2*, and *COL9A3*) or in matrilin-3, another extracellular protein found in cartilage. With misfolding mutations in COMP and matrilin-3, the activation of the UPR has been described, providing further evidence that the UPR is a component of pathology of these conditions.

Some CDs are caused by mutations in genes that affect early development of cartilage and related structures. The most common form of short-limbed dwarfism, achondroplasia, is caused by mutations in the gene for a receptor for a fibroblastic growth factor (*FGFR3*). The mutations in the *FGFR3* gene causing achondroplasias are unusual in several respects. The same single-base mutation in the gene that converts glycine to arginine at position 380 in the *FGFR3* gene is present in >90% of patients. Most patients harbor sporadic new mutations, and therefore this nucleotide change must be one of the most common recurring mutations in the human genome. The mutation causes unregulated signal transduction through the receptor and inappropriate development of cartilage. Mutations that alter other domains of *FGFR3* have been found in patients with the more severe disorders of hypochondroplasia and thanatophoric dysplasia and in a few families with a variant of craniosynostosis. However, most patients with craniosynostosis appear to have mutations in the related *FGFR2* gene. The similarities between the phenotypes produced by mutations in genes for FGF receptors and mutations in structural proteins of cartilage are probably explained by the observation that the activity of FGFs is regulated in part by binding of FGFs to proteins sequestered in the extracellular matrix. Therefore, the situation parallels the interactions between transforming growth factors (TGFs) and fibrillin in MFS (see below).

Other mutations involve the proteoglycans of cartilage, aggrecan (*AGC1*) and perlecan (*HSPG2*), and in the proteoglycan posttranslational sulphation pathway (*DTDST*, *PAPSS2* and *CHST3*). Mutations in >45 other genes have been defined in CDs.

Diagnosis The diagnosis of CDs is made on the basis of the physical appearance, slit-lamp eye examinations, x-ray findings, histologic changes, and clinical course. Targeted gene and exome sequencing or more global sequencing strategies are used for molecular diagnosis. Given the wide spectrum of CD phenotypes, these genes tests are becoming critical diagnostic tools. For Stickler syndrome, more precise diagnostic criteria have made it possible to identify Type I variants with mutations in the *COL2A1* gene with a high degree of accuracy. It has been suggested that the Type II variant with mutations in the *COL11A1* gene can be identified on the basis of a "beaded" vitreous phenotype, and the Type III variant with mutations in the *COL11A2* gene can be identified on the basis of the characteristic systemic features without the ocular involvement. Prenatal diagnosis based on analysis of DNA obtained from chorionic villus or amniotic fluid is possible.

TREATMENT

Chondrodysplasias

The treatment is symptomatic and is directed to secondary features such as degenerative arthritis. Many patients require joint replacement surgery and corrective surgery for cleft palate. The eyes should be monitored carefully for the development of cataracts and the need for laser therapy to prevent retinal detachment. In general, patients should be advised to avoid obesity and contact sports. Counseling for the psychological problems of short stature is critical. Several clinical trials therapeutically targeting the *FGFR3* pathway in achondroplasia are underway.

MARFAN SYNDROME (MFS)

MFS includes features that primarily affect the skeleton, the cardiovascular system, and the eyes. Most patients have mutations in the gene for fibrillin-1 (*FBN1*).

Classification MFS was initially characterized by a triad of features: (1) skeletal changes that include long, thin extremities, frequently associated with loose joints; (2) reduced vision as the result of dislocations of the lenses (ectopia lentis); and (3) aortic aneurysms. An international panel has developed a series of revised "Ghent criteria" that are useful in classifying patients.

Incidence and Inheritance The incidence of MFS is among the highest of any heritable disorder: about 1 in 3000/5000 births in most racial and ethnic groups. The related syndromes are less common. Mutations are generally inherited as autosomal dominant traits, but about one-fourth of patients have sporadic new mutations.

Skeletal Effects Patients have long limbs and are usually tall compared to other members of the same family. The ratio of the upper segment (top of the head to the top of the pubic ramus) to the lower segment (top of the pubic ramus to the floor) is usually 2 SDs below mean for age, race, and sex. The fingers and hands are long and slender and have a spider-like appearance (arachnodactyly). Many patients have severe chest deformities, including depression (pectus excavatum), protrusion (pectus carinatum), or asymmetry. Scoliosis is frequent and usually accompanied by kyphosis. High-arched palate and high pedal arches or pes cavus are common. A few patients have severe joint hypermobility similar to EDS. CT or MRI examinations of the lumbar sacral region frequently reveals enlargement of the neural canal, thinning of the pedicles and laminae, widening of the foramins, or anterior meningocele (dural ectasia).

Cardiovascular Features Cardiovascular abnormalities are the major source of morbidity and mortality (Chap. 274). Mitral valve prolapse develops early in life and progresses to mitral valve regurgitation of increasing severity in about one-quarter of patients. Dilation of the root of the aorta and the sinuses of Valsalva are characteristic and ominous features of the disease that can develop at any age. The rate of dilation is unpredictable, but it can lead to aortic regurgitation, dissection of the aorta, and rupture. Dilation is probably accelerated by physical and emotional stress, as well as by pregnancy. Patients usually differ from patients with familial aortic aneurysms who tend to develop aneurysms in the abdominal aorta. The location of the aneurysms, however, is somewhat variable, and the high incidence of aortic aneurysms in the general population (1 in 100) makes the differential diagnosis difficult unless other features of MFS are clearly present.

Ocular Features Upward displacement of the lens is common. It is usually not progressive but may contribute to the formation of cataracts. The ocular globe is frequently elongated, and most patients are myopic, but with adequate vision. Retinal detachment can occur.

Other Features Striae may occur over the shoulders and buttocks. A number of patients develop spontaneous pneumothorax. Inguinal and incisional hernias are common. Patients are typically thin with little subcutaneous fat, but adults may develop centripetal obesity.

Molecular Defects More than 90% of patients clinically classified as having MFS by the "Ghent criteria" have a mutation in the gene for *FBN1*. Mutations in the same gene are found in a few patients who do not meet the Ghent criteria. Also, a few MFS patients without mutations in the *FBN1* gene have mutations in the gene for TGF-β receptor 2 (*TGFBR2*). In addition, mutations in either *TGFBR2* or *TGFBR1* are found in the related Loeys-Dietz syndrome which is characterized by aortic aneurysms, cleft palate, and hypertelorism. Mutations in the *FBN2* gene, which is structurally similar to the *FBN1* gene, are found in patients with MFS-like syndrome of congenital contractual arachnodactyly.

FBN1 gene mutations are scattered throughout its 65 coding exons. Most are private mutations, but ~10% are recurrent new mutations that are largely located in CpG sequences known to be "hot spots." Most severe mutations are located in the central codons (24–32). About one-third of the mutations introduce premature termination codons, and about two-thirds are missense mutations that alter calcium-binding domains in the repetitive epidermal growth factor–like domains of the protein. Rarer mutations alter the processing of the protein. As in many genetic diseases, the severity of the phenotype cannot be predicted from the nature of the mutation.

The discovery that syndromes similar to MFS are caused by mutations in *TGFBR1* and *TGFBR2* refocused attention on structural similarity between *FBN1* and TGF-β binding proteins that sequester TGF-β in the extracellular matrix. As a result, some of the manifestations of MFS have been shown to arise from alterations in binding sites that modulate TGF-β bioavailability during development of the skeleton and other tissues. Likewise, *TGFBR1* and *TGFBR2* mutations in Loeys-Dietz syndrome alter TGF-β signaling. In both MFS and Loey-Dietz syndrome, the pathogenic mechanisms involve increased TGF-β signaling which contributes to aneurysm formation.

Diagnosis All patients with a suspected diagnosis of MFS should have a slit-lamp examination and an echocardiogram. Also, homocystinuria should be ruled out by amino acid analysis of plasma (Chap. 413). The diagnosis of MFS according to the international Ghent standards places emphasis on major criteria that include presence of at least four skeletal abnormalities: ectopia lentis; dilation of the ascending aorta with or without dissection; dural ectasia; and a blood relative who meets the same criteria, with or without a DNA diagnosis. A final diagnosis is based on a balanced assessment of the major criteria together with several minor criteria. The absence of ocular changes suggests the Loeys-Dietz syndrome, and the presence of contractures with some of the signs of OI suggests congential contractual arachnodactyly.

Diagnostic tests based on gene sequencing or detection of protein defects are available. These results are unlikely to alter the treatment or prognosis, but are helpful to inform the patients and families and to rapidly exclude the diagnosis in unaffected family members.

TREATMENT

Marfan Syndrome

Patients should be advised that the risks are increased by severe physical exertion, emotional stress, and pregnancy. Surgical correction of the aorta, aortic valve, and mitral valve has been successful in many patients, but tissues are frequently friable. The scoliosis tends to be progressive and should be treated by mechanical bracing and physical therapy if >20° or by surgery if it progresses to >45°. Dislocated lenses rarely require surgical removal, but patients should be followed closely for retinal detachment.

Propranolol or other β-adrenergic blocking agents are used to lower blood pressure and thereby delay or prevent aortic dilation. The finding that MFS pathophysiology involves alterations in TGF-β signaling has raised the possibility of new therapeutic strategies. Attenuation of TGF-β signaling with agents such as angiotensin II receptor blockers (e.g., Losartan) was effective reducing aortic enlargement in animal studies and clinical trials are still in progress.

ELASTIN-RELATED DISEASES

Mutations in the elastin gene (*ELN*) have been found in patients with supravalvular aortic stenosis and skin that hangs in loose and redundant folds (cutis laxa). As indicated in Table 406-3, patients with several forms of EDS have similar changes in skin that were initially thought to reflect changes in elastin.

EPIDERMOLYSIS BULLOSA (EB)

EB has been defined as the category of heritable disorders involving skin that is specifically characterized by blistering as a result of friction. Using this criterion, it was possible to define subtypes by the ultrastructural layer of skin in which the cleavage and blistering occurred. These functional and anatomical criteria made it possible to establish that most patients with a specific subtype have mutations in genes coding for a structural protein or a cell adherence protein expressed in the corresponding layer of skin.

Classification and Incidence The four major types of EB are: (1) EB simplex in which cleavage occurs within the epidermis, (2) junctional EB in which cleavage occurs within the lamina lucida, (3) dystrophic EB in which cleavage occurs within the sub-lamina densa, and (4) Kindler syndrome with a mixed level of cleavage in different layers. Patients are then separated into major and minor subtypes based on clinical features and analysis of mutations.

The incidence of EB in the United States is about 1 in 50,000.

Molecular Defects The distinctive anatomic locations in skin have made it possible to relate the clinical subtypes of EB to mutations for specific components. In EB simplex, mutations are found primarily in the genes for the major keratins of basal epithelial cells (keratins 5 and 14), and the cell adhesion proteins plectin, plakophilin-1, desmoplakin, and dystonin. Mutations in exophilin-5 and transglutaminase 5, both of which impact keratin filaments, have been reported. Patients with the related syndrome, epidermolytic ichthyosis, have mutations in keratin 1 and keratin 10. In junctional EB, mutations occur in Type XVII collagen, a laminin (laminin-332) and α6β4 integrin. In the severe syndrome of dystrophic EB, mutations are found in the gene that codes for Type VII collagen that forms long loops anchoring the epidermis to the dermis. Patients with more complex features of what is classified as the Kindler syndrome have mutations in Kindlin-1, a. focal adhesion protein involved in integrin activation.

Diagnosis and Treatment The diagnosis is based on skin that readily breaks and forms blisters from minor trauma. EB simplex is generally milder than junctional EB or dystrophic EB. Dystrophic EB variants usually have large and prominent scars. Precise classification

within subtypes usually requires immunofluorescent mapping. DNA diagnostic tests have been developed as research tools but are not readily available. The treatment is symptomatic. Novel therapeutic approaches such as gene therapy, protein replacement therapy and cell therapy are being explored.

ALPORT SYNDROME

AS is an inherited disorder characterized by hematuria and several associated features. It was not initially considered as a disorder of connective tissue. However, the search for mutations in the genes coding for collagens found that most patients had mutations in collagen found in basement membranes (Type IV). Four forms of the AS are now recognized: (1) classic AS, which is inherited as an X-linked disorder with hematuria, sensorineural deafness, and conical deformation of the anterior surface of the lens (lenticonus); (2) an X-linked form associated with diffuse leiomyomatosis; (3) an autosomal recessive form; and (4) an autosomal dominant form. Both autosomal recessive and dominant forms can cause renal disease without deafness or lenticonus.

Incidence The incidence of AS is about 1 in 10,000 births in the general population and as high as 1 in 5000 in some ethnic groups. About 80% of AS patients have the classical X-linked variant.

Molecular Defects Most patients have mutations in four of the six genes for the chains of Type IV collagen (*COL4A3*, *COL4A4*, *COL4A5*, and *COL4A6*). The genes for the proteins are arranged in tandem pairs on different chromosomes in an unusual head-to-head orientation and with overlapping promoters; i.e., the *COL4A1* and *COL4A2* genes are head-to-head on chromosome 13q34, the *COL4A3* and *COL4A4* genes are on chromosome 2q35–37, and the *COL4A5* and *COL4A6* genes are on chromosome Xq22. The X-linked variants are caused by either mutations in the *COL4A5* gene or by partial deletions of both of the adjacent *COL4A4* and *COL4A5* genes. The autosomal recessive variants are caused by mutations in either the *COL4A3* or *COL4A4* genes. The mutations responsible for the autosomal dominant variants are still unknown, but they have been mapped to the same locus as the *COL4A3* and *COL4A4* genes.

Diagnosis and Treatment The diagnosis of classic AS is based on X-linked inheritance of hematuria, sensorineural deafness, and lenticonus. The lenticonus together with hematuria is pathognomonic of classic AS. The sensorineural deafness is primarily in the high-tone range. It can frequently be detected only by an audiogram and is usually not progressive. Because of the X-linked transmission, women are generally underdiagnosed and are usually less severely affected than men. The hematuria usually progresses to nephritis and may cause renal failure in late adolescence in affected males and at older ages in some women. Renal transplantation is usually successful.

FURTHER READING

Bateman JF et al: Genetic diseases of connective tissues: Cellular and extracellular effects of ECM mutations. Nat Rev Genet 10:173, 2009.

Bonafe L et al: Nosology and classification of genetic skeletal disorders: 2015 revision. Am J Med Genet A 167A:2869, 2015.

Forlino A, Marini JC: Osteogenesis imperfecta. Lancet 387:1657, 2016.

Hynes RO: The evolution of metazoan extracellular matrix. J Cell Biol 196:671, 2012.

Loeys BL et al: The revised Ghent nosology for the Marfan syndrome. J Med Genet 47:476, 2010.

Malfait F, de Paepe A: The Ehlers-Danlos syndrome. Adv Exp Med Biol 802:129, 2014.

Marom R et al: Pharmacological and biological therapeutic strategies for osteogenesis imperfecta. Am J Med Genet C Semin Med Genet 172:367, 2016.

Prockop DJ, Kivirikko KI: Collagens: molecular biology, diseases, and potentials for therapy. Annu Rev Biochem 64:403–434, 1995. Review.

Ramirez F, Dietz HC: Marfan's syndrome: From molecular pathogenesis to clinical treatment. Curr Opin Genet Dev 17:252, 2007.

Ramirez F, Sakai LY: Biogenesis and function of fibrillin assemblies. Cell Tissue Res 339:71, 2010.

Shapiro JR et al (eds): *Osteogenesis Imperfecta. A Translational Approach to Brittle Bone Disease*. London, Academic Press, 2014.

Uitto J et al: Molecular pathology of the basement membrane zone in heritable blistering diseases: The paradigm of epidermolysis bullosa. Matrix Biology 57-58:76, 2017.

407 Hemochromatosis

Lawrie W. Powell

DEFINITION

Hemochromatosis results from a relatively common inherited genetic mutation in European populations. Once thought to be a single disease entity it is now known to be an iron-storage disorder with genetic heterogeneity but with a final common metabolic pathway resulting in inappropriately low production of the hormone hepcidin. This leads to an increase in intestinal iron absorption and the deposition of excessive amounts of iron in parenchymal cells with eventual tissue damage and organ failure. Thus, the term *hemochromatosis* now refers to a group of genetic diseases that predispose to iron overload, potentially leading to fibrosis and organ failure. Cirrhosis of the liver, diabetes mellitus, arthritis, cardiomyopathy, and hypogonadotropic hypogonadism are the major clinical manifestations.

The following terminology is widely accepted.

1. *Hereditary hemochromatosis* is most often caused by a mutant gene, termed *HFE*, which is tightly linked to the HLA-A locus on chromosome 6p. Persons who are homozygous for the mutation are at increased risk of iron overload and account for 80–90% of clinical hereditary hemochromatosis in persons of northern European descent. In such subjects, the presence of hepatic fibrosis, cirrhosis, arthropathy, or hepatocellular carcinoma constitutes iron overload–related disease. Rarer forms of non-*HFE* hemochromatosis are caused by mutations in other genes involved in iron metabolism **(Table 407-1)**. The disease can be recognized during its early stages when iron overload and organ damage are minimal. At this stage,

TABLE 407-1 Classification of Iron Overload States

Hereditary Hemochromatosis
Hemochromatosis, *HFE*-related (type 1)
C282Y homozygosity
C282Y/H63D compound heterozygosity
Hemochromatosis, non-*HFE*-related
Juvenile hemochromatosis (type 2A) (hemojuvelin mutations)
Juvenile hemochromatosis (type 2B) (hepcidin mutation)
Mutated transferrin receptor 2, *TFR2* (type 3)
Mutated ferroportin 1 gene, *SLC11A3* (type 4)

Acquired Iron Overload	
Iron-loading anemias	Chronic liver disease
Thalassemia major	Hepatitis C
Sideroblastic anemia	Alcoholic cirrhosis, especially when advanced
Chronic hemolytic anemias	Nonalcoholic steatohepatitis
Transfusional and parenteral iron overload	Porphyria cutanea tarda
Dietary iron overload	Dysmetabolic iron overload syndrome
	Post-portacaval shunting

Miscellaneous
Iron overload in sub-Saharan Africa
Neonatal iron overload
Aceruloplasminemia
Congenital atransferrinemia

the disease is best referred to as *early hemochromatosis* or *precirrhotic hemochromatosis*.

2. *Secondary iron overload* occurs as a result of an iron-loading anemia, such as thalassemia or sideroblastic anemia, in which erythropoiesis is increased but ineffective. In the acquired iron-loading disorders, massive iron deposits in parenchymal tissues can lead to the same clinical and pathologic features as in hemochromatosis.

PREVALENCE

Although *HFE*-associated hemochromatosis mutations are common, the prevalence varies in different ethnic groups. It is most common in populations of northern European extraction in whom ~1 in 10 persons are heterozygous carriers and 0.3–0.5% are homozygotes. However, expression of the disease is variable and modified by several factors, especially alcohol consumption, dietary iron intake, blood loss associated with menstruation and pregnancy, and blood donation. Recent population studies indicate that ~30% of homozygous men develop iron overload–related disease and about 6% develop hepatic cirrhosis; for women, the figure is closer to 1%. Presumably there are as yet unidentified modifying genes responsible for expression. Nearly 70% of untreated patients develop the first symptoms between ages 40 and 60. The disease is rarely evident before age 20, although with family screening (see "Screening for Hemochromatosis," below) and periodic health examinations, asymptomatic subjects with iron overload can be identified, including young menstruating women.

In contrast to *HFE*-associated hemochromatosis, the non-*HFE*-associated forms of hemochromatosis (Table 407-1) are rare, but they affect all races and young people (juvenile hemochromatosis).

These result from mutations in one or more of the genes for proteins in the hepcidin pathway **(Fig. 407-1)**, i.e., hemojuvelin, transferrin receptor 2 (TfR2), or ferroportin. The resultant clinical disease is very similar to HFE-related disease because they all lead to hepcidin deficiency, which is the final common pathway (Fig. 407-1).

A rare autosomal dominant form of hemochromatosis results from two types of mutations in the gene for the iron transporter ferroportin. Firstly, loss of function mutations decrease the cell surface localization of ferroportin, thereby reducing its ability to export iron ("ferroportin disease"). A second mutation abolishes the hepcidin-induced ferroportin internalization and degradation resulting in a "gain-of-function." Here the tissue iron distribution is similar to that in HFE-related disease (e.g., in parenchymal cells).

GENETIC BASIS

A homozygous G to A mutation in the *HFE* gene resulting in a cysteine to tyrosine substitution at position 282 (C282Y) is the most common mutation. It is identified in 85–90% of patients with hereditary hemochromatosis in populations of northern European descent but is found in only 60% of cases from Mediterranean populations. A second, relatively common *HFE* mutation (H63D) results in a substitution of histidine to aspartic acid at codon 63. Homozygosity for H63D is not associated with clinically significant iron overload. Some compound heterozygotes (e.g., one copy each of C282Y and H63D) have mild to moderately increased body-iron stores but develop clinical disease only in association with cofactors such as heavy alcohol intake or hepatic steatosis. Thus, *HFE*-associated hemochromatosis is inherited as an autosomal recessive trait; heterozygotes have no, or minimal, increase in iron stores. However, this slight increase in hepatic iron can act as a cofactor that may modify the expression of other diseases such as porphyria cutanea tarda (PCT) or nonalcoholic steatohepatitis (NASH).

Thus, mutations in other genes involved in iron metabolism are responsible for non-*HFE*-associated hemochromatosis, including

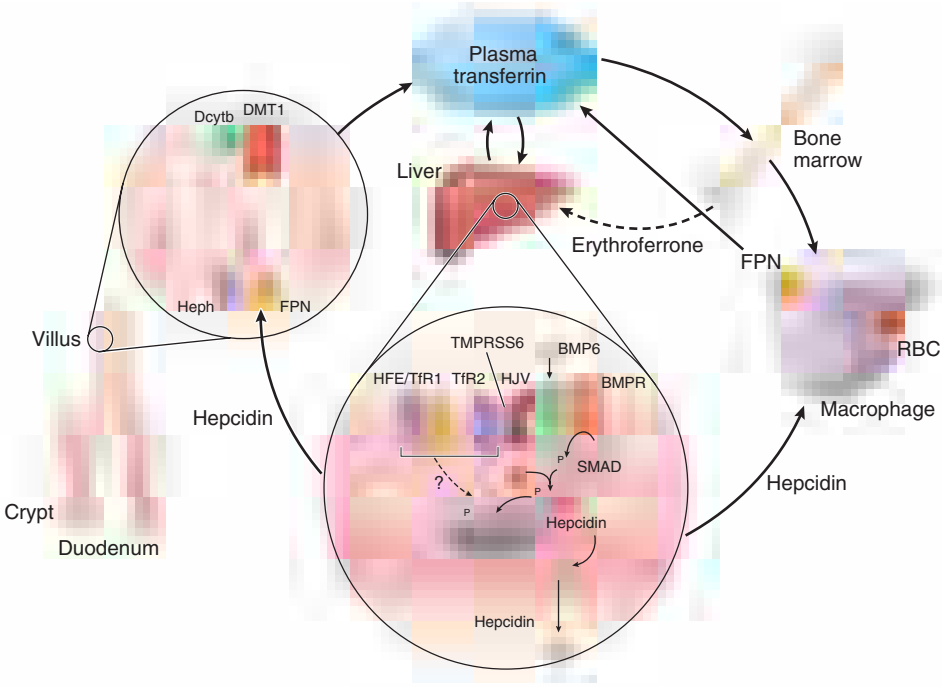


FIGURE 407-1 Pathways of normal iron homeostasis. Dietary inorganic iron traverses the brush border membrane of duodenal enterocytes via the divalent metal-ion transporter 1 (DMT1) after reduction of ferric (Fe^{3+}) iron to the ferrous (Fe^{2+}) state by duodenal cytochrome B (DcytB). Iron then moves from the enterocyte to the circulation via a process requiring the basolateral iron exporter ferroportin (FPN) and the iron oxidase hephaestin (Heph). In the circulation, iron binds to plasma transferrin and is thereby distributed to sites of iron utilization and storage. Much of the diferric transferrin supplies iron to immature erythrocyte cells in the bone marrow for hemoglobin synthesis. At the end of their life, senescent red blood cells (RBCs) are phagocytosed by macrophages, and iron is returned to the circulation after export through ferroportin. The liver-derived peptide hepcidin represses basolateral iron transport in the gut as well as iron released from macrophages and other cells and serves as a central regulator of body-iron traffic. At least three separate signals regulate hepcidin production in response to changes in body-iron requirements. The first involves the detection of circulating diferric transferrin by HFE and TfR2. A second relies on hepatic iron stores activating the hemojuvelin (HJV)-dependent bone morphogenetic protein (BMP)/SMAD pathway. The third involves signaling molecules released from erythroid precursor cells and there is strong evidence that erythroferrone fulfills this role. TMPRSS6 is a protease that modulates HJV activity. Heme is metabolized by heme oxygenase within the enterocytes, and the released iron then follows the same pathway. Mutations in the genes encoding HFE, TfR2, hemojuvelin, and hepcidin all lead to decreased hepcidin release and increased iron absorption, resulting in hemochromatosis (Table 407-1).

juvenile hemochromatosis, which affects persons in the second and third decades of life (Table 407-1). Mutations in the genes encoding hepcidin, transferrin receptor 2 (TfR2), and hemojuvelin (Fig. 407-1) result in clinicopathologic features that are indistinguishable from HFE-associated hemochromatosis. However, mutations in ferroportin, responsible for the efflux of iron from enterocytes and most other cell types, result in iron loading of reticuloendothelial cells and macrophages as well as parenchymal cells.

PATHOPHYSIOLOGY AND THE ROLE OF HEPCIDIN

Normally, the body-iron content of 3–4 g is maintained such that intestinal mucosal absorption of iron is equal to iron loss. This amount is ~1 mg/d in men and 1.5 mg/d in menstruating women. In hemochromatosis, mucosal absorption is greater than body requirements and amounts to ≥4 mg/d. The progressive accumulation of iron increases plasma iron and saturation of transferrin and results in a progressive increase of plasma ferritin **(Fig. 407-2)**. The key regulatory hormone that allows the liver to communicate with the bone marrow was discovered quite serendipitously and has transformed our understanding of the coordination of absorption, mobilization, and storage of iron to meet the requirements of erythropoiesis. It was called hepcidin based upon its anti-bacterial activity ("**HEPatic** bacterio**CID**al prote**IN**"). This liver-derived peptide represses basolateral iron transport in the intestine and iron release from macrophages and other cells by binding to ferroportin. Hepcidin, in turn, responds to signals in the liver mediated by HFE, TfR2, and hemojuvelin (Fig. 407-1). The development of minihepcidins, i.e., small peptides that mimic the action of hepcidin, is promising for the development of new therapeutic approaches for iron overload disorders caused by low hepcidin levels.

The *HFE* gene encodes a 343-amino-acid protein that is structurally related to MHC class I proteins (HFE). The basic defect in *HFE*-associated hemochromatosis is a lack of cell surface expression of HFE (due to the C282Y mutation). The normal (wild-type) HFE protein forms a complex with β_2-microglobulin and transferrin receptor 1 (TfR1). The C282Y mutation completely abrogates this interaction. As a result, the mutant HFE protein remains trapped intracellularly, reducing TfR1-mediated iron uptake by the intestinal crypt cell. This impaired TfR1-mediated iron uptake leads to upregulation of the divalent metal transporter (DMT1) on the brush border of the villus cells, causing inappropriately increased intestinal iron absorption (Fig. 407-1). In advanced disease, the body may contain 20 g or more of iron that is deposited mainly in parenchymal cells of the liver, pancreas, and heart. Iron deposition in the pituitary causes hypogonadotropic hypogonadism in both men and women. Tissue injury may result from disruption of iron-laden lysosomes, from lipid peroxidation of subcellular organelles by excess iron, or from stimulation of collagen synthesis by activated stellate cells.

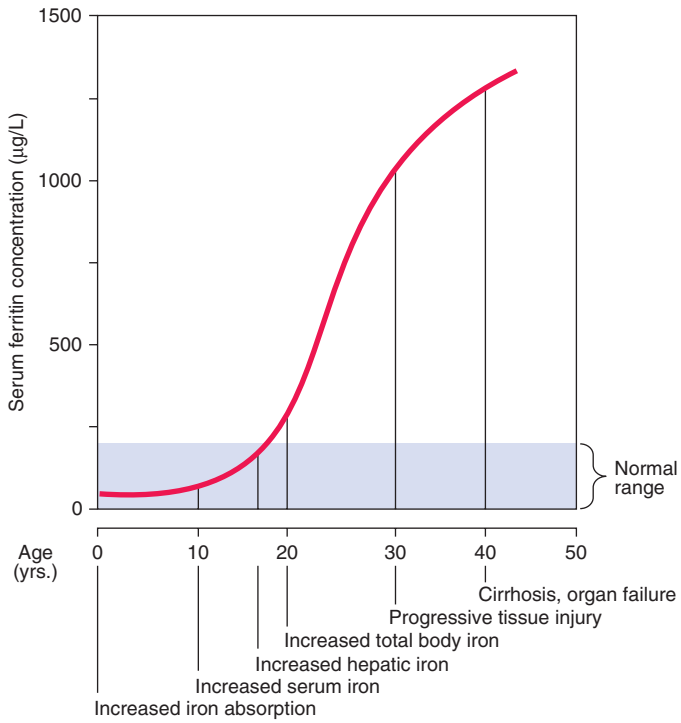


FIGURE 407-2 Sequence of events in genetic hemochromatosis and their correlation with the serum ferritin concentration. Increased iron absorption is present throughout life. Overt, symptomatic disease usually develops between ages 40 and 60, but latent disease can be detected long before this.

Secondary iron overload with deposition in parenchymal cells occurs in chronic disorders of erythropoiesis, particularly in those due to defects in hemoglobin synthesis or ineffective erythropoiesis such as sideroblastic anemia and thalassemia **(Chap. 94)**. In these disorders, iron absorption is increased. Moreover, these patients require blood transfusions and are frequently treated inappropriately with iron. PCT, a disorder characterized by a defect in porphyrin biosynthesis **(Chap. 409)**, can also be associated with excessive parenchymal iron deposits. The magnitude of the iron load in PCT is usually insufficient to produce tissue damage. However, some patients with PCT also have mutations in the *HFE* gene, and some have associated hepatitis C virus (HCV) infection. Although the relationship between these disorders remains to be clarified, iron overload accentuates the inherited enzyme deficiency in PCT and should be avoided along with other agents (alcohol, estrogens, haloaromatic compounds) that may exacerbate PCT. Another cause of hepatic parenchymal iron overload is hereditary aceruloplasminemia. In this disorder, impairment of iron mobilization due to deficiency of ceruloplasmin (a ferroxidase) causes iron overload in hepatocytes.

Excessive iron ingestion over many years rarely results in hemochromatosis. An important exception has been reported in South Africa among groups who brew fermented beverages in vessels made of iron (see later). Hemochromatosis has been described in apparently normal persons who have taken medicinal iron over many years, but such individuals probably had genetic disorders.

The common denominator in all patients with hemochromatosis is *excessive amounts of iron in parenchymal tissues*. Parenteral administration of iron in the form of blood transfusions or iron preparations results predominantly in reticuloendothelial cell iron overload. This appears to lead to less tissue damage than iron loading of parenchymal cells.

In the liver, parenchymal iron is in the form of ferritin and hemosiderin. In the early stages, these deposits are seen in the periportal parenchymal cells, especially within lysosomes in the pericanalicular cytoplasm of the hepatocytes. This stage progresses to perilobular fibrosis and to fibrous septa due to activation of stellate cells. In the advanced stage, a macronodular or mixed macro- and micronodular cirrhosis develops. Hepatic fibrosis and cirrhosis correlate significantly with hepatic iron concentration.

Histologically, iron is increased in many organs, particularly in the liver, heart, and pancreas, and, to a lesser extent, in the endocrine glands. The epidermis of the skin is thin, and melanin is increased in the cells of the basal layer and dermis. Deposits of iron are present around the synovial lining cells of the joints.

CLINICAL MANIFESTATIONS

C282Y homozygotes can be characterized by the stage of progression as follows: (1) a genetic predisposition without abnormalities; (2) iron overload without symptoms; (3) iron overload with symptoms (e.g., arthritis and fatigue); and (4) iron overload with organ damage—in particular, cirrhosis. Thus, many subjects with significant iron overload are asymptomatic. For example, in a study of 672 asymptomatic C282Y homozygous subjects—identified by either family screening or routine health examinations—there was hepatic iron overload (grades 2–4) in 56% and 34.5% of male and female subjects, respectively; hepatic fibrosis (stages 2–4) in 18.4% and 5.4%, respectively; and cirrhosis in 5.6% and 1.9%, respectively.

Initial symptoms are often nonspecific and include lethargy, arthralgia, skin pigmentation, loss of libido, and features of diabetes mellitus. Hepatomegaly, increased pigmentation, spider angiomas, splenomegaly, arthropathy, ascites, cardiac arrhythmias, congestive heart failure,

loss of body hair, testicular atrophy, and jaundice are prominent in advanced disease.

The *liver* is usually the first organ to be affected, and hepatomegaly is present in >95% of symptomatic patients.

Manifestations of portal hypertension and esophageal varices occur less commonly than in cirrhosis from other causes. Hepatocellular carcinoma develops in ~30% of patients with cirrhosis, and it is the most common cause of death in treated patients—hence the importance of early diagnosis and therapy. The incidence increases with age, it is more common in men, and it occurs almost exclusively in cirrhotic patients.

Excessive skin pigmentation is present in patients with advanced disease. The characteristic metallic or slate-gray hue is sometimes referred to as *bronzing* and results from increased melanin and iron in the dermis. Pigmentation usually is diffuse and generalized.

Diabetes mellitus occurs in ~65% of patients with advanced disease and is more likely to develop in those with a family history of diabetes, suggesting that direct damage to the pancreatic islets by iron deposition occurs in combination with other risk factors. The management is similar to that of other forms of diabetes.

Arthropathy develops in 25–50% of symptomatic patients. It usually occurs after age 50 but may occur as a first manifestation or long after therapy. The joints of the hands, especially the second and third metacarpophalangeal joints, are usually the first joints involved, a feature that helps to distinguish the chondrocalcinosis associated with hemochromatosis from the idiopathic form **(Chap. 365)**. A progressive polyarthritis involving wrists, hips, ankles, and knees may also ensue. Acute brief attacks of synovitis may be associated with deposition of calcium pyrophosphate (chondrocalcinosis or pseudogout), mainly in the knees. Radiologic manifestations include cystic changes of the subchondral bones, loss of articular cartilage with narrowing of the joint space, diffuse demineralization, hypertrophic bone proliferation, and calcification of the synovium. The arthropathy tends to progress despite removal of iron by phlebotomy. Although the relation of these abnormalities to iron metabolism is not known, the fact that similar changes occur in other forms of iron overload suggests that iron is directly involved.

Cardiac involvement is the presenting manifestation in ~15% of symptomatic patients. The most common manifestation is congestive heart failure, which occurs in ~10% of young adults with the disease, especially those with juvenile hemochromatosis. Symptoms of congestive heart failure may develop suddenly, with rapid progression to death if untreated. The heart is diffusely enlarged; this may be misdiagnosed as idiopathic cardiomyopathy if other overt manifestations are absent. Cardiac arrhythmias include premature supraventricular beats, paroxysmal tachyarrhythmias, atrial flutter, atrial fibrillation, and varying degrees of atrioventricular block.

Hypogonadism occurs in both sexes and may antedate other clinical features. Manifestations include loss of libido, impotence, amenorrhea, testicular atrophy, gynecomastia, and sparse body hair. These changes are primarily the result of decreased production of gonadotropins due to impairment of hypothalamic-pituitary function by iron deposition.

DIAGNOSIS

The association of (1) hepatomegaly, (2) skin pigmentation, (3) diabetes mellitus, (4) heart disease, (5) arthritis, and (6) hypogonadism should suggest the diagnosis. However, as stated above, significant iron overload may exist with none or only some of these manifestations. Therefore, a high index of suspicion is needed to make the diagnosis early. Treatment before permanent organ damage occurs can reverse the iron toxicity and restore life expectancy to normal.

The history should be particularly detailed in regard to disease in other family members; alcohol ingestion; iron intake; and ingestion of large doses of ascorbic acid, which promotes iron absorption **(Chap. 326)**. Appropriate tests should be performed to exclude iron deposition due to hematologic disease. The presence of liver, pancreatic, cardiac, and joint disease should be confirmed by physical examination, radiography, and standard function tests of these organs.

The degree of increase in total body iron stores can be assessed by (1) measurement of serum iron and the percent saturation of transferrin (or the unsaturated iron-binding capacity), (2) measurement of serum ferritin concentration, (3) liver biopsy with measurement of the iron concentration and calculation of the hepatic iron index **(Table 407-2)**, and (4) magnetic resonance imaging (MRI) of the liver. In addition, a retrospective assessment of body-iron storage is also provided by performing weekly phlebotomy and calculating the amount of iron removed before iron stores are exhausted (1 mL blood = ~0.5 mg iron).

Each of these methods for assessing iron stores has advantages and limitations. The serum iron level and percent saturation of transferrin are elevated early in the course, but their specificity is reduced by significant false-positive and false-negative rates. For example, serum iron concentration may be increased in patients with alcoholic liver disease without iron overload; in this situation, however, the hepatic iron index is usually not increased as in hemochromatosis (Table 407-1). In otherwise healthy persons, a fasting serum transferrin saturation >45% is abnormal and suggests homozygosity for hemochromatosis.

The serum ferritin concentration is usually a good index of body-iron stores, whether decreased or increased. In fact, an increase of 1 μg/L in serum ferritin level reflects an increase of ~5 mg in body stores. In most untreated patients with hemochromatosis, the serum ferritin level is significantly increased (Fig. 407-2 and Table 407-1), and a serum ferritin level >1000 μg/L is the strongest predictor of disease expression among individuals homozygous for the C282Y mutation. However, in patients with inflammation and hepatocellular necrosis, serum ferritin levels may be elevated out of proportion to body-iron stores due to increased release from tissues. Therefore, a repeat determination of serum ferritin should be carried out after acute hepatocellular damage has subsided (e.g., in alcoholic liver disease). Ordinarily, the combined measurements of the percent transferrin saturation and serum ferritin level provide a simple and reliable screening test for hemochromatosis, including the precirrhotic phase of the disease. If either of these tests is abnormal, genetic testing for hemochromatosis should be performed **(Fig. 407-3)**.

The role of liver biopsy in the diagnosis and management of hemochromatosis has been reassessed as a result of the widespread availability of genetic testing for the C282Y mutation. The absence of

TABLE 407-2 Representative Iron Values in Normal Subjects, Patients with Hemochromatosis, and Patients with Alcoholic Liver Disease

DETERMINATION	NORMAL	SYMPTOMATIC HEMOCHROMATOSIS	HOMOZYGOTES WITH EARLY, ASYMPTOMATIC HEMOCHROMATOSIS	HETEROZYGOTES	ALCOHOLIC LIVER DISEASE
Plasma iron, μmol/L (μg/dL)	9–27 (50–150)	32–54 (180–300)	Usually elevated	Elevated or normal	Often elevated
Total iron-binding capacity, μmol/L (μg/dL)	45–66 (250–370)	36–54 (200–300)	36–54 (200–300)	Elevated or normal	45–66 (250–370)
Transferrin saturation, %	22–45	50–100	50–100	Normal or elevated	27–60
Serum ferritin, μg/L		1000–6000	200–500	Usually <500	10–500
Men	20–250				
Women	15–150				
Liver iron, μg/g dry wt	300–1400	6000–18,000	2000–4000	300–3000	300–2000
Hepatic iron index	<1.0	>2	1.5–2	<2	<2

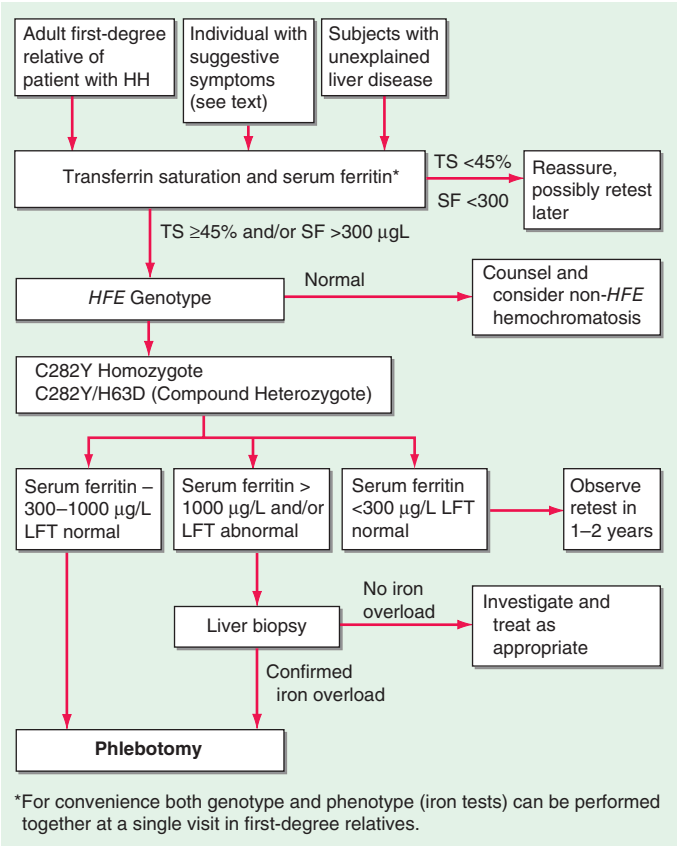


FIGURE 407-3 Algorithm for screening for *HFE*-associated hemochromatosis. HH, hereditary hemochromatosis, homozygous subject (C282Y +/+); LFT, liver function tests; SF, serum ferritin concentration; TS, transferrin saturation.

severe fibrosis can be accurately predicted in most patients using clinical and biochemical variables. Thus, there is virtually no risk of severe fibrosis in a C282Y homozygous subject with (1) serum ferritin level <1000 µg/L, (2) normal serum alanine aminotransferase values, (3) no hepatomegaly, and (4) no excess alcohol intake. However, it should be emphasized that liver biopsy is the only reliable method for establishing or excluding the presence of hepatic cirrhosis, which is the critical factor determining prognosis and the risk of developing hepatocellular carcinoma. Biopsy also permits histochemical estimation of tissue iron and measurement of hepatic iron concentration. Increased density of the liver due to iron deposition can be demonstrated by computed tomography (CT) or MRI, and with improved technology, MRI has become more accurate in determining hepatic iron concentration.

SCREENING FOR HEMOCHROMATOSIS

When the diagnosis of hemochromatosis is established, it is important to counsel and screen other family members **(Chap. 457)**. Asymptomatic and symptomatic family members with the disease usually have an increased saturation of transferrin and an increased serum ferritin concentration. These changes occur even before the iron stores are greatly increased (Fig. 407-2). All adult first-degree relatives of patients with hemochromatosis should be tested for the C282Y and H63D mutations and counseled appropriately (Fig. 407-3). In affected individuals, it is important to confirm or exclude the presence of cirrhosis and begin therapy as early as possible. For children of an identified proband, testing for *HFE* of the other parent is helpful because if normal, the child is merely an obligate heterozygote and at no risk. Otherwise, for practical purposes, children need not be checked before they are 18 years old.

The role of population screening for hemochromatosis is controversial. Recent studies indicate that it is highly effective for primary care physicians to screen subjects using transferrin saturation and serum ferritin levels. Such screening also detects iron deficiency. Genetic screening of the normal population is feasible but is probably not cost effective.

TREATMENT

Hemochromatosis

The therapy of hemochromatosis involves removal of the excess body iron and supportive treatment of damaged organs. Iron removal is best accomplished by weekly or twice-weekly phlebotomy of 500 mL. Although there is an initial modest decline in the volume of packed red blood cells to about 35 mL/dL, the level stabilizes after several weeks. The plasma transferrin saturation remains increased until the available iron stores are depleted. In contrast, the plasma ferritin concentration falls progressively, reflecting the gradual decrease in body-iron stores. One 500-mL unit of blood contains 200–250 mg of iron, and ≥25 g of iron may have to be removed. Therefore, in patients with advanced disease, weekly phlebotomy may be required for 1–2 years, and it should be continued until the serum ferritin level is ≤100 µg/L. Thereafter, phlebotomies are performed at appropriate intervals to maintain ferritin levels at ≤100 µg/L. The transferrin saturation fluctuates and may still be elevated but should not dictate further therapy unless it is persistently at 100% when free unbound iron may circulate. Usually one phlebotomy every 3 months will suffice. It is important, however, not to overtreat and render the patient iron-deficient.

Chelating agents such as deferoxamine, when given parenterally, remove 10–20 mg of iron per day, which is much less than that mobilized by once-weekly phlebotomy. Phlebotomy is also less expensive, more convenient, and safer for most patients. However, chelating agents are indicated when anemia or hypoproteinemia is severe enough to preclude phlebotomy. Subcutaneous infusion of deferoxamine using a portable pump is the most effective means of its administration.

An effective oral iron chelating agent, deferasirox (Exjade), is now available. This agent is effective in thalassemia and secondary iron overload, but it is expensive and carries the risk of significant side effects.

Alcohol consumption should be severely curtailed or eliminated because it increases the risk of cirrhosis in hereditary hemochromatosis nearly tenfold. Dietary adjustments are unnecessary, although vitamin C and iron supplements should be avoided. The management of hepatic failure, cardiac failure, and diabetes mellitus is similar to conventional therapy for these conditions. Loss of libido and change in secondary sex characteristics are managed with testosterone replacement or gonadotropin therapy **(Chap. 384)**.

End-stage liver disease may be an indication for liver transplantation, although results are improved if the excess iron can be removed beforehand. The available evidence indicates that the fundamental metabolic abnormality in hemochromatosis is reversed by successful liver transplantation.

PROGNOSIS

The principal causes of death are cardiac failure, hepatocellular failure or portal hypertension and hepatocellular carcinoma.

Life expectancy is improved by removal of the excessive stores of iron and maintenance of these stores at near-normal levels. The 5-year survival rate with therapy increases from 33 to 89%. With repeated phlebotomy, the liver decreases in size, liver function improves, pigmentation of skin decreases, and cardiac failure may be reversed. Diabetes improves in ~40% of patients, but removal of excess iron has little effect on hypogonadism or arthropathy. Hepatic fibrosis may decrease, but established cirrhosis is irreversible. Hepatocellular carcinoma occurs as a late sequela in patients who are cirrhotic at presentation. The apparent increase in its incidence in treated patients is probably related to their increased life span. Hepatocellular carcinoma rarely develops if the disease is treated in the precirrhotic stage. Indeed, the life expectancy of homozygotes treated before the development of cirrhosis is normal.

The importance of family screening and early diagnosis and treatment cannot be overemphasized. Asymptomatic individuals detected by family studies should have phlebotomy therapy if iron stores are

moderately to severely increased. Assessment of iron stores at appropriate intervals is also important. With this management approach, most manifestations of the disease can be prevented.

ROLE OF *HFE* MUTATIONS IN OTHER LIVER DISEASES

There is considerable interest in the role of *HFE* mutations and hepatic iron in several other liver diseases. Several studies have shown an increased prevalence of *HFE* mutations in PCT patients. Iron accentuates the inherited enzyme deficiency in PCT and clinical manifestations of PCT. The situation in NASH is less clear, but some studies have shown an increased prevalence of *HFE* mutations in NASH patients. The role of phlebotomy therapy, however, is unproven despite an intriguing fall in liver enzyme levels. In chronic HCV infection, *HFE* mutations are not more common, but some subjects have increased hepatic iron. Before initiating antiviral therapy in these patients, it is reasonable to perform phlebotomy therapy to remove excess iron stores, because this reduces liver enzyme levels.

HFE mutations are not increased in frequency in alcoholic liver disease. Hemochromatosis in a heavy drinker can be distinguished from alcoholic liver disease by the presence of the C282Y mutation.

End-stage liver disease may also be associated with iron overload of the degree seen in hemochromatosis. The mechanism is uncertain, although studies have shown that alcohol suppresses hepatic hepcidin secretion. Hemolysis also plays a role. *HFE* mutations are uncommon.

Whether subjects homozygous for C282Y are at increased risk of breast and colorectal cancer is controversial.

GLOBAL CONSIDERATIONS

The *HFE* mutation is of northern European origin (Celtic or Nordic) with a heterozygous carrier rate of ~1 in 10 (1 in 8 in Ireland). Thus, *HFE*-associated hemochromatosis is quite rare in non-European populations, e.g., Asia. However, non-*HFE*-associated hemochromatosis resulting from mutations in other genes involved in iron metabolism (Fig. 407-1) is ubiquitous and should be considered when one encounters iron overload.

African iron overload occurs primarily in sub-Saharan Africa and was previously thought to be due to the consumption of an iron-rich fermented maize beverage. However, recent evidence suggests that it is primarily the result of a non-*HFE*-related genetic trait that is exacerbated by the dietary iron loading. A similar form of iron-overload has been described in African Americans. Further research is needed to clarify this condition.

FURTHER READING

Bardou-Jacquet E et al: Decreased cardiovascular and extrahepatic cancer-related mortality in treated patients with mild HFE hemochromatosis. J Hepatol 62:682, 2015.

Guo S et al: Iron homeostasis: Transport, metabolism, and regulation. Curr Opin Clin Nutr Metab Care 19:276, 2016.

Powell LW et al: Haemochromatosis. Lancet 388:706, 2016.

408 Wilson's Disease

George J. Brewer

Wilson's disease is an autosomal recessive disorder caused by mutations in the *ATP7B* gene, which encodes a membrane-bound, copper-transporting ATPase. Clinical manifestations are caused by copper toxicity and primarily involve the liver and the brain. Because effective treatment is available, it is important to make this diagnosis early.

The frequency of Wilson's disease in most populations is about 1 in 30,000–40,000, and the frequency of carriers of *ATP7B* mutations is ~1%. Siblings of a diagnosed patient have a 1 in 4 risk of Wilson's disease, whereas children of an affected patient have about a 1 in 200 risk. Although a large number of inactivating mutations have been reported in the *ATP7B* gene, mutation screening for diagnosis is now available, and a definitive diagnosis can be obtained in 50–75% of patients, depending on the population. DNA haplotype analysis can be used to genotype siblings of an affected patient. A rare multisystem disorder of copper metabolism with features of both Menke's and Wilson's diseases has been reported. It is termed the MEDNIK (mental retardation, enteropathy, deafness, neuropathy, ichthyosis, keratodermia) syndrome and is caused by mutations in the *AP1S1* gene, which encodes an adaptor protein necessary for intracellular trafficking of copper pump proteins ATP7A (Menke's disease) and ATP7B (Wilson's disease).

PATHOGENESIS

ATP7B protein deficiency impairs biliary copper excretion, resulting in positive copper balance, hepatic copper accumulation, and copper toxicity from oxidant damage. Excess hepatic copper is initially bound to metallothionein; liver damage begins as this storage capacity is exceeded, sometimes by 3 years of age. Defective copper incorporation into apoceruloplasmin leads to excess catabolism and low blood levels of ceruloplasmin. Serum copper levels are usually lower than normal because of low blood levels of ceruloplasmin, which normally binds >90% of serum copper. As the disease progresses, nonceruloplasmin serum copper ("free" copper) levels increase, resulting in copper buildup in other parts of the body (e.g., in the brain, with consequent neurologic and psychiatric disease).

CLINICAL PRESENTATION

Hepatic Features Wilson's disease may present as hepatitis, cirrhosis, or hepatic decompensation. Patients typically present in the mid- to late teenage years in Western countries, although the age of presentation is quite broad and extends into the fifth decade of life. An episode of hepatitis may occur—with elevated serum aminotransferase levels, with or without jaundice—and then spontaneously regress. Hepatitis often recurs, and most of these patients eventually develop cirrhosis. Hepatic decompensation is associated with elevated serum bilirubin, reduced serum albumin and coagulation factors, ascites, peripheral edema, and hepatic encephalopathy. In severe hepatic failure, hemolytic anemia may develop because large amounts of copper derived from hepatocellular necrosis are released into the bloodstream. The association of hemolysis and liver disease makes Wilson's disease a likely diagnosis.

Neurologic Features The neurologic manifestations of Wilson's disease typically occur in patients in their early twenties, although the age of onset extends into the sixth decade of life. MRI and CT scans reveal damage in the basal ganglia and occasionally in the pons, medulla, thalamus, cerebellum, and subcortical areas. The three main movement disorders include dystonia, incoordination, and tremor. Dysarthria and dysphagia are common. In some patients, the clinical picture closely resembles that of Parkinson's disease. Dystonia can involve any part of the body and eventually leads to grotesque positions of the limbs, neck, and trunk. Autonomic disturbances may include orthostatic hypotension and sweating abnormalities as well as bowel, bladder, and sexual dysfunction. Memory loss, migraine-type headaches, and seizures may occur. Patients have difficulty focusing on tasks, but cognition usually is not grossly impaired. Sensory abnormalities and muscular weakness are not features of the disease.

Psychiatric Features Half of patients with neurologic disease have a history of behavioral disturbances with onset in the 5 years before diagnosis. The features are diverse and may include loss of emotional control (temper tantrums, crying bouts), depression, hyperactivity, or loss of sexual inhibition.

Other Manifestations Some female patients have repeated spontaneous abortions, and most become amenorrheic prior to diagnosis. Cholelithiasis and nephrolithiasis occur with increased frequency. Some patients have osteoarthritis, particularly of the knee. Microscopic hematuria is common, and levels of urinary excretion of phosphates, amino acids, glucose, or urates may increase; however, a full-blown

TABLE 408-1 Useful Tests for Wilson's Disease

TEST	USEFULNESS[a]	NORMAL VALUE	HETEROZYGOUS CARRIERS	WILSON'S DISEASE
Serum ceruloplasmin	+	180–350 mg/L (18–35 mg/dL)	Low in 20%	Low in 90%
Kayser- Fleischer rings	++	Absent	Absent	Present in >99% if neurologic or psychiatric symptoms are present Present in 30–50% in hepatic presentation and presymptomatic state
Urine copper (24 h)	+++	0.3–0.8 µmol (20–50 µg)	Normal to 1.3 µmol (80 µg)	>1.6 µmol (>100 µg) in symptomatic patients; 0.9 to >1.6 µmol (60 to >100 µg) in presymptomatic patients
Liver copper	++++	0.3–0.8 µmol/g (20–50 µg/g of tissue)	Normal to 2.0 µmol (125 µg)	>3.1 µmol (>200 µg) (Obstructive liver disease can cause false-positive results.)
DNA testing of *ATP7B* gene	+++	No mutations	One mutation	Two mutations
Haplotype analysis	++++ (siblings only)	0 matches	1 match	2 matches

[a]Usefulness range: + (somewhat useful) to ++++ (very useful).

Fanconi's syndrome is rare. Sunflower cataracts and Kayser-Fleischer rings (copper deposits in the outer rim of the cornea) may be seen. Electrocardiographic and other cardiac abnormalities have been reported but are not common.

DIAGNOSIS

Diagnostic tests for Wilson's disease are listed in Table 408-1. Serum ceruloplasmin levels should not be used for definitive diagnosis, because they are normal in up to 10% of affected patients and are reduced in 20% of carriers. Kayser-Fleischer rings (Fig. 408-1) can be definitively diagnosed only by an ophthalmologist using a slit lamp. They are present in >99% of patients with neurologic/psychiatric forms of the disease and have been described very rarely in the absence of Wilson's disease. Kayser-Fleischer rings are present in only ~30–50% of patients diagnosed in the hepatic or presymptomatic state; thus, the absence of rings does not exclude the diagnosis.

Urine copper measurement is an important diagnostic tool, but urine must be collected carefully to avoid contamination. Symptomatic patients invariably have urine copper levels >1.6 µmol (>100 µg) per 24 h. Heterozygotes have values <1.3 µmol (<80 µg) per 24 h. About half of presymptomatic patients who are ultimately affected have diagnostically elevated urine copper values, but the other half have levels that are in an intermediate range between 0.9 and 1.6 µmol (60–100 µg) per 24 h. Because heterozygotes may have values up to 1.3 µmol (80 µg) per 24 h, patients in this range may require a liver biopsy for definitive diagnosis.

Testing for mutations in *ATP7B* can be done and is recommended for a definitive diagnosis. Several hundred inactivating mutations have been described and genetic testing laboratories can now make a definitive diagnosis in about 50–75% of patients, depending on the population.

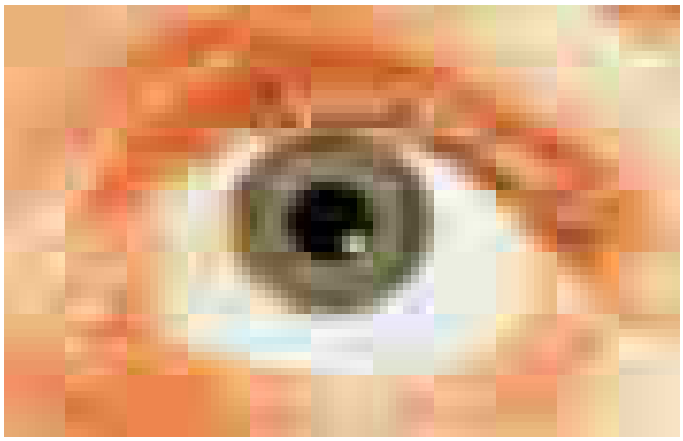


FIGURE 408-1 A Kayser-Fleischer ring. Although in this case, the brownish ring rimming the cornea is clearly visible to the naked eye, confirmation is usually made by slit-lamp examination.

The gold standard for diagnosis remains liver biopsy with quantitative copper assays. Affected patients have values >3.1 µmol/g (>200 µg/g [dry weight] of liver). Copper stains are not reliable. False-positive results can occur with long-standing obstructive liver disease, which can elevate hepatic and urine copper concentrations and rarely causes Kayser-Fleischer rings.

TREATMENT

Wilson's Disease

Recommended anticopper treatments are listed in Table 408-2. Penicillamine was previously the primary anticopper treatment but now plays only a minor role because of its toxicity and because it often worsens existing neurologic disease if used as initial therapy. If penicillamine is given, it should always be accompanied by pyridoxine (25 mg/d). Trientine is a less toxic chelator and is supplanting penicillamine when a chelator is indicated.

For patients with hepatitis or cirrhosis but without evidence of hepatic decompensation or neurologic/psychiatric symptoms, zinc is the therapy of choice although some experts advocate therapy with trientine. Zinc has proven efficacy in Wilson's disease and is essentially nontoxic. It produces a negative copper balance by blocking intestinal absorption of copper, and it induces hepatic metallothionein synthesis, thereby sequestering additional toxic copper.

TABLE 408-2 Recommended Anticopper Drugs for Wilson's Disease

DISEASE STATUS	FIRST CHOICE	SECOND CHOICE
Initial hepatic		
Hepatitis or cirrhosis without decompensation	Zinc[a]	Trientine
Hepatic decompensation		
Mild	Trientine[b] and zinc	Penicillamine[b] and zinc
Moderate	Trientine and zinc	Hepatic transplantation
Severe	Hepatic transplantation	Trientine and zinc
Initial neurologic/ psychiatric	Tetrathiomolybdate[c] and zinc	Zinc
Maintenance	Zinc	Trientine
Presymptomatic	Zinc	Trientine
Pediatric	Zinc	Trientine
Pregnant	Zinc	Trientine

[a]Zinc acetate is supplied as Galzin, manufactured by Gate Pharmaceutical. The recommended adult dose for all the above indications is 50 mg of elemental zinc three times daily, with each dose separated by at least 1 h from consumption of food and beverages other than water as well as from trientine or penicillamine doses. [b]Trientine is supplied as Syprine and penicillamine as Cuprimine, both manufactured by Merck. The recommended adult dosage for both drugs is 500 mg twice daily, with each dose at least 0.5 h before or 2 h after meals and separated by at least 1 h from zinc administration. [c]Tetrathiomolybdate is being studied in clinical trials.

TABLE 408-3 Prognostic Index of Nazer

LABORATORY MEASUREMENT	NORMAL VALUE	SCORE (IN POINTS)				
		0	1	2	3	4
Serum bilirubin[a]	0.2–1.2 mg/dL	<5.8	5.8–8.8	8.8–11.7	11.7–17.5	>17.5
Serum aspartate aminotransferase	10–35 IU/L	<100	100–150	151–200	201–300	>300
Prolongation of prothrombin time (s)	—	<4	4–8	9–12	13–20	>20

[a]If hemolysis is present, serum bilirubin cannot be used as a measure of liver function until the hemolysis subsides.

Source: Modified from H Nazer et al: Gut 27:1377, 1986; with permission from BMJ Publishing Group.

All presymptomatic patients should be treated prophylactically because the disease is close to 100% penetrant.

The first step in evaluating patients presenting with hepatic decompensation is to establish disease severity, which can be estimated with the Nazer prognostic index **(Table 408-3)**. Patients with scores <7 can usually be managed with medical therapy. Patients with scores >9 should be considered immediately for liver transplantation. For patients with scores between 7 and 9, clinical judgment is required in deciding whether to recommend transplantation or medical therapy. A combination of trientine and zinc has been used to treat patients with Nazer scores as high as 9, but such patients should be watched carefully for indications of hepatic deterioration, which mandates transplantation.

For initial medical treatment of patients with hepatic decompensation, the recommended regimen is a chelator (preferably trientine) plus zinc (Table 408-2). Zinc should not, however, be ingested simultaneously with trientine, which chelates zinc and forms therapeutically ineffective complexes. Administration of the two drugs should be separated by at least 1 h.

For initial neurologic therapy, tetrathiomolybdate is emerging as the drug of choice because of its rapid control of free copper, preservation of neurologic function, and low toxicity. Penicillamine and trientine should be avoided because both have a high risk of worsening the neurologic condition. Until tetrathiomolybdate is commercially available, zinc therapy is recommended. Although it is relatively slow-acting, zinc itself does not exacerbate neurologic abnormalities. Although hepatic transplantation may alleviate neurologic symptoms, it does so only by copper removal, which can be done more safely and inexpensively with anticopper drugs. Pregnant patients should be treated with zinc or trientine throughout pregnancy but without tight copper control because copper deficiency can be teratogenic.

Anticopper therapy must be lifelong. With treatment, liver function usually recovers after about a year although residual liver damage is usually present. Neurologic and psychiatric symptoms usually improve after 6–24 months of treatment.

MONITORING ANTICOPPER THERAPY

When trientine or penicillamine is first used, it is necessary to monitor for drug toxicity, particularly bone marrow suppression and proteinuria. Complete blood counts, standard biochemical profiles, and a urinalysis should be performed at weekly intervals for 1 month, then at twice-weekly intervals for 2 or 3 months, then at monthly intervals for 3 or 4 months, and at 4- to 6-month intervals thereafter.

The anticopper effects of trientine and penicillamine can be monitored by following "free" serum copper levels. Changes in urine copper levels are more difficult to interpret because excretion reflects the effect of the drug as well as body loading with copper. Free serum copper is calculated by subtracting the ceruloplasmin copper from the total serum copper. Each 10 mg/L (1 mg/dL) of ceruloplasmin contributes 0.5 μmol/L (3 μg/dL) of serum copper. The normal serum-free copper value is 1.6–2.4 μmol/L (10–15 μg/dL); the level is often as high as 7.9 μmol/L (50 μg/dL) in untreated Wilson's disease. With treatment, the serum-free copper should be <3.9 μmol/L (<25 μg/dL).

Zinc treatment does not require monitoring of blood or urine for toxicity. Its only significant side effect is gastric burning or nausea in ~10% of patients, usually with the first morning dose. This effect can be mitigated if the first dose is taken an hour after breakfast or if zinc is taken with a small amount of protein. Because zinc mainly affects stool copper, 24-h urine copper can be used to reflect body loading. The typical value in untreated symptomatic patients is >3.1 μmol (>200 μg) per 24 h. This level should decrease during the first 1–2 years of therapy to <2.0 μmol (<125 μg) per 24 h. A normal value (0.3–0.8 μmol [20–50 μg]) is rarely reached during the first decade of therapy and should raise concern about overtreatment (copper deficiency), the first sign of which is anemia and/or leukopenia.

GLOBAL CONSIDERATIONS

The age of onset of clinical disease may be considerably younger in India and the Far East; in these regions, onset often occurs in children at only 5 or 6 years of age. The incidence of the disease may be increased in certain populations as a result of founder effects. For example, in Sardinia, the incidence may be 1 in 3000. In countries where penicillamine, trientine, and zinc acetate (as Galzin) are not available or are unaffordable, zinc salts such as gluconate or sulfate provide an alternative treatment option.

FURTHER READING

Askari FK et al: Treatment of Wilson's disease with zinc. XVIII. Initial treatment of the hepatic decompensation presentation with trientine and zinc. J Lab Clin Med 142:385, 2003.

Brewer GJ: *Wilson's Disease: A Clinician's Guide to Recognition, Diagnosis, and Management*. Boston, Kluwer Academic Publishers, 2001.

Brewer GJ et al: Treatment of Wilson disease with ammonium tetrathiomolybdate: III. Initial therapy in a total of 55 neurologically affected patients and follow-up with zinc therapy. Arch Neurol 60:379, 2003.

Brewer GJ et al: Treatment of Wilson's disease with zinc: XV long-term follow-up studies. J Lab Clin Med 132:264, 1998.

Martinelli D et al: MEDNIK syndrome: A novel defect of copper metabolism treatable by zinc acetate therapy. Brain 136:872, 2013.

Nazer H et al: Wilson's disease: Clinical presentation and use of prognostic index. Gut 27:1377, 1986.

Scheinberg IH, Sternlieb I: *Wilson's Disease (Major Problems in Internal Medicine, Volume 23)*. LHJ Smith (ed). Philadelphia, W.B. Saunders Company, 1984.

Schilsky ML: Wilson disease: Genetic basis of copper toxicity and natural history. Semin Liver Dis 16:83, 1996.

409 The Porphyrias

Robert J. Desnick, Manisha Balwani

THE PORPHYRIAS: INTRODUCTION

The porphyrias are metabolic disorders, each resulting from the deficiency or increased activity of a specific enzyme in the heme biosynthetic pathway **(Fig. 409-1 and Table 409-1)**. These enzyme disorders are inherited as autosomal dominant, autosomal recessive, or X-linked traits, with the exception of porphyria cutanea tarda (PCT), which is

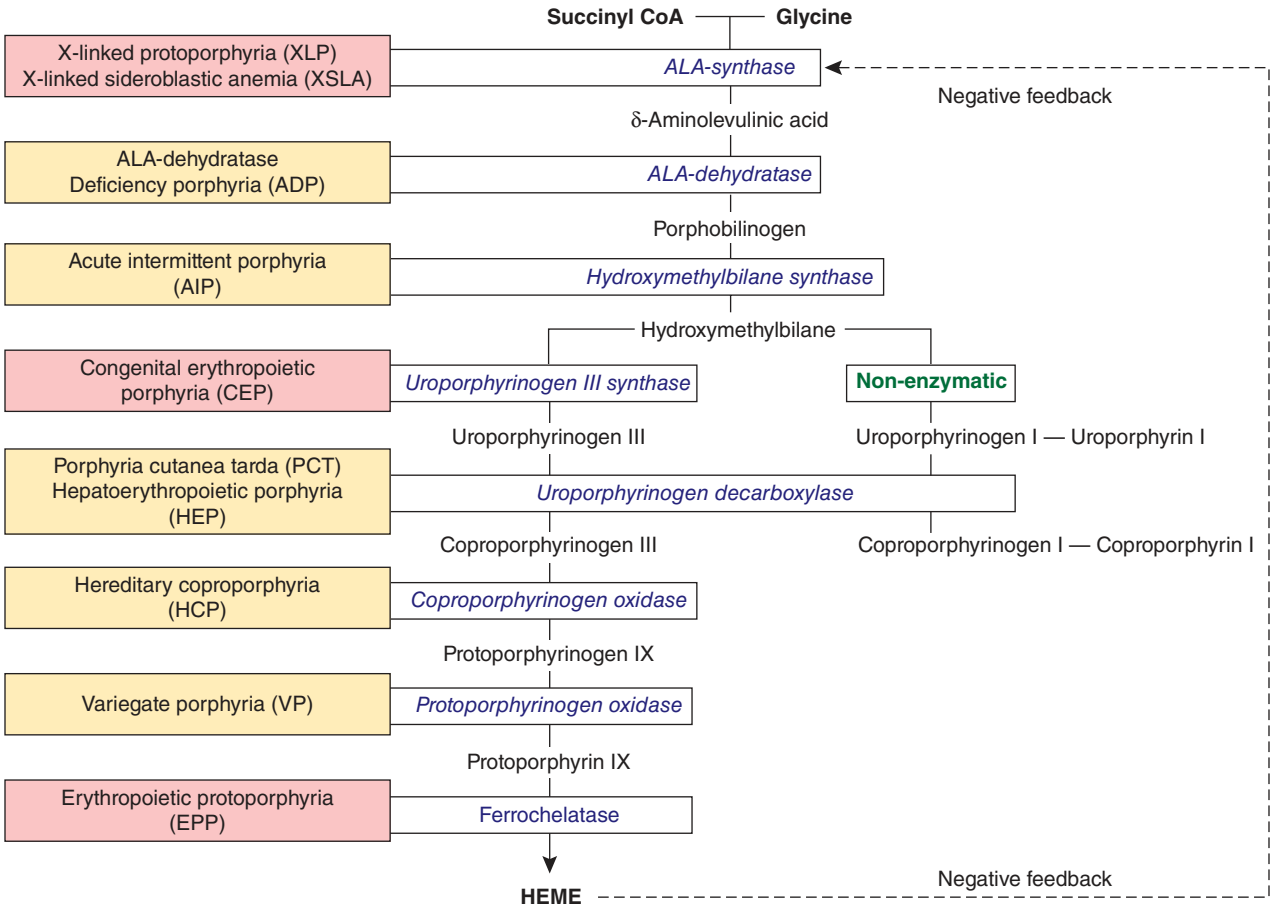


FIGURE 409-1 The human heme biosynthetic pathway indicating in *linked boxes* the enzyme that, when deficient or overexpressed, causes the respective porphyria. Hepatic porphyrias are shown in *yellow boxes* and erythropoietic porphyrias in *pink boxes*.

TABLE 409-1 Human Porphyrias: Major Clinical and Laboratory Features

PORPHYRIA	DEFICIENT ENZYME	INHERITANCE	PRINCIPAL SYMPTOMS: NV OR CP+	ENZYME ACTIVITY % OF NORMAL	INCREASED PORPHYRIN PRECURSORS AND/OR PORPHYRINS		
					ERYTHROCYTES	URINE	STOOL
Hepatic Porphyrias							
5-ALA-dehydratase-deficient porphyria (ADP)	ALA-dehydratase	AR	NV	~5	Zn-Protoporphyrin	ALA, coproporphyrin III	—
Acute intermittent porphyria (AIP)	HMB-synthase	AD	NV	~50	—	ALA, PBG, uroporphyrin	—
Porphyria cutanea tarda (PCT)	URO-decarboxylase	AD	CP	~20	—	Uroporphyrin, 7-carboxylate porphyrin	Isocoproporphyrin
Hereditary coproporphyria (HCP)	COPRO-oxidase	AD	NV and CP	~50	—	ALA, PBG, coproporphyrin III	Coproporphyrin III
Variegate porphyria (VP)	PROTO-oxidase	AD	NV and CP	~50	—	ALA, PBG, coproporphyrin III	Coproporphyrin III, protoporphyrin
Erythropoietic Porphyrias							
Congenital erythropoietic porphyria (CEP)	URO-synthase	AR	CP	1–5	Uroporphyrin I Coproporphyrin I	Uroporphyrin I[a] Coproporphyrin I[a]	Coproporphyrin I
Erythropoietic protoporphyria (EPP)	Ferrochelatase	AR	CP	~20–30	Protoporphyrin	—	Protoporphyrin
X-linked protoporphyria (XLP)	ALA-synthase 2	XL	CP	>100[b]	Protoporphyrin	—	Protoporphyrin

[a]Type I isomers. [b]Increased activity due to "gain-of-function" mutations in *ALAS2* exon 11.

Abbreviations: AD, autosomal dominant; ALA, 5-aminolevulinic acid; AR, autosomal recessive; COPRO I, coproporphyrin I; COPRO III, coproporphyrin III; CP, cutaneous photosensitivity; ISOCOPRO, isocoproporphyrin; + Nv, neurovisceral; PBG, porphobilinogen; PROTO, protoporphyrin IX; URO I, uroporphyrin I; URO III, uroporphyrin III; XL, X-linked.

TABLE 409-2 Human HEME Biosynthetic Enzymes and Genes

ENZYME	GENE SYMBOL	CHROMOSOMAL LOCATION	cDNA (bp)	GENE		PROTEIN (aa)	SUBCELLULAR LOCATION	KNOWN MUTATIONS[b]	THREE-DIMENSIONAL STRUCTURE[c]
				SIZE (kb)	EXONS[a]				
ALA-synthase									
Housekeeping	*ALAS1*	3p21.1	2199	17	11	640	M	—	
Erythroid-specific	*ALAS2*	Xp11.2	1937	22	11	587	M	>30	—
ALA-dehydratase									
Housekeeping	*ALAD*	9q32	1149	15.9	12 (1A + 2 – 12)	330	C	12	Y
Erythroid-specific	*ALAD*	9q32	1154	15.9	12 (1B + 2 – 12)	330	C	—	
HMB-synthase									
Housekeeping	*HMBS*	11q23.3	1086	11	15 (1 + 3 – 15)	361	C	400	E
Erythroid-specific	*HMBS*	11q23.3	1035	11	15 (2 – 15)	344	C	10	
URO-synthase									
Housekeeping	*UROS*	10q26.2	1296	34	10 (1 + 2B – 10)	265	C	45	H
Erythroid-specific	*UROS*	10q26.2	1216	34	10 (2A + 2B – 10)	265	C	4	
URO-decarboxylase	*UROD*	1p34.1	1104	3	10	367	C	122	H
COPRO-oxidase	*CPOX*	3q12.1	1062	14	7	354	M	70	H
PROTO-oxidase	*PPOX*	1q23.3	1431	5.5	13	477	M	181	—
Ferrochelatase	*FECH*	18q21.31	1269	45	11	423	M	192	B

[a]Number of exons and those encoding separate housekeeping and erythroid-specific forms indicated in parentheses. [b]Number of known mutations from the Human Gene Mutation Database (*www.hgmd.org*). [c]Crystallized from human (H), murine (M), *Escherichia coli* (E), *Bacillus subtilis* (B), or yeast (Y) purified enzyme; references in Protein Data Bank (*www.rcsb.org*).

Abbreviations: C, cytoplasm; M, mitochondria.

Source: From KE Anderson et al: Disorders of heme biosynthesis: X-linked sideroblastic anemia and the porphyrias, in *The Metabolic and Molecular Bases of Inherited Diseases*, CR Scriver et al (eds). New York, McGraw-Hill, 2001, pp 2991–3062.

usually sporadic (Table 409-1). The porphyrias are classified as either *hepatic* or *erythropoietic*, depending on the primary site of overproduction and accumulation of their respective porphyrin precursors or porphyrins **(Tables** 409-1 **and 409-2)**, although some have overlapping features. For example, PCT, the most common porphyria, is hepatic and presents with blistering cutaneous photosensitivity, which is typically characteristic of the erythropoietic porphyrias (EPPs).

The major manifestations of the acute hepatic porphyrias are neurologic, including neuropathic abdominal pain, peripheral motor neuropathy, and mental disturbances, with attacks often precipitated by dieting, certain porphyrinogenic drugs, and hormonal changes. While hepatic porphyrias are symptomatic primarily in adults, rare homozygous variants of the autosomal dominant hepatic porphyrias usually manifest clinically prior to puberty. In contrast, the erythropoietic porphyrias usually present at birth or in early childhood with cutaneous photosensitivity, or in the case of congenital erythropoietic porphyria (CEP), even *in utero* as nonimmune hydrops fetalis. Cutaneous sensitivity to sunlight results from excitation of excess porphyrins in the skin by long-wave ultraviolet light, leading to cell damage, scarring, and disfigurement. Thus, the porphyrias are metabolic disorders in which environmental, physiologic, and genetic factors interact to cause disease.

Because many symptoms of the porphyrias are nonspecific, diagnosis is often delayed. Laboratory measurement of porphyrin precursors (5′-aminolevulinic acid [ALA] and porphobilinogen [PBG]) in the urine or porphyrins in the urine, plasma, erythrocytes, or feces is required to confirm or exclude the various types of porphyria (see below). However, a definite diagnosis requires demonstration of the specific gene defect **(Table 409-3)**. The genes encoding all the heme biosynthetic enzymes have been characterized, permitting identification of the mutations causing each porphyria (Table 409-2). Molecular genetic analyses now make it possible to provide precise heterozygote or homozygote identification and prenatal diagnoses in families with known mutations.

In addition to recent reviews of the porphyrias, informative and up-to-date websites are sponsored by the American Porphyria Foundation (*www.porphyriafoundation.com*) and the European Porphyria Initiative (*www.porphyria-europe.org*). An extensive list of unsafe and safe drugs for individuals with acute porphyrias is provided at the Drug Database for Acute Porphyrias (*www.drugs-porphyria.com*).

GLOBAL CONSIDERATIONS

The porphyrias are panethnic metabolic diseases that affect individuals around the globe. The acute hepatic porphyrias—acute intermittent porphyria (AIP), hereditary coproporphyria (HCP), and variegate porphyria (VP)—are autosomal dominant disorders. The frequency of symptomatic AIP, the most common acute hepatic porphyria, is ~1 in 20,000 among Caucasian individuals of Western European ancestry, and it is particularly frequent in Scandinavians, with a frequency of ~1 in 10,000 in Sweden. However, a recent study using genomic/exomic databases showed an estimated frequency of pathogenic variants in the *HMBS* gene as ~1 in 1,700. Thus, the penetrance of AIP, and likely the other acute hepatic porphyrias, is low, about 1–10% of those with pathogenic mutations experiencing acute attacks (see below).

VP is particularly frequent in South Africa, where its high prevalence (>10,000 affected patients) is in part due to a genetic "founder effect." The autosomal recessive acute hepatic porphyria, ALA-dehydratase-deficient porphyria (ADP), is very rare, and <20 patients have been reported worldwide.

The EPPs—CEP, EPP, and X-linked protoporphyria (XLP)—also are panethnic. EPP is the most common porphyria in children, whereas CEP is very rare, with about 200 reported cases worldwide. The frequency of EPP varies globally because most patients have the common low expression ferrochelatase (*FECH*) mutation that varies in frequency in different populations. It rarely occurs in Africans, is present in about 10% of whites, and is frequent (~30%) in the Japanese. The reported prevalence of EPP in the Caucasian population ranges from 1 in 75,000 to 1 in 152,000.

The autosomal recessive porphyrias—ADP, CEP, hepatoerythropoietic porphyria (HEP)—are more frequent in regions with high rates of consanguineous unions. PCT, which is typically sporadic, occurs more

TABLE 409-3 Diagnosis of Acute and Cutaneous Porphyrias

SYMPTOMS	FIRST-LINE TEST: ABNORMALITY	POSSIBLE PORPHYRIA	SECOND-LINE TESTING IF FIRST-LINE TESTING IS POSITIVE: TO INCLUDE: URINE (U), PLASMA (P), AND FECAL (F) PORPHYRINS; FOR ACUTE PORPHYRIAS, ADD RED BLOOD CELL (RBC) HMB-SYNTHASE; FOR BLISTERING SKIN LESIONS, ADD P AND RBC PORPHYRINS	CONFIRMATORY TEST: ENZYME ASSAY AND/OR MUTATION ANALYSIS
Neurovisceral	**Spot U:** ↑↑ALA and normal PBG	ADP	**U porphyrins:** ↑↑, mostly COPRO III **P & F porphyrins:** normal or slightly ↑ **RBC HMB-synthase:** normal	Rule out other causes of elevated ALA; ↓↓RBC ALA-dehydratase activity (<10%); ALA-dehydratase mutation analysis
	Spot U: ↑↑PBG	AIP	**U porphyrins:** ↑↑, mostly URO and COPRO **P & F porphyrins:** normal or slightly ↑ **RBC HMB-synthase:** usually ↓	HMB-synthase mutation analysis
	"	HCP	**U porphyrins:** ↑↑, mostly COPRO III **P porphyrins:** normal or slightly ↑(↑ if skin lesions present) **F porphyrins:** ↑↑, mostly COPRO III	Measure RBC HMB-synthase: normal activity COPRO-oxidase mutation analysis
	"	VP	**U porphyrins:** ↑↑, mostly COPRO III **P porphyrins:** ↑↑(characteristic fluorescence peak at neutral pH) **F porphyrins:** ↑↑, mostly COPRO and PROTOO	Measure RBC HMB-synthase: normal activity PROTO-oxidase mutation analysis
Blistering skin lesions	**P:** ↑ porphyrins	PCT and HEP	**U porphyrins:** ↑↑, mostly URO and heptacarboxylate porphyrin **P porphyrins:** ↑↑ **F porphyrins:** ↑↑, including increased isocoproporphyrin **RBC porphyrins:** ↑↑ zinc PROTO in HEP[a]	RBC URO-decarboxylase activity: half-normal in familial PCT (~20% of all PCT cases); substantially deficient in HEP URO-decarboxylase mutation analysis: mutation(s) present in familial PCT (heterozygous) and HEP (homozygous)
	"	HCP and VP	See HCP and VP above. Also, U ALA and PBG: may be ↑	
	"	CEP	**RBC and U porphyrins:** ↑↑, mostly URO I and COPRO I **F porphyrins:** ↑↑; mostly COPRO I	↓↓ RBC URO-synthase activity (<15%) URO-synthase mutation analysis
Nonblistering photosensitivity	**P:** porphyrins usually ↑	EPP	**RBC porphyrins:** ↓↓, mostly free PROTO **U porphyrins:** normal **F porphyrins:** normal or ↓, mostly PROTO	*FECH* mutation analysis
	P: porphyrins usually ↑	XLP	**RBC porphyrins:** ↑↑, approximately equal free and zinc PROTO **U porphyrins:** normal **F porphyrins:** normal or ↑, mostly PROTO	*ALAS2* mutation analysis

[a]Nonspecific increases in zinc protoporphyrins are common in other porphyrias.

Abbreviations: ADP, 5-ALA-dehydratase-deficient porphyria; AIP, acute intermittent porphyria; ALA, 5-aminolevulinic acid; CEP, congenital erythropoietic porphyria; COPRO I, coproporphyrin I; COPRO III, coproporphyrin III; EPP, erythropoietic protoporphyria; F, fecal; HCP, hereditary coporphyria; HEP, hepatoerythropoietic porphyria; ISOCOPRO, isocoproporphyrin; P, plasma; PBG, porphobilinogen; PCT, porphyria cutanea tarda; PROTO, protoporphyrin IX; RBC, erythrocytes; U, urine; URO I, uroporphyrin I; URO III, uroporphyrin III; VP, variegate porphyria; XLP, X-linked protoporphyria.

Source: Based on KE Anderson et al: Ann Intern Med 142:439, 2005.

frequently in countries in which its predisposing risk factors such as hepatitis C and HIV are more prevalent.

HEME BIOSYNTHESIS

Heme biosynthesis involves eight enzymatic steps in the conversion of glycine and succinyl-CoA to heme (Fig. 409-2 and Table 409-2). These eight enzymes are encoded by nine genes, as the first enzyme in the pathway, ALA-synthase, has two genes that encode unique housekeeping (*ALAS1*) and erythroid-specific (*ALAS2*) isozymes. The first and last three enzymes in the pathway are located in the mitochondrion, whereas the other four are in the cytosol. Heme is required for a variety of hemoproteins such as hemoglobin, myoglobin, respiratory cytochromes, and the cytochrome P450 (CYPs) enzymes. Hemoglobin synthesis in erythroid precursor cells accounts for ~85% of daily heme synthesis in humans. Hepatocytes account for most of the rest, primarily for the synthesis of CYPs, which are especially abundant in the liver endoplasmic reticulum, and turn over more rapidly than many other hemoproteins, such as the mitochondrial respiratory cytochromes. As shown in Fig. 409-2, the pathway intermediates are the porphyrin precursors, ALA and PBG, and porphyrins (mostly in their reduced forms, known as *porphyrinogens*). At least in humans, these intermediates do not accumulate in significant amounts under normal conditions or have important physiologic functions.

The first enzyme, ALA-synthase, catalyzes the condensation of glycine, activated by pyridoxal phosphate and succinyl-coenzyme A, to form ALA. In the liver, this rate-limiting enzyme can be induced by a variety of drugs, steroids, and other chemicals. Distinct nonerythroid (e.g., housekeeping) and erythroid-specific forms of ALA-synthase are encoded by separate genes located on chromosome 3p21.1 (*ALAS1*) and Xp11.2 (*ALAS2*), respectively. Defects in the erythroid gene *ALAS2* that decrease its activity cause an X-linked sideroblastic anemia (XLSA). Gain-of-function mutations in the last exon (11) of *ALAS2* that increase its activity cause an X-linked form of EPP, known as XLP.

The second enzyme, ALA-dehydratase, catalyzes the condensation of two molecules of ALA to form PBG. Hydroxymethylbilane synthase (HMB-synthase; also known as PBG-deaminase) catalyzes the head-to-tail condensation of four PBG molecules by a series of deaminations to form the linear tetrapyrrole, HMB. Uroporphyrinogen III synthase (URO-synthase) catalyzes the rearrangement and rapid cyclization of HMB to form the asymmetric, physiologic, octacarboxylate porphyrinogen, uroporphyrinogen (URO'gen) III.

The fifth enzyme in the pathway, uroporphyrinogen decarboxylase (URO-decarboxylase), catalyzes the sequential removal of the four carboxyl groups from the acetic acid side chains of URO'gen III to form coproporphyrinogen (COPRO'gen) III, a tetracarboxylate porphyrinogen. This compound then enters the mitochondrion via a specific transporter, where COPRO-oxidase, the sixth enzyme, catalyzes the decarboxylation of two of the four propionic acid groups to form the two vinyl groups of protoporphyrinogen (PROTO'gen) IX, a decarboxylate porphyrinogen. Next, PROTO-oxidase oxidizes PROTO'gen to protoporphyrin IX by the removal of six hydrogen atoms. The product of the reaction is a porphyrin (oxidized form), in contrast to

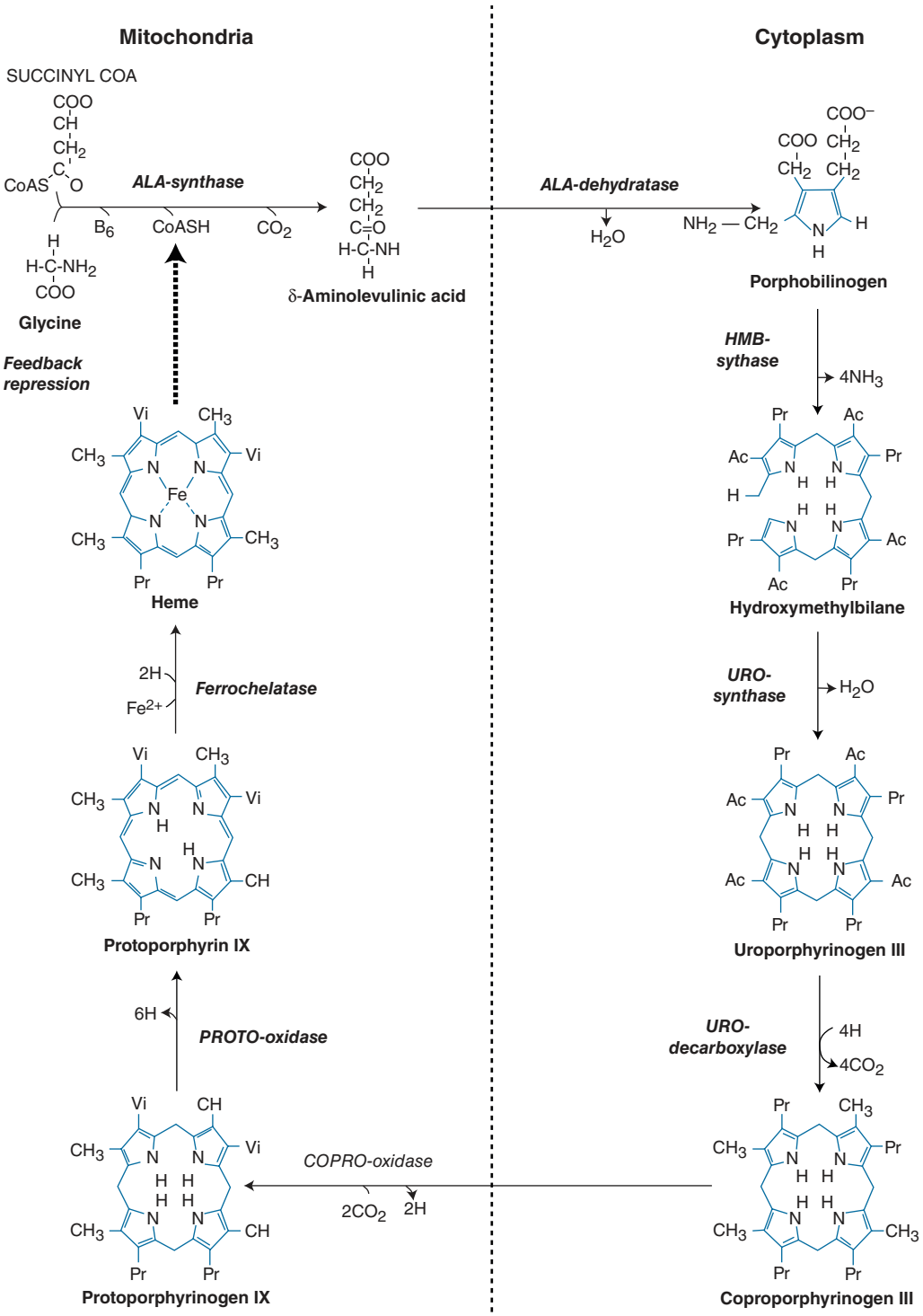


FIGURE 409-2 The heme biosynthetic pathway showing the eight enzymes and their substrates and products. Four of the enzymes are localized in the mitochondria and four in the cytosol.

the preceding tetrapyrrole intermediates, which are porphyrinogens (reduced forms). Finally, ferrous iron is inserted into protoporphyrin to form heme, a reaction catalyzed by the eighth enzyme in the pathway, FECH (also known as heme synthase or protoheme ferrolyase).

REGULATION OF HEME BIOSYNTHESIS

Regulation of heme synthesis differs in the two major heme-forming tissues, the liver and erythron. In the liver, the concentration of "free" heme regulates the synthesis and mitochondrial translocation of the housekeeping form of ALA-synthase 1. Heme represses the synthesis of the ALA-synthase 1 messenger RNA (mRNA) and interferes with the transport of the enzyme from the cytosol into mitochondria. Hepatic ALA-synthase 1 is increased by many of the same chemicals that induce the CYPs enzymes in the endoplasmic reticulum of the liver. Because most of the heme in the liver is used for the synthesis of CYPs enzymes, hepatic ALA-synthase 1 and the CYPs are regulated in a coordinated fashion, and many drugs that induce hepatic ALA-synthase 1 also induce the CYP genes. The other hepatic heme biosynthetic enzymes are presumably expressed at constant levels, although their relative activities and kinetic properties differ. For example, normal individuals have high activities of ALA-dehydratase, but low activities of HMB-synthase, the latter being the second rate-limiting step in the pathway.

In the erythron, novel regulatory mechanisms allow for the production of the very large amounts of heme needed for hemoglobin synthesis. The response to stimuli for hemoglobin synthesis occurs

during cell differentiation, leading to an increase in cell number. In contrast, the erythroid-specific ALA-synthase 2 is expressed at higher levels than the housekeeping enzyme, and erythroid-specific control mechanisms regulate other pathway enzymes as well as iron transport into erythroid cells. Separate erythroid-specific and nonerythroid or "housekeeping" transcripts are known for the first four enzymes in the pathway. As noted above, housekeeping- and erythroid-specific ALA-synthases are encoded by genes on different chromosomes, but for each of the next three genes in the pathway, both erythroid and nonerythroid transcripts are transcribed by alternative promoters from their single respective genes (Table 409-2).

CLASSIFICATION OF THE PORPHYRIAS

As mentioned above, the porphyrias can be classified as either *hepatic* or *erythropoietic*, depending on whether the heme biosynthetic intermediates that accumulate arise initially from the liver or developing erythrocytes, or as *acute* or *cutaneous*, based on their clinical manifestations. Table 409-1 lists the porphyrias, their principal symptoms, and major biochemical abnormalities. Three of the five hepatic porphyrias—AIP, HCP, and VP—usually present during adult life with acute attacks of neurologic manifestations and elevated levels of one or both of the porphyrin precursors, ALA and PBG, and are thus classified as *acute porphyrias*. Patients with ADP have presented in infancy and adolescence, and typically have elevated ALA with normal or slightly elevated PBG levels. The fifth hepatic disorder, PCT, presents with blistering skin lesions. HCP and VP also may have cutaneous manifestations similar to PCT.

The erythropoietic porphyrias—CEP, EPP, and XLP—are characterized by elevations of porphyrins in bone marrow and erythrocytes and present with cutaneous photosensitivity. The skin lesions in CEP resemble PCT but are usually much more severe, whereas EPP and XLP cause a more immediate, severe, painful, and nonblistering type of photosensitivity. EPP is the most common porphyria to cause symptoms before puberty. Around 20% of EPP patients develop minor abnormalities of liver function, with up to about 5% developing hepatic complications that can become life-threatening. XLP has a clinical presentation similar to EPP causing photosensitivity and liver disease.

DIAGNOSIS OF PORPHYRIA

A few specific and sensitive first-line laboratory tests should be used whenever symptoms or signs suggest the diagnosis of porphyria (Table 409-3). If a first-line test is significantly abnormal, more comprehensive testing should follow to establish the type of porphyria, including the specific causative gene mutation.

Acute Porphyrias An acute porphyria should be suspected in patients with neurovisceral symptoms after puberty. Symptoms include acute abdominal pain, nausea, vomiting, tachycardia, hypertension, and motor neuropathy. As these symptoms are common, other causes should be ruled out. The diagnosis is made by measuring urinary porphyrin precursors (ALA and PBG) in a spot sample of urine (Fig. 409-2). Urinary PBG is always increased during acute attacks of AIP, HCP, and VP and is not substantially increased in any other medical condition. Therefore, this measurement is both sensitive and specific. Results from spot (single-void) urine specimens are highly informative because very substantial increases in PBG are expected during acute attacks of porphyria. A 24-h collection can unnecessarily delay diagnosis. The same spot urine specimen should be saved for quantitative determination of ALA, PBG, and creatinine, in order to confirm the qualitative PBG result and also to detect patients with ADP. Urinary porphyrins may remain increased longer than porphyrin precursors in HCP and VP. Therefore, it is useful to measure total urinary porphyrins in the same sample, keeping in mind that urinary porphyrin increases are often nonspecific. Measurement of urinary porphyrins alone should be avoided for screening, because these may be increased in disorders other than porphyrias, such as chronic liver disease, and misdiagnoses of porphyria can result from minimal increases in urinary porphyrins that have no diagnostic significance. Measurement of erythrocyte HMB-synthase is not useful as a first-line test. Moreover, the enzyme activity is not decreased in all AIP patients, a borderline low normal value is not diagnostic, and the enzyme is not deficient in other acute porphyrias.

Cutaneous Porphyrias Blistering skin lesions due to porphyria are virtually always accompanied by increases in total plasma porphyrins. A fluorometric method is preferred, because the porphyrins in plasma in VP are mostly covalently linked to plasma proteins and may be less readily detected by high-performance liquid chromatography (HPLC). The normal range for plasma porphyrins is somewhat increased in patients with end-stage renal disease.

Although a total plasma porphyrin determination will usually detect EPP and XLP, an erythrocyte protoporphyrin determination is more sensitive. Increases in erythrocyte protoporphyrin occur in many other conditions. Therefore, the diagnosis of EPP must be confirmed by showing a predominant increase in free protoporphyrin rather than zinc protoporphyrin. In XLP, both free and zinc protoporphyrin are markedly increased. Interpretation of laboratory reports can be difficult, because the term *free erythrocyte protoporphyrin* sometimes actually represents zinc protoporphyrin.

More extensive testing is justified when an initial test is positive. A substantial increase in PBG may be due to AIP, HCP, or VP. These acute porphyrias can be distinguished by measuring urinary porphyrins (using the same spot urine sample), fecal porphyrins, and plasma porphyrins. Assays for COPRO-oxidase or PROTO-oxidase are not available for clinical testing. More specifically, mutation analysis by sequencing the genes encoding HMB-synthase, COPRO-oxidase, and PROTO-oxidase will detect almost all disease-causing mutations, and will be diagnostic even when the levels of urinary ALA and PBG have returned to normal or near normal. The various porphyrias that cause blistering skin lesions can be differentiated by measuring porphyrins in urine, feces, and plasma. These porphyrias also should be confirmed at the DNA level by the demonstration of the causative gene mutation(s). It is often difficult to diagnose or "rule out" porphyria in patients who have had suggestive symptoms months or years in the past, and in relatives of patients with acute porphyrias, because porphyrin precursors and porphyrins may be normal. In those situations, detection of the specific gene mutation in the index case can make the diagnosis and facilitate the diagnosis and genetic counseling of at-risk relatives. Consultation with a specialist laboratory and physician will assist in selecting the heme biosynthetic gene or genes to be sequenced.

THE HEPATIC PORPHYRIAS

Markedly elevated plasma and urinary concentrations of the porphyrin precursors, ALA and/or PBG, which originate from the liver, are especially evident during attacks of neurologic manifestations of the four acute porphyrias—ADP, AIP, HCP, and VP. In PCT, excess porphyrins also accumulate initially in the liver and cause chronic blistering of sun-exposed areas of the skin.

ALA-DEHYDRATASE-DEFICIENT PORPHYRIA

ADP is a rare, autosomal recessive, acute hepatic porphyria caused by a severe deficiency of ALA-dehydratase activity. To date, there are only a few documented cases, some in children or young adults, in which specific gene mutations have been identified. These affected homozygotes had <10% of normal ALA-dehydratase activity in erythrocytes, but their clinically asymptomatic parents and heterozygous relatives had about half-normal levels of activity and did not excrete increased levels of ALA. The frequency of ADP is unknown, but the frequency of heterozygous individuals with <50% normal ALA-dehydratase activity was ~2% in a screening study in Sweden. Because there are multiple causes for deficient ALA-dehydratase activity, it is important to confirm the diagnosis of ADP by mutation analysis.

Clinical Features The clinical presentation depends on the amount of residual ALA-dehydratase activity. Four of the documented patients were male adolescents with symptoms resembling those of AIP, including abdominal pain and neuropathy. One patient was an infant with more severe disease, including failure to thrive beginning at birth. The earlier age of onset and more severe manifestations in this patient reflect a more significant deficiency of ALA-dehydratase activity. Another patient developed an acute motor polyneuropathy at age 63 that was associated with a myeloproliferative disorder. He

was heterozygous for an δ-aminolevulinic acid dehydratase (*ALAD*) mutation that presumably was present in erythroblasts that underwent clonal expansion due to the bone marrow malignancy.

Diagnosis All patients had significantly elevated levels of plasma and urinary ALA and urinary coproporphyrin (COPRO) III; ALAD activities in erythrocytes were <10% of normal. Hereditary tyrosinemia type 1 (fumarylacetoacetase deficiency) and lead intoxication should be considered in the differential diagnosis because either succinylacetone (which accumulates in hereditary tyrosinemia and is structurally similar to ALA) or lead can inhibit ALA-dehydratase, increase urinary excretion of ALA and COPRO III, and cause manifestations that resemble those of the acute porphyrias. Heterozygotes are clinically asymptomatic and do not excrete increased levels of ALA but can be detected by demonstration of intermediate levels of erythrocyte ALA-dehydratase activity or a specific mutation in the *ALAD* gene. To date, molecular studies of ADP patients have identified point mutations, splice-site mutations, and a two-base deletion in the *ALAD* gene (Human Gene Mutation Database; *www.hgmd.org*). The parents in each case were not consanguineous, and the index cases had inherited a different *ALAD* mutation from each parent. Prenatal diagnosis of this disorder is possible by determination of ALA-dehydratase activity and/or gene mutations in cultured chorionic villi or amniocytes.

TREATMENT

ALA-Dehydratase-Deficient Porphyria

The treatment of ADP acute attacks is similar to that of AIP (see below). The severely affected infant referred to above was supported by hyperalimentation and periodic blood transfusions but did not respond to intravenous hemin and died after liver transplantation.

ACUTE INTERMITTENT PORPHYRIA

This hepatic porphyria is an autosomal dominant condition resulting from the half-normal level of HMB-synthase activity. The disease is widespread but is especially common in Scandinavia and Great Britain. Clinical expression is highly variable, and activation of the disease is often related to environmental or hormonal factors, such as drugs, diet, and steroid hormones. Attacks can be prevented by avoiding known precipitating factors. Rare homozygous dominant AIP also has been described in children (see below).

Clinical Features Induction of the rate-limiting hepatic enzyme ALA-synthase in heterozygotes who have half-normal HMB-synthase activity is thought to underlie the acute attacks in AIP. The disorder remains latent (or asymptomatic) in the great majority of those who are heterozygous for *HMBS* mutations, and this is almost always the case prior to puberty. In patients with no history of acute symptoms, porphyrin precursor excretion is usually normal, suggesting that half-normal hepatic HMB-synthase activity is sufficient and that hepatic ALA-synthase activity is not increased. However, under conditions where heme synthesis is increased in the liver, half-normal HMB-synthase activity may become limiting, and ALA, PBG, and other heme pathway intermediates may accumulate and be excreted in the urine. Common precipitating factors include endogenous and exogenous steroids, porphyrinogenic drugs, alcohol ingestion, and low-calorie diets, usually instituted for weight loss.

The fact that AIP is almost always latent before puberty suggests that adult levels of steroid hormones are important for clinical expression. Symptoms are more common in women, suggesting a role for estrogens or progestins. Premenstrual attacks are probably due to endogenous progesterone. Acute porphyrias are sometimes exacerbated by exogenous steroids, including oral contraceptive preparations containing progestins. Surprisingly, pregnancy is usually well tolerated, suggesting that beneficial metabolic changes may ameliorate the effects of high levels of progesterone. **Table 409-4** provides a partial list of the major drugs that are harmful in AIP (and also in HCP and VP). Extensive lists of unsafe and safe drugs are available on websites sponsored by the American Porphyria Foundation (*www.porphyriafoundation.com*) and the European Porphyria Initiative (*www.porphyria-europe.org*), and at the Drug Database for Acute Porphyrias website (*www.drugs-porphyria.com*). Reduced intake of calories and carbohydrate, as may occur with illness or attempts to lose weight, can also increase porphyrin precursor excretion and induce attacks of porphyria. Increased carbohydrate intake may ameliorate attacks. Studies in a knockout AIP mouse model indicate that the hepatic *ALAS1* gene is regulated by the peroxisome proliferator-activated receptor γ coactivator 1α (PGC-1α). Hepatic PGC-1α is induced by fasting, which in turn activates *ALAS1* transcription, resulting in increased heme biosynthesis. This finding suggests an important link between nutritional status and the attacks in acute porphyrias. Attacks also can be provoked by infections, surgery, and ethanol.

Because the neurovisceral symptoms rarely occur before puberty and are often nonspecific, a high index of suspicion is required to make the diagnosis. The disease can be disabling but is rarely fatal. Abdominal pain, the most common symptom, is poorly localized, but may be associated with cramping, ileus, abdominal distention, and decreased bowel sounds. However, increased bowel sounds and diarrhea may occur. Abdominal tenderness, fever, and leukocytosis are usually absent or mild because the symptoms are neurologic rather than inflammatory. Nausea; vomiting; constipation; tachycardia; hypertension; mental symptoms; pain in the limbs, head, neck, or chest; muscle weakness; sensory loss; dysuria; and urinary retention are characteristic. Tachycardia, hypertension, restlessness, tremors, and excess sweating are due to sympathetic overactivity.

The peripheral neuropathy is due to axonal degeneration (rather than demyelinization) and primarily affects motor neurons. Significant neuropathy does not occur with all acute attacks; abdominal symptoms are usually more prominent. Motor neuropathy affects the proximal muscles initially, more often in the shoulders and arms. The course and degree of involvement are variable and sometimes may be focal and involve cranial nerves. Deep tendon reflexes initially may be normal or hyperactive but become decreased or absent as the neuropathy advances. Sensory changes such as paresthesia and loss of sensation are less prominent. Progression to respiratory and bulbar paralysis and death occurs especially when the diagnosis and treatment are delayed. Sudden death may result from sympathetic overactivity and cardiac arrhythmia.

Mental symptoms such as anxiety, insomnia, depression, disorientation, hallucinations, and paranoia can occur in acute attacks. Seizures can be due to neurologic effects or to hyponatremia. Treatment of seizures is difficult because most anti-seizure drugs can exacerbate AIP (clonazepam may be safer than phenytoin or barbiturates). Hyponatremia results from hypothalamic involvement and inappropriate vasopressin secretion or from electrolyte depletion due to vomiting, diarrhea, poor intake, or excess renal sodium loss. Persistent hypertension and impaired renal function may occur. When an attack resolves, abdominal pain may disappear within hours, and paresis begins to improve within days and may continue to improve over several years.

Homozygous dominant AIP is a rare form of AIP in which patients inherit *HMBS* mutations from each of their heterozygous parents and, therefore, have very low (<2%) enzyme activity. The disease has been described in a Dutch girl, two young British siblings, and a Spanish boy. In these homozygous affected patients, the disease presented in infancy with failure to thrive, developmental delay, bilateral cataracts, and/or hepatosplenomegaly. Urinary ALA and PBG concentrations were markedly elevated. All of these patients' *HMBS* mutations (R167W, R167Q, and R172Q) were in exon 10 within five bases of each other. Studies of the brain magnetic resonance images (MRIs) of children with homozygous AIP have suggested damage primarily in white matter that was myelinated postnatally, while tracks that myelinated prenatally were normal. Most children with homozygous AIP die at an early age.

Diagnosis ALA and PBG levels are substantially increased in plasma and urine, especially during acute attacks. For example, urinary PBG excretion during an attack is usually 50–200 mg/24 h (220–880

TABLE 409-4 Unsafe Drugs in Porphyria

DOCUMENTED PORPHYRINOGENIC	PROBABLY PORPHYRINOGENIC	POSSIBLY PORPHYRINOGENIC	
Carbamazepine	Altretamine	Aceclofenac	Parecoxib
Carisoprodol	Aminophylline	Acitretin	Pentifylline
Chloramphenicol	Amiodarone	Acrivastine	Pentoxyverine
Clindamycin	Amitriptyline	Alfuzosin	Phenylpropanolamine + cinnarizine
Dextropropoxyphene	Amlodipine	Anastrozole	Pizotifen
Dihydralazine	Amprenavir	Auranofin	Polidocanol
Dihydroergotamine	Aprepitant	Azelastine	Polyestradiol
Drospirenone + estrogen	Atorvastatin	Benztropine	Phosphate
Dydrogesterone	Azathioprine	Benzydamine	Potassium canrenoate
Etonogestrel	Bosentan	Betaxolol	Pravastatin
Fosphenytoin sodium	Bromocriptine	Bicalutamide	Prednisolone
Hydralazine	Buspirone	Biperiden	Prilocaine
Hydroxyzine	Busulfan	Bupropion	Proguanil
Indinavir	Butylscopolamine	Carvedilol	Propafenone
Ketamine	Cabergoline	Chlorambucil	Pseudoephedrine + dexbrompheniramine
Ketoconazole	Ceftriaxone + lidocaine	Chlorcyclizine + guaifenesin	Quillaia extract
Lidocaine	Cerivastatin	Chloroquine	Quinagolide
Lynestrenol	Cetirizine	Chlorprothixene	Quinine
Lynestrenol + estrogen	Cholinetheophyllinate	Chlorzoxazone	Quinupristin + Dalfopristin
Mecillinam	Clarithromycin	Chorionic	Reboxetine
Medroxyprogesterone	Clemastine	Gonadotropin	Repaglinide
Megestrol	Clonidine	Ciclosporin	Rizatriptan
Methylergometrine	Cyclizine	Cisapride	Rofecoxib
Methyldopa	Cyproterone	Citalopram	Ropinirole
Mifepristone	Danazol	Clomethiazole	Ropivacaine
Nicotinic acid/meclozine/	Delavirdine	Clomiphene	Roxithromycin
hydroxyzine	Desogestrel + estrogen	Clomipramine	Sertraline
Nitrofurantoin	Diazepam	Clopidogrel	Sevoflurane
Norethisterone	Dienogest + estrogen	Clotrimazole	Sibutramine
Norgestimate + estrogen	Diclofenac	Cortisone	Sildenafil
Orphenadrine	Diltiazem	Cyclandelate	Sirolimus
Phenobarbital	Diphenhydramine	Cyclophosphamide	Sodium aurothiomalate
Phenytoin	Disopyramide	Cyproheptadine	Sodium oleate + chlorocymol
Pivampicillin	Disulfiram	Dacarbazine	Stavudine
Pivmecillinam	Drospirenone + estrogen	Daunorubicin	Sulindac
Primidone	Dydrogesterone	Desogestrel	Sumatriptan
Rifampicin	Ergoloid mesylate	Dichlorobenzyl alcohol	Tacrolimus
Ritonavir	Erythromycin	Dithranol	Tadalafil
Spironolactone	Estramustine	Docetaxel	Tegafur + uracil
Sulfadiazine + trimethoprim	Ethosuximide	Donepezil	Telmisartan
Tamoxifen	Etoposide	Doxycycline	Thioridazine
Testosterone, injection	Exemestane	Ebastine	Thioguanine
Thiopental	Felbamate	Econazole	Tolfenamic acid
Trimethoprim	Felodipine	Efavirenz	Tolterodine
Valproic acid	Fluconazole	Escitalopram	Torsemide
Venlafaxine	Flunitrazepam	Esomeprazole	Triamcinolone
Vinblastine	Fluvastatin	Estradiol/tablets	Trihexyphenidyl
Vincristine	Glibenclamide	Estriol/tablets	Trimipramine
Vindesine	Halothane	Estrio/vaginal crème,	Valerian
Vinorelbine	Hyoscyamine	tablet	Venlafaxine
Xylometazoline	Ifosfamide	Estrogen, conjugate	Vinblastine
Zaleplon	Imipramine	Finasteride	Vincristine
Ziprasidone	Irinotecan	Flecainide	Vindesine
Zolmitriptan	Isoniazid	Flucloxacillin	Vinorelbine
Zolpidem	Isradipine	Fluoxetine	Xylometazoline
Zuclopenthixol	Itraconazole	Flupentixol	Zaleplon
	Lamivudine + zidovudine	Flutamide	Ziprasidone
	Lansoprazole	Fluvoxamine	Zolmitriptan
	Lercanidipine	Follitropin alfa and beta	Zolpidem
	Levonorgestrel	Galantamine	Zuclopenthixol
	Lidocaine	Glimepiride	

(*Continued*)

TABLE 409-4 Unsafe Drugs in Porphyria (Continued)

DOCUMENTED PORPHYRINOGENIC	PROBABLY PORPHYRINOGENIC	POSSIBLY PORPHYRINOGENIC
	Lopinavir	Glipizide
	Lutropin alfa	Gonadorelin
	Lymecycline	Gramicidin
	Meclozine	Guaifenesin
	Medroxyprogesterone + estrogen	Hydrocortisone
	Metoclopramide	Hydroxycarbamide
	Metronidazole	Hydroxychloroquine
	Metyrapone	Ibutilide
	Moxonidine	Imatinib
	Nandrolone	Indomethacin
	Nefazodone	Ketobemidone + DDBA
	Nelfinavir	Ketoconazole
	Nevirapine	Ketorolac
	Nifedipine	Lamotrigine
	Nimodipine	Letrozole
	Nitrazepam	Levodopa + benserazide
	Norethisterone	Levonorgestrel intra-
	Nortriptyline	uterine
	Oxcarbazepine	Levosimendan
	Oxytetracycline	Lidocaine
	Paclitaxel	Linezolid
	Paroxetine	Lofepramine
	Phenazone + caffeine	Lomustine
	Pioglitazone	Malathion
	Probenecid	Maprotiline
	Progesterone, vaginal gel	Mebendazole
	Quinidine	Mefloquine
	Rabeprazole	Melperone
	Raloxifene	Melphalan
	Rifabutin	Mepenzolate
	Riluzole	Mepivacaine
	Risperidone	Mercaptopurine
	Rosiglitazone	Methadone
	Saquinavir	Methylprednisolone
	Selegiline	Methixene
	Simvastatin	Metolazone
	Sulfasalazine	Metronidazole
	Telithromycin	Mexiletine
	Terbinafine	Mianserin
	Terfenadine	Midazolam
	Testosterone, transdermal patch	Minoxidil
	Tetracycline	Mirtazapine
	Theophylline	Mitomycin
	Thiamazole	Mitoxantrone
	Tibolone	Moclobemide
	Ticlopidine	Montelukast
	Tinidazole	Morphine + scopolamine
	Thiotepa	Multivitamins
	Topiramate	Mupirocin
	Topotecan	Nabumetone
	Toremifene	Nafarelin
	Tramadol	Naltrexone
	Trimegestone + estrogen	Nateglinide
	Verapamil	Nilutamide
	Voriconazole	Noscapine
	Zidovudine/AZT	Omeprazole
		Oxybutynin
		Oxycodone
		Pantoprazole
		Papaverine

Note: Based on list in "Patient's and Doctor's Guide to Medication in Acute Porphyria," Swedish Porphyria Association and Porphyria Centre Sweden. Also see the website Drug Database for Acute Porphyrias (*www.drugs-porphyria.com*) for a searchable list of safe and unsafe drugs.

μmol/24 h) (normal, 0–4 mg/24 h, [0–18 μmol/24 h]), and urinary ALA excretion is 20–100 mg/24 h (150–760 μmol/24 h) (normal, 1–7 mg/24 h [8–53 μmol/24 h]). Because levels often remain high after symptoms resolve, the diagnosis of an acute attack in a patient with biochemically proven AIP is based primarily on clinical features. Excretion of ALA and PBG decreases over a few days after intravenous hemin administration. A normal urinary PBG level before hemin effectively excludes AIP as a cause for current symptoms. Fecal porphyrins are usually normal or minimally increased in AIP, in contrast to HCP and VP. Most AIP heterozygotes with no history of symptoms have normal urinary excretion of ALA and PBG and are classified as latent. Patients can also have high levels of urine PBG and ALA with no clinical symptoms. These patients may have a previous history of an acute attack. These patients are classified as "asymptomatic high excretors" (ASHE). Therefore, the detection of the family's *HMBS* mutation will diagnose asymptomatic family members.

Patients with *HMBS* mutations in the initiation of translation codon in exon 1 and in the intron 15′-splice donor site have normal enzyme levels in erythrocytes and deficient activity only in nonerythroid tissues. This occurs because the erythroid and housekeeping forms of HMB-synthase are encoded by a single gene, which has two promoters. Thus, the enzyme assay may not be diagnostic, and genetic testing should be used to confirm the diagnosis.

More than 410 *HMBS* mutations have been identified in AIP, including missense, nonsense, and splicing mutations and insertions and deletions, with most mutations found in only one or a few families (Human Gene Mutation Database, *www.hgmd.org*). The prenatal diagnosis of a fetus at risk can be made with cultured amniotic cells or chorionic villi. However, this is seldom done, because the prognosis of individuals with *HMBS* mutations is generally favorable.

TREATMENT

Acute Intermittent Porphyria

During acute attacks, narcotic analgesics may be required for abdominal pain, and phenothiazines are useful for nausea, vomiting, anxiety, and restlessness. Chloral hydrate can be given for insomnia, and benzodiazepines are probably safe in low doses if a minor tranquilizer is required. Carbohydrate loading, usually with intravenous glucose (at least 300 g daily), may be effective in milder acute attacks of porphyria (without paresis, hyponatremia, etc.) if hemin is not available. Intravenous hemin is more effective and should be used as first-line therapy for all acute attacks. The standard regimen is 3–4 mg/kg of heme, in the form of lyophilized hematin (Recordati Rare Diseases), heme albumin (hematin reconstituted with human albumin), or heme arginate (Orphan Europe), infused daily for 4 days. Heme arginate and heme albumin are chemically stable and are less likely than hematin to produce phlebitis or an anticoagulant effect. Recovery depends on the degree of neuronal damage and usually is rapid if therapy is started early. Recovery from severe motor neuropathy may require months or years. Identification and avoidance of inciting factors can hasten recovery from an attack and prevent future attacks. Inciting factors are usually multiple, and removal of one or more hastens recovery and helps prevent future attacks. Frequent attacks that occur during the luteal phase of the menstrual cycle may be prevented with a gonadotropin-releasing hormone analogue, which prevents ovulation and progesterone production, or by prophylactic hematin administration.

The long-term risk of hypertension and chronic renal disease is increased in AIP; a number of patients have undergone successful renal transplantation. Chronic, low-grade abnormalities in liver function tests are common, and the risk of hepatocellular carcinoma is increased. Hepatic imaging is recommended at least yearly for early detection of these tumors. Other long-term complications include neuropathy, chronic pain, nausea, depression, and/or anxiety.

Orthotopic liver transplantation (OLT) has been successful and is curative in patients with severe, disabling, intractable attacks that are refractory to hemin therapy. Reports from both the UK and the U.S. show a marked improvement with no subsequent attacks, an improvement in the neuropathic manifestations, and normalization of the urinary PBG and ALA levels after liver transplantation. OLT is associated with morbidity and mortality and should be considered a treatment of last resort in these patients. In addition, patients who already have advanced neuropathy are considered poor risks for transplantation. Some patients with both recurrent attacks and end-stage renal disease have benefitted from combined liver and kidney transplantation.

Liver-directed gene therapy has proven successful in the prevention of drug-induced biochemical attacks in a murine model of human AIP, and clinical trials of adeno-associated virus vector (AAV)-*HMBS* gene transfer have been initiated. Although the therapy was safe, there was essentially no biochemical evidence of its effectiveness, nor did it prevent the recurrent attacks in the treated patients. Recent data from Phase 1 clinical trials of a hepatic-targeted RNA interference (RNAi) therapy directed to inhibit the markedly elevated levels of the hepatic *ALAS1* mRNA in patients with high levels of ALA and PBG showed reduced levels of the *ALAS1* mRNA and markedly decreased urinary ALA and PBG concentrations. In AIP patients with recurrent attacks, early studies indicate that the RNAi therapy reduced the frequency of acute attacks.

PORPHYRIA CUTANEA TARDA

PCT, the most common of the porphyrias, can be either sporadic (type 1) or familial (type 2) and can also develop after exposure to halogenated aromatic hydrocarbons. Hepatic URO-decarboxylase is deficient in all types of PCT, and for clinical symptoms to manifest, this enzyme deficiency must be substantial (~20% of normal activity or less); it is currently attributed to generation of an URO-decarboxylase inhibitor in the liver, which forms uroporphomethene in the presence of iron and under conditions of oxidative stress. The majority of PCT patients (~80%) have no *UROD* mutations and are said to have sporadic (type 1) disease. PCT patients heterozygous for *UROD* mutations have familial (type 2) PCT. In these patients, inheritance of a *UROD* mutation from one parent results in half-normal enzyme activity in liver and all other tissues, which is a significant predisposing factor, but is insufficient by itself to cause symptomatic PCT. As discussed below, other genetic and environmental factors contribute to susceptibility for both types of PCT. Because penetrance of the genetic trait is low, many patients with familial (type 2) PCT have no family history of the disease. HEP is an autosomal recessive form of porphyria that results from the marked systemic deficiency of URO-decarboxylase activity with clinical symptoms in childhood.

Clinical Features Blistering skin lesions that appear most commonly on the backs of the hands are the major clinical feature **(Fig. 409-3)**. These rupture and crust over, leaving areas of atrophy and scarring. Lesions may also occur on the forearms, face, legs, and feet. Skin friability and small white papules termed milia are common, especially on the backs of the hands and fingers. Hypertrichosis and hyperpigmentation, especially of the face, are especially troublesome in women. Occasionally, the skin over sun-exposed areas becomes severely thickened, with scarring and calcification that resembles systemic sclerosis. Neurologic features are absent.

A number of susceptibility factors, in addition to inherited *UROD* mutations in type 2 PCT, can be recognized clinically and can affect management. These include hepatitis C, HIV, excess alcohol, elevated iron levels, and estrogens. The importance of excess hepatic iron as a precipitating factor is underscored by the finding that the incidence of the common hemochromatosis-causing mutations, hemochromatosis gene (*HFE*) mutations C282Y and H63D, are increased in patients with types 1 and 2 PCT **(Chap. 407)**. Excess alcohol is a long-recognized contributor, as is estrogen use in women. HIV is probably an independent but less common risk factor that, like hepatitis C, does not cause PCT in isolation. Multiple susceptibility factors that appear to act synergistically can be identified in the individual PCT patient. Patients

FIGURE 409-3 Typical cutaneous lesions in a patient with porphyria cutanea tarda. Chronic, crusted lesions resulting from blistering due to photosensitivity on the dorsum of the hand of a patient with porphyria cutanea tarda. *(Courtesy of Dr. Karl E. Anderson; with permission.)*

with PCT characteristically have chronic liver disease and sometimes cirrhosis and are at risk for hepatocellular carcinoma. Various chemicals can also induce PCT; an epidemic of PCT occurred in eastern Turkey in the 1950s as a consequence of wheat contaminated with the fungicide hexachlorobenzene. PCT also occurs after exposure to other chemicals, including di- and trichlorophenols and 2,3,7,8-tetrachlorodibenzo-(*p*)-dioxin (TCDD, dioxin).

Diagnosis Porphyrins are increased in the liver, plasma, urine, and stool. The urinary ALA level may be slightly increased, but the PBG level is normal. Urinary porphyrins consist mostly of uroporphyrins and heptacarboxylate porphyrin, with lesser amounts of coproporphyrin and hexa- and pentacarboxylate porphyrins. Plasma porphyrins are also increased, and fluorometric scanning of diluted plasma at neutral pH can rapidly distinguish VP and PCT (Table 409-3). Isocoproporphyrins, which are increased in feces and sometimes in plasma and urine, are diagnostic for hepatic URO-decarboxylase deficiency.

Type 2 PCT and HEP can be distinguished from type 1 by finding decreased URO-decarboxylase in erythrocytes. URO-decarboxylase activity in liver, erythrocytes, and cultured skin fibroblasts in type 2 PCT is ~50% of normal in affected individuals and in family members with latent disease. In HEP, the URO-decarboxylase activity is markedly deficient, with typical levels of 3–10% of normal. Over 120 mutations have been identified in the *UROD* gene (Human Gene Mutation Database; *www.hgmd.org*). Of the mutations listed in the database, ~65% are missense or nonsense and ~10% are splice-site mutations. Most *UROD* mutations have been identified in only one or two families.

TREATMENT

Porphyria Cutanea Tarda

Alcohol, estrogens, iron supplements, and, if possible, any drugs that may exacerbate the disease should be discontinued, but this step does not always lead to improvement. A complete response can almost always be achieved by the standard therapy, repeated phlebotomy, to reduce hepatic iron. A unit (450 mL) of blood can be removed every 1–2 weeks. The aim is to gradually reduce excess hepatic iron until the serum ferritin level reaches the lower limits of normal. Because iron overload is not marked in most cases, remission may occur after only five or six phlebotomies; however, PCT patients with hemochromatosis may require more treatments to bring their iron levels down to the normal range. To document improvement in PCT, it is most convenient to follow the total plasma porphyrin concentration, which becomes normal sometime after the target ferritin level is reached. Hemoglobin levels or hematocrits and serum ferritin should be followed closely to prevent development of iron deficiency and anemia. After remission, continued phlebotomy may not be needed. Plasma porphyrin levels are followed at 6- to 12-month intervals for early detection of recurrences, which are treated by additional phlebotomy.

An alternative when phlebotomy is contraindicated or poorly tolerated is a low-dose regimen of chloroquine or hydroxychloroquine, both of which complex with the excess porphyrins and promote their excretion. Small doses (e.g., 125 mg chloroquine phosphate twice weekly) should be given, because standard doses can induce transient, sometimes marked increases in photosensitivity and hepatocellular damage. Recent studies indicate that low-dose hydroxychloroquine is as safe and effective as phlebotomy in PCT. Hepatic imaging can diagnose or exclude complicating hepatocellular carcinoma. Treatment of PCT in patients with end-stage renal disease is facilitated by administration of erythropoietin.

HEREDITARY COPROPORPHYRIA

HCP is an autosomal dominant hepatic porphyria that results from the half-normal activity of COPRO-oxidase. The disease presents with acute attacks, as in AIP. Cutaneous photosensitivity also may occur, but much less commonly than in VP. HCP patients may have acute attacks and cutaneous photosensitivity together or separately. HCP is less common than AIP and VP. Homozygous dominant HCP and harderoporphyria, a biochemically distinguishable variant of HCP, present with clinical symptoms in children (see below).

Clinical Features HCP is influenced by the same factors that cause attacks in AIP. The disease is latent before puberty, and symptoms, which are virtually identical to those of AIP, are more common in women. HCP is generally less severe than AIP. Blistering skin lesions are identical to PCT and VP and begin in childhood in rare homozygous cases.

Diagnosis COPRO III is markedly increased in the urine and feces in symptomatic patients, and often persists, especially in feces, when there are no symptoms. Urinary ALA and PBG levels are increased (but less than in AIP) during acute attacks, but may revert to normal more quickly than in AIP when symptoms resolve. Plasma porphyrins are usually normal or only slightly increased, but they may be higher in cases with skin lesions. The diagnosis of HCP is readily confirmed by increased fecal porphyrins consisting almost entirely of COPRO III, which distinguishes it from other porphyrias.

Although the diagnosis can be confirmed by measuring COPRO-oxidase activity, the assays for this mitochondrial enzyme are not available and require cells other than erythrocytes. To date, >65 mutations have been identified in the *CPOX* gene, 67% of which are missense or nonsense (Human Gene Mutation Database; *www.hgmd.org*). Detection of a *CPOX* mutation in a symptomatic individual permits the identification of asymptomatic family members.

TREATMENT

Hereditary Coproporphyria

Neurologic symptoms are treated as in AIP (see above). Phlebotomy and chloroquine are not effective for the cutaneous lesions.

VARIEGATE PORPHYRIA

VP is an autosomal dominant hepatic porphyria that results from the deficient activity of PROTO-oxidase, the seventh enzyme in the heme biosynthetic pathway, and can present with neurologic symptoms, photosensitivity, or both. VP is particularly common in South Africa, where 3 of every 1000 whites have the disorder. Most are descendants of a couple who emigrated from Netherlands to South Africa in 1688. In other countries, VP is less common than AIP. Rare cases of homozygous dominant VP, presenting in childhood with cutaneous symptoms, also have been reported.

Clinical Features VP can present with skin photosensitivity, acute neurovisceral crises, or both. In two large studies of VP patients, ~60% had only skin lesions, 20% had only acute attacks, and ~20% had both. Acute attacks are identical to those in AIP and are precipitated by the same factors as AIP (see above). Blistering skin manifestations are similar to those in PCT, but are more difficult to treat and usually are of longer duration. Homozygous VP is associated with photosensitivity, neurologic symptoms, and developmental disturbances, including growth retardation, in infancy or childhood; all cases had increased erythrocyte levels of zinc protoporphyrin, a characteristic finding in all homozygous porphyrias so far described.

Diagnosis Urinary ALA and PBG levels are increased during acute attacks, but may return to normal more quickly than in AIP. Increases in fecal protoporphyrin and COPRO III and in urinary COPRO III are more persistent. Plasma porphyrin levels also are increased, particularly when there are cutaneous lesions. VP can be distinguished rapidly from all other porphyrias by examining the fluorescence emission spectrum of porphyrins in plasma since VP has a unique fluorescence peak at neutral pH.

Assays of PROTO-oxidase activity in cultured fibroblasts or lymphocytes are not widely available. Over 180 mutations have been identified in the *PPOX* gene from unrelated VP patients (Human Gene Mutation Database; *www.hgmd.org*). The missense mutation R59W is the common mutation in most South Africans with VP of Dutch descent. Five missense mutations were common in English and French VP patients; however, most mutations have been found in only one or two families.

TREATMENT

Variegate Porphyria

Acute attacks are treated as in AIP, and hemin should be started early in most cases. Other than avoiding sun exposure, there are few effective measures for treating the skin lesions. β-Carotene, phlebotomy, and chloroquine are not helpful.

THE ERYTHROPOIETIC PORPHYRIAS

In the erythropoietic porphyrias, excess porphyrins from bone marrow erythrocyte precursors are transported via the plasma to the skin and lead to cutaneous photosensitivity.

X-LINKED SIDEROBLASTIC ANEMIA

XLSA results from the deficient activity of the erythroid form of ALA-synthase (ALA-synthase 2) and is associated with ineffective erythropoiesis, weakness, and pallor.

Clinical Features Typically, males with XLSA develop refractory hemolytic anemia, pallor, and weakness during infancy. They have secondary hypersplenism, become iron overloaded, and can develop hemosiderosis. The severity depends on the level of residual erythroid ALA-synthase activity and on the responsiveness of the specific mutation to pyridoxal 5′-phosphate supplementation (see below). Peripheral blood smears reveal a hypochromic, microcytic anemia with striking anisocytosis, poikilocytosis, and polychromasia; the leukocytes and platelets appear normal. Hemoglobin content is reduced, and the mean corpuscular volume and mean corpuscular hemoglobin concentration are decreased. Patients with milder, late-onset disease have been reported recently.

Diagnosis Bone marrow examination reveals hypercellularity with a left shift and megaloblastic erythropoiesis with an abnormal maturation. A variety of Prussian blue-staining sideroblasts are observed. Levels of urinary porphyrin precursors and of both urinary and fecal porphyrins are normal. The activity of erythroid ALA-synthase 2 is decreased in bone marrow, but this enzyme is difficult to measure in the presence of the normal ALA-synthase 1 housekeeping enzyme. Definitive diagnosis requires the demonstration of mutations in the erythroid *ALAS2* gene.

TREATMENT

X-Linked Sideroblastic Anemia

The severe anemia may respond to pyridoxine supplementation. This cofactor is essential for ALA-synthase activity, and mutations in the pyridoxine binding site of the enzyme have been found in several responsive patients. Cofactor supplementation may make it possible to eliminate or reduce the frequency of transfusion. Unresponsive patients may be transfusion dependent and require chelation therapy.

CONGENITAL ERYTHROPOIETIC PORPHYRIA

CEP, also known as Günther's disease, is an autosomal recessive disorder. It is due to the markedly deficient, but not absent, activity of URO-synthase and the resultant accumulation of URO I and COPRO I isomers. CEP is associated with hemolytic anemia and cutaneous lesions.

Clinical Features Severe cutaneous photosensitivity typically begins from birth. The skin over light-exposed areas is friable, and bullae and vesicles are prone to rupture and infection. Skin thickening, focal hypo- and hyperpigmentation, and hypertrichosis of the face and extremities are characteristic. Secondary infection of the cutaneous lesions can lead to disfigurement of the face and hands. Porphyrins are deposited in teeth and in bones. As a result, the teeth are brownish and fluoresce on exposure to long-wave ultraviolet light. Hemolysis is due to the marked increase in erythrocyte porphyrins and leads to splenomegaly. Adults with a milder later-onset form of the disease also have been described.

Diagnosis URO and COPRO (mostly type I isomers) accumulate in the bone marrow, erythrocytes, plasma, urine, and feces. The predominant porphyrin in feces is COPRO I. The diagnosis of CEP can be confirmed by demonstration of markedly deficient URO-synthase activity and/or by the identification of specific mutations in the *UROS* gene. The disease can be detected in utero by measuring porphyrins in amniotic fluid and URO-synthase activity in cultured amniotic cells or chorionic villi, or by the detection of the family's specific gene mutations. Molecular analyses of the mutant alleles from unrelated patients have revealed the presence of >50 mutations in the *UROS* gene, including four in the erythroid-specific promoter of the *UROS* gene. Genotype/phenotype correlations can predict the severity of the disease. The CEP phenotype may be modulated by sequence variations in the erythroid-specific ALA-synthase 2, mutation of which typically causes XLP. One mutation (p.ArgR216WTrp) in *GATA1*, encoding the X-linked erythroid-specific transcription factor GATA binding protein 1 (*GATA1*), has been identified in an individual with CEP, thrombocytopenia, and β-thalassemia.

TREATMENT

Congenital Erythropoietic Porphyria

Severe cases often require transfusions for anemia. Chronic transfusions of sufficient blood to suppress erythropoiesis are effective in reducing porphyrin production but result in iron overload. Splenectomy may reduce hemolysis and decrease transfusion requirements. Protection from sunlight and from minor skin trauma is important. Complicating bacterial infections should be treated promptly. Recently, bone marrow and cord blood transplantation has proven curative in several transfusion-dependent children, providing the rationale for stem cell gene therapy.

ERYTHROPOIETIC PROTOPORPHYRIA

EPP is an autosomal recessive disorder resulting from the deficient activity of FECH, the last enzyme in the heme biosynthetic pathway. EPP is the most common erythropoietic porphyria in children and, after PCT, the second most common porphyria in adults. EPP patients have FECH activities as low as 15–25% of normal in lymphocytes and

cultured fibroblasts. Protoporphyrin accumulates in bone marrow reticulocytes and then appears in plasma, is taken up in the liver, and is excreted in bile and feces. Protoporphyrin transported to the vessels in the skin causes the nonblistering phototoxicity. In most symptomatic patients (~90%) with this disorder, a deleterious mutation in one *FECH* allele was inherited with the relatively common (~10% of Caucasians) intronic 3 (IVS3) alteration (IVS3–48T>C) on the other allele that results in the low expression of the normal enzyme. In about 2% of EPP families, two *FECH* deleterious mutations have been found.

XLP is a less common condition with the same phenotype in affected males, including increased erythrocyte protoporphyrin levels resulting from gain-of-function mutations in the last exon of the erythroid-specific form of 5-aminolevulinate-synthase 2 (*ALAS2*). These mutations delete ALAS2 C-terminal amino acids resulting in its increased activity and the subsequent accumulation of protoporphyrin. Manifestations in female heterozygotes with XLP can range from asymptomatic to as severe as their affected male relatives. The variation in the presence and severity of manifestations in XLP heterozygotes results primarily from random X-chromosomal inactivation. XLP accounts for ~2–10% of cases with the EPP phenotype in Europe and North America.

Clinical Features In EPP and male XLP patients, skin photosensitivity, which differs from that in other porphyrias, usually begins in childhood and consists of pain, tingling, and itching occurring within minutes of sunlight exposure **(Fig. 409-4)**. Photosensitivity is associated with substantial elevations in erythrocyte protoporphyrin and occurs only in patients with genotypes that result in FECH activities below ~35% of normal. Vesicular lesions are uncommon. Redness, swelling, burning, and itching can develop shortly after sun exposure and resemble angioedema. Pain symptoms may seem out of proportion to the visible skin involvement. Sparse vesicles and bullae occur in ~10% of cases. Chronic skin changes may include lichenification, leathery pseudovesicles, labial grooving, and nail changes. Severe scarring is rare, as are pigment changes, friability, and hirsutism. Unless hepatic or other complications develop, protoporphyrin levels and symptoms of photosensitivity remain remarkably stable over many years in most patients. Factors that exacerbate the hepatic porphyrias play little or no role in EPP or XLP.

The primary source of excess protoporphyrin is the bone marrow reticulocytes. Erythrocyte protoporphyrin is free (not complexed with zinc) and is mostly bound to hemoglobin. In plasma, protoporphyrin is bound to albumin. Hemolysis and anemia are usually absent or mild.

Although EPP is an erythropoietic porphyria, up to 20% of EPP patients may have minor abnormalities of liver function, and in about 5% of these patients the accumulation of protoporphyrins causes chronic liver disease that can progress to liver failure requiring transplantation. Protoporphyrin is insoluble, and excess amounts form crystalline structures in liver cells (Fig. 409-4) and can decrease hepatic bile flow. Studies in the mouse model of EPP have shown that the bile duct epithelium may be damaged by toxic bile, leading to biliary fibrosis. Thus, rapidly progressive liver disease appears to be related to the cholestatic effects of protoporphyrins and is associated with increasing hepatic protoporphyrin levels due to impaired hepatobiliary excretion and increased photosensitivity. The hepatic complications also are often characterized by increasing levels of protoporphyrins in erythrocytes and plasma as well as severe abdominal and back pains, especially in the right upper quadrant. Gallstones composed at least in part of protoporphyrin occur in some patients. Hepatic complications appear to be higher in autosomal recessive EPP due to two *FECH* mutations and in males with XLP.

Diagnosis A substantial increase in erythrocyte protoporphyrin, which is predominantly free and not complexed with zinc, is the hallmark of EPP. Protoporphyrin levels are also variably increased in bone marrow, plasma, bile, and feces. Erythrocyte protoporphyrin concentrations are increased in other conditions such as lead poisoning, iron deficiency, various hemolytic disorders, all homozygous forms of other porphyrias, and sometimes even in acute porphyrias. In all these conditions, however, in contrast to EPP, protoporphyrin is complexed with zinc. Therefore, after an increase in erythrocyte protoporphyrin is found in a suspected EPP patient, it is important to confirm the diagnosis by an assay that distinguishes free and zinc-complexed protoporphyrin. Erythrocytes in EPP also exhibit red fluorescence under a fluorescence microscopy at 620 nm. Urinary levels of porphyrins and porphyrin precursors are normal. FECH activity in cultured lymphocytes or fibroblasts is decreased (<30% of normal mean). DNA diagnosis by mutation analysis is recommended to detect the causative *FECH* mutation(s) and/or the presence of the IVS3–48T>C low expression allele. To date, >190 mutations have been identified in the *FECH* gene, many of which result in an unstable or absent enzyme protein (null alleles) (Human Gene Mutation Database; *www.hgmd.org*).

In XLP, the erythrocyte protoporphyrin levels appear to be higher than in EPP and the proportions of free and zinc protoporphyrins may reach 50%. To date, four *ALAS2* mutations, three deletions of one to four bases, and one novel nonsense mutation have been described, which markedly increase ALA-synthase 2 activity and cause XLP. XLP accounts for about 2% of patients with the EPP phenotype in Western Europe. Recent studies show that about 10% of North American patients with the EPP phenotype have XLP.

FIGURE 409-4 Erythema and edema of the hands due to acute photosensitivity in a 10-year-old boy with erythropoietic protoporphyria. *(From P Poblette-Gutierrez et al: Eur J Dermatol 16:230, 2006.)*

TREATMENT

Erythropoietic Protoporphyria

Avoiding sunlight exposure and wearing clothing designed to provide protection for conditions with chronic phototoxicity are essential. Various other treatments, including oral β-Carotene, have proven of little benefit. Afamelanotide, an α-melanocyte-stimulating hormone (MSH) analogue, that stimulates tanning, has been approved for the treatment of EPP and XLP in the European Union by the European Medicines Agency. Approval by the U.S. Food and Drug Administration is pending at this time.

Treatment of hepatic complications, which may be accompanied by motor neuropathy, is difficult. Cholestyramine and other porphyrin absorbents such as activated charcoal may interrupt the enterohepatic circulation of protoporphyrin and promote its fecal excretion, leading to some improvement. Splenectomy may be helpful when the disease is accompanied by hemolysis and significant splenomegaly. Plasmapheresis and intravenous hemin are sometimes beneficial.

Liver transplantation has been carried out in some EPP and XLP patients with severe liver complications and is often successful in the short term. However, the disease often recurs in the transplanted liver due to continued bone marrow production of excess

protoporphyrin. In a retrospective study of 17 liver-transplanted EPP patients, 11 (65%) had recurrent EPP liver disease. Posttransplantation treatment with hematin and plasmapheresis should be considered to prevent the recurrence of liver disease. However, bone marrow transplantation, which has been successful in human EPP and which prevented liver disease in a mouse model, should be considered after liver transplantation, if a suitable donor can be found.

ACKNOWLEDGMENT

The authors thank Dr. Karl E. Anderson for his review of the manuscript and helpful comments and suggestions. This work is supported in part by the Porphyrias Consortium (U54 DK083909), a part of the National Institutes of Health (NIH) Rare Disease Clinical Research Network (RDCRN), supported through collaboration between the NIH Office of Rare Diseases Research (ORDR) at the National Center for Advancing Translational Science (NCATS), and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. MB is supported by a career development award K23-DK-095946.

FURTHER READING

ANDERSON KE et al: Recommendations for the diagnosis and treatment of the acute porphyrias. Ann Intern Med 142:439, 2005.

BALWANI M, DESNICK RJ: The porphyrias: Advances in diagnosis and treatment. Blood 120:4496, 2012.

BONKOVSKY HL et al: Acute porphyrias in the USA: Features of 108 subjects from porphyrias consortium. Am J Med 127:1233, 2015.

CHEN B et al: Acute intermittent porphyria: Predicted pathogenicity of HMBS variants indicates extremely low penetrance of the autosomal dominant disease. Hum Mutat 37:1215, 2016.

PUY H et al: Porphyrias. Lancet 375:924, 2010.

LANGENDONK JG et al: Afamelanotide for erythropoietic protoporphyria. N Engl J Med 373:48, 2015.

PISCHIK E, KAUPPINEN R: An update of clinical management of acute intermittent porphyria. Appl Clin Genet 8:201, 2015.

SCHULENBURG-BRAND D et al: The cutaneous porphyrias. Dermatol Clin 32:369, 2014.

STEIN PE et al: Update review of the acute porphyrias. Br J Haematol 176:527, 2017.

YASUDA M et al: AAV8-mediated gene therapy prevents induced biochemical attacks of intermittent porphyria and improves neuromotor function. Mol Ther 18:17, 2010.

YASUDA M et al: Liver transplantation for acute intermittent porphyria: Biochemical and pathologic studies of the explanted liver. Mol Med 21:487, 2015.

410 Disorders of Purine and Pyrimidine Metabolism

John N. Mecchella, Christopher M. Burns

Purines (adenine and guanine) and pyrimidines (cytosine, thymine, uracil) serve fundamental roles in the replication of genetic material, gene transcription, protein synthesis, and cellular metabolism. Disorders that involve abnormalities of nucleotide metabolism range from relatively common diseases such as hyperuricemia and gout, in which there is increased production or impaired excretion of a metabolic end product of purine metabolism (uric acid), to rare enzyme deficiencies that affect purine and pyrimidine synthesis or degradation. Understanding these biochemical pathways has led, in some instances, to the development of specific forms of treatment, such as the use of allopurinol and febuxostat to reduce uric acid production.

URIC ACID METABOLISM

Uric acid is the final breakdown product of purine degradation in humans. It is a weak diprotic acid with pK_a values of 5.75 and 10.3. Urates, the ionized forms of uric acid, predominate in plasma, extracellular fluid, and synovial fluid, with ~98% existing as monosodium urate at pH 7.4.

Plasma is saturated with monosodium urate at a concentration of 405 µmol/L (6.8 mg/dL) at 37°C. At higher concentrations, plasma is therefore supersaturated—a situation that creates the potential for urate crystal precipitation. However, plasma urate concentrations can reach 4800 µmol/L (80 mg/dL) without precipitation, perhaps because of the presence of solubilizing substances.

The pH of urine greatly influences the solubility of uric acid. At pH 5.0, urine is saturated with uric acid at concentrations ranging from 360 to 900 µmol/L (6–15 mg/dL). At pH 7.0, saturation is reached at concentrations from 9840 to 12,000 µmol/L (158–200 mg/dL). Ionized forms of uric acid in urine include monosodium, disodium, potassium, ammonium, and calcium urates.

Although purine nucleotides are synthesized and degraded in all tissues, urate is produced only in tissues that contain xanthine oxidase, primarily the liver and small intestine. Urate production varies with the purine content of the diet and with rates of purine biosynthesis, degradation, and salvage **(Fig. 410-1)**. Normally, two-thirds to three-fourths of urate is excreted by the kidneys, and most of the remainder is eliminated through the intestines.

The kidneys clear urate from the plasma and maintain physiologic balance by utilizing specific organic anion transporters (OATs), including urate transporter 1 (URAT1, SLC22A12) **(Fig. 410-2)**. In humans, OAT1 (SLC22A6), OAT2 (SLC22A7), and OAT3 (SLC22A8) are located on the basolateral membrane of renal proximal tubule cells. OAT4 (SLC22A11), OAT10 (SLC22A13), and URAT1 are located on the apical brush-border membrane of these cells. The latter transporters carry urate and other organic anions into the tubular cells from the lumen in exchange for intracellular organic anions. Once inside the cell, urate must pass to the basolateral side of the lumen in a process controlled by voltage-dependent carriers, including glucose transporter 9 (GLUT9, SLC2A9). *Uricosuric* compounds **(Table 410-1)** directly inhibit URAT1 on the apical side of the tubular cell (so-called *cis*-inhibition). In contrast, *antiuricosuric* compounds (those that promote hyperuricemia), such as nicotinate, pyrazinoate, lactate, and other aromatic organic acids, serve as the exchange anion inside the cell, thereby stimulating anion exchange and urate reabsorption (*trans*-stimulation). The

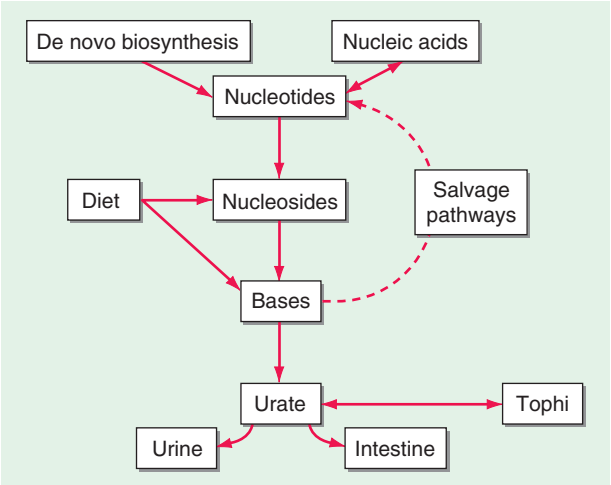


FIGURE 410-1 The total-body urate pool is the net result between urate production and excretion. Urate production is influenced by dietary intake of purines and the rates of de novo biosynthesis of purines from nonpurine precursors, nucleic acid turnover, and salvage by phosphoribosyltransferase activities. The formed urate is normally excreted by urinary and intestinal routes. Hyperuricemia can result from increased production, decreased excretion, or a combination of both mechanisms. When hyperuricemia exists, urate can precipitate and deposit in tissues as tophi.

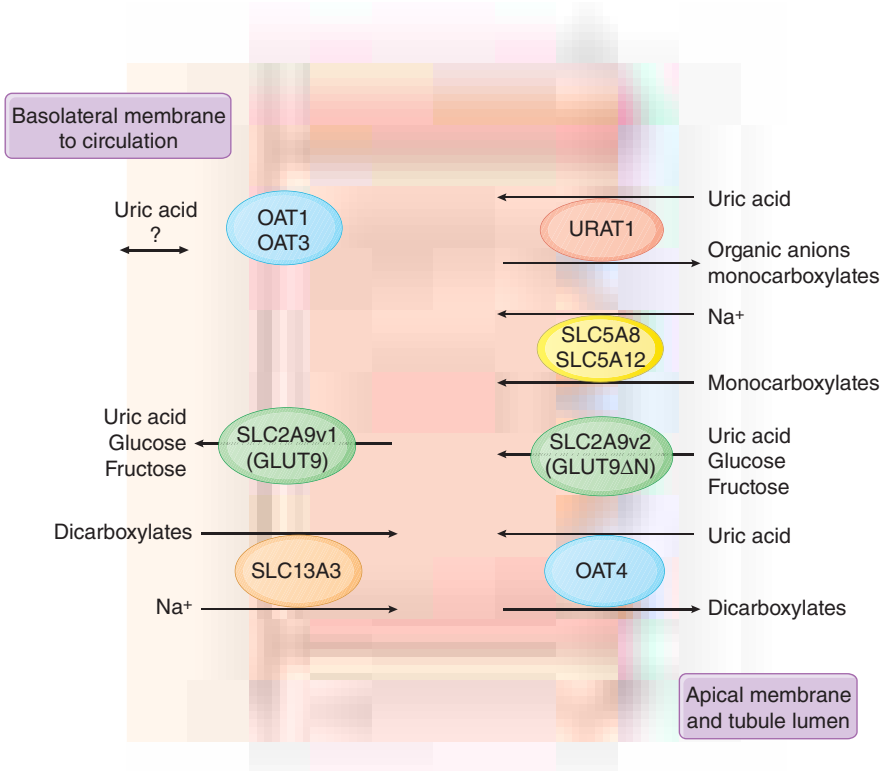


FIGURE 410-2 Schematic for handling of uric acid by the kidney. A complex interplay of transporters on both the apical and basolateral aspects of the renal tubule epithelial cell is involved in the reabsorption of uric acid. See text for details. Most uricosuric compounds inhibit URAT1 on the apical side, as well as OAT1, OAT3, and GLUT9 on the basolateral side.

activities of URAT1, other OATs, and sodium anion transporters result in excretion of 8–12% of the filtered urate as uric acid.

Most children have serum urate concentrations of 180–240 µmol/L (3–4 mg/dL). Levels begin to rise in males during puberty but remain low in females until menopause. The most recent mean serum urate values for men and premenopausal women in the United States are 415 and 360 µmol/L (6.14 and 4.87 mg/dL), respectively, according to National Health and Nutrition Evaluation Survey (NHANES) data for 2007–2008. After menopause, values for women increase to approximately those for men. In adulthood, concentrations rise steadily over time and vary with height, body weight, blood pressure, renal function, and alcohol intake.

HYPERURICEMIA

Hyperuricemia can result from increased production or decreased excretion of uric acid or from a combination of the two processes. Sustained hyperuricemia predisposes some individuals to develop clinical manifestations including gouty arthritis (Chap. 365), urolithiasis, and renal dysfunction (see below).

TABLE 410-1 Medications with Uricosuric Activity

Acetohexamide	Glyceryl guaiacolate
Adrenocorticotropic hormone	Glycopyrrolate
Ascorbic acid	Halofenate
Azauridine	Losartan
Benzbromarone	Meclofenamate
Calcitonin	Phenolsulfonphthalein
Chlorprothixene	Phenylbutazone
Citrate	Probenecid
Dicumarol	Radiographic contrast agents
Diflunisal	Salicylates (>2 g/d)
Estrogens	Sulfinpyrazone
Fenofibrate	Tetracycline that is outdated
Glucocorticoids	Zoxazolamine

In general, hyperuricemia is defined as a plasma (or serum) urate concentration >405 µmol/L (>6.8 mg/dL). The risk of developing gouty arthritis or urolithiasis increases with higher urate levels and escalates in proportion to the degree of elevation. The prevalence of hyperuricemia is increasing among ambulatory adults and even more markedly among hospitalized patients. The prevalence of gout in the United States more than doubled between the 1960s and the 1990s. Based on NHANES data from 2007 to 2008, these trends continue, with an approximate prevalence of gout among men of 5.9% (6.1 million) and among women of 2.0% (2.2 million). Mean serum urate levels rose to 6.14 mg/dL among men and 4.87 mg/dL among women, with consequent hyperuricemia prevalences of 21.2 and 21.6%, respectively (with *hyperuricemia* defined as a serum urate level of >7.0 mg/dL [415 µmol/L] for men and >5.7 mg/dL [340 µmol/L] for women). These numbers represent a 1.2% increase in the prevalence of gout, a 0.15-mg/dL increase in the serum urate level, and a 3.2% increase in the prevalence of hyperuricemia over figures reported in NHANES-III (1988–1994). These rises are thought to be driven by increased obesity and hypertension and perhaps also by better medical care and increased longevity.

CAUSES OF HYPERURICEMIA

Hyperuricemia may be classified as primary or secondary, depending on whether the cause is innate or an acquired disorder. However, it is more useful to classify hyperuricemia in relation to the underlying pathophysiology—i.e., whether it results from increased production, decreased excretion, or a combination of the two (Fig. 410-1, Table 410-2).

Increased Urate Production Diet contributes to the serum urate concentration in proportion to its purine content. Strict restriction of purine intake reduces the mean serum urate level by ~60 µmol/L (~1 mg/dL) and urinary uric acid excretion by ~1.2 mmol/d (~200 mg/d).

TABLE 410-2 Classification of Hyperuricemia by Pathophysiology

Urate Overproduction		
Primary idiopathic	Myeloproliferative diseases	Rhabdomyolysis
HPRT deficiency	Polycythemia vera	Exercise
PRPP synthetase overactivity	Psoriasis	Alcohol
Hemolytic processes	Paget's disease	Obesity
Lymphoproliferative diseases	Glycogenosis III, V, and VII	Purine-rich diet
Decreased Uric Acid Excretion		
Primary idiopathic	Starvation ketosis	Drug ingestion
Renal insufficiency	Berylliosis	Salicylates (<2 g/d)
Polycystic kidney disease	Sarcoidosis	Diuretics
Diabetes insipidus	Lead intoxication	Alcohol
Hypertension	Hyperparathyroidism	Levodopa
Acidosis	Hypothyroidism	Ethambutol
Lactic acidosis	Toxemia of pregnancy	Pyrazinamide
Diabetic ketoacidosis	Bartter's syndrome	Nicotinic acid
	Down syndrome	Cyclosporine
Combined Mechanism		
Glucose-6-phosphatase deficiency	Fructose-1-phosphate aldolase deficiency	Alcohol
		Shock

Abbreviations: HPRT, hypoxanthine phosphoribosyltransferase; PRPP, phosphoribosylpyrophosphate.

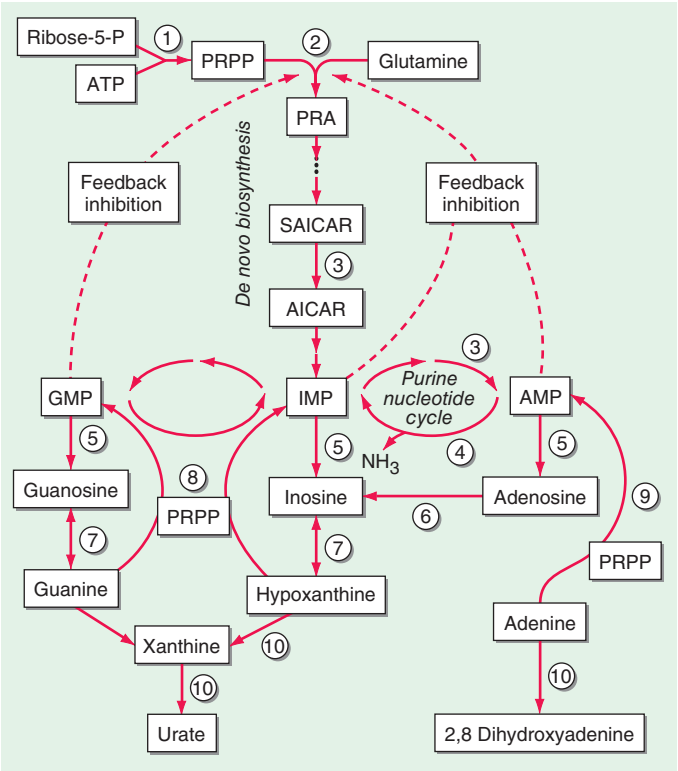


FIGURE 410-3 Abbreviated scheme of purine metabolism. (1) Phosphoribosylpyrophosphate (PRPP) synthetase, (2) amidophosphoribosyltransferase (amidoPRT), (3) adenylosuccinate lyase, (4) (myo-)adenylate (AMP) deaminase, (5) 5′-nucleotidase, (6) adenosine deaminase, (7) purine nucleoside phosphorylase, (8) hypoxanthine phosphoribosyltransferase (HPRT), (9) adenine phosphoribosyltransferase (APRT), and (10) xanthine oxidase. AICAR, aminoimidazole carboxamide ribotide; ATP, adenosine triphosphate; GMP, guanylate; IMP, inosine monophosphate; PRA, phosphoribosylamine; SAICAR, succinylaminoimidazole carboxamide ribotide.

Foods high in nucleic acid content include liver, "sweetbreads" (i.e., thymus and pancreas), kidney, and anchovy.

Endogenous sources of purine production also influence the serum urate level **(Fig. 410-3)**. De novo purine biosynthesis is a multistep process that forms inosine monophosphate (IMP). The rates of purine biosynthesis and urate production are predominantly determined by amidophosphoribosyltransferase (amidoPRT), which combines phosphoribosylpyrophosphate (PRPP) and glutamine. A secondary regulatory pathway is the salvage of purine bases by hypoxanthine phosphoribosyltransferase (HPRT). HPRT catalyzes the combination of the purine bases hypoxanthine and guanine with PRPP to form the respective ribonucleotides IMP and guanosine monophosphate (GMP).

Serum urate levels are closely coupled to the rates of de novo purine biosynthesis, which is driven in part by the level of PRPP, as evidenced by two X-linked inborn errors of purine metabolism **(Table 410-3)**. Both increased PRPP synthetase activity and HPRT deficiency are associated with overproduction of purines, hyperuricemia, and hyperuricaciduria (see below for clinical descriptions).

Accelerated purine nucleotide degradation can also cause hyperuricemia—i.e., with conditions of rapid cell turnover, proliferation, or cell death, as in leukemic blast crises, cytotoxic therapy for malignancy, hemolysis, or rhabdomyolysis. Hyperuricemia can result from excessive degradation of skeletal muscle ATP after strenuous physical exercise or status epilepticus and in glycogen storage disease types III, V, and VII **(Chap. 412)**. The hyperuricemia of myocardial infarction, smoke inhalation, and acute respiratory failure may also be related to accelerated breakdown of ATP.

Decreased Uric Acid Excretion More than 90% of individuals with sustained hyperuricemia have a defect in the renal handling of uric acid. For any given plasma urate concentration, patients who have gout excrete ~40% less uric acid than those who do not. When plasma urate levels are raised by purine ingestion or infusion, uric acid excretion increases in patients with and without gout; however, in those with gout, plasma urate concentrations must be 60–120 μmol/L (1–2 mg/dL) higher than normal to achieve equivalent uric acid excretion rates.

Diminished uric acid excretion could theoretically result from decreased glomerular filtration, decreased tubular secretion, or enhanced tubular reabsorption. Decreased urate filtration does not appear to cause primary hyperuricemia but does contribute to the hyperuricemia of renal insufficiency. Although hyperuricemia is invariably present in chronic renal disease, the correlation among serum creatinine, urea nitrogen, and urate concentrations is poor. Extrarenal clearance of uric acid increases as renal damage becomes more severe.

Many agents that cause hyperuricemia exert their effects by stimulating reabsorption rather than inhibiting secretion. This stimulation appears to occur through a process of "priming" renal urate reabsorption through the sodium-dependent loading of proximal tubular epithelial cells with anions capable of *trans*-stimulating urate reabsorption. The sodium-coupled monocarboxyl transporters SMCT1 and 2 (SLC5A8, SLC5A12) in the brush border of the proximal tubular cells mediate sodium-dependent loading of these cells with

TABLE 410-3 Inborn Errors of Purine Metabolism

ENZYME	ACTIVITY	INHERITANCE	CLINICAL FEATURES	LABORATORY FEATURES
Hypoxanthine phosphoribosyltransferase	Complete deficiency	X-linked	Self-mutilation, choreoathetosis, gout, and uric acid lithiasis	Hyperuricemia, hyperuricosuria
	Partial deficiency	X-linked	Gout and uric acid lithiasis	Hyperuricemia, hyperuricosuria
Phosphoribosylpyrophosphate synthetase	Overactivity	X-linked	Gout, uric acid lithiasis, and deafness	Hyperuricemia, hyperuricosuria
Adenine phosphoribosyltransferase	Deficiency	Autosomal recessive	2,8-Dihydroxyadenine lithiasis	—
Xanthine oxidase	Deficiency	Autosomal recessive	Xanthinuria and xanthine lithiasis	Hypouricemia, hypouricosuria
Adenylosuccinate lyase	Deficiency	Autosomal recessive	Autism and psychomotor retardation	—
Myoadenylate deaminase	Deficiency	Autosomal recessive	Myopathy with exercise intolerance or asymptomatic	—
Adenosine deaminase	Deficiency	Autosomal recessive	Severe combined immunodeficiency disease and chondro-osseous dysplasia	—
Purine nucleoside phosphorylase	Deficiency	Autosomal recessive	T cell–mediated immunodeficiency	—

monocarboxylates. A similar transporter, SLC13A3, mediates sodium-dependent influx of dicarboxylates into the epithelial cell from the basolateral membrane. Some of these carboxylates are well known to cause hyperuricemia, including pyrazinoate (from pyrazinamide treatment), nicotinate (from niacin therapy), and the organic acids lactate, β-hydroxybutyrate, and acetoacetate. The mono- and divalent anions then become substrates for URAT1 and OAT4, respectively, and are exchanged for uric acid from the proximal tubule. Increased blood levels of these anions result in their increased glomerular filtration and greater reabsorption by proximal tubular cells. The increased intraepithelial cell concentrations lead to increased uric acid reabsorption by promoting URAT1-, OAT4-, and OAT10-dependent anion exchange. Low doses of salicylates also promote hyperuricemia by this mechanism. Sodium loading of proximal tubular cells also provokes urate retention by reducing extracellular fluid volume and increasing angiotensin II, insulin, and parathyroid hormone release. Additional OAT1, OAT2, and OAT3 are involved in the movement of uric acid through the basolateral membrane, although the detailed mechanisms are still being elucidated.

GLUT9 (SLC2A9) is an electrogenic hexose transporter with splicing variants that mediate co-reabsorption of uric acid along with glucose and fructose at the apical membrane (GLUT9ΔN/SLC2A9v2) as well as through the basolateral membrane (SLC2A9v1) and thus into the circulation. GLUT9 has recently been identified as a high-capacity urate transporter, with rates 45–60 times faster than its glucose/fructose transport activity. GLUT9 may be responsible for the observed association of the consumption of fructose-sweetened soft drinks with an increased risk of hyperuricemia and gout. Genome-wide association scanning suggests that polymorphisms in SLC2A9 may play an important role in susceptibility to gout in the Caucasian population. The presence of one predisposing variant allele increases the relative risk of developing gout by 30–70%, most likely by increasing expression of the shorter isoform, SLC2A9v2 (GLUT9ΔN). Notably, though, genetic polymorphisms explain only ~6% of the differences in serum uric acid levels in Caucasians. Clearly, gout is polygenic and complex, and at this time the utility of genetic testing for relevant polymorphisms remains investigational and of no clinical utility.

Alcohol promotes hyperuricemia because of increased urate production and decreased uric acid excretion. Excessive alcohol consumption accelerates hepatic breakdown of ATP to increase urate production. Alcohol consumption can also induce hyperlacticacidemia, which blocks uric acid secretion. The higher purine content in some alcoholic beverages may also be a factor. Consumption of beer confers a greater risk of gout than liquor, and moderate wine intake does not increase gout risk. Intake of red meat and fructose increases the risk of gout, whereas intake of low-fat dairy products, purine-rich vegetables, whole grains, nuts and legumes, less sugary fruits, coffee, and vitamin C reduces the risk.

EVALUATION

Hyperuricemia does not necessarily represent a disease, nor is it a specific indication for therapy. The decision to treat depends on the cause and the potential consequences of hyperuricemia in each individual.

Quantification of uric acid excretion can be used to determine whether hyperuricemia is caused by overproduction or decreased excretion. On a purine-free diet, men with normal renal function excrete <3.6 mmol/d (600 mg/d). Thus, the hyperuricemia of individuals who excrete uric acid above this level while on a purine-free diet is due to purine overproduction; for those who excrete lower amounts on the purine-free diet, it is due to decreased excretion. If the assessment is performed while the patient is on a regular diet, the level of 4.2 mmol/d (800 mg/d) can be used as the discriminating value.

COMPLICATIONS

The most recognized complication of hyperuricemia is *gouty arthritis*. NHANES 2007–2008 found a prevalence of gout among U.S. adults of 3.9%, with figures of ~6% for men and ~2% for women. The higher the serum urate level, the more likely an individual is to develop gout. In one study, the incidence of gout was 4.9% among individuals with serum urate concentrations >540 μmol/L (>9.0 mg/dL) as opposed to only 0.5% among those with values between 415 and 535 μmol/L (7.0 and 8.9 mg/dL). The complications of gout correlate with both the duration and the severity of hyperuricemia. **For further discussion of gout, see Chap. 365**.

Hyperuricemia also causes several renal problems: (1) nephrolithiasis; (2) urate nephropathy, a rare cause of renal insufficiency attributed to monosodium urate crystal deposition in the renal interstitium; and (3) uric acid nephropathy, a reversible cause of acute renal failure resulting from deposition of large amounts of uric acid crystals in the renal collecting ducts, pelvis, and ureters.

Nephrolithiasis Uric acid nephrolithiasis occurs most commonly, but not exclusively, in individuals with gout. In gout, the prevalence of nephrolithiasis correlates with the serum and urinary uric acid levels, reaching ~50% with serum urate levels of 770 μmol/L (13 mg/dL) or urinary uric acid excretion >6.5 mmol/d (1100 mg/d).

Uric acid stones can develop in individuals with no evidence of arthritis, only 20% of whom are hyperuricemic. Uric acid can also play a role in other types of kidney stones. Some individuals who do not have gout but have calcium oxalate or calcium phosphate stones have hyperuricemia or hyperuricaciduria. Uric acid may act as a nidus on which calcium oxalate can precipitate or lower the formation product for calcium oxalate crystallization.

Urate Nephropathy Urate nephropathy, sometimes referred to as *urate nephrosis*, is a late manifestation of severe gout and is characterized histologically by deposits of monosodium urate crystals surrounded by a giant-cell inflammatory reaction in the medullary interstitium and pyramids. The disorder is now rare and cannot be diagnosed in the absence of gouty arthritis. The lesions may be clinically silent or cause proteinuria, hypertension, and renal insufficiency.

Uric Acid Nephropathy This reversible cause of acute renal failure is due to precipitation of uric acid in renal tubules and collecting ducts that obstructs urine flow. Uric acid nephropathy develops following sudden urate overproduction and marked hyperuricaciduria. Factors that favor uric acid crystal formation include dehydration and acidosis. This form of acute renal failure occurs most often during an aggressive "blastic" phase of leukemia or lymphoma prior to or coincident with cytolytic therapy but has also been observed in individuals with other neoplasms, following epileptic seizures, and after vigorous exercise with heat stress. Autopsy studies have demonstrated intraluminal precipitates of uric acid, dilated proximal tubules, and normal glomeruli. The initial pathogenic events are believed to include obstruction of collecting ducts with uric acid and obstruction of the distal renal vasculature.

If recognized, uric acid nephropathy is potentially reversible. Appropriate therapy has reduced the mortality rate from ~50% to near zero. Serum levels cannot be relied on for diagnosis because this condition has developed in the presence of urate concentrations varying from 720 to 4800 μmol/L (12–80 mg/dL). The distinctive feature is the urinary uric acid concentration. In most forms of acute renal failure with decreased urine output, urinary uric acid content is either normal or reduced, and the ratio of uric acid to creatinine is <1. In acute uric acid nephropathy, the ratio of uric acid to creatinine in a random urine sample or a 24-h specimen is >1, and a value that high is essentially diagnostic.

HYPERURICEMIA AND METABOLIC SYNDROME

Metabolic syndrome **(Chap. 401)** is characterized by abdominal obesity with visceral adiposity, impaired glucose tolerance due to insulin resistance with hyperinsulinemia, hypertriglyceridemia, increased low-density lipoprotein cholesterol, decreased high-density lipoprotein cholesterol, and hyperuricemia. Hyperinsulinemia reduces the renal excretion of uric acid and sodium. Not surprisingly, hyperuricemia resulting from euglycemic hyperinsulinemia may precede the onset

of type 2 diabetes, hypertension, coronary artery disease, and gout in individuals with metabolic syndrome.

TREATMENT

Hyperuricemia

ASYMPTOMATIC HYPERURICEMIA

Hyperuricemia is present in ~21% of the population and in at least 25% of hospitalized individuals. The vast majority of hyperuricemic persons are at no clinical risk. In the past, the association of hyperuricemia with cardiovascular disease and renal failure led to the use of urate-lowering agents for patients with asymptomatic hyperuricemia. This practice is no longer recommended except for individuals receiving cytolytic therapy for neoplastic disease, who are treated with urate-lowering agents in an effort to prevent uric acid nephropathy. Because hyperuricemia can be a component of the metabolic syndrome, its presence is an indication to screen for and aggressively treat any accompanying obesity, hyperlipidemia, diabetes mellitus, or hypertension.

Hyperuricemic individuals, especially those with higher serum urate levels, are at risk for the development of gouty arthritis. However, most hyperuricemic persons never develop gout, and prophylactic treatment is not indicated. Furthermore, neither structural kidney damage nor tophi are identifiable before the first attack. Reduced renal function cannot be attributed to asymptomatic hyperuricemia, and treatment of asymptomatic hyperuricemia does not alter the progression of renal dysfunction in patients with renal disease. An increased risk of stone formation in those with asymptomatic hyperuricemia has not been established.

Thus, because treatment with specific antihyperuricemic agents entails inconvenience, cost, and potential toxicity, routine treatment of asymptomatic hyperuricemia cannot be justified other than for prevention of acute uric acid nephropathy. In addition, routine screening for asymptomatic hyperuricemia is not recommended. If hyperuricemia is diagnosed, however, the cause should be determined. Causal factors should be corrected if the condition is secondary, and associated problems such as hypertension, hypercholesterolemia, diabetes mellitus, and obesity should be treated.

SYMPTOMATIC HYPERURICEMIA

See Chap. 365 for treatment of gout, including urate nephrosis.

Nephrolithiasis Antihyperuricemic therapy is recommended for the individual who has both gouty arthritis and either uric acid– or calcium-containing stones, both of which may occur in association with hyperuricaciduria. Regardless of the nature of the calculi, fluid ingestion should be sufficient to produce a daily urine volume >2 L. Alkalinization of the urine with sodium bicarbonate or acetazolamide may be justified to increase the solubility of uric acid. Specific treatment of uric acid calculi requires reducing the urine uric acid concentration with a xanthine oxidase inhibitor, such as allopurinol or febuxostat. These agents decrease the serum urate concentration and the urinary excretion of uric acid in the first 24 h, with a maximal reduction within 2 weeks. Allopurinol can be given once a day because of the long half-life (18 h) of its active metabolite, oxypurinol. In the febuxostat trials, the generally recommended dose of allopurinol (300 mg/d) was effective at achieving a target serum urate concentration <6.0 mg/dL (357 μmol/L) in <50% of patients; this result suggested that higher doses should be considered. Allopurinol is effective in patients with renal insufficiency, but the dose should be reduced. Allopurinol is also useful in reducing the recurrence of calcium oxalate stones in patients with gout and in individuals with hyperuricemia or hyperuricaciduria who do not have gout. Febuxostat (40–80 mg/d) is also taken once daily, and doses do not need to be adjusted in the presence of mild to moderate renal dysfunction. Potassium citrate (30–80 mmol/d orally in divided doses) is an alternative therapy for patients with uric acid stones alone or mixed calcium/uric acid stones. A xanthine oxidase inhibitor is also indicated for the treatment of 2,8-dihydroxyadenine kidney stones.

Uric Acid Nephropathy Uric acid nephropathy is often preventable, and immediate appropriate therapy has greatly reduced the mortality rate. Vigorous IV hydration and diuresis with furosemide dilute the uric acid in the tubules and promote urine flow to ≥100 mL/h. The administration of acetazolamide (240–500 mg every 6–8 h) and sodium bicarbonate (89 mmol/L) IV enhances urine alkalinity and thereby solubilizes more uric acid. It is important to ensure that the urine pH remains >7.0 and to watch for circulatory overload. In addition, antihyperuricemic therapy in the form of allopurinol in a single dose of 8 mg/kg is administered to reduce the amount of urate that reaches the kidney. If renal insufficiency persists, subsequent daily doses should be reduced to 100–200 mg because oxypurinol, the active metabolite of allopurinol, accumulates in renal failure. Despite these measures, hemodialysis may be required. Urate oxidase (rasburicase) can also be administered IV to prevent or to treat tumor lysis syndrome.

HYPOURICEMIA

Hypouricemia, defined as a serum urate concentration <120 μmol/L (<2.0 mg/dL), can result from decreased production of urate, increased excretion of uric acid, or a combination of both mechanisms. This condition occurs in <0.2% of the general population and <0.8% of hospitalized individuals. Hypouricemia causes no symptoms or pathology and therefore requires no therapy.

Most hypouricemia results from increased renal uric acid excretion. The finding of normal amounts of uric acid in a 24-h urine collection from an individual with hypouricemia is evidence for a renal cause. Medications with uricosuric properties (Table 410-1) include aspirin (at doses >2.0 g/d), losartan, fenofibrate, x-ray contrast materials, and glyceryl guaiacolate. Total parenteral hyperalimentation can also cause hypouricemia, possibly a result of the high glycine content of the infusion formula. Other causes of increased urate clearance include conditions such as neoplastic disease, hepatic cirrhosis, diabetes mellitus, and inappropriate secretion of vasopressin; defects in renal tubular transport such as primary Fanconi syndrome and Fanconi syndromes caused by Wilson's disease, cystinosis, multiple myeloma, and heavy metal toxicity; and isolated congenital defects in the bidirectional transport of uric acid. Hypouricemia can be a familial disorder that is generally inherited in an autosomal recessive manner. Most cases are caused by a loss of function mutation in *SLC22A12*, the gene that encodes URAT-1, resulting in increased renal urate clearance. Individuals with normal *SLC22A12* most likely have a defect in other urate transporters. Although hypouricemia is usually asymptomatic, some patients suffer from urate nephrolithiasis or exercise-induced renal failure.

SELECTED INBORN ERRORS OF PURINE AND PYRIMIDINE METABOLISM

(See also Table 410-3, **Table 410-4**, Fig. 410-3, and **Fig. 410-4)** More than 30 defects in human purine and pyrimidine metabolic pathways have been identified thus far. Many are benign, but about half are associated with clinical manifestations, some causing major morbidity and mortality. Advances in genetics, along with high-performance liquid chromatography and tandem mass spectrometry, have facilitated diagnosis.

PURINE DISORDERS

HPRT Deficiency The HPRT gene is located on the X chromosome. Affected males are hemizygous for the mutant gene; carrier females are asymptomatic. A complete deficiency of HPRT, the Lesch-Nyhan syndrome, is characterized by hyperuricemia, self-mutilative behavior, choreoathetosis, spasticity, and mental retardation. A partial deficiency of HPRT, the Kelley-Seegmiller syndrome, is associated with hyperuricemia but no central nervous system manifestations. In both disorders, the hyperuricemia results from urate overproduction and can cause

TABLE 410-4 Inborn Errors of Pyrimidine Metabolism

ENZYME	ACTIVITY	INHERITANCE	CLINICAL FEATURES	LABORATORY FEATURES
Uridine-5′-monophosphate synthetase	Deficiency	Autosomal recessive	Orotic acid crystalluria; obstructive uropathy, hypochromic megaloblastic anemia	Orotic aciduria
Pyrimidine 5′-nucleotidase	Deficiency	Autosomal recessive	Hemolytic anemia	Basophilic stippling of erythrocytes; high levels of cytidine and uridine ribonucleotides
Pyrimidine 5′-nucleotidase	Superactivity	Uncertain	Developmental delay, seizures, ataxia, language deficit	Hypouricosuria
Thymidine phosphorylase	Deficiency	Autosomal recessive	Mitochondrial neurogastrointestinal encephalopathy	Hypouricosuria
Dihydropyrimidine dehydrogenase	Deficiency	Autosomal recessive	Seizures, motor and mental retardation	High levels of uracil, thymine, and 5-hydroxymethyluracil and low levels of dihydropyrimidines in urine
Dihydropyrimidinase	Deficiency	Uncertain	Seizures, mental retardation	Dihydropyrimidinuria
Ureidopropionase	Deficiency	Uncertain	Hypotonia, dystonia, developmental delay	High urinary excretion of *N*-carbamyl-β-alanine and *N*-carbamyl β-aminoisobutyric acid

uric acid crystalluria, nephrolithiasis, obstructive uropathy, and gouty arthritis. Early diagnosis and appropriate therapy with allopurinol can prevent or eliminate all the problems attributable to hyperuricemia without affecting behavioral or neurologic abnormalities.

Increased PRPP Synthetase Activity Like the HPRT deficiency states, PRPP synthetase overactivity is X-linked and results in gouty arthritis and uric acid nephrolithiasis. Neurologic hearing loss occurs in some families.

Adenine Phosphoribosyltransferase (APRT) Deficiency APRT deficiency is inherited as an autosomal recessive trait. Affected individuals develop kidney stones composed of 2,8-dihydroxyadenine. Caucasians with the disorder have a complete deficiency (type I), whereas Japanese individuals have some measurable enzyme activity (type II). Expression of the defect is similar in the two populations, as is the frequency of the heterozygous state (0.4–1.1 per 100). Allopurinol treatment prevents stone formation.

Hereditary Xanthinuria A deficiency of xanthine oxidase causes all purine in the urine to occur in the form of hypoxanthine and xanthine. About two-thirds of deficient individuals are asymptomatic. The remainder develop kidney stones composed of xanthine.

Myoadenylate Deaminase Deficiency Primary (inherited) and secondary (acquired) forms of myoadenylate deaminase deficiency have been described. The primary form is inherited as an autosomal recessive trait. Clinically, some persons may have relatively mild myopathic symptoms with exercise or other triggers, but most individuals with this defect are asymptomatic. Therefore, another explanation for the myopathy should be sought in symptomatic patients with this deficiency. The acquired deficiency occurs in association with a wide variety of neuromuscular diseases, including muscular dystrophies, neuropathies, inflammatory myopathies, and collagen vascular diseases.

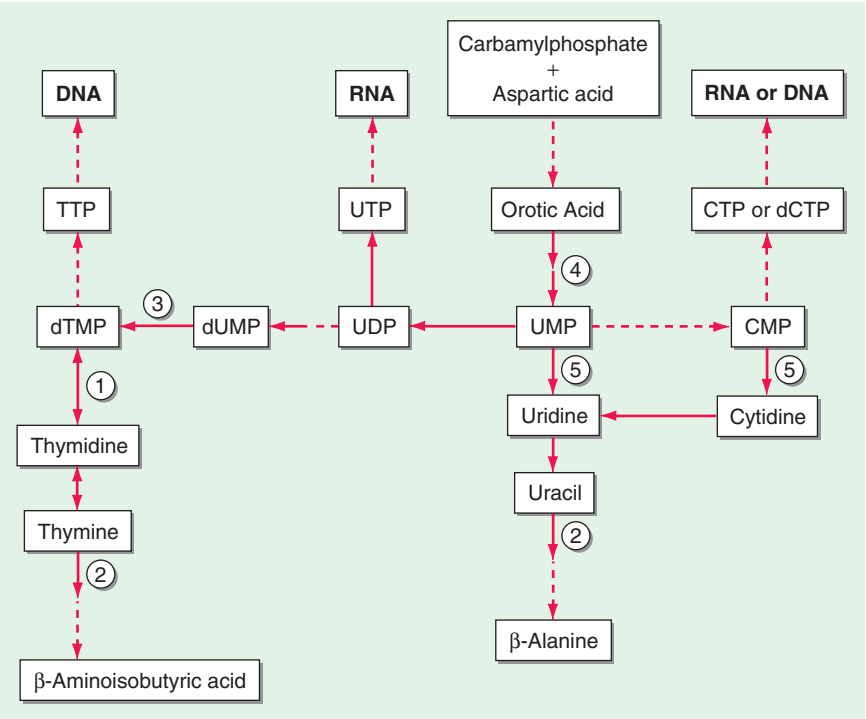


FIGURE 410-4 Abbreviated scheme of pyrimidine metabolism. (1) Thymidine kinase, (2) dihydropyrimidine dehydrogenase, (3) thymidylate synthase, (4) UMP synthase, (5) 5′-nucleotidase. CMP, cytidine-5′-monophosphate; dTMP, deoxythymidine-5′-monophosphate; dUMP, deoxyuridine-5′-monophosphate; TTP, thymidine triphosphate; UDP, uridine-5′-diphosphate; UMP, uridine-5′-monophosphate; UTP, uridine triphosphate.

Adenylosuccinate Lyase Deficiency Deficiency of this enzyme is due to an autosomal recessive trait and causes profound psychomotor retardation, seizures, and other movement disorders. All individuals with this deficiency are mentally retarded, and most are autistic.

Adenosine Deaminase Deficiency and Purine Nucleoside Phosphorylase Deficiency See Chap. 344.

PYRIMIDINE DISORDERS

The pyrimidine cytidine is found in both DNA and RNA; it is a complementary base pair for guanine. Thymidine is found only in DNA, where it is paired with adenine. Uridine is found only in RNA and can pair with either adenine or guanine in RNA secondary structures. Pyrimidines can be synthesized by a de novo pathway (Fig. 410-4) or reused in a salvage pathway. Although >25 different enzymes are involved in pyrimidine metabolism, disorders of these pathways are rare. Seven disorders of pyrimidine metabolism have been discovered (Table 410-4), three of which are discussed below.

Orotic Aciduria Hereditary orotic aciduria is caused by mutations in a bifunctional enzyme, uridine-5′-monophosphate (UMP) synthase, which converts orotic acid to UMP in the de novo

synthesis pathway (Fig. 410-4). The disorder is characterized by hypochromic megaloblastic anemia that is unresponsive to vitamin B_{12} and folic acid, growth retardation, and neurologic abnormalities. Increased excretion of orotic acid causes crystalluria and obstructive uropathy. Replacement of uridine (100–200 mg/kg per day) corrects anemia, reduces orotic acid excretion, and improves the other sequelae of the disorder.

Pyrimidine 5′-nucleotidase Deficiency Pyrimidine 5′-nucleotidase catalyzes the removal of the phosphate group from pyrimidine ribonucleoside monophosphates (cytidine-5′-monophosphate or UMP) (Fig. 410-4). An inherited deficiency of this enzyme causes hemolytic anemia with prominent basophilic stippling of erythrocytes. The accumulation of pyrimidines or cytidine diphosphate choline is thought to induce hemolysis. There is no specific treatment. Acquired pyrimidine 5′-nucleotidase deficiency has been reported in lead poisoning and in thalassemia.

Dihydropyrimidine Dehydrogenase Deficiency Dihydropyrimidine dehydrogenase is the rate-limiting enzyme in the pathway of uracil and thymine degradation (Fig. 410-4). Deficiency of this enzyme causes excessive urinary excretion of uracil and thymine. In addition, this deficiency causes nonspecific cerebral dysfunction with convulsive disorders, motor retardation, and mental retardation. No specific treatment is available.

Medication Effect on Pyrimidine Metabolism A variety of medications can influence pyrimidine metabolism. The anticancer agents fluorodeoxyuridine and 5-fluorouracil and the antimicrobial agent fluorocytosine cause cytotoxicity when converted to fluorodeoxyuridylate, a specific suicide inhibitor of thymidylate synthase. Fluorocytosine must be converted to 5-fluorouracil to be effective. This conversion is catalyzed by cytosine deaminase activity. Fluorocytosine's action is selective because cytosine deaminase is present in bacteria and fungi but not in human cells. Dihydropyrimidine dehydrogenase is involved in the degradation of 5-fluorouracil. Consequently, deficiency of this enzyme is associated with 5-fluorouracil neurotoxicity.

Leflunomide, which is used to treat rheumatoid arthritis, inhibits de novo pyrimidine synthesis by inhibiting dihydroorotate dehydrogenase, resulting in an antiproliferative effect on T cells. Allopurinol, which inhibits xanthine oxidase in the purine metabolic pathway, also inhibits the activity of orotidine-5′-phosphate decarboxylase, a step in UMP synthesis. Consequently, allopurinol use is associated with increased excretion of orotidine and orotic acid. There are no known clinical effects of this inhibition.

Acknowledgment

The authors are grateful to Robert L. Wortmann for contributions to this chapter in previous editions of the book.

FURTHER READING

Balasubramaniam S et al: Inborn errors of purine metabolism: Clinical update and therapies. J Inherit Metab Dis 37:669, 2014.

Balasubramaniam S et al: Inborn errors of pyrimidine metabolism: Clinical update and therapy. J Inherit Metab Dis 37:687, 2014.

Ben Salem C et al: Drug-induced hyperuricaemia and gout. Rheumatology (Oxford) 266:56:679, 2017.

Bhole V, Krishnan E: Gout and the heart. Rheum Dis Clin North Am 40:125, 2014.

Burns CM, Wortman RL: Clinical features and treatment of gout, in *Kelley and Firestein's Textbook of Rheumatology*, 10th ed, GS Firestein et al (eds). Philadelphia, Elsevier, 2017, pp 1620–1644.

Hirano M, Peters GJ: Advances in purine and pyrimidine metabolism in health and diseases. Nucleosides Nucleotides Nucleic Acids 35:495, 2016.

Terkeltaub R (ed): *Gout and Other Crystal Arthropathies*. Philadelphia, Elsevier Health Sciences, 2012.

411 Lysosomal Storage Diseases

Robert J. Hopkin, Gregory A. Grabowski

Lysosomes are heterogeneous subcellular organelles containing specific hydrolyses that allow selective processing or degradation of proteins, nucleic acids, carbohydrates, and lipids. There are more than 50 different lysosomal storage diseases (LSDs), classified according to the nature of the stored material (Table 411-1). Several of the most prevalent disorders are reviewed here: Tay-Sachs disease, Fabry disease, Gaucher disease, Niemann-Pick disease, lysosomal acid lipase deficiency (LALD), the mucopolysaccharidoses, and Pompe disease. LSDs should be considered in the differential diagnosis of patients with neurologic, renal, or muscular degeneration and/or unexplained hepatomegaly, splenomegaly, cardiomyopathy, or skeletal dysplasias and deformations. Physical findings are disease specific, and enzyme assays or genetic testing can be used to make a definitive diagnosis. Although the nosology of LSDs segregates the variants into distinct phenotypes, these are heuristic; in the clinic, each disease exhibits—to varying degrees—a continuous spectrum of manifestations, from severe to attenuated variants.

PATHOGENESIS

Lysosomal biogenesis involves ongoing synthesis of lysosomal hydrolases, membrane constitutive proteins, and new membranes. Lysosomes originate from the fusion of trans-Golgi network vesicles with late endosomes. Progressive vesicular acidification accompanies the maturation of these vesicles; this gradient facilitates the pH-dependent dissociation of receptors and ligands and also activates lysosomal hydrolases. Lysosomes are components of the lysosome/autophagy/mitophagy system, which are disrupted in the LSDs.

Abnormalities at any biosynthetic step can impair enzyme activation and lead to a lysosomal storage disorder. After leader sequence clipping, remodeling of complex oligosaccharides (including the lysosomal targeting ligand mannose-6-phosphate as well as high-mannose oligosaccharide chains of many soluble lysosomal hydrolases) occurs during transit through the Golgi. Lysosomal integral or associated membrane proteins are sorted to the membrane or interior of the lysosome by several different peptide signals. Phosphorylation, sulfation, additional proteolytic processing, and macromolecular assembly of heteromers occur concurrently. Such posttranslational modifications are critical to enzyme function, and defects can result in multiple enzyme/protein deficiencies.

The final common pathway for LSDs is the accumulation of specific macromolecules within tissues and cells that normally have a high flux of these substrates. The majority of lysosomal enzyme deficiencies result from point mutations or genetic rearrangements at a locus that encodes a single lysosomal hydrolase. However, some mutations cause deficiencies of several different lysosomal hydrolases by alteration of the enzymes/proteins involved in targeting, active site modifications, or macromolecular association or trafficking. All LSDs are inherited as autosomal recessive disorders except for Hunter (mucopolysaccharidosis type II), Danon, and Fabry diseases, which are X-linked. Substrate accumulation leads to lysosomal distortion/dysfunction, which has significant pathologic consequences. In addition, abnormal amounts of metabolites may also have pharmacologic effects important to disease pathophysiology and propagation, particularly activation of the innate immune responses.

For many LSDs, the accumulated substrates are endogenously synthesized within particular tissue sites of pathology. Other diseases have greater exogenous substrate supplies. For example, substrates are delivered by low-density lipoprotein receptor–mediated uptake in Fabry and LALD or by phagocytosis in Gaucher disease type 1. The *threshold hypothesis* refers to a level of enzyme activity below which disease develops; small changes in enzyme activity near the threshold can lead to or modify disease. A critical element of this model is that

TABLE 411-1 Selected Lysosomal Storage Diseases

DISORDER[a]	ENZYME DEFICIENCY [SPECIFIC THERAPY]	STORED MATERIAL	CLINICAL TYPES (ONSET)	INHERITANCE	CLINICAL FEATURES					
					NEUROLOGIC	LIVER, SPLEEN ENLARGEMENT	SKELETAL DYSPLASIA	OPHTHALMOLOGIC	HEMATOLOGIC	UNIQUE FEATURES
Mucopolysaccharidoses CTMucopolysaccharidoses (MPS)										
MPS I H, Hurler (136) MPS I H/S, Hurler/ Scheie	α-L-Iduronidase [ET, HSCT]	Dermatan sulfate Heparan sulfate	Infantile Intermediate Childhood/ Adult	AR	Cognitive degeneration Cognitive degeneration	+++	++++	Corneal clouding	Vacuolated lymphocytes	Coarse facies; cardiovascular involvement; joint stiffness
MPS I S, Scheie					None					
MPS II, Hunter (136)	Iduronate sulfatase [ET]	Dermatan sulfate Heparan sulfate	Severe infantile Mild juvenile	X-linked	Cognitive degeneration, less in mild form	+++	++++	Retinal degeneration, no corneal clouding	Granulated lymphocytes	Coarse facies; cardiovascular involvement; joint stiffness; distinctive pebbly skin lesions
MPS III A, Sanfilippo A (136)	Heparan-*N*-sulfatase	Heparan sulfate	Late infantile	AR	Severe Cognitive degeneration	+	+	None	Granulated lymphocytes	Mild coarse facies
MPS III B, Sanfilippo B (136)	*N*-Acetyl-α-glucosaminidase	Heparan sulfate	Late infantile	AR	Severe Cognitive degeneration	+	+	None	Granulated lymphocytes	Mild coarse facies
MPS III C, Sanfilippo C (136)	Acetyl-CoA: α-glucosaminide *N*-acetyltransferase	Heparan sulfate	Late infantile	AR	Severe Cognitive degeneration	+	+	None	Granulated lymphocytes	Mild coarse facies
MPS III D, Sanfilippo D (136)	*N*-Acetylglucosamine-6-sulfate sulfatase	Heparan sulfate	Late infantile	AR	Severe Cognitive degeneration	+	+	None	Granulated lymphocytes	Mild coarse facies
MPS IV A, Morquio A (136)	*N*-Acetylgalactosamine-6-sulfate sulfatase [ET—trials]	Keratan sulfate Chondroitin-6 sulfate	Childhood	AR	None	+	++++	Corneal clouding	Granulated neutrophils	Distinctive skeletal deformity; odontoid hypoplasia; aortic valve disease
MPS IV B, Morquio (136)	β-Galactosidase		Childhood	AR	None	±	++++			
MPS VI, Maroteaux-Lamy (136)	Arylsulfatase B [ET, BMT]	Dermatan sulfate	Late infantile	AR	None	++	++++	Corneal clouding	Granulated neutrophils and lymphocytes	Coarse facies; valvular heart disease
MPS VII (136)	β-Glucuronidase	Dermatan sulfate Heparan sulfate	Neonatal Infantile Adult	AR	Cognitive degeneration, absent in some adults	+++	+++	Corneal clouding	Granulated neutrophils	Coarse facies; vascular involvement; hydrops fetalis in neonatal form
GM_2 Gangliosidoses										
Tay-Sachs disease (153)	β-Hexosaminidase A	GM_2 gangliosides	Infantile Juvenile	AR	Cognitive degeneration; seizures; later juvenile form	None	None	Cherry red spot in infantile form	None	Macrocephaly; hyperacusis in infantile form
Sandhoff disease (153)	β-Hexosaminidases A and B	GM_2 gangliosides	Infantile	AR	Cognitive degeneration; seizures	++	±	Cherry red spot	None	Macrocephaly; hyperacusis

Neutral Glycosphingolipidoses										
Fabry disease (150)	α-Galactosidase A [ET]	Globotriaosylceramide	Childhood	X-linked	Painful acroparesthesias	None	None	Corneal dystrophy, vascular lesions	None	Cutaneous angiokeratomas; hypo-hydrosis
Gaucher disease (146)	Acid β-glucosidase [ET, SRT]	Glucosylceramide, glycosylsphingosine	Type 1 Type 2 Type 3	AR	None ++++ ++	++++ +++ ++++	++++ + ++++	None Eye movements Eye movements	Gaucher cells in bone marrow; cytopenias	Adult form highly variable
Niemann-Pick disease (144) A and B	Sphingomyelinase [ET—trials]	Sphingomyelin	Neuronopathic, type A Nonneuronopathic, type B	AR	Cognitive degeneration; seizures	++++	None Osteoporosis	Macular degeneration	Foam cells in bone marrow	Pulmonary infiltrates Lung failure
Glycoproteinoses										
Fucosidosis (140)	α-Fucosidase	Glycopeptides; oligosaccharides	Infantile Juvenile	AR	Cognitive degeneration	++	++	None	Vacuolated lymphocytes; foam cells	Coarse facies; angiokeratomas in juvenile form
α-Mannosidosis (140)	α-Mannosidase	Oligosaccharides	Infantile Milder variant	AR	Cognitive degeneration	+++	+++	Cataracts, corneal clouding	Vacuolated lymphocytes, granulated neutrophils	Coarse facies; enlarged tongue
β-Mannosidosis (140)	β-Mannosidase	Oligosaccharides		AR	Seizures; Cognitive degeneration		++	None	Vacuolated lymphocytes, foam cells	Angiokeratomas
Aspartylglucosaminuria (141)	Aspartylglucosaminidase	Aspartylglucosamine; glycopeptides	Young adult	AR	Cognitive degeneration	±	++	None	Vacuolated lymphocytes, foam cells	Coarse facies
Sialidosis (140)	Neuraminidase	Sialyloligosaccharides	Type I, congenital Type II, infantile and juvenile	AR	Myoclonus; Cognitive degeneration	++, less in type I	++, less in type I	Cherry red spot	Vacuolated lymphocytes	MPS phenotype in type II
Mucolipidoses (ML)										
ML-II, I-cell disease (138)	UDP-*N*-Acetylglucosamine-1-phosphotransferase	Glycoprotein; glycolipids	Infantile	AR	Cognitive degeneration	+	++++	Corneal clouding	Vacuolated and granulated neutrophils	Coarse facies; absence of mucopoly-sacchariduria; gingival hypoplasia
ML-III, pseudo-Hurler polydystrophy (138)	UDP-*N*-Acetylglucosamine-1-phosphotransferase	Glycoprotein; glycolipids	Late infantile	AR	Mild Cognitive degeneration	None	+++	Corneal clouding, mild retinopathy, hyperopic astigmatism		Coarse facies; stiffness of hands and shoulders
Leukodystrophies										
Krabbe disease (147)	Galactosylceramidase [BMT/HSCT]	Galactosylceramide Galactosylsphingosine	Infantile	AR	Cognitive degeneration	None	None	None	None	White matter globoid cells
Metachromatic leukodystrophy (148)	Arylsulfatase A	Cerebroside sulfate	Infantile Juvenile Adult	AR	Cognitive degeneration; dementia; psychosis in adult	None	None	Optic atrophy	None	Gait abnormalities in late infantile form
Multiple sulfatase deficiency (149)	Active site cysteine to C_{α}-formylglycine-converting enzyme	Sulfatides; mucopoly-saccharides	Late infantile	AR	Cognitive degeneration	+	++	Retinal degeneration	Vacuolated and granulated cells	Absent activity of all known cellular sulfatases

(*Continued*)

TABLE 411-1 Selected Lysosomal Storage Diseases (Continued)

DISORDER[a]	ENZYME DEFICIENCY [SPECIFIC THERAPY]	STORED MATERIAL	CLINICAL TYPES (ONSET)	INHERITANCE	CLINICAL FEATURES					
					NEUROLOGIC	LIVER, SPLEEN ENLARGEMENT	SKELETAL DYSPLASIA	OPHTHALMOLOGIC	HEMATOLOGIC	UNIQUE FEATURES
Disorders of Neutral Lipids										
Infantile-onset LALD (142)	Acid lysosomal lipase [ET]	Cholesteryl esters; triglycerides	Infantile	AR	None	+++	None	None	None	Adrenal calcification
Childhood/ Adult-onset LALD (142)	Acid lysosomal lipase [ET]	Cholesteryl esters	Childhood	AR	None	Hepatomegaly	None	None	None	Fatty liver disease; cirrhosis
Farber disease (143)	Acid ceramidase	Ceramide	Infantile Juvenile	AR	Occasional Cognitive degeneration	±	None	Macular degeneration	None	Arthropathy, subcutaneous nodules
Disorders of Glycogen										
Pompe disease (135)	Acid α-glucosidase [ET]	Glycogen	Infantile, late onset	AR	Neuromuscular	±	None	None	None	Myocardiopathy
Late onset GAA deficiency (135)	Acid α-glucosidase [ET]	Glycogen	Variable: juvenile to adulthood	AR	Neuromuscular	None	None	None	None	Respiratory insufficiency; neuromuscular disease
Danon disease (154)	LAMP-2 (lysosomal associated membrane protein-2)	Glycogen	Variable: childhood to adulthood	X-linked (?Dominant)	Cardiomyopathy Neuromuscular Inconsistent Cognitive degeneration	None	None	None	None	Myocardial vacuolar degeneration

[a]Numbers in parentheses refer to the chapters in CR Scriver et al: *The Online Metabolic and Molecular Bases of Inherited Disease*, New York, McGraw-Hill, *ommbid.mhmedical.com*, which provide comprehensive reviews.

Abbreviations: AR, autosomal recessive; BMT/HSCT, bone marrow or stem cell transplantation; ET, enzyme therapy; SRT, substrate reduction therapy.

enzymatic activity can be challenged by changes in substrate flux based on genetic background, cell turnover, recycling, or metabolic demands. Thus, a set level of residual enzyme may be adequate for substrate in some tissues or cells but not in others. In addition, several variants of each LSD exist at a clinical level. These disorders therefore represent a continuum of manifestations that are not easily dissociated into discrete entities. The molecular/genetic bases for such variations have not been elucidated in any detail.

SELECTED DISORDERS

TAY-SACHS DISEASE

About 1 in 30 Ashkenazi Jews is a carrier for Tay-Sachs disease (total hexosaminidase A [Hex A] deficiency), resulting from defective α-chains. The infantile form is a fatal neurodegenerative disease with macrocephaly, loss of motor skills, increased startle reaction, and a macular cherry red spot. The juvenile-onset form presents as ataxia and dementia, with death by age 10–15 years. The adult-onset disorder is characterized by clumsiness in childhood; progressive motor weakness in adolescence; and additional spinocerebellar and lower-motor-neuron signs and dysarthria in adulthood. Intelligence declines slowly, and psychiatric disorders are common. Screening for Tay-Sachs disease carriers is recommended in the Ashkenazi Jewish population. Sandhoff disease, due to a deficiency in both Hex A and Hex B resulting from defective β-chains, is phenotypically similar to Tay-Sachs disease, but also includes hepatosplenomegaly and bony dysplasias.

FABRY DISEASE

Fabry disease, an X-linked disorder, results from mutations in *GALA* that encodes α-galactosidase A. The estimated prevalence of hemizygous males ranges from 1 in 40,000 to 1 in 3500 in selected populations. Clinically, the disease manifests with angiokeratomas (telangiectatic skin lesions), hypohidrosis, corneal and lenticular opacities, acroparesthesia; and progressive small-vessel disease of the kidney, heart, and brain.

Angiokeratomas and acroparesthesias may appear in childhood. Angiokeratomas are punctate, dark red to blue-black, flat or slightly raised, and usually symmetric; they do not blanch with pressure. They are often small and can be easily overlooked. They usually are most dense between the umbilicus and the knees—the "bathing suit area"—but may occur anywhere, including the mucosal surfaces. Angiokeratomas also occur in several other very rare LSDs. Corneal and lenticular lesions, detectable on slit-lamp examination, may help in establishing a diagnosis of Fabry disease. Debilitating episodic burning pain of the hands, feet, and proximal extremities (acroparesthesia) can last from minutes to days and can be precipitated by changes in temperature, exercise, fatigue, or fever. Abdominal pain can resemble that from appendicitis or renal colic. Proteinuria, isosthenuria, and progressive renal dysfunction occur in the second to fourth decades; ~5% of male patients with idiopathic renal failure have *GALA* mutations. Hypertension, left ventricular hypertrophy, anginal chest pain, and congestive heart failure can occur in the third to fourth decades. About 1–3% of patients with idiopathic hypertrophic myocardiopathy have Fabry disease. Similarly, ~3–5% of male patients with idiopathic stroke at 35–50 years of age have *GALA* mutations. Leg lymphedema without hypoproteinemia and episodic diarrhea also occur. Death is due to renal failure or cardiovascular or cerebrovascular disease in untreated male patients. Variants with residual α-galactosidase A activity may have late-onset manifestations that are limited to the cardiovascular system and resemble hypertrophic cardiomyopathy. Variants with predominant cardiac, renal, or central nervous system (CNS) manifestations are becoming better defined. Up to 70% of heterozygous females exhibit clinical manifestations. However, in females, heart disease is the most common life-threatening manifestation, followed in frequency by stroke and renal disease.

Gabapentin and carbamazepine diminish chronic and episodic acroparesthesia. Chronic hemodialysis or kidney transplantation can be lifesaving in patients with renal failure. Intravenous enzyme therapy clears stored lipids from a variety of cells, particularly those of the renal, cardiac, and skin vascular endothelium. Renal insufficiency appears to be irreversible. Early institution of enzyme therapy may prevent or slow the progression of life-threatening complications.

GAUCHER DISEASE

Gaucher disease, an autosomal recessive disorder, results from defective activity of acid β-glucosidase; ~600 *GBA1* mutations have been described in such patients. Disease variants are classified by the absence or presence and progression of neuronopathic involvement.

Gaucher disease type 1 is a nonneuronopathic disease (i.e., absence of early-onset or progressive CNS disease) presenting in childhood to adulthood as slowly to rapidly progressive visceral disease. About 55–60% of patients are diagnosed at <20 years of age in white populations and at even younger ages in other groups. This pattern of presentation is distinctly bimodal, with peaks at <10–15 years and at ~25 years. Younger patients tend to have greater degrees of hepatosplenomegaly and accompanying blood cytopenias. In contrast, the older patients have a greater tendency for chronic bone disease. Hepatosplenomegaly occurs in virtually all symptomatic patients and can be minor or massive. Accompanying anemia and thrombocytopenia are variable and are not directly related to liver or spleen volumes. Severe liver dysfunction is unusual. Splenic infarctions can resemble an acute abdomen. Pulmonary hypertension and alveolar Gaucher cell accumulation are uncommon but life-threatening and can occur at any age. *GBA1* mutations in the hetero- or homozygous state are a significantly increased life-time risk for developing Parkinson disease.

All patients with Gaucher disease have nonuniform infiltration of bone marrow by lipid-laden macrophages termed *Gaucher cells*. This phenomenon can lead to marrow packing with subsequent infarction, ischemia, necrosis, and cortical bone destruction. Bone marrow involvement spreads from proximal to distal in the limbs and can involve the axial skeleton extensively, causing vertebral collapse. In addition to bone marrow involvement, bone remodeling is defective, with loss of total bone calcium leading to osteopenia, osteonecrosis, avascular infarction, and vertebral compression fractures with spinal cord involvement. Aseptic necrosis of the femoral head is common, as is fracture of the femoral neck. The mechanism by which diseased bone marrow macrophages interact with osteoclasts and/or osteoblasts to cause bone disease is not well understood. Chronic, ill-defined bone pain can be debilitating and poorly correlated with radiographic findings. "Bone crises" are associated with localized excruciating pain and, on occasion, local erythema, fever, and leukocytosis. These crises represent acute infarctions of bone, as evidenced in nuclear scans by localized absent uptake of pyrophosphate agents. Decreased acid β-glucosidase activity (0–20% of normal) in nucleated cells establishes the diagnosis. The enzyme is not normally present in bodily fluids. The sensitivity of enzyme testing is poor for heterozygote detection; molecular testing by *GBA1* sequencing is preferred. The disease frequency varies from about 1 in 1000 among Ashkenazi Jews to <1 in 100,000 in other populations; ~1 in 12–15 Ashkenazi Jews carries a Gaucher disease allele. Four common mutations account for ~85% of the mutations in that population of affected patients: N370S (1226G), 84GG (a G insertion at cDNA position 84), L444P (1448C), and IVS-2 (an intron 2 splice junction mutation).

Genotype/phenotype studies indicate a significant, though not absolute, correlation between disease type and severity and the *GBA1* genotype. The most common mutation in the Ashkenazi Jewish population (N370S) shares a 100% association with nonneuronopathic or type 1 Gaucher disease. The N370S/N370S and N370S/other mutant allele genotypes are associated with later-onset/less severe disease and with earlier-onset/severe disease, respectively. As many as 50–60% of individuals with the N370S/N370S genotype are asymptomatic. Other alleles include L444P (very low activity), 84GG (null), or IVS-2 (null) and rare/private or uncharacterized alleles. The L444P/L444P patients frequently have life-threatening to very severe/early-onset disease, and many, though not all, develop CNS involvement in the first two decades of life.

Symptom-based treatment of blood cytopenias and joint replacement surgeries continue to have important roles in management. However, regular intravenous enzyme therapy has been the first-line

treatment for significantly affected patients and is highly efficacious and safe in diminishing hepatosplenomegaly and improving hematologic values. An oral substrate reduction therapy (eliglustat tartrate), which inhibits glycolipid synthesis, is approved as a first-line therapy for adults. Bone disease is decreased and can be prevented, but irreversible damage cannot be reversed, by enzyme therapy. Adult patients may benefit from adjunctive treatment with bisphosphonates to improve bone density. Patients who cannot be treated with enzyme, either because it is not effective or because they have developed an allergy or other hypersensitivities to the enzyme, may receive *substrate reduction therapy* with either eliglustat tartrate or miglustat; the latter is approved as a second-line oral therapy.

Gaucher disease type 2 is a rare, severe, progressive CNS disease that leads to death by 2 years of age. Gaucher disease type 3 has highly variable manifestations in the CNS and viscera. It can present in early childhood with rapidly progressive, massive visceral disease and slowly progress to static CNS involvement; in adolescence with dementia; or in early adulthood with rapidly progressive, uncontrollable myoclonic seizures and mild visceral disease. Visceral disease in type 3 is nearly identical to that in type 1 but is generally more severe. Early CNS findings may be limited to defects in lateral gaze tracking, which may remain static for decades. Cognitive degeneration can be slowly progressive or static. Type 3 is much more frequent among individuals of non-Western World descent. Visceral—but not CNS—involvement responds to enzyme therapy.

NIEMANN-PICK DISEASE

Niemann-Pick diseases are autosomal recessive disorders that result from defects in acid sphingomyelinase. Types A and B are distinguished by the early age of onset and progressive CNS disease in type A. Type A typically has its onset in the first 6 months of life, with rapidly progressive CNS deterioration, spasticity, failure to thrive, and massive hepatosplenomegaly. Type B has a later, more variable onset and is characterized by a progression of hepatosplenomegaly, with eventual development of cirrhosis and hepatic replacement by foam cells. Affected patients develop progressive pulmonary disease with dyspnea, hypoxemia, and a reticular infiltrative pattern on chest x-ray. Foam cells are present in alveoli, lymphatic vessels, and pulmonary arteries. Progressive hepatic or lung disease can lead to death in adolescence or early adulthood.

The diagnosis is established by markedly decreased (1–10% of normal) sphingomyelinase activity in nucleated cells. There is no approved specific treatment for Niemann-Pick disease, but intravenous enzyme therapy clinical trials are in phase 3. The efficacy of hepatic or bone marrow transplantation has not been clearly established.

Niemann-Pick C diseases are progressive CNS diseases due to mutations in either *NPC1* or *NPC2*, lysosomal proteins involved in cholesterol transport out of the lysosome. They can present with liver or splenic disease, but their major manifestations are progressive CNS disease over one to two decades. Treatment with substrate inhibition agents (e.g., Miglustat) has shown some promise and substrate depletion with cyclodextrin is in clinical trials for NPC1 disease.

MUCOPOLYSACCHARIDOSES

Mucopolysaccharidosis type I (MPS I) is an autosomal recessive disorder caused by deficiency of α-L-iduronidase. The continuum of involvement traditionally has been divided into three categories: (1) Hurler disease (MPS I H) for severe deficiency with neurodegeneration, (2) Scheie disease (MPS I S) for later-onset disease without neurologic involvement and with relatively less severe disease in other organ systems, and (3) Hurler-Scheie syndrome (MPS I H/S) for patients intermediate between these extremes. MPS I H/S is characterized by severe somatic disease, usually without overt neurologic deterioration.

MPS I often presents in infancy or early childhood as chronic rhinitis, clouding of the corneas, and hepatosplenomegaly. As the disease progresses, nearly every organ system can be affected. In the more severe forms, cardiac and respiratory diseases become life threatening in childhood. Skeletal disease can be quite severe, resulting in very limited mobility.

There are two current treatments for the MPS I diseases. Hematopoietic stem cell transplantation (HSCT) is the standard treatment for patients presenting at <2 years of age who appear to have or are at risk for neurologic degeneration. HSCT results in stabilization of CNS disease and reverses hepatosplenomegaly. It also beneficially affects cardiac and respiratory disease. HSCT does not eliminate corneal disease or result in the resolution of progressive skeletal disease. Enzyme therapy effectively addresses hepatosplenomegaly and alleviates cardiac and respiratory disease. The enzyme does not effectively penetrate the CNS and does not directly affect CNS disease. Enzyme therapy and HSCT appear to have similar effects on visceral signs and symptoms. Enzyme therapy poses a lower risk of life-threatening complications and may therefore be advantageous for patients who have attenuated manifestations without CNS disease. A combination of enzyme therapy and HSCT has been used, with enzyme therapy initiated prior to transplantation in an attempt to reduce the disease burden. The experience with this approach is not well documented, but it appears to have advantages over HSCT alone.

Hunter disease (MPS II) is an X-linked disorder due to deficiency in iduronate sulfate sulfatase and has manifestations similar to those of MPS I, including neurologic degeneration. There is no corneal clouding or other eye disease. Like MPS I, MPS II is clinically variable, with CNS and non-CNS variants. HSCT has not been successful in treating CNS disease associated with MPS II. The FDA and the European Medicines Agency (EMA) have approved enzyme therapy for the visceral manifestations of MPS II.

MPS IV or Morquio syndrome is a rare autosomal recessive condition (1 to 200,000–300,000) and is different than the other mucopolysaccharidoses in presenting as a spondyloepiphyseal skeletal dysplasia. There are also important heart and respiratory complications. This disorder often presents in childhood, but the age of onset and rate of progression are quite variable. Two variants, type A and type B, are caused by deficiencies in *N*-acetyl-galactosamine-6-sulfatase (GALNS) and an acid β–galactosidase, respectively. A recombinant human GALNS enzyme replacement therapy (elosulfase alfa) is approved for the treatment of MPS IVA, making it is essential to confirm the specific diagnosis. Treatment has been shown to improve ambulatory mobility and decrease pain. There is no current specific treatment for MPS IVB.

Enzyme therapy for Maroteaux-Lamy disease (MPS VI), arylsulfatase B deficiency, has received U.S. Food and Drug Administration (FDA) approval as well as by similar agencies in other countries. This very rare autosomal recessive disorder is characterized by hepatosplenomegaly, bone disease, heart disease, and respiratory compromise. Short stature is also an important manifestation. Visceral signs and symptoms are similar to those in MPS I; however, MPS VI is not associated with neurologic degeneration.

MPS VII, Sly syndrome, is due to mutations in the *GUSB* gene, which codes for the β-glucuronidase enzyme. Severe deficiency in this enzyme may present with fetal hydrops which can lead to stillbirth or perinatal demise. Other patients with MPS VII may present later with short stature coarse facial features and hepatosplenomegaly. There is current research on enzyme replacement therapy for this disorder.

POMPE DISEASE

Acid maltase (acid α-glucosidase, GAA) deficiency, also called Pompe disease, is the only LSD leading to primary glycogen storage. The classic severe infantile form presents with hypotonia, myocardiopathy, and hepatosplenomegaly. This variant is rapidly progressive and generally results in death in the first year of life. However, as with other LSDs, there are early- and late-onset forms of this disorder. The late-onset variants may be as common as 1 in 40,000; patients typically present with a slowly progressive myopathy that may resemble limb-girdle muscular dystrophy. Respiratory insufficiency may be the presenting sign or may develop with advancing disease. In late stages of the disease, patients may require mechanical ventilation, report swallowing difficulties, and experience loss of bowel and bladder control. Myocardiopathy is not usually seen in late-onset variants of Pompe disease.

The FDA, EMA, and similar agencies have approved enzyme therapy for Pompe disease patients of all ages. This treatment clearly

prolongs life in the infantile form, consistently resulting in improved cardiac function. Respiratory function is also improved in most treated infants. Some infants demonstrate marked improvement in motor functions, while others have minor changes in muscle tone or strength. Prevention of deterioration has been shown with GAA enzyme therapy in the late-onset forms. Early intervention with GAA enzyme therapy in such patients may limit or prevent deterioration, but very advanced disease will have significant irreversible components.

LYSOSOMAL ACID LIPASE DEFICIENCY

Wolman syndrome (now infantile-onset LALD) and cholesterol ester storage disease (now childhood/adult-onset LALD) are caused by deficiency of lysosomal acid lipase due to autosomal recessive mutations in the *LIPA*. The enzyme is responsible for hydrolysis of cholesterol esters and triglycerides delivered to the lysosome via the LDLR pathway. Accumulation of these in the tissues leads to progressive organ dysfunction including liver disease, intestinal malabsorption, heart dysfunction, and other manifestations. The most severe form presents in early infancy with failure to thrive, vomiting, and hepatosplenomegaly. The infantile-onset LALD patients die without specific treatment by age 1 year (median age of death 3.7 months). Childhood/adult-onset LALD can have a variable age of initial presentation with nonspecific signs, but often involves elevated liver enzymes, nonalcoholic fatty liver disease, cryptogenic cirrhosis, and varying severities of hepatosplenomegaly. Disease progresses throughout life and may result in early (adolescence) liver cirrhosis and (early adulthood) atherosclerosis or early death without treatment. Importantly, statins can decrease the hypercholesterolemia, but do not alter the basic progressive tissue, (e.g., liver) pathology. Enzyme replacement therapy for LALD has major effects in reversing disease manifestations and was approved for patients at all ages by the EMA, FDA, and several other country agencies in 2015 and 2016.

FURTHER READING

Aldenhoven M et al: Long-term outcome of Hurler syndrome patients after hematopoietic cell transplantation: An international multicenter study. Blood 125:2164, 2015.

Balwani M et al: Recommendations for the use of eliglustat in the treatment of adults with Gaucher disease type 1 in the United States. Mol Genet Metab 117:95, 2016.

Hopkin RJ et al: The management and treatment of children with Fabry disease: A United States-based perspective; Fabry Pediatric Expert Panel. Mol Genet Metab 117:104, 2016.

Ortiz A et al: Time to treatment benefit for adult patients with Fabry disease receiving agalsidase β: Data from the Fabry Registry. J Med Genet. 53:495, 2016.

Schoser B et al: Survival and long-term outcomes in late-onset Pompe disease following alglucosidase alfa treatment: A systematic review and meta-analysis. J Neurol. 264:621, 2017.

412 Glycogen Storage Diseases and Other Inherited Disorders of Carbohydrate Metabolism

Priya S. Kishnani, Yuan-Tsong Chen

Carbohydrate metabolism plays a vital role in cellular function by providing the energy required for most metabolic processes. The relevant biochemical pathways involved in the metabolism of these carbohydrates are shown in Fig. 412-1. Glucose is the principal substrate of energy metabolism in humans. Metabolism of glucose generates ATP through glycolysis and mitochondrial oxidative phosphorylation. The body obtains glucose through the ingestion of polysaccharides (primarily starch) and disaccharides (e.g., lactose, maltose, and sucrose). Lactose and fructose are two other monosaccharides that serve as sources of fuel for cellular metabolism; however, their role as fuel sources is much less significant than that of glucose. Lactose is derived from galactose + glucose, which is found in milk products, and is an important component of certain glycolipids, glycoproteins, and glycosaminoglycans. Fructose is found in fruits, vegetables, and honey. Sucrose (fructose + glucose) is another dietary source of fructose and is a commonly used sweetener.

Glycogen, the storage form of glucose in animal cells, is composed of glucose residues joined in straight chains by α1-4 linkages and branched at intervals of 4–10 residues by α1-6 linkages. Glycogen forms a treelike molecule and can have a molecular weight of many millions. Glycogen may aggregate to form structures recognizable by electron microscopy. With the exception of type 0 disease, defects in glycogen metabolism typically cause an accumulation of glycogen in the tissues—hence the designation *glycogen storage diseases* (GSDs). The structure of stored glycogen can be normal or abnormal in the various disorders. Defects in gluconeogenesis or glycolytic pathways including galactose and fructose metabolism usually do not result in glycogen accumulation.

Clinical manifestations of the various disorders of carbohydrate metabolism differ markedly. The symptoms range from minimally harmful to lethal. Unlike disorders of lipid metabolism, mucopolysaccharidoses, or other storage diseases, many carbohydrate disorders have been managed with dietary therapy, yet despite these strides, long-term complications result, and there is a need for definitive therapies. Genes responsible for inherited defects of carbohydrate metabolism have been cloned, and mutations have been identified. With the use of tools such as DNA sequencing panels, whole exome sequencing, and whole genome sequencing, new disorders of glycogen storage continue to be identified and the phenotype of known disorders continues to expand, as is seen in the case of GSD type II and III. Advances in our understanding of the molecular basis of these diseases are now being used to improve diagnosis and management. Some of these disorders are candidates for enzyme replacement therapy, substrate reduction therapy, gene therapy, and other genomic tools, such as SiRNA technology and CRISPR genome editing technology.

Historically, the GSDs were categorized numerically in the order in which the enzymatic defects were identified. They are also classified by the organs involved (liver, muscle, and/or heart) and clinical manifestations. The latter is the system followed in this chapter (Table 412-1). The overall frequency of all forms of GSDs is ~1 in 20,000 live births. Most are inherited as autosomal recessive traits; however, phosphoglycerate kinase deficiency—one form of liver phosphorylase kinase (PhK) deficiency—and lysosomal-associated membrane protein 2 (LAMP2) deficiency are X-linked disorders. The most common childhood disorders are glucose-6-phosphatase deficiency (type I), lysosomal acid α-glucosidase deficiency (type II), debrancher deficiency (type III), and liver PhK deficiency (type IX). The most common adult disorder is myophosphorylase deficiency (type V, or McArdle disease).

SELECTED LIVER GLYCOGENOSES

DISORDERS WITH HEPATOMEGALY AND HYPOGLYCEMIA

Type I GSD (Glucose-6-Phosphatase or Translocase Deficiency, Von Gierke Disease) Type I GSD is an autosomal recessive disorder caused by glucose-6-phosphatase deficiency in liver, kidney, and intestinal mucosa. There are two subtypes of GSD I: type Ia, in which the glucose-6-phosphatase enzyme is defective, and type Ib, in which the translocase that transports glucose-6-phosphate across the microsomal membrane is defective. The defects in both subtypes lead to inadequate conversion of glucose-6-phosphate to glucose in the liver and thus make affected individuals susceptible to fasting hypoglycemia.

CLINICAL AND LABORATORY FINDINGS Persons with type I GSD may develop hypoglycemia and lactic acidosis during the neonatal period; however, more commonly, they exhibit hepatomegaly at 3–4 months

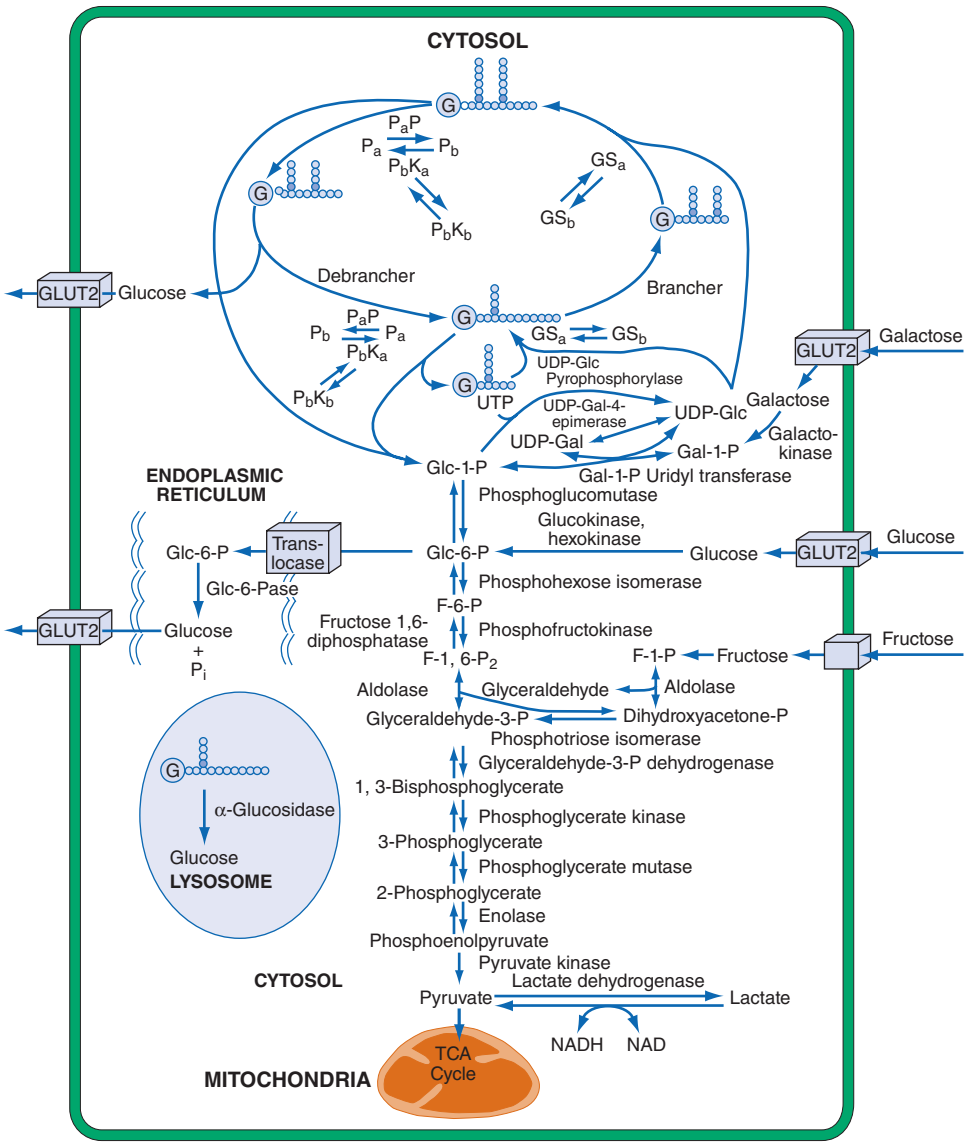


FIGURE 412-1 Metabolic pathways related to glycogen storage diseases and galactose and fructose disorders. Nonstandard abbreviations are as follows: GS_a, active glycogen synthase; GS_b, inactive glycogen synthase; P_a, active phosphorylase; P_b, inactive phosphorylase; P_aP, phosphorylase α phosphatase; P_bK_a, active phosphorylase β kinase; P_bK_b, inactive phosphorylase β kinase; G, glycogenin, the primer protein for glycogen synthesis. *(Modified from AR Beaudet, in KJ Isselbacher et al [eds]: Harrison's Principles of Internal Medicine, 13th ed. New York, McGraw-Hill, 1994, p 1855.)*

of age. Hypoglycemia, hypoglycemic seizures, and lactic acidosis can develop after a short fast. These children usually have doll-like facies with fat cheeks, relatively thin extremities, short stature, and a protuberant abdomen that is due to massive hepatomegaly. The kidneys are enlarged, but the spleen and heart are of normal size. The hepatocytes are distended by glycogen and fat, with large and prominent lipid vacuoles. Despite hepatomegaly, liver enzyme levels are usually normal or near normal. Easy bruising and epistaxis are associated with prolonged bleeding time as a result of impaired platelet aggregation/adhesion. Hyperuricemia is present. Hyperlipidemia includes elevation of triglycerides, low-density lipoproteins, and phospholipids. Type Ib patients have additional findings of neutropenia and impaired neutrophil function, which result in recurrent bacterial infections and oral and intestinal mucosal ulceration. GSD I patients may experience intermittent diarrhea, which can worsen with age. In GSD Ib, diarrhea is largely due to loss of mucosal barrier function caused by inflammation.

LONG-TERM COMPLICATIONS Gout usually becomes symptomatic at puberty as a result of long-term hyperuricemia. Puberty is often delayed. Some female patients have ultrasound findings consistent with polycystic ovaries; however, the other clinical features of polycystic ovary syndrome, such as acne and hirsutism, are not seen. Several reports of successful pregnancy in women with GSD I suggest that fertility is not affected. Increased bleeding during menstrual cycles, including life-threatening menorrhagia, has been reported. Secondary to lipid abnormalities, there is an increased risk of pancreatitis. Patients with GSD I may be at increased risk for cardiovascular disease. In adult patients, frequent fractures can occur and radiographic evidence of osteopenia/osteoporosis can be found; in prepubertal patients, radial bone mineral content is significantly reduced. Pulmonary hypertension—although rare—has been reported. By the second or third decade of life, many patients with type I GSD develop hepatic adenomas that can hemorrhage and, in some cases, become malignant. Renal disease is a serious late complication. Almost all patients aged >20 years have proteinuria, and many have hypertension, kidney stones, nephrocalcinosis, and altered creatinine clearance. In some patients, renal function deteriorates and progresses to complete failure, requiring dialysis or transplantation.

DIAGNOSIS Clinical presentation and abnormal plasma lactate and lipid values suggest that a patient may have GSD I, and gene-based mutation analysis provides a noninvasive means of reaching a definitive diagnosis for most patients with types Ia and Ib disease. Before the glucose-6-phosphatase and glucose-6-phosphate translocase genes were cloned, a definitive diagnosis required a liver biopsy to demonstrate a deficiency.

Type III GSD (Debrancher Deficiency, Limit Dextrinosis)

Type III GSD is an autosomal recessive disorder caused by a deficiency of glycogen debranching enzymes. Debranching and phosphorylase enzyme are responsible for the complete degradation of glycogen into glucose. When debranching enzyme is defective, glycogen breakdown is incomplete, resulting in abnormal glycogen accumulation with short outer chains, resembling limit dextrin.

CLINICAL AND LABORATORY FINDINGS Patients with GSD III present with hepatomegaly, hypoglycemia, short stature, variable skeletal myopathy, and cardiomyopathy. GSD type IIIa involves both liver and muscle. However, ~15% of patients have only liver involvement and is classified as *type IIIb*. Hypoglycemia and hyperlipidemia occur in children. In type III disease (as opposed to type I disease), fasting ketosis can be prominent, aminotransferase levels are elevated, and blood lactate and uric acid concentrations are usually normal. Serum creatine kinase (CK) levels can sometimes be used to identify patients with muscle involvement, but normal levels do not rule out muscle enzyme deficiency. In most patients with type III disease, hepatomegaly improves with age; however, liver fibrosis, cirrhosis progressing to liver failure, and hepatocellular carcinoma, are noted in many in late adulthood. Hepatic adenomas may occur, although less commonly than in GSD I. Left ventricular hypertrophy, significant scarring of the myocardium, and life-threatening arrhythmias have been reported. Patients with type IIIa disease may experience muscle weakness in childhood that can become severe after the third or fourth decade of life. The pattern

TABLE 412-1 Features of Glycogen Storage Diseases and Galactose and Fructose Disorders

TYPE/COMMON NAME	BASIC DEFECT	CLINICAL FEATURES	COMMENTS
Liver Glycogenoses			
Disorders with Hepatomegaly and Hypoglycemia			
Ia/von Gierke	Glucose-6-phosphatase	Growth retardation, enlarged liver and kidney, hypoglycemia, elevated blood lactate, cholesterol, triglycerides, and uric acid	Common, severe hypoglycemia. Complications in adulthood include hepatic adenomas, hepatic carcinoma, osteoporosis, pulmonary hypertension and renal failure.
Ib	Glucose-6-phosphate translocase	As for Ia, with additional findings of neutropenia and neutrophil dysfunction, increased risk for mucosal ulceration, and periodontal disease, inflammatory bowel disease, hypothyroidism	~10% of type I
IIIa/Cori or Forbes	Liver and muscle debranching enzyme	*Childhood:* Hepatomegaly, growth retardation, muscle weakness, hypoglycemia, hyperlipidemia, elevated liver aminotransferases	Common, intermediate severity of hypoglycemia, yet severe cases are seen. Hepatic adenomas, liver cirrhosis, and hepatic carcinoma can occur.
		Adulthood: Proximal and distal muscle atrophy and weakness; peripheral neuropathy with preferential median nerve involvement; variable cardiomyopathy, liver cirrhosis, progressive liver failure, risk for HCC in some	Muscle weakness can progress to need for ambulatory aids such as wheel chair.
IIIb	Liver debranching enzyme (normal muscle debrancher activity)	Liver symptoms same as in type IIIa; no muscle symptoms	~15% of type III
IV/Andersen	Branching enzyme	Failure to thrive, hypotonia, hepatomegaly, splenomegaly, progressive liver cirrhosis and failure (death usually before fifth year); a small subset do not have liver progression. *Adult form:* Isolated myopathy, neurogenic bladder, peripheral neuropathy, cognitive impairment.	One of the rarer glycogenoses. Other neuromuscular variants exist.
VI/Hers	Liver phosphorylase	Hepatomegaly, variable hypoglycemia, hyperlipidemia, and ketosis	Often underdiagnosed, severe cases being recognized
IX/phosphorylase kinase deficiency		As for VI, progressive liver failure is seen in some patients.	Common, X-linked, typically less severe than autosomal forms; clinical variability within and between subtypes; severe cases being recognized across different subtypes
IXa (*PHKA2*)	Liver PhK	Hypoglycemia, hyperketosis hepatomegaly, chronic liver disease, hyperlipidemia, elevated liver enzymes, growth retardation.	X-linked
IXb(*PHKB*)	Liver and muscle PhK	Hepatomegaly, growth retardation	Autosomal recessive
IXc (*PHKG2*)	Liver PhK	more severe than IXa; marked hepatomegaly, recurrent hypoglycemia, liver cirrhosis.	Autosomal recessive
IXd (*PHKA2*)	Muscle PhK	Exercise intolerance, cramps, myalgia, myoglobinuria; no hepatomegaly.	X-linked or Autosomal recessive
0/liver glycogen synthase deficiency	Glycogen synthase	Fasting hypoglycemia and ketosis, elevated lactic acid, alanine levels and hyperglycemia after glucose load, no hepatomegaly	Decreased liver glycogen stores
XI/Fanconi-Bickel	Glucose transporter 2	Failure to thrive, short stature, hypophosphatemic rickets, metabolic acidosis, hepatomegaly, proximal renal tubular dysfunction, impaired glucose and galactose utilization	Rare, consanguinity in 70%
Muscle Glycogenoses			
Disorders with Muscle-Energy Impairment			
V/McArdle	Muscle phosphorylase	Exercise intolerance, muscle cramps, myoglobinuria on strenuous exercise, increased CK, "second-wind" phenomenon	Common, male predominance
VII/Tarui	Phosphofructokinase—M subunit	As for type V, with additional findings of compensated hemolysis, myalgia	Prevalent in Ashkenazi Jews and Japanese
Phosphoglycerate kinase deficiency	Phosphoglycerate kinase	As for type V, with additional findings of hemolytic anemia and CNS dysfunction	Rare, X-linked
Phosphoglycerate mutase deficiency	Phosphoglycerate mutase—M subunit	As for type V	Rare, most patients African American
Lactate dehydrogenase deficiency	Lactic acid dehydrogenase—M subunit	As for type V, with additional findings of erythematous skin eruption and uterine stiffness resulting in childbirth difficulty in females	Rare
Fructose 1,6-bisphosphate aldolase A deficiency	Fructose 1,6-bisphosphate aldolase A	As for type V, with additional finding of hemolytic anemia, splenomegaly, rhabdomyolysis, jaundice	Rare
Pyruvate kinase deficiency	Pyruvate kinase—muscle isozyme	Muscle cramps and/or fixed muscle weakness	Rare
Muscle phosphorylase kinase deficiency	Muscle-specific phosphorylase kinase	As for type V. Some patients may have muscle weakness and atrophy.	Rare, autosomal recessive

(*Continued*)

TABLE 412-1 Features of Glycogen Storage Diseases and Galactose and Fructose Disorders (Continued)

TYPE/COMMON NAME	BASIC DEFECT	CLINICAL FEATURES	COMMENTS
β-Enolase deficiency	Muscle β-enolase	Exercise intolerance	Rare
Disorders with Progressive Skeletal Muscle Myopathy and/or Cardiomyopathy			
II/Pompe	Lysosomal acid α-glucosidase	*Infantile:* Hypotonia, muscle weakness, cardiac enlargement and failure, fatal early. *Late onset (juvenile and adult):* Progressive skeletal muscle weakness and atrophy, proximal muscles and respiratory muscles seriously affected.	Common, undetectable or very low level of enzyme activity in infantile form; variable residual enzyme activity in late-onset form
PRKAG2 deficiency	AMP-activated gamma 2 protein kinase	Severe cardiomyopathy and early heart failure (9–55 years). Congenital fetal form is rapidly fatal with hypertrophic cardiomyopathy and Wolff-Parkinson-White syndrome. Other involvement includes myalgia, myopathy, and seizures.	Autosomal dominant
Danon disease	Lysosomal-associated membrane protein 2 (LAMP2)	Severe cardiomyopathy and heart failure (8–15 years)	Very rare, X-linked
Late-onset polyglucosan body myopathy	Glycogenin-1	Adult-onset proximal muscle weakness, nervous system involvement uncommon	Autosomal recessive, rare
Galactose Disorders			
Galactosemia with uridyltransferase deficiency	Galactose 1-phosphate uridyltransferase	Vomiting, hepatomegaly, jaundice, cataracts, amino aciduria, failure to thrive	Long-term complications exist despite early diagnosis and treatment.
Galactokinase deficiency	Galactokinase	Cataracts	Benign
Uridine diphosphate galactose 4-epimerase deficiency	Uridine diphosphate galactose 4-epimerase	Similar to transferase deficiency with additional findings of hypotonia and nerve deafness	Benign variant exists.
Fructose Disorders			
Essential fructosuria	Fructokinase	Asymptomatic, positive urine reducing substance	Benign, autosomal recessive
Hereditary fructose intolerance	Fructose 1,6-bisphosphate aldolase B	Vomiting, lethargy, failure to thrive, hepatic failure, aversion to sweets, severity of symptoms depending on age/quantity of sugar ingested	Prognosis good with early diagnosis and fructose restriction, autosomal recessive
Fructose 1,6-diphosphatase deficiency	Fructose 1,6-diphosphatase	Episodic hypoglycemia, hyperlactic acidemia, and ketoacidosis usually following illness, hepatomegaly	Avoid fasting, good prognosis.

Abbreviations: CK, creatine kinase; CNS, central nervous system; M, muscle; PhK, phosphorylase kinase.

of muscle weakness is variable and both proximal and distal muscle weakness are seen. Peripheral neuropathy may become discernible later in life with preferential median nerve involvement. Polycystic ovaries are common in GSD III, and some patients develop features of polycystic ovarian syndrome, such as hirsutism and irregular menstrual cycles. Reports of successful pregnancy in women with GSD III suggest that fertility is normal.

DIAGNOSIS Deficient debranching enzyme activity can be demonstrated in liver, skeletal muscle, and heart in type IIIa GSD. In type IIIb, debranching enzyme deficiency is seen in the liver but not in muscle. The liver has distended hepatocytes due to glycogen buildup; areas of periportal fibrosis are also noted very early in the disease course. With the availability of molecular genetic testing, reliance on invasive tests such as liver and muscle biopsies is declining. DNA-based analyses now provide a noninvasive way of subtyping these disorders in most patients.

Type IX GSD (Liver Phosphorylase Kinase Deficiency)

Defects of PhK cause a heterogeneous group of glycogenoses. The PhK enzyme complex consists of four subunits (α, β, γ, and δ). Each subunit is encoded by different genes (X chromosome as well as autosomes) that are differentially expressed in various tissues. PhK deficiency can be divided into several subtypes on the basis of the gene/subunit involved, the tissues primarily affected, and the mode of inheritance. The most common subtype is GSD IXa, an X-linked liver PhK deficiency caused by mutations in the *PHKA2* gene, which is also one of the most common liver glycogenoses. PhK activity may also be deficient in erythrocytes and leukocytes but is normal in muscle. Typically, a child between the ages of 1 year and 5 years presents with growth retardation and hepatomegaly. Children tend eventually to exhibit normal growth patterns initiated by a delayed growth spurt during puberty. Liver fibrosis has been identified in some patients, including children. Levels of cholesterol, triglycerides, and liver enzymes are mildly elevated. Fasting ketosis is a feature of the disease. Lactic and uric acid levels are usually normal. Hypoglycemia may be mild in some but severe and recurrent in others. Phenotypic variability is being increasingly recognized, with significant disease involvement in some cases of the X-linked form. Liver histology shows distention of hepatocytes due to excess glycogen accumulation. There is a broad clinical spectrum of presentations identified with GSD IX. Hepatomegaly and abnormal blood chemistries gradually return to normal with age. Many adults reach a normal final height and are practically asymptomatic, despite persistent PhK deficiency, yet liver involvement can progress to cirrhosis, fibrosis, and liver failure. Some patients have significant ketosis. It is recommended that patients be monitored for hepatic complications with regular CT or MRI scans. Though previously thought to be a mild disease, the understanding is evolving with more severe cases coming to light, even in the X-linked form. Further research is needed to completely understand the natural history and long-term complications of GSD IX.

Treatment is symptom-based. A diet rich in complex carbohydrates and proteins and in small, frequent feedings are effective in preventing hypoglycemia. Blood ketones and glucose should be evaluated during times of stress. Liver transplantation may be considered in those with severe hepatic involvement.

Other subtypes of type IX GSD include GSD IXb, an autosomal recessive form of liver and muscle PhK deficiency caused by *PHKB* mutations. GSD IXc, an autosomal recessive form of liver PhK deficiency that often develops into liver cirrhosis, is due to *PHKG2* mutations. GSD IXd, a muscle-specific PhK deficiency that causes cramps and myoglobinuria with exercise, is caused by *PHKA1* mutations. The previous reports of cardiac-specific PhK deficiency is now considered to be a secondary phenomenon, as these patients have mutations in the *PRKAG2* gene. Patients with cardiac PRKAG2 syndrome often present with cardiomyopathy during infancy. The condition is lethal because of massive glycogen deposition in the myocardium. Details about this condition are described under the section about PRKAG2 deficiency.

Type IV GSD (Branching enzyme deficiency, Amylopectinosis, Polyglucosan disease or Andersen disease) is caused by deficiency of branching enzyme activity leading to accumulation of an abnormal glycogen with poor solubility. The disease is clinically heterogeneous. Individuals typically present in the first 18 months of life with failure to thrive, hepatosplenomegaly, and progressive liver cirrhosis leading to death in early childhood. Some patients may develop hepatic adenomas and hepatocellular carcinoma. GSD IV has extrahepatic manifestations involving the central and peripheral nervous system as well as cardiac and skeletal muscles. The adult form is known as adult polyglucosan body disease (APBD) and may present as an isolated myopathy or with systemic involvement of the central and peripheral nervous system characterized by neurogenic bladder, peripheral neuropathy, leukodystrophy, and mild cognitive impairment. Definitive diagnosis requires demonstration of branching enzyme deficiency in liver, muscle, cultured skin fibroblasts or leukocytes, or genetic testing of the *GBE1* gene. It is likely that life expectancy is shortened in APBD patients though it is yet to be confirmed by long-term natural history studies. Good supportive care is crucial to improve clinical outcomes.

Treatment for the adult form of GSD IV includes symptomatic support for gait abnormalities, bladder dysfunction, as well as periodic monitoring to uncover any new neurological deficits. Liver transplantation may be performed for progressive hepatic failure. However, caution should be exercised in selecting patients for liver transplant as a nonprogressive hepatic form of the disease exists in some whereas in others, cardiac and nervous system involvement may occur after transplantation.

Other Liver Glycogenoses with Hepatomegaly and Hypoglycemia These disorders include hepatic phosphorylase deficiency (Hers disease, type VI) and hepatic glycogenosis with renal Fanconi syndrome (type XI). Patients with GSD type VI can have growth retardation, hyperlipidemia, and hyperketosis in addition to hepatomegaly and hypoglycemia. Some patients have a less severe clinical course. GSD XI is caused by defects in the facilitative glucose transporter 2 (GLUT-2), which transports glucose and galactose in and out of hepatocytes, pancreatic cells, and the basolateral membranes of intestinal and renal epithelial cells. The disease is characterized by proximal renal tubular dysfunction, impaired glucose and galactose utilization, and accumulation of glycogen in liver and kidney.

SELECTED MUSCLE GLYCOGENOSES

DISORDERS WITH MUSCLE-ENERGY IMPAIRMENT

Type V GSD (Muscle Phosphorylase Deficiency, McArdle Disease) Type V GSD is an autosomal recessive disorder caused by deficiency of muscle phosphorylase. McArdle disease is a prototypical muscle-energy disorder as the enzyme deficiency limits ATP generation by glycogenolysis and results in glycogen accumulation.

CLINICAL AND LABORATORY FINDINGS There can be a broad, heterogeneous spectrum of clinical presentations with the neonatal form, which is rapidly fatal at one extreme, and the classical form with myalgia, cramps, and dark-colored urine at the other. Symptoms can be precipitated by: (1) brief, high intensity activity, such as sprinting or carrying heavy loads; and/or (2) less intense but sustained activity, such as climbing stairs or walking uphill. Most patients can engage in moderate exercise, such as walking on level ground, for long periods. Patients often exhibit the "second-wind" phenomenon, in which, after a short break from the initiation of strenuous physical effort, they are able to continue the activity without pain. Although most patients experience episodic muscle pain and cramping as a result of exercise, 35% report permanent pain that seriously affects sleep and other activities. Burgundy-colored urine is reported after exercise; resulting from myoglobinuria secondary to rhabdomyolysis. Renal failure can result from intense myoglobinuria after vigorous exercise. Symptom onset as late as the eighth decade has been reported.

Although cardiac involvement is not usually associated with muscle phosphorylase deficiency, hypertrophic cardiomyopathy has been observed in an adult patient with GSD V. In rare cases, electromyographic findings may suggest inflammatory myopathy, a diagnosis that may be confused with polymyositis. These patients may be at risk for statin-induced myopathy and rhabdomyolysis.

At rest, the serum CK level is usually elevated; after exercise, the CK level increases even more. Exercise leads to an increase in levels of blood ammonia, inosine, hypoxanthine, and uric acid; these abnormalities reflect residues of accelerated muscle purine nucleotide recycling as a result of insufficient ATP production. NADH is underproduced during physical exertion.

DIAGNOSIS Lack of an increase in blood lactate and exaggerated blood ammonia elevations after an ischemic exercise test are indicative of a muscle glycogenosis and suggest a defect in the conversion of glycogen or glucose to lactate. This abnormal exercise response, however, can also occur with other defects in glycogenolysis or glycolysis, such as deficiencies of muscle phosphofructokinase or debranching enzyme (when the test is done after fasting). A noninvasive, nonischemic forearm exercise test has been developed. Although this test has high sensitivity, is easy to perform and is cost-effective, the abnormal exercise response does not exclude other muscle glycogenosis. The cycle test detects the hallmark heart rate observed during the second-wind phenomenon. A diagnostic confirmation is established by enzymatic assay in muscle tissue or by mutation analysis of the myophosphorylase gene.

DISORDERS WITH PROGRESSIVE SKELETAL MUSCLE MYOPATHY AND/OR CARDIOMYOPATHY

Pompe Disease, Type II GSD (Acid α-1,4 Glucosidase Deficiency) Pompe disease is an autosomal recessive disorder caused by a deficiency of lysosomal acid α-1,4 glucosidase, an enzyme responsible for the degradation of glycogen in the lysosomes. This disease is characterized by the accumulation of glycogen in the lysosomes as opposed to accumulation in cytoplasm (as in the other glycogenoses).

CLINICAL AND LABORATORY FINDINGS The disorder encompasses a range of phenotypes. Each includes myopathy but differs in the age of onset, extent of organ involvement, and clinical severity. The most severe is the infantile form, with cardiomegaly, hypotonia, and death before 2 years of age. Infants often present with cardiomyopathy at birth, and develop a generalized muscle weakness with feeding difficulties, macroglossia, hepatomegaly, and congestive heart failure due to the rapidly progressive hypertrophic cardiomyopathy.

The late-onset form (juvenile/late-childhood or adult form, LOPD) is characterized primarily by skeletal muscle manifestations and respiratory muscle involvement, and a more slowly progressive course. The juvenile form typically presents as delayed motor milestones (if age of onset is early enough) and difficulty in walking. With disease progression, patients often develop swallowing difficulties, proximal muscle weakness, and respiratory muscle involvement. Death may occur before the end of the second decade.

Adults typically present between the second and seventh decades with slowly progressive myopathy without overt cardiac involvement. The clinical picture is dominated by slowly progressive, predominantly proximal limb girdle muscle weakness. The pelvic girdle, paraspinal muscles, and diaphragm are most seriously affected. Respiratory symptoms include somnolence, morning headache, orthopnea, and exertional dyspnea. Respiratory failure causes significant morbidity and mortality in the late-onset form. In rare instances, patients present with respiratory insufficiency as the initial symptom. Basilar artery

aneurysms and dilation of the ascending aorta have been observed in patients with Pompe disease. Ptosis, lingual weakness, gastrointestinal dysmotility, and incontinence due to poor sphincter tone are now being recognized as part of the clinical spectrum. Small-fiber neuropathy, which presents with painful paresthesia or pins-and-needles sensations, is also seen in some patients with LOPD. Individuals with advanced disease often require some form of ventilatory support and are dependent on a walking aid or wheelchair.

Laboratory findings include elevated levels of serum CK, aspartate aminotransferase, alanine aminotransferase, and lactate dehydrogenase. Levels of urine glucose tetrasaccharide (Glc_4), a breakdown product of glycogen, are elevated, especially on the severe end of the disease spectrum and can be used as a biomarker to monitor disease progression and treatment responsiveness. In infants, chest x-ray shows massive cardiomegaly, and electrocardiographic findings include a high-voltage QRS complex and a shortened PR interval. Muscle biopsy shows vacuoles that stain positive for glycogen; the muscle acid phosphatase level is increased, presumably from a compensatory increase of lysosomal enzymes. Electromyography reveals myopathic features, with irritability of muscle fibers and pseudomyotonic discharges, which appears early in the paraspinal muscles. Serum CK is not always elevated in adults and, depending on the muscle biopsied or tested, muscle histology or electromyography may not be abnormal.

DIAGNOSIS The confirmatory step for a diagnosis of Pompe disease is enzyme assay demonstrating deficient acid α-glucosidase or a gene sequence with two pathogenic mutations in the *GAA* gene. Enzyme activity can be measured in muscle, cultured skin fibroblasts, or blood. The latter is increasingly being used and is very reliable when performed in laboratories with experience. Prenatal diagnosis using mutation analysis of DNA extracted from fetal cells obtained by amniocentesis or by measuring GAA enzyme activity in chorionic villi or amniocytes is available. Carrier detection and prenatal diagnosis, using DNA-based targeted mutation analysis, are also possible if disease-causing family mutations are already known.

The approval of enzyme replacement therapy with alglucosidase alfa in 2006 has changed the natural history and clinical course of Pompe disease. Other adjunctive treatment options include dietary modifications, submaximal aerobic exercise, and respiratory muscle strength training. Gene therapy is under study as another treatment modality. Early diagnosis with early enzyme replacement initiation is the key to treatment efficacy.

Pompe disease is now part of the recommended uniform screening panel (RUSP) for newborns in the United States and newborn screening (NBS) has been initiated in several states. In Taiwan, where NBS for Pompe disease is performed routinely for all infants, early disease detection and treatment initiation has led to better treatment outcomes in infantile Pompe patients.

Late-Onset Polyglucosan Body Myopathy due to GYG1 Mutations This is an autosomal recessive, slowly progressive skeletal myopathy caused by mutations in the *GYG1* gene blocking glycogenin-1 biosynthesis. GYG1 mutations results in a reduced or complete absence of glyogenin-1 which is necessary for glycogen synthesis in muscles. Affected individuals commonly present with adult-onset proximal muscle weakness prominently affecting the hip and shoulder girdles. Cardiomyopathy and cardiac failure necessitating cardiac transplantation is seen. Compared to GSD IV APBD, nervous system involvement has not been reported although both disorders cause polyglucosan deposition.

GSD Mimicking Hypertrophic Cardiomyopathy Danon disease is an X-linked glycogen storage disorder caused by mutations in the LAMP2 gene. This results in deficiency of lysosomal-associated membrane protein 2 (LAMP2), leading to accumulation of glycogen in the heart and skeletal muscle. Patients present primarily with hypertrophic cardiomyopathy, but can be distinguished from the usual causes of hypertrophic cardiomyopathy by their electrophysiological abnormalities, particularly ventricular pre-excitation and conduction defects. The onset of cardiac symptoms such as chest pain, palpitations, syncope, and cardiac arrest may occur between the ages of 8 and 15 years. Ocular manifestations are often under-recognized and include peripheral pigmentary retinopathy, lens changes, and abnormal electroretinograms. The prognosis for LAMP2 deficiency is poor, with progressive end-stage heart failure early in adulthood. Treatment is mainly symptomatic and involves management of heart failure, correction of conduction abnormalities, and physical therapy, among others. Cardiac transplantation can be considered for refractory cases of heart failure.

AMP-ACTIVATED PROTEIN KINASE GAMMA 2 DEFICIENCY (PRKAG2 DEFICIENCY) AMP-activated protein kinase gamma 2 (PRKAG2) deficiency is caused by mutations in the *PRKAG2* gene. This gene encodes the γ2 subunit of AMP-activated protein kinase (AMPK) which is important in many cellular ATP metabolic pathways. Affected individuals present with cardiac abnormalities including hypertrophic cardiomyopathy and conduction system abnormalities, particularly Wolff-Parkinson-White syndrome. The extent of cardiac involvement is variable and includes supraventricular tachycardia, sinus bradycardia, left ventricular dysfunction or even sudden cardiac death in some cases. In addition to cardiac involvement, there is a broad spectrum of phenotypic presentations including myalgia, myopathy and seizures. Unlike Danon disease, cardiomyopathy due to *PRKAG2* mutations is compatible with long-term survival except for a congenital form that presents in early infancy with a rapid fatal course. PRKAG2 syndrome should be considered as a differential diagnosis in infants presenting with severe hypertrophic cardiomyopathy. In rare instances, PRKAG2 cases may be misdiagnosed as infantile Pompe disease due to phenotypical similarity. Treatment is usually symptomatic and supportive, as in Danon disease.

TREATMENT

Glycogen Storage Disease Mimicking Hypertrophic Cardiomyopathy

Heart transplantation has been suggested as a preventive measure for LAMP2 deficiency and noncongenital PRKAG2 deficiency.

SELECTED DISORDERS OF GALACTOSE METABOLISM

"Classic" galactosemia is caused by galactose 1-phosphate uridyltransferase (GALT) deficiency. It is a serious disease with an incidence of 1 in 60,000 and an early onset of symptoms. The newborn infant normally receives up to 40% of caloric intake as lactose (glucose + galactose). Without the transferase, the infant is unable to metabolize galactose 1-phosphate (Fig. 412-1), which consequently accumulates, resulting in injury to parenchymal cells of the kidney, liver, and brain. After the first feeding, infants can present with vomiting, diarrhea, hypotonia, jaundice, and hepatomegaly. There is an increased risk for *Escherichia coli* neonatal sepsis in galactosemic infants; often with the onset of sepsis preceding the diagnosis of galactosemia.

Widespread newborn screening for galactosemia has identified these infants early and allowed them to be placed on dietary restriction. Elimination of galactose from the diet reverses growth failure as well as renal and hepatic dysfunction, improving the prognosis. However, on long-term follow-up, some patients still have ovarian failure manifesting as primary or secondary amenorrhea as well as developmental delays and learning disabilities that increase in severity with age. Of women with classic galactosemia, 80–90% or more report hypergonadotropic hypogonadism. While most female patients are infertile when they reach childbearing age, a few have given birth. Several mutations appear to be protective, particularly the p.Ser135Leu mutation, which is more common in the African-American population. Methods for fertility preservation, such as cryopreservation, are available. In addition, most patients have speech disorders, and a smaller proportion demonstrate poor growth, impaired motor function, and balance (with or without overt ataxia). Adults on dairy-free diets have developed cataracts, tremors, and low bone density. The treatment of galactosemia to prevent long-term complications remains a challenge.

Deficiency of *galactokinase* (Fig. 412-1) causes cataracts. Deficiency of *uridine diphosphate galactose 4-epimerase* can be benign when the enzyme deficiency is limited to blood cells, but can be as severe as classic galactosemia when the enzyme deficiency is generalized.

SELECTED DISORDERS OF FRUCTOSE METABOLISM

Fructokinase deficiency, or essential fructosemia (Fig. 412-1), causes a benign condition that is incidentally diagnosed from the presence of fructose as a reducing substance in the urine.

Deficiency of *fructose 1, 6-bisphosphate aldolase* (aldolase B; hereditary fructose intolerance) is a serious disease in infants. These patients are healthy and symptom-free until fructose or sucrose (table sugar) is ingested (usually from fruit, sweetened cereal, or sucrose-containing formula). Clinical manifestations may include jaundice, hepatomegaly, vomiting, lethargy, irritability, and convulsions. The incidence of celiac disease is higher among patients with hereditary fructose intolerance (>10%) than in the general population (1–3%). Laboratory findings show prolonged clotting time, hypoalbuminemia, elevation of bilirubin and aminotransferase levels, and proximal renal tubular dysfunction. If the disease goes undiagnosed and the deleterious intake of sugar continues, hypoglycemic episodes recur, and eventually death can occur from progressive liver and renal failure. The mainstay of treatment is the elimination of all sources of sucrose, fructose, and sorbitol from the diet. Once dietary control is established, liver and kidney dysfunction improve, and catch-up growth is common; intellectual development is usually not affected. Over time, the patient's symptom intensity improves, even after fructose ingestion. The long-term prognosis is good.

Fructose 1,6-diphosphatase deficiency is characterized by childhood life-threatening episodes of hypoglycemia, acidosis, hyperventilation, convulsions, and coma. These episodes are often triggered by foods that contain fructose, and include febrile infections and gastroenteritis when oral food intake is low. Laboratory findings include low blood glucose levels, high lactate and uric acid levels, and metabolic acidosis. Renal tubular and liver functions are normal and aversion to sweets is usually not seen, unlike hereditary fructose intolerance. Treatment of acute episodes requires the correction of hypoglycemia and acidosis by IV infusion of dextrose. Further episodes can be prevented by avoidance of fasting and elimination of fructose and sucrose from the diet. A complex carbohydrate such as cornstarch, which provides slow and sustained levels of glucose, is useful for the long-term prevention of hypoglycemia. With proper treatment, prognosis is good, and patients who survive childhood develop normally.

GLOBAL CONSIDERATIONS

The GSDs and other inherited disorders of carbohydrate metabolism, although individually rare, are reported in most ethnic populations. The prevalent genetic mutations for each disease may vary in different ethnic populations, but clinical symptoms are remarkably similar and treatment guidelines apply to all. Symptomatic treatment is available for these disorders, and today, advances in the field have resulted in more definitive treatment approaches. Availability of NBS for Pompe disease has shown that the frequency of Pompe disease is much higher than previously estimated. This has allowed for early treatment initiation, and improved outcomes. NBS also mitigates the long diagnostic delays and misdiagnoses often associated with Pompe disease. The lessons learned from Pompe disease have bearing on the other GSDs.

Acknowledgment

The authors are grateful to Mrudu Herbert for her contributions to this chapter.

FURTHER READING

Bosch AM: Classical galactosaemia revisited. J Inherit Metab Dis 29:516, 2006.

Bouteldja N, Timson DJ: The biochemical basis of hereditary fructose intolerance. J Inherit Metab Dis 33:105, 2010.

Kishnani PS et al: Diagnosis and management of glycogen storage disease type I: A practice guideline of the American College of Medical Genetics and Genomics. Genet Med 16:e1, 2014.

Kishnani PS, Beckemeyer AA, Mendelsohn NJ: The new era of Pompe disease: Advances in the detection, understanding of the phenotypic spectrum, pathophysiology, and management. Am J Med Genet C Semin Med Genet 160c:1, 2012.

Kishnani PS et al: Glycogen storage disease type III diagnosis and management guidelines. ACMG. Genet Med 12:446, 2010.

Lévesque S et al: Diagnosis of late-onset Pompe disease and other muscle disorders by next-generation sequencing. Orphanet J Rare Dis 11:8, 2016.

Martin MA et al: Glycogen storage disease type V, in GeneReviews® [Internet], RA Pagon et al (eds). Seattle (WA): University of Washington, Seattle, 2006.

Porto AG et al: Clinical Spectrum of PRKAG2 Syndrome. Circ Arrhythm Electrophysiol 9: e003121, 2016.

Yang CF et al: Very early treatment for infantile-onset Pompe disease contributes to better outcomes. J Pediatr 169:174.e1, 2016.

413 Inherited Disorders of Amino Acid Metabolism in Adults

Nicola Longo

Amino acids are the building blocks of proteins and serve as neurotransmitters (glycine, glutamate, γ-aminobutyric acid) or as precursors of hormones, coenzymes, pigments, purines, or pyrimidines. Eight amino acids, referred to as *essential* (histidine, isoleucine, leucine, lysine, methionine, phenylalanine, valine, threonine, and tryptophan), cannot be synthesized by humans and must be obtained from dietary sources. The others are formed endogenously. Each amino acid has a unique degradative pathway by which its nitrogen and carbon components are used for the synthesis of other amino acids, carbohydrates, and lipids. Disorders of amino acid metabolism and transport **(Chap. 414)** are individually rare—the incidences range from 1 in 10,000 for cystinuria or phenylketonuria to 1 in 200,000 for homocystinuria or alkaptonuria—but collectively, they affect perhaps 1 in 1000 newborns. Almost all are transmitted as autosomal recessive traits.

The features of inherited disorders of amino acid catabolism are summarized in **Table 413-1**. In general, these disorders are named for the compound that accumulates to highest concentration in blood (*-emias*) or urine (*-urias*). In the aminoacidopathies, the parent amino acid is found in excess, whereas products in the catabolic pathway accumulate in organic acidemias. Which compound(s) accumulates depends on the site of the enzymatic block, the reversibility of the reactions proximal to the lesion, and the availability of alternative pathways of metabolic "runoff." Biochemical and genetic heterogeneity are common. Five distinct forms of hyperphenylalaninemia, nine forms of homocystinuria, and methylmalonic acidemia are recognized. Such heterogeneity reflects the presence of a large array of molecular defects.

The manifestations of these conditions differ widely (Table 413-1). Some, such as sarcosinemia, produce no clinical consequences. At the other extreme, complete deficiency of ornithine transcarbamylase is lethal in the untreated neonate. Central nervous system (CNS) dysfunction, in the form of developmental retardation, seizures, alterations in sensorium, or behavioral disturbances, is present in more than half the disorders. Protein-induced vomiting, neurologic dysfunction, and hyperammonemia occur in many disorders of urea cycle intermediates. Metabolic ketoacidosis, often accompanied by hyperammonemia, is

TABLE 413-1 Inherited Disorders of Amino Acid Metabolism

AMINO ACID(S)	CONDITION	ENZYME DEFECT	CLINICAL FINDINGS	INHERITANCE
Phenylalanine	DNAJC12 Deficiency	Hydroxylase Co-Chaperone	Dystonia, parkinsonism, intellectual disability	AR
	Phenylketonuria	Phenylalanine hydroxylase	Intellectual disability, microcephaly, hypopigmented skin and hairs, eczema, "mousy" odor	AR
	DHPR deficiency	Dihydropteridine reductase	Intellectual disability, hypotonia, spasticity, myoclonus	AR
	PTPS deficiency	6-Pyruvoyl-tetrahydropterin synthase	Dystonia, neurologic deterioration, seizures, intellectual disability	AR
	GTP cyclohydrolase I deficiency	GTP cyclohydrolase I	Intellectual disability, seizures, dystonia, temperature instability	AR
	Carbinolamine dehydratase deficiency	Pterin-4α-carbinolamine dehydratase	Transient hyperphenylalaninemia (benign)	AR
Tyrosine	Tyrosinemia type I (hepatorenal)	Fumarylacetoacetate hydrolase	Liver failure, cirrhosis, rickets, failure to thrive, peripheral neuropathy, "boiled cabbage" odor	AR
	Tyrosinemia type II (oculocutaneous)	Tyrosine transaminase	Palmoplantar keratosis, painful corneal erosions with photophobia, learning disability	AR
	Tyrosinemia type III	4-Hydroxyphenylpyruvate dioxygenase	Hypertyrosinemia with normal liver function, occasional mental delay	AR
	Hawkinsinuria	4-Hydroxyphenylpyruvate dioxygenase	Transient failure to thrive, metabolic acidosis in infancy	AD
	Alkaptonuria	Homogentisic acid oxidase	Ochronosis, arthritis, cardiac valve involvement, coronary artery calcification	AR
	Albinism (oculocutaneous)	Tyrosinase	Hypopigmentation of hair, skin, and optic fundus; visual loss; photophobia	AR
	Albinism (ocular)	Different enzymes or transporters	Hypopigmentation of optic fundus, visual loss	AR, XL
	DOPA-responsive dystonia	Tyrosine hydroxylase	Rigidity, truncal hypotonia, tremor, intellectual disability	AR
GABA	4-Hydroxybutyric aciduria	Succinic semialdehyde dehydrogenase	Seizures, intellectual disability, hypotonia	AR
	ABAT deficiency	GABA transaminase	Seizures, intellectual disability, hypotonia	AR
Tryptophan	Hydroxykynureninuria	Kynureninase	Intellectual disability, spasticity	AR
Histidine	Histidinemia	Histidine-ammonia lyase	Benign	AR
	Urocanic aciduria	Urocanase	Occasional intellectual disability	AR
	Formiminoglutamic aciduria	Formiminotransferase	Occasional intellectual disability	AR
Glycine	Glycine encephalopathy	Glycine cleavage (4 enzymes)	Infantile seizures, lethargy, apnea, profound intellectual disability	AR
	Sarcosinemia	Sarcosine dehydrogenase	Benign	AR
	Hyperoxaluria type I	Alanine:glyoxylate aminotransferase	Calcium oxalate nephrolithiasis, renal failure	AR
	Hyperoxaluria type II	D-Glyceric acid dehydrogenase/ glyoxylate reductase	Calcium oxalate nephrolithiasis, renal failure	AR
Serine	Phosphoglycerate dehydrogenase deficiency	Phosphoglycerate dehydrogenase	Seizures, microcephaly, intellectual disability	AR
	PSAT1 deficiency	Phosphoserine aminotransferase	Seizures, microcephaly, intellectual disability	AR
	PSPHP deficiency	Phosphoserine phosphatase	Seizures, microcephaly, intellectual disability	AR
Proline	Hyperprolinemia type I	Proline oxidase	Benign	AR
	Hyperprolinemia type II	Δ^1-Pyrroline-5-carboxylate dehydrogenase	Febrile seizures, intellectual disability	AR
	Hyperhydroxyprolinemia	Hydroxyproline oxidase	Benign	AR
	Prolidase deficiency	Prolidase	Mild intellectual disability, chronic dermatitis	AR
Methionine	Hypermethioninemia	Methionine adenosyltransferase	Usually benign	AR
	S-Adenosylhomocysteine hydrolase deficiency	S-Adenosylhomocysteine hydrolase	Hypotonia, intellectual disability, absent tendon reflexes, delayed myelination	AR
	Glycine N-methyltransferase deficiency	Glycine N-methyltransferase	Elevated liver transaminases	AR
	Adenosine kinase deficiency	Adenosine kinase	Intellectual disability, seizures, liver dysfunction	AR
Homocystine	Homocystinuria	Cystathionine β-synthase	Lens dislocation, thrombotic vascular disease, intellectual disability, osteoporosis	AR
	Homocystinuria	5,10-Methylenetetrahydrofolate reductase	Intellectual disability, gait and psychiatric abnormalities, recurrent strokes	AR
	Homocystinuria	Methionine synthase (cblE, G)	Intellectual disability, hypotonia, seizures, megaloblastic anemia	AR
	Homocystinuria and methylmalonic acidemia	Vitamin B_{12} lysosomal efflux and metabolism (cblC, -D, -F, -J, -X)	Intellectual disability, lethargy, failure to thrive, hypotonia, seizures, megaloblastic anemia	AR
Cystathionine	Cystathioninuria	β-Cystathioninase	Benign	AR
Cystine	Cystinosis	Cystinosin CTNS (lysosomal efflux)	Renal Fanconi's syndrome, rickets, photophobia, hypotonia, renal failure	AR

(*Continued*)

TABLE 413-1 Inherited Disorders of Amino Acid Metabolism (Continued)

AMINO ACID(S)	CONDITION	ENZYME DEFECT	CLINICAL FINDINGS	INHERITANCE
S-Sulfo-L- cysteine	Sulfocysteinuria	Sulfate oxidase or molybdenum cofactor deficiency	Seizures, intellectual disability, dislocated lenses	AR
Lysine	Hyperlysinemia, saccharopinuria	α-Aminoadipic semialdehyde synthase	Benign	AR
	Pyridoxine-dependent seizures	L-Δ^1-Piperideine-6-carboxilate dehydrogenase	Seizures, intellectual disability	AR
Lysine, tryptophan	α-Ketoadipic acidemia	α-Ketoadipic acid dehydrogenase	Benign	?
	Glutaric acidemia type I	Glutaryl-CoA dehydrogenase	Progressive severe dystonia and athetosis, motor delays	AR
	Glutaric acidemia type II	Electron transfer flavoproteins (ETF) or ETF:ubiquinone oxidoreductase	Hypoglycemia, metabolic acidosis, "sweaty feet" odor, hypotonia, cardiomyopathy	AR
Ornithine	Gyrate atrophy of the choroid and retina	Ornithine-Δ-aminotransferase	Myopia, night blindness, loss of peripheral vision, cataracts, chorioretinal degeneration	AR
Urea cycle	Carbamoylphosphate synthase-1 deficiency	Carbamoylphosphate synthase-1	Lethargy progressing to coma, protein aversion, intellectual disability, hyperammonemia	AR
	N-Acetylglutamate synthase deficiency	*N*-Acetylglutamate synthase	Lethargy progressing to coma, protein aversion, intellectual disability, hyperammonemia	AR
	Ornithine transcarbamylase deficiency	Ornithine transcarbamylase	Lethargy progressing to coma, protein aversion, intellectual disability, hyperammonemia	XL
	Citrullinemia type I	Argininosuccinate synthase	Lethargy progressing to coma, protein aversion, intellectual disability, hyperammonemia, liver failure	AR
	Argininosuccinic acidemia	Argininosuccinate lyase	Lethargy progressing to coma, protein aversion, intellectual disability, hyperammonemia, trichorrhexis nodosa	AR
	Arginase deficiency	Arginase	Spastic tetraparesis, microcephaly, intellectual disability, mild hyperammonemia	AR
	Hyperornithinemia, hyperammonemia, homocitrullinuria	Mitochondrial ornithine carrier ORNT1	Vomiting, lethargy, failure to thrive, intellectual disability, episodic confusion, hyperammonemia, protein intolerance	AR
	Citrullinemia type 2	Mitochondrial aspartate/ glutamate carrier CTLN2	Neonatal intrahepatic cholestasis, adult presentation with sudden behavioral changes and stupor, coma, hyperammonemia	AR
Proline, ornithine, arginine	Δ^1-Pyrroline-5-carboxylate synthase deficiency	Δ^1-Pyrroline-5-carboxylate synthase	Hypotonia, seizures, neurodegeneration, peripheral neuropathy, joint laxity, skin hyperelasticity, subcapsular cataracts, hyperammonemia	AR
Glutamine	Glutamine synthase deficiency	Glutamine synthase	Brain malformations, pachygyria, seizures, hypotonia, dysmorphic features	AR
Valine	Hypervalinemia	Branched chain aminotransferase-2	Headache, memory impairment, failure to thrive, hypotonia, developmental delays	AR
	Isobutyryl-CoA dehydrogenase deficiency	Isobutyryl-CoA dehydrogenase	Benign	AR
Valine, leucine, isoleucine	Maple syrup urine disease	Branched chain ketoacid dehydrogenase (E1α, E1β, E2, E3 deficiency)	Lethargy, vomiting, encephalopathy, seizures, intellectual disability, "maple syrup" odor, protein intolerance	AR
Leucine	Isovaleric acidemia	Isovaleryl-CoA dehydrogenase	Acidosis, ketosis, vomiting, coma, hyperammonemia, "sweaty feet" odor, protein intolerance	AR
	3-Methylcrotonyl glycinuria	3-Methylcrotonyl-CoA carboxylase	Stress-induced metabolic acidosis, hypotonia, hypoglycemia, "cat's urine" odor	AR
	3-Methylglutaconic aciduria type I	3-Methylglutaconyl-CoA hydratase deficiency	Stress-induced acidosis, leukodystrophy, hypotonia, hepatomegaly	AR
	3-Hydroxy-3-methylglutaric aciduria	3-Hydroxy-3-methylglutaryl-CoA lyase	Stress-induced hypoketotic hypoglycemia and acidosis, encephalopathy, hyperammonemia	AR
Isoleucine	2-Methylbutyryl-glycinuria	2-Methylbutyryl-CoA dehydrogenase	Benign	AR
	2-Methyl-3-hydroxybutyryl-CoA dehydrogenase deficiency	2-Methyl-3-hydroxybutyryl-CoA dehydrogenase	Developmental regression, seizures, and rigidity sometimes triggered by illnesses	XL
	3-Oxothiolase deficiency	3-Oxothiolase	Fasting-induced acidosis and ketosis, vomiting, lethargy	AR
Valine, isoleucine, methionine, threonine	Propionic acidemia (pccA, -B, -C)	Propionyl-CoA carboxylase	Metabolic ketoacidosis, hyperammonemia, hypotonia, lethargy, coma, protein intolerance, intellectual disability, hyperglycinemia	AR
	Multiple carboxylase/ biotinidase deficiency	Holocarboxylase synthase or biotinidase	Metabolic ketoacidosis, diffuse rash, alopecia, seizures, intellectual disability	AR
	Methylmalonic acidemia (mutase, cblA, B, racemase)	Methylmalonyl-CoA mutase/ racemase or cobalamin reductase/adenosyltransferase	Metabolic ketoacidosis, hyperammonemia, hypertonia, lethargy, coma, protein intolerance, intellectual disability, hyperglycinemia	AR

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; Cbl, cobalamin; DOPA, dihydroxyphenylalanine; GABA, γ-aminobutyric acid; GTP, guanosine 5′-triphosphate; XL, X-linked.

frequent in organic acidemias. Some disorders produce focal tissue or organ involvement such as liver disease, renal failure, cutaneous abnormalities, or ocular lesions.

The analysis of plasma amino acids (by ion-exchange chromatography or liquid chromatography/tandem mass spectrometry), urine organic acids (by gas chromatography/mass spectrometry), and plasma acylcarnitine profile (by tandem mass spectrometry) is commonly used to diagnose and monitor most of these disorders. The diagnosis is confirmed by enzyme assay on cells or tissues from the patients or, more commonly, by DNA testing. The clinical manifestations in many of these conditions can be prevented or mitigated if a diagnosis is achieved early and appropriate treatment (e.g., dietary protein or amino acid restriction or vitamin supplementation) is instituted promptly. For this reason, newborn screening programs seek to identify several of these disorders. Infants with a positive screening test need additional metabolic testing (usually suggested by the newborn screening program) to confirm or exclude the diagnosis. Confirmed cases should be referred to a metabolic center for initiation of therapy. The parents need to be counseled about the natural history the disease and its recurrence risk in future pregnancies. In some cases, parents need testing because they might have a disorder themselves (such as glutaric acidemia type 1, methylcrotonyl coenzyme A carboxylase deficiency, primary carnitine deficiency, or fatty acid oxidation defects) since mothers with these conditions can sometimes be identified by abnormal newborn screening results in their offspring. Some metabolic disorders can remain asymptomatic until adult age, presenting only when fasting or severe stress require full activity of affected metabolic pathways to provide energy.

Selected disorders that illustrate the principles, properties, and problems presented by the disorders of amino acid metabolism are discussed in this chapter.

THE HYPERPHENYLALANINEMIAS

The hyperphenylalaninemias (Table 413-1) result from impaired conversion of phenylalanine to tyrosine. The most common and clinically important is *phenylketonuria* (frequency 1:16,500), which is an autosomal recessive disorder characterized by an increased concentration of phenylalanine and its by-products in body fluids and by severe intellectual disability if untreated in infancy. It results from reduced activity of phenylalanine hydroxylase. The accumulation of phenylalanine inhibits the transport of other amino acids required for protein or neurotransmitter synthesis, reduces synthesis and increases degradation of myelin, and leads to inadequate formation of norepinephrine and serotonin. Phenylalanine is a competitive inhibitor of tyrosinase, a key enzyme in the pathway of melanin synthesis, and accounts for the hypopigmentation of hair and skin. Untreated children with classic phenylketonuria are normal at birth but fail to attain early developmental milestones, develop microcephaly, and demonstrate progressive impairment of cerebral function. Hyperactivity, seizures, and severe intellectual disability are major clinical problems later in life. Electroencephalographic abnormalities; "mousy" odor of skin, hair, and urine (due to phenylacetate accumulation); and a tendency to develop hypopigmentation and eczema complete the devastating clinical picture. In contrast, affected children who are detected and treated at birth show none of these abnormalities.

TREATMENT

Phenylketonuria

To prevent intellectual disability, diagnosis and initiation of dietary treatment of classic phenylketonuria must occur before the child is 2 weeks of age. For this reason, newborns in North America, Australia, and Europe are screened by determinations of blood phenylalanine levels. Abnormal values are confirmed using quantitative analysis of plasma amino acids. Dietary phenylalanine restriction is usually instituted if blood phenylalanine levels are >360 μmol/L (6 mg/dL). Treatment consists of a special diet low in phenylalanine and supplemented with tyrosine, since tyrosine becomes an essential amino acid in phenylalanine hydroxylase deficiency. With therapy, plasma phenylalanine concentrations should be maintained between 120 and 360 μmol/L (2 and 6 mg/dL). Dietary restriction should be continued and monitored indefinitely. Some patients with milder forms of phenylketonuria (phenylalanine <1200 μmol/L at presentation) show increased tolerance to dietary proteins and improved metabolic control when treated with tetrahydrobiopterin (5–20 mg/kg per day), an essential cofactor of phenylalanine hydroxylase.

A number of women with phenylketonuria who have been treated since infancy will reach adulthood and become pregnant. If maternal phenylalanine levels are not strictly controlled before and during pregnancy, their offspring are at increased risk for congenital defects and microcephaly (*maternal phenylketonuria*). After birth, these children have severe intellectual disability and growth retardation. Pregnancy risks can be minimized by continuing lifelong phenylalanine-restricted diets and assuring strict phenylalanine restriction 2 months prior to conception and throughout gestation.

THE HOMOCYSTINURIAS (HYPERHOMOCYSTEINEMIAS)

The homocystinurias are nine biochemically and clinically distinct disorders (Table 413-1) characterized by increased concentration of the sulfur-containing amino acid homocystine in blood and urine.

Classic homocystinuria, the most common (frequency 1:400,000), results from reduced activity of cystathionine β-synthase **(Fig. 413-1)**, the pyridoxal phosphate–dependent enzyme that condenses homocysteine with serine to form cystathionine. Most patients present between 3 and 5 years of age with dislocated optic lenses and intellectual disability (in about half of cases). Some patients develop a marfanoid habitus and radiologic evidence of osteoporosis.

Life-threatening vascular complications (affecting coronary, renal, and cerebral arteries) can occur during the first decade of life and are the major cause of morbidity and mortality. Classic homocystinuria can be diagnosed with analysis of plasma amino acids, showing elevated methionine and presence of free homocystine. Total plasma homocysteine is also extremely elevated (usually >100 μ*M*). Treatment consists of a special diet restricted in protein and methionine. In approximately half of patients, oral pyridoxine (25–500 mg/d) produces a fall in plasma methionine and homocystine concentration in body fluids. Folate and vitamin B_{12} deficiency should be prevented by adequate supplementation. Betaine is also effective in reducing homocystine levels by favoring its remethylation to methionine.

The other forms of homocystinuria are the result of impaired remethylation of homocysteine to methionine. This can be caused by defective methionine synthase or reduced availability of two essential cofactors, 5-methyltetrahydrofolate and methylcobalamin (methyl-vitamin B_{12}). In contrast to cystathionine β-synthase, elevated levels of free homocystine are associated with low levels of methionine in the plasma amino acid profile in remethylation defects. Therapy in these cases requires administration of methylfolate, hydroxycobalamin (an activated form of vitamin B_{12}), and betaine.

Hyperhomocysteinemia refers to increased total plasma concentration of homocysteine with or without an increase in free homocystine (disulfide form). Hyperhomocysteinemia, in the absence of significant homocystinuria, is found in some heterozygotes for the genetic defects noted above or in homozygotes for milder variants. Changes of homocysteine levels are also observed with increasing age; with smoking; in postmenopausal women; in patients with renal failure, hypothyroidism, leukemias, inflammatory bowel disease, or psoriasis; and during therapy with drugs such as methotrexate, nitrous oxide, isoniazid, and some antiepileptic agents. Homocysteine can act as an atherogenic and thrombophilic agent and increased total plasma homocysteine have been associated with an increased risk for coronary, cerebrovascular, and peripheral arterial disease as well as for deep-vein thrombosis. In addition, hyperhomocysteinemia and folate and vitamin B_{12} deficiencies have been associated with an increased risk of neural tube defects

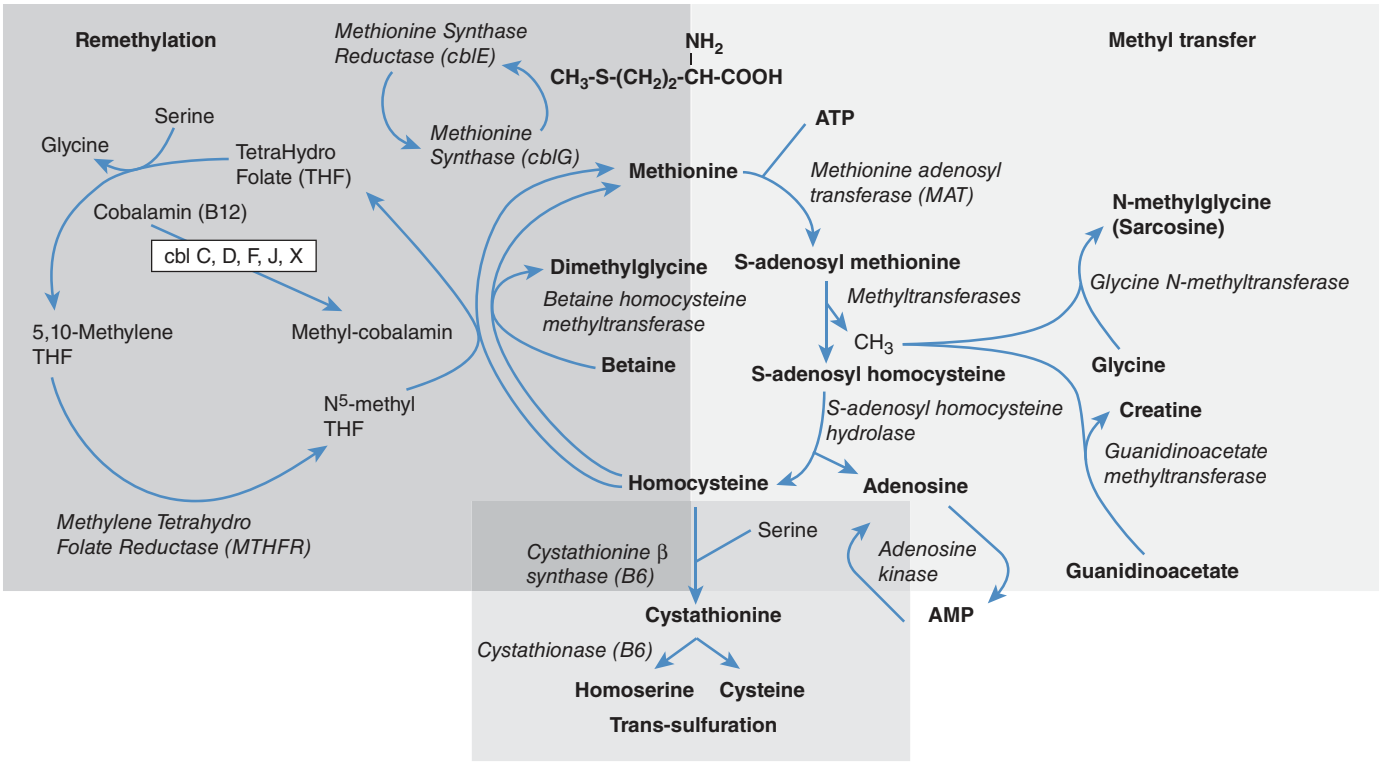


FIGURE 413-1 Pathways, enzymes, and coenzymes involved in the homocystinurias. Methionine transfers a methyl group during its conversion to homocysteine. Defects in methyl transfer or in the subsequent metabolism of homocysteine by the pyridoxal phosphate (vitamin B_6)-dependent cystathionine β-synthase increase plasma methionine levels. Homocysteine is transformed into methionine via remethylation. This occurs through methionine synthase, a reaction requiring methylcobalamin and folic acid. Deficiencies in these enzymes or lack of cofactors is associated with decreased or normal methionine levels. In an alternative pathway, homocysteine can be remethylated by betaine:homocysteine methyl transferase.

in pregnant women and dementia (Alzheimer's type) in the general population. Vitamin supplements are effective in reducing plasma homocysteine levels in these cases, although there are limited effects on cardiovascular disease.

ALKAPTONURIA

Alkaptonuria is a rare (frequency 1:200,000) disorder of tyrosine catabolism in which deficiency of homogentisate 1,2-dioxygenase (also known as *homogentisic acid oxidase*) leads to excretion of large amounts of homogentisic acid in urine and accumulation of oxidized homogentisic acid pigment in connective tissues (*ochronosis*). Alkaptonuria may go unrecognized until middle life, when degenerative joint disease develops. Prior to this time, about half of patients might be diagnosed for the presence of urine that become dark with standing or addition of alkali. Foci of gray-brown scleral pigment and generalized darkening of the concha, anthelix, and, finally, helix of the ear usually develop after age 30. Low back pain usually starts between 30 and 40 years of age. *Ochronotic arthritis* is heralded by pain, stiffness, and some limitation of motion of the hips, knees, and shoulders. Acute arthritis may resemble rheumatoid arthritis, but small joints are usually spared. Pigmentation of heart valves, larynx, tympanic membranes, and skin occurs, and occasional patients develop pigmented renal or prostatic calculi. Pigment deposition in the heart and blood vessels leads to aortic stenosis necessitating valve replacement, especially after 60 years of age. The diagnosis should be suspected in a patient whose urine darkens to blackness. Homogentisic acid in urine is identified by urine organic acid analysis. Ochronotic arthritis is treated symptomatically with pain medications, spinal surgery, and arthroplasty **(Chap. 364)**. Ascorbic acid and protein restriction are not effective in reducing homogentisic acid production. By contrast, nitisinone (2-[2-nitro-4-trifluoromethylbenzoyl]-1,3-cyclohexanedione), a drug used in tyrosinemia type I, reduces urinary excretion of homogentisic acid and, in conjunction with a low-protein diet, might prevent the long-term complications of alkaptonuria.

UREA CYCLE DEFECTS

Excess ammonia generated from protein nitrogen is removed by the urea cycle, a process mediated by several enzymes and transporters (**Fig. 413-2**, Table 413-1). Complete absence of any of these enzymes usually causes severe hyperammonemia in newborns, while milder variants can be seen in adults. The accumulation of ammonia and glutamine leads to direct neuronal toxicity and brain edema. Deficiencies in urea cycle enzymes are individually rare, but as a group, they affect about 1:35,000 individuals. They are all transmitted as autosomal recessive traits, with the exception of ornithine transcarbamylase deficiency, which is X-linked and the most frequent urea cycle defect. Hepatocytes of females with ornithine transcarbamylase deficiency express either the normal or the mutant allele due to random X-inactivation and may be unable to remove excess ammonia if mutant cells are predominant.

Infants with classic urea cycle defects present at 1–4 days of life with refusal to eat and lethargy progressing to coma and death. Milder enzyme deficiencies present with protein avoidance, recurrent vomiting, migraine, mood swings, chronic fatigue, irritability, and disorientation that can progress to coma. Some cases have presented with acute or chronic hepatic dysfunction. Females with ornithine transcarbamylase deficiency can present at time of childbirth due to the combination of involuntary fasting and stress that favors catabolism. Administration of systemic corticosteroids can precipitate hyperammonemia and can be fatal in previously asymptomatic individuals. These patients may be misdiagnosed as having gastrointestinal disorders, food allergies, behavioral problems, or nonspecific hepatitis. The diagnosis requires measurement of plasma ammonia, plasma amino acids, and urine orotic acid, useful for differentiating ornithine transcarbamylase deficiency from carbamyl phosphate synthase-1 and *N*-acetylglutamate synthase deficiency. Increased plasma glutamine is seen with all urea cycle defects since ammonia not removed by the urea cycle in periportal hepatocytes is conjugated to glutamate by glutamine synthase in perivenous hepatocytes. Citrulline is low or undetectable in proximal defects of the urea cycle (*N*-acetylglutamate

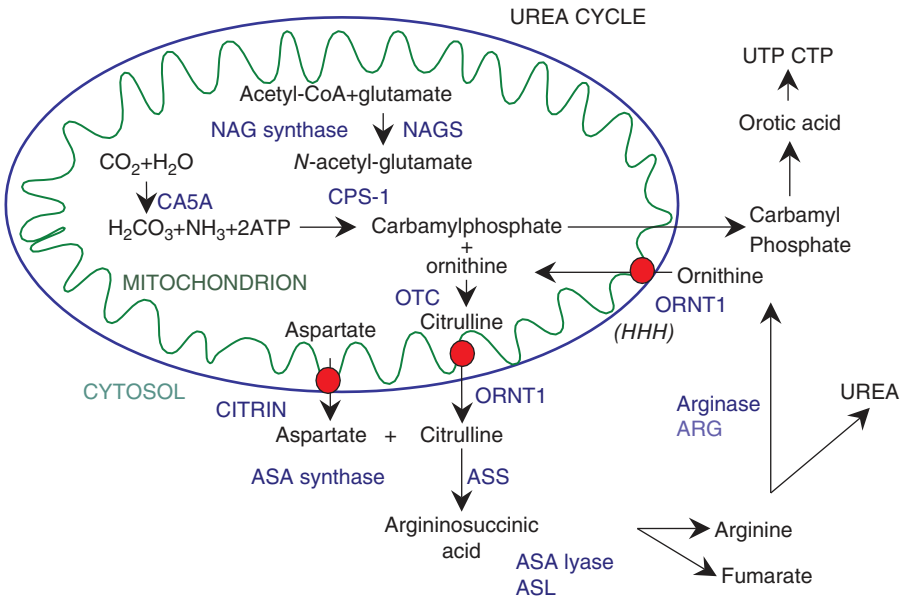


FIGURE 413-2 The urea cycle. This cycle, which is fully expressed only in the liver, forms urea starting from ammonia (NH_3) derived from the nitrogen group of all amino acids. It requires many enzymes and mitochondrial transporters, any of which can be defective and may impair the function of the urea cycle. Ammonia escaping the urea cycle in periportal hepatocytes is conjugated with glutamate by glutamine synthase in perivenous hepatocytes to generate glutamine. ARG, arginase; ASA, Argininosuccinic acid; ASL, Arginino succinate lyase; ASS, argininosuccinate synthase; CA5A, carbonic anhydrase 5a; citrin (*SLC25A13*), aspartate/glutamate exchanger; CP, carbamylphosphate; CPS-1, carbamyl phosphate synthase 1; CTP, cytidine triphosphate; HHH, hyperammonemia, hyperornithinemia, homocitrullinuria syndrome; NAG, *N*-acetylglutamate; NAGS, N-acetylglutamate synthase; ORNT1 (*SLC25A15*), ornithine/citrulline mitochondrial transporter; OTC, ornithine transcarbamylase; UTP, uridine triphosphate.

synthase, carbamylphosphate synthase 1, and ornithine transcarbamylase deficiency), with urine orotic acid being increased only in ornithine transcarbamylase deficiency. Plasma citrulline is markedly increased in argininosuccinic acid synthase deficiency (citrullinemia type 1), with a milder elevation in argininosuccinic acid lyase deficiency in the presence of argininosuccinic acid (argininosuccinic aciduria). Arginine levels are usually normal to low in these conditions and become markedly elevated only in patients with arginase deficiency. In addition to urea cycle defects, hyperammonemia can also be caused by liver disease from any cause and several organic acidemias and fatty acid oxidation defects (the latter two excluded by the analysis of urine organic acids and plasma acylcarnitine profile).

TREATMENT

Urea Cycle Defects

Therapy is aimed at stopping catabolism and ammonia production by providing adequate calories (as IV glucose and lipids in the comatose patient) and, if needed, insulin. Excess nitrogen is removed by IV phenylacetate and benzoate (0.25 g/kg for the priming dose and subsequently as an infusion over 24 h) that conjugate with glutamine and glycine, respectively, to form phenylacetylglutamine and hippuric acid, water-soluble molecules efficiently excreted in urine. Arginine (200 mg/kg per day) becomes an essential amino acid (except in arginase deficiency) and should be provided intravenously to resume protein synthesis. If these measures fail to reduce ammonia, hemodialysis should be initiated promptly. Chronic therapy consists of a protein-restricted diet, phenylbutyrate, glycerol phenylbutyrate (a liquid drug better tolerated by most patients), arginine, or citrulline supplements, depending on the specific diagnosis. Oral carglumic acid can restore a functional urea cycle in patients with *N*-acetylglutamate synthase deficiency and can render other therapies unnecessary. Liver transplantation should be considered in patients with severe urea cycle defects that are difficult to control medically.

Hyperammonemia due to a functional deficiency of glutamine synthase can occur in patients receiving chemotherapy for different malignancies or undergoing solid organ transplants. It can also be seen with hepatic cirrhosis. Several of these patients have been successfully rescued from hyperammonemia using the protocol described above for urea cycle defects.

FURTHER READING

Camp KM et al: Phenylketonuria Scientific Review Conference: State of the science and future research needs. Mol Genet Metab 112:87, 2014.

Ranganath LR et al: Suitability Of Nitisinone In Alkaptonuria 1 (SONIA 1): An international, multicentre, randomised, open-label, no-treatment controlled, parallel-group, dose-response study to investigate the effect of once daily nitisinone on 24-h urinary homogentisic acid excretion in patients with alkaptonuria after 4 weeks of treatment. Ann Rheum Dis 75:362, 2016.

Smith AD, Refsum H: Homocysteine, B vitamins, and cognitive impairment. Annu Rev Nutr 36:211, 2016.

414 Inherited Defects of Membrane Transport

Nicola Longo

Specific membrane transporters mediate the passage of a wide variety of substances across cellular membranes. Classes of substrates include amino acids, sugars, cations, anions, vitamins, and water. The number of inherited disorders of membrane transport continues to increase with the identification of new transporters on the plasma membrane or intracellular organelles and the clarification of the molecular basis of diseases with previously unknown pathophysiology. The first transport disorders identified affected the gut or the kidney, but transport processes are now proving essential for the normal function of every organ. Mutations in transporter molecules cause disorders of the heart, muscle, brain, and endocrine and sensory organs **(Table 414-1)**. Inherited defects impairing the transport of selected amino acids that can

TABLE 414-1 Genetic Disorders of Membrane Transport (Selected Examples)

CLASS OF SUBSTANCE AND DISORDER	INDIVIDUAL SUBSTRATES	TISSUES MANIFESTING TRANSPORT DEFECT	MOLECULAR DEFECT	MAJOR CLINICAL MANIFESTATIONS	INHERITANCE
Amino Acids					
Cystinuria	Cystine, lysine, arginine, ornithine	Proximal renal tubule, jejunal mucosa	Shared dibasic-cystine transporter *SLC3A1, SLC7A9*	Cystine nephrolithiasis	AR
Lysinuric protein intolerance	Lysine, arginine, ornithine	Proximal renal tubule, jejunal mucosa	Dibasic transporter *SLC7A7*	Protein intolerance, hyperammonemia, intellectual disability	AR
Hartnup disease	Neutral amino acids	Proximal renal tubule, jejunal mucosa	Neutral amino acid transporter *SLC6A19*	Constant neutral aminoaciduria, intermittent symptoms of pellagra	AR
Brain branched-chain amino acid deficiency	Leucine, Isoleucine, Valine	Plasma membrane of blood brain barrier	Branched-chain amino acid transporter *SLC7A5*	Microcephaly, intellectual disability, seizures	AR
Citrullinemia type 2	Aspartate, glutamate, malate	Inner mitochondrial membrane	Mitochondrial aspartate/ glutamate carrier 2 *SLC25A13*	Sudden behavioral changes with stupor, coma, hyperammonemia	AR
Hyperornithinemia, hyperammonemia, homocitrullinuria	Ornithine, citrulline	Inner mitochondrial membrane	Mitochondrial ornithine carrier *SLC25A15*	Vomiting, lethargy, failure to thrive, intellectual disability, episodic confusion, hyperammonemia, protein intolerance	AR
Histidinuria	Histidine	Proximal renal tubule, jejunal mucosa	Histidine transporter	Intellectual disability	AR
Iminoglycinuria	Glycine, proline, hydroxyproline	Proximal renal tubule, jejunal mucosa	Shared glycine–imino acid transporter *SLC36A2, SLC6A19, SLC6A20*	None	AR
Dicarboxylic aminoaciduria	Glutamic acid, aspartic acid	Proximal renal tubule, jejunal mucosa	Shared dicarboxylic amino acid transporter *SLC1A1*	None	AR
Cystinosis	Cystine	Lysosomal membranes	Lysosomal cystine transporter	Renal failure, hypothyroidism, blindness	AR
Hexoses					
Glucose-galactose malabsorption	D-Glucose D-Galactose	Proximal renal tubule, jejunal mucosa	Sodium-dependent glucose/galactose transporter SGLT1	Watery diarrhea on feeding glucose, lactose, sucrose, or galactose	AR
Glucose-transport defect	D-Glucose	Ubiquitous blood-brain barrier	Facilitative glucose transporter GLUT1	Seizures, intellectual disability	AD
Fanconi-Bickel syndrome	D-Glucose	Liver, kidney, pancreas, intestine	Facilitative glucose transporter GLUT2	Growth retardation, rickets, hepatorenal glycogenosis, hypo- and hyperglycemia	AR
Urate					
Hypouricemia	Uric acid	Proximal renal tubule	Urate transporter *SLC22A12*	Hypouricemia, uric acid urolithiasis	AR
Vitamins					
Thiamine-responsive megaloblastic anemia	Thiamine	Ubiquitous	Thiamine transporter *SLC19A2*	Megaloblastic anemia, deafness, diabetes mellitus	AR
Biotin-thiamine-responsive basal ganglia disease	Biotin, thiamine	Ubiquitous	Biotin-thiamine transporter *SLC19A3*	Dystonia, seizures, psychomotor delay, Wernicke-like encephalopathy	AR
Riboflavin transporter deficiencies	Riboflavin	Blood brain barrier	Riboflavin transporters *SLC52A2, SLC52A3*	Ataxia, weakness, neuropathy, hearing loss,	AR
Other					
Carnitine deficiency	Carnitine	Kidney, muscle, heart	Carnitine transporter OCTN2	Hypoketotic hypoglycemia, cardiomyopathy, sudden death	AR
Creatine deficiency	Creatine	Brain	Creatine transporter *SLC6A8*	Intellectual disability, seizures, hypotonia	XL

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; XL, X-linked recessive.

present in adults are discussed here as examples of the abnormalities encountered; others are considered elsewhere in this text.

CYSTINURIA

Cystinuria (frequency of 1 in 10,000 to 1 in 15,000) is an autosomal recessive disorder caused by defective transporters in the apical brush border of proximal renal tubule and small intestinal cells. It is characterized by impaired reabsorption and excessive urinary excretion of the dibasic amino acids lysine, arginine, ornithine, and cystine. Because cystine is poorly soluble, its excess excretion predisposes to the formation of renal, ureteral, and bladder stones. Such stones are responsible for the signs and symptoms of the disorder.

There are two variants of cystinuria. Homozygotes for both variants have high urinary excretion of cystine, lysine, arginine, and ornithine. Type I heterozygotes usually have normal urinary amino acid excretion, whereas most non–type I (formerly type II and type III) heterozygotes have moderately increased urinary excretion of each of the four amino acids. The gene for type I cystinuria (*SLC3A1*, chromosome 2p16.3) encodes a membrane glycoprotein. Non–type I cystinuria is caused by mutations in *SLC7A9* (chromosome 19q13) that encodes the $b^{0,+}$ amino acid transporter. The glycoprotein encoded by *SLC3A1* favors the correct processing of the $b^{0,+}$ membrane transporter and explains why mutations in two different genes cause a similar disease.

Cystine stones account for 1–2% of all urinary tract calculi but are the most common cause of stones in children. Cystinuria homozygotes regularly excrete 2400–7200 µmol (600–1800 mg) of cystine daily. Since the maximum solubility of cystine in the physiologic urinary pH range of 4.5–7.0 is about 1200 µmol/L (300 mg/L), cystine needs to be diluted to 2.5–7 L of water to prevent crystalluria. Stone formation usually manifests in the second or third decade but may occur in the first year of life. Symptoms and signs are those typical of urolithiasis: hematuria, flank pain, renal colic, obstructive uropathy, and infection **(Chap. 312)**. Recurrent urolithiasis may lead to progressive renal insufficiency.

Cystinuria is suspected after observing typical hexagonal crystals in the sediment of acidified, concentrated, chilled urine, or after performing a urinary nitroprusside test. Quantitative urine amino acid analysis confirms the diagnosis of cystinuria by showing selective overexcretion of cystine, lysine, arginine, and ornithine. Quantitative measurements are important for differentiating heterozygotes from homozygotes and for following free cystine excretion during therapy.

Management is aimed at preventing cystine crystal formation by increasing urinary volume and by maintaining an alkaline urine pH. Fluid ingestion in excess of 4 L/d is essential, and 5–7 L/d is optimal. Urinary cystine concentration should be <1000 µmol/L (250 mg/L). The daily fluid ingestion necessary to maintain this dilution of excreted cystine should be spaced over 24 h, with one-third of the total volume ingested between bedtime and 3 A.M. Cystine solubility rises sharply above pH 7.5, and urinary alkalinization (with bicarbonate or potassium citrate) can be therapeutic. Penicillamine (1–3 g/d) and tiopronin (α-mercaptopropionylglycine, 800–1200 mg/d in four divided doses) undergo sulfhydryl-disulfide exchange with cystine to form mixed disulfides. Because these disulfides are much more soluble than cystine, pharmacologic therapy can prevent and promote dissolution of calculi. Penicillamine can have significant side effects and should be reserved for patients who fail to respond to hydration alone or who are in a high-risk category (e.g., one remaining kidney, renal insufficiency). When medical management fails, shock waves lithotripsy, ureteroscopy, and percutaneous nephrolithotomy are effective for most stones. Open urologic surgery is considered only for complex staghorn stones or when the patient has concomitant renal or ureteral abnormalities. Occasional patients progress to renal failure and require kidney transplantation.

LYSINURIC PROTEIN INTOLERANCE

This disorder is characterized by a defect in renal tubular reabsorption of the three dibasic amino acids lysine, arginine, and ornithine but *not* cystine (*lysinuric protein intolerance*). Homozygotes show defective intestinal transport of dibasic amino acids as well as exaggerated renal losses. Lysinuric protein intolerance is most common in Finland (1 in 60,000), southern Italy, and Japan, but is rare elsewhere. The transport defect affects basolateral rather than luminal membrane transport and is associated with impairment of the urea cycle. The defective gene (*SLC7A7*, chromosome 14q11.2) encodes the y^+LAT membrane transporter, which associates with the cell-surface glycoprotein 4F2 heavy chain to form the complete sodium-independent transporter y^+L.

Manifestations are related to impairment of the urea cycle and to immune dysfunction potentially attributable to nitric oxide overproduction secondary to arginine intracellular trapping within macrophages. Affected patients present in childhood with hepatosplenomegaly, protein intolerance, and episodic ammonia intoxication. Older patients may present with severe osteoporosis, impairment of kidney function, pulmonary alveolar proteinosis, various autoimmune disorders, and an incompletely characterized immune deficiency. Plasma concentrations of lysine, arginine, and ornithine are reduced, whereas urinary excretion of lysine and orotic acid are increased. Hyperammonemia may develop after the ingestion of protein loads or with infections, probably because of insufficient amounts of arginine and ornithine to maintain proper function of the urea cycle. Therapy consists of dietary protein restriction and supplementation of citrulline (2–8 g/d), a neutral amino acid that fuels the urea cycle when metabolized to arginine and ornithine. Pulmonary disease responds to glucocorticoids or recombinant human GM-CSF in some patients. Women with lysinuric protein intolerance who become pregnant have an increased risk of anemia, toxemia, and bleeding complications during delivery. These can be minimized by aggressive nutritional therapy and control of blood pressure. Their infants can have intrauterine growth restriction but have normal neurologic function.

CITRULLINEMIA TYPE 2 (CITRIN DEFICIENCY)

Citrullinemia type 2 is a recessive condition caused by deficiency of the mitochondrial aspartate-glutamate carrier AGC2 (citrin). A defect in this transporter reduces the availability of cytoplasmic aspartate to combine with citrulline to form argininosuccinate **(see Fig. 413-1)**, impairing the urea cycle and decreasing the transfer of reducing equivalents from the cytosol to the mitochondria through the malate–aspartate NADH shuttle. Mutations in the *SLC25A13* gene on chromosome 7q21.3 that encodes for this transporter are rare in Caucasians, but affect about 1:20,000 people with ancestry from Japan, China, and Southeast Asia with variable penetrance.

The disease can present in children with neonatal intrahepatic cholestasis, failure to thrive, and dyslipidemia, but usually presents with sudden onset between 20 and 50 years of age with recurring episodes of hyperammonemia with associated neuropsychiatric symptoms such as altered mental status, irritability, seizures, or coma-resembling hepatic encephalopathy. Some patients might come to medical attention for hypertriglyceridemia, pancreatitis, hepatoma, or fatty liver histologically similar to nonalcoholic steatohepatitis. Without therapy, most patients die with cerebral edema within a few years of onset. Episodes are usually triggered by medications (such as acetaminophen), surgery, alcohol consumption or high sugar intake, the latter conditions causing excess NADH production. NADH is not generated by the metabolism of proteins or fats, and many individuals with citrullinemia type 2 spontaneously prefer foods such as meat, eggs, and fish, and avoid carbohydrates.

Laboratory studies during an acute attack include elevated ammonia, citrulline, and arginine with low or normal levels of glutamine (the latter is usually increased in classic urea cycle defects). Levels of galactose-1-phosphate in red blood cells are also increased, reflecting defective transfer of reducing equivalents from the cytosol to mitochondria. The diagnosis is confirmed by demonstrating mutations in the *SLC25A13* gene. Liver transplantation prevents progression of the disease and normalizes biochemical parameters. A diet high in fats and proteins and low in carbohydrates with supplements of medium chain triglycerides, arginine, and pyruvate is also effective in preventing further episodes, at least in the short term.

HARTNUP DISEASE

Hartnup disease (frequency 1 in 24,000) is an autosomal recessive disorder characterized by pellagra-like skin lesions, variable neurologic

manifestations, and neutral and aromatic aminoaciduria. Alanine, serine, threonine, valine, leucine, isoleucine, phenylalanine, tyrosine, tryptophan, glutamine, asparagine, and histidine are excreted in urine in quantities 5–10 times greater than normal, and intestinal transport of these same amino acids is defective. The defective neutral amino acid transporter, B°AT1 encoded by the *SLC6A19* gene on chromosome 5p15, requires either collectrin or angiotensin-converting enzyme 2 for surface expression in the kidney and intestine, respectively.

The clinical manifestations result from nutritional deficiency of the essential amino acid tryptophan, caused by its intestinal and renal malabsorption, and of niacin, which derives in part from tryptophan metabolism. Only a small fraction of patients with the chemical findings of this disorder develop a pellagra-like syndrome, implying that manifestations depend on other factors in addition to the transport defect. The diagnosis of Hartnup disease should be suspected in any patient with clinical features of pellagra who does not have a history of dietary niacin deficiency **(Chap. 326)**. The neurologic and psychiatric manifestations range from attacks of cerebellar ataxia to mild emotional lability to frank delirium, and they are usually accompanied by exacerbations of the erythematous, eczematoid skin rash. Fever, sunlight, stress, and sulfonamide therapy provoke clinical relapses. Diagnosis is made by detection of the neutral aminoaciduria, which does not occur in dietary niacin deficiency. Treatment is directed at niacin repletion and includes a high-protein diet and daily nicotinamide supplementation (50–250 mg).

CYSTINOSIS

Cystinosis (frequency 1:100,000–1:200,000) is an autosomal recessive disorder caused by mutations in the *CTNS* gene encoding the lysosomal cystine/proton transporter (cystinosin). In this condition, cystine derived from protein degradation accumulates inside lysosomes and forms crystals due to its poor solubility. Depending on the degree of impairment of transporter function, three clinical forms are recognized. The most severe form, classic nephropathic cystinosis, causes renal Fanconi syndrome during the first year of life and, without treatment, evolves to renal failure usually by ten years of age. Intermediate nephropathic cystinosis leads to kidney failure between 15 and 25 years of age while photophobia, caused by deposition of cystine crystals in the cornea, is the only manifestation of ocular non-nephropathic cystinosis. Cystinosis is suspected by the identification of cystine crystals in the cornea by slit lamp examination and diagnosed by measuring cystine content in white blood cells. DNA testing (including deletion analysis) of the *CTNS* gene can further confirm the diagnosis. Therapy consists in the administration of cysteamine that enters lysosomes, forms a mixed disulfide with cysteine, and is exported from the lysosome using a cationic amino acid transporter. Oral cysteamine therapy (60–90 mg/kg per day up to 2 g per day in adults, 0.2 to 0.3 grams/m^2 per day divided into two doses given every 12 hours for the extended release formulation) can delay renal failure and is more effective if started early in the course of the disease. Therapy with cysteamine reduces intracellular cystine accumulation in white blood cells, but compliance with therapy is difficult due to the unpleasant odor of the drug and the need for frequent administration. Cysteamine eye drops can relieve photophobia. Renal replacement therapy with salts, alkali, and activated vitamin D is necessary for renal Fanconi syndrome. Cystine accumulation occurs in virtually all organs and tissues, causing additional complications such as hypothyroidism, hypohydrosis, diabetes, delayed puberty in both males and females with primary hypogonadism in males. Growth hormone replacement, L-thyroxine for hypothyroidism, insulin for diabetes mellitus, and testosterone for hypogonadism in males may be necessary. Despite therapy, many patients with cystinosis progress to end-stage renal failure and require kidney transplantation. Late-onset complications include hepatomegaly and splenomegaly that occur in approximately one-third of subjects and a vacuolar myopathy causing weakness (initially involving the distal extremities), swallowing difficulties, gastrointestinal dysmotility and pulmonary insufficiency. Before the availability of cystine-depleting therapy and renal transplantation, the life span in nephropathic cystinosis was less than ten years. With current therapies, affected individuals can survive into the late forties with satisfactory quality of life.

FURTHER READING

EMMA F et al: Nephropathic cystinosis: An international consensus document. Nephrol Dial Transplant. 29 Suppl 4:87, 2014.

TANNER LM et al: Inhaled sargramostim induces resolution of pulmonary alveolar proteinosis in lysinuric protein intolerance. JIMD Rep 34:97, 2017.

TĂRLUNGEANU DC et al: Impaired amino acid transport at the blood brain barrier is a cause of autism spectrum disorder. Cell 167:1481, 2016.

Section 1 Diagnosis of Neurologic Disorders

415 Approach to the Patient with Neurologic Disease

Daniel H. Lowenstein, Joseph B. Martin, Stephen L. Hauser

Neurologic diseases are common and costly. According to estimates by the World Health Organization, neurologic disorders affect over 1 billion people worldwide, constitute 12% of the global burden of disease, and cause 14% of global deaths (Table 415-1). These numbers are only expected to increase as the world's population ages. Because therapies now exist for many neurologic disorders, a skillful approach to diagnosis is essential. Errors commonly result from an overreliance on costly neuroimaging procedures and laboratory tests, which, while useful, do not substitute for an adequate history and examination. The proper approach begins with the patient and focuses the clinical problem first in anatomic and then in pathophysiologic terms; only then should a specific neurologic diagnosis be entertained. This method ensures that technology is judiciously applied, a correct diagnosis is established in an efficient manner, and treatment is promptly initiated.

THE NEUROLOGIC METHOD

DEFINE THE ANATOMY

The first priority is to identify the region of the nervous system that is likely to be responsible for the symptoms. Can the disorder be mapped to one specific location, is it multifocal, or is a diffuse process present? Are the symptoms restricted to the nervous system, or do they arise in the context of a systemic illness? Is the problem in the central nervous system (CNS), the peripheral nervous system (PNS), or both? In the CNS, is the cerebral cortex, basal ganglia, brainstem, cerebellum, or spinal cord responsible? Are the pain-sensitive meninges involved? In the PNS, could the disorder be located in peripheral nerves and, if so, are motor or sensory nerves primarily affected, or is a lesion in the neuromuscular junction or muscle more likely?

The first clues to defining the anatomic area of involvement appear in the history, and the examination is then directed to confirm or rule out these impressions and to clarify uncertainties. A more detailed examination of a particular region of the CNS or PNS is often indicated. For example, the examination of a patient who presents with a history of ascending paresthesias and weakness should be directed toward deciding, among other things, if the lesion is in the spinal cord or peripheral nerves. Focal back pain, a spinal cord sensory level, and incontinence suggest a spinal cord origin, whereas a stocking-glove pattern of sensory loss suggests peripheral nerve disease; areflexia usually indicates peripheral neuropathy but may also be present with spinal shock in acute spinal cord disorders.

Deciding "where the lesion is" accomplishes the task of limiting the possible etiologies to a manageable, finite number. In addition, this strategy safeguards against making serious errors. Symptoms of recurrent vertigo, diplopia, and nystagmus should not trigger "multiple sclerosis" as an answer (etiology) but "brainstem" or "pons" (location); then a diagnosis of brainstem arteriovenous malformation will not be missed for lack of consideration. Similarly, the combination of optic neuritis and spastic ataxic paraparesis suggests optic nerve and spinal cord disease; multiple sclerosis (MS), CNS syphilis, and vitamin B_{12} deficiency are treatable disorders that can produce this syndrome. Once the question, "Where is the lesion?" is answered, then the question "What is the lesion?" can be addressed.

TABLE 415-1 Global Disability-Adjusted Life-Years (DALYs) and Number of Annual Deaths for Selected Neurologic Disorders in 2015

DISORDER	DALYs	DEATHS
Low back and neck pain	94,941,000	—
Cerebrovascular diseases	118,627,000	6,326,000
Meningitis and encephalitis	33,848,000	529,000
Migraine	32,898,000	—
Epilepsy	12,418,000	125,000
Dementia	23,779,000	1,908,000
Parkinson's disease	2,059,000	117,000
% of total DALYs or deaths for all causes that are neurologic	9.7%	16.3%
% change of DALYs for neurologic disorders between 2005 and 2015	17.0%	35.2%

Source: GBD 2015 DALYs and HALE Collaborators: Global, Regional, and National Disability-Adjusted Life-Years (Dalys) for 315 Diseases and Injuries and Healthy Life Expectancy (HALE), 1990–2015: A Systematic Analysis for the Global Burden of Disease Study 2015. Lancet 388:1603, 2016; and GBD 2015 Mortality and Causes of Death Collaborators: Global, Regional, and National Life Expectancy, All-Cause Mortality, and Cause-Specific Mortality for 249 Causes of Death, 1980–2015: A Systematic Analysis for the Global Burden of Disease Study 2015. Lancet 388:1459, 2016.

IDENTIFY THE PATHOPHYSIOLOGY

Clues to the pathophysiology of the disease process may also be present in the history. Primary neuronal (gray matter) disorders often present as early cognitive disturbances, movement disorders, or seizures, whereas white matter involvement produces "long tract" disorders of motor, sensory, visual, and cerebellar pathways. Progressive and symmetric symptoms often have a metabolic or degenerative origin; in such cases lesions are usually not sharply circumscribed. Thus, a patient with paraparesis and a clear spinal cord sensory level is unlikely to have vitamin B_{12} deficiency as the explanation. A Lhermitte symptom (electric shock–like sensations evoked by neck flexion) is due to ectopic impulse generation in white matter pathways and occurs with demyelination in the cervical spinal cord; among many possible causes, this symptom may indicate MS in a young adult or compressive cervical spondylosis in an older person. Symptoms that worsen after exposure to heat or exercise may indicate conduction block in demyelinated axons, as occurs in MS. A patient with recurrent episodes of diplopia and dysarthria associated with exercise or fatigue may have a disorder of neuromuscular transmission such as myasthenia gravis. Slowly advancing visual scotoma with luminous edges, termed *fortification spectra*, indicates spreading cortical depression, typically with migraine.

THE NEUROLOGIC HISTORY

Attention to the description of the symptoms experienced by the patient and substantiated by family members and others often permits an accurate localization and determination of the probable cause, even before the neurologic examination is performed. The history also helps focus the neurologic examination that follows. Each complaint should be pursued as far as possible to identify the location of the lesion, the likely underlying pathophysiology, and potential etiologies. For example, a patient complains of weakness of the right arm. What are the associated features? Does the patient have difficulty with brushing hair or reaching upward (proximal) or buttoning buttons or opening a twist-top bottle (distal)? Negative associations may also be crucial. A patient with a right hemiparesis without a language deficit likely has a lesion (internal capsule, brainstem, or spinal cord) different from that of a patient with a right hemiparesis and aphasia (left hemisphere). Other pertinent features of the history include the following:

1. *Temporal course of the illness*. It is important to determine the precise time of appearance and rate of progression of the symptoms experienced by the patient. The rapid onset of a neurologic complaint, occurring within seconds or minutes, usually indicates a vascular event, a seizure, or migraine. The onset of sensory symptoms located in one extremity that spread over a few seconds to adjacent portions of that extremity and then to the other regions of the body suggests a seizure. A similar but slower temporal march of symptoms accompanied by headache, nausea, or visual disturbance suggests migraine. Less well-localized symptoms that are either sudden or more gradual in onset point to the possibility of a transient ischemic attack (TIA). The presence of "positive" sensory symptoms (e.g., tingling or sensations that are difficult to describe) or involuntary motor movements suggests a seizure; in contrast, transient loss of function (negative symptoms) suggests a TIA. A stuttering onset where symptoms appear, stabilize, and then progress over hours or days also suggests cerebrovascular disease; an additional history of transient remission or regression indicates that the process is more likely due to ischemia rather than hemorrhage. A gradual evolution of symptoms over hours or days suggests a toxic, metabolic, infectious, or inflammatory process. Progressing symptoms associated with the systemic manifestations of fever, stiff neck, and altered level of consciousness imply an infectious process. Relapsing and remitting symptoms involving different levels of the nervous system suggest MS or other inflammatory processes. Slowly progressive symptoms without remissions are characteristic of neurodegenerative disorders, chronic infections, gradual intoxications, and neoplasms.
2. *Patients' descriptions of the complaint*. The same words often mean different things to different patients. "Dizziness" may imply impending syncope, a sense of disequilibrium, or true spinning vertigo. "Numbness" may mean a complete loss of feeling, a positive sensation such as tingling, or even weakness. "Blurred vision" may be used to describe unilateral visual loss, as in transient monocular blindness, or diplopia. The interpretation of the true meaning of the words used by patients to describe symptoms obviously becomes even more complex when there are differences in primary languages and cultures.
3. *Corroboration of the history by others*. It is almost always helpful to obtain additional information from family, friends, or other observers to corroborate or expand the patient's description. Memory loss, aphasia, loss of insight, intoxication, and other factors may impair the patient's capacity to communicate normally with the examiner or prevent openness about factors that have contributed to the illness. Episodes of loss of consciousness necessitate that details be sought from observers to ascertain precisely what has happened during the event.
4. *Family history*. Many neurologic disorders have an underlying genetic component. The presence of a Mendelian disorder, such as Huntington's disease or Charcot-Marie-Tooth neuropathy, is often obvious if family data are available. More detailed questions about family history are often necessary in polygenic disorders such as MS, migraine, and many types of epilepsy. It is important to elicit family history about all illnesses, in addition to neurologic and psychiatric disorders. A familial propensity to hypertension or heart disease is relevant in a patient who presents with a stroke. There are numerous inherited neurologic diseases that are associated with multisystem manifestations that may provide clues to the correct diagnosis (e.g., neurofibromatosis, Wilson's disease, mitochondrial disorders).
5. *Medical illnesses*. Many neurologic diseases occur in the context of systemic disorders. Diabetes mellitus, hypertension, and abnormalities of blood lipids predispose to cerebrovascular disease. A solitary mass lesion in the brain may be an abscess in a patient with valvular heart disease, a primary hemorrhage in a patient with a coagulopathy, a lymphoma or toxoplasmosis in a patient with AIDS, or a metastasis in a patient with underlying cancer. Patients with malignancy may also present with a neurologic paraneoplastic syndrome **(Chap. 90)** or complications from chemotherapy or radiotherapy. Marfan's syndrome and related collagen disorders predispose to dissection of the cranial arteries and aneurysmal subarachnoid hemorrhage; the latter may also occur with polycystic kidney disease. Various neurologic disorders occur with dysthyroid states or other endocrinopathies. It is especially important to look for the presence of systemic diseases in patients with peripheral neuropathy. Most patients with coma in a hospital setting have a metabolic, toxic, or infectious cause.
6. *Drug use and abuse and toxin exposure*. It is essential to inquire about the history of drug use, both prescribed and illicit. Sedatives, antidepressants, and other psychoactive medications are frequently associated with acute confusional states, especially in the elderly. Aminoglycoside antibiotics may exacerbate symptoms of weakness in patients with disorders of neuromuscular transmission, such as myasthenia gravis, and may cause dizziness secondary to ototoxicity. Vincristine and other antineoplastic drugs can cause peripheral neuropathy, and immunosuppressive agents such as cyclosporine can produce encephalopathy. Excessive vitamin ingestion can lead to disease; examples include vitamin A and pseudotumor cerebri or pyridoxine and peripheral neuropathy. Many patients are unaware that over-the-counter sleeping pills, cold preparations, and diet pills are actually drugs. Alcohol, the most prevalent neurotoxin, is often not recognized as such by patients, and other drugs of abuse such as cocaine and heroin can cause a wide range of neurologic abnormalities. A history of environmental or industrial exposure to neurotoxins may provide an essential clue; consultation with the patient's coworkers or employer may be required.
7. *Formulating an impression of the patient*. Use the opportunity while taking the history to form an impression of the patient. Is the information forthcoming, or does it take a circuitous course? Is there evidence of anxiety, depression, or hypochondriasis? Are there any clues to problems with language, memory, insight, comportment, or behavior? The neurologic assessment begins as soon as the patient comes into the room and the first introduction is made.

THE NEUROLOGIC EXAMINATION

The neurologic examination is challenging and complex; it has many components and includes a number of skills that can be mastered only through repeated use of the same techniques on a large number of individuals with and without neurologic disease. Mastery of the complete neurologic examination is usually important only for physicians in neurology and associated specialties. However, knowledge of the basics of the examination, especially those components that are effective in screening for neurologic dysfunction, is essential for all clinicians, especially generalists.

There is no single, universally accepted sequence of the examination that must be followed, but most clinicians begin with assessment of mental status followed by the cranial nerves (CN), motor system, reflexes, sensory system, coordination, and gait. Whether the examination is basic or comprehensive, it is essential that it is performed in an orderly and systematic fashion to avoid errors and serious omissions. Thus, the best way to learn and gain expertise in the examination is to choose one's own approach and practice it frequently and do it in the same exact sequence each time.

The detailed description that follows describes the more commonly used parts of the neurologic examination, with a particular emphasis on the components that are considered most helpful for the assessment of common neurologic problems. Each section also includes a brief description of the minimal examination necessary to adequately screen for abnormalities in a patient who has no symptoms suggesting neurologic dysfunction. A screening examination done in this way can be completed in 3–5 min. *Video demonstrations of the neurologic screening examination (V6) and the detailed neurologic examination (V7) can be found in the Harrison's Video Collection included in this textbook.*

Several additional points about the examination are worth noting. First, in recording observations, it is important to describe what is found rather than to apply a poorly defined medical term (e.g., "patient groans to sternal rub" rather than "obtunded"). Second, subtle CNS abnormalities are best detected by carefully comparing a patient's

performance on tasks that require simultaneous activation of both cerebral hemispheres (e.g., eliciting a pronator drift of an outstretched arm with the eyes closed; extinction on one side of bilaterally applied light touch, also with eyes closed; or decreased arm swing or a slight asymmetry when walking). Third, if the patient's complaint is brought on by some activity, reproduce the activity in the office. If the complaint is of dizziness when the head is turned in one direction, have the patient do this and also look for associated signs on examination (e.g., nystagmus or dysmetria). If pain occurs after walking two blocks, have the patient leave the office and walk this distance and immediately return, and repeat the relevant parts of the examination. Finally, the use of tests that are individually tailored to the patient's problem can be of value in assessing changes over time. Tests of walking a 7.5-m (25-ft) distance (normal, 5–6 s; note assistance, if any), repetitive finger or toe tapping (normal, 20–25 taps in 5 s), or handwriting are examples.

MENTAL STATUS EXAMINATION

- *The bare minimum: During the interview, look for difficulties with communication and determine whether the patient has recall and insight into recent and past events.*

The mental status examination is under way as soon as the physician begins observing and speaking with the patient. If the history raises any concern for abnormalities of higher cortical function or if cognitive problems are observed during the interview, then detailed testing of the mental status is indicated. The patient's ability to understand the language used for the examination, cultural background, educational experience, sensory or motor problems, or comorbid conditions needs to be factored into the applicability of the tests and interpretation of results.

The Mini-Mental State Examination (MMSE) is a standardized screening examination of cognitive function that is extremely easy to administer and takes <10 min to complete (Chap. 25). Using age-adjusted values for defining normal performance, the test is ~85% sensitive and 85% specific for making the diagnosis of dementia that is moderate or severe, especially in educated patients. When there is sufficient time available, the MMSE is one of the best methods for documenting the current mental status of the patient, and this is especially useful as a baseline assessment to which future scores of the MMSE can be compared.

Individual elements of the mental status examination can be subdivided into level of consciousness, orientation, speech and language, memory, fund of information, insight and judgment, abstract thought, and calculations.

Level of consciousness is the patient's relative state of awareness of the self and the environment, and ranges from fully awake to comatose. When the patient is not fully awake, the examiner should describe the responses to the minimum stimulus necessary to elicit a reaction, ranging from verbal commands to a brief, painful stimulus such as a squeeze of the trapezius muscle. Responses that are directed toward the stimulus and signify some degree of intact cerebral function (e.g., opening the eyes and looking at the examiner or reaching to push away a painful stimulus) must be distinguished from reflex responses of a spinal origin (e.g., triple flexion response—flexion at the ankle, knee, and hip in response to a painful stimulus to the foot).

Orientation is tested by asking the person to state his or her name, location, and time (day of the week and date); time is usually the first to be affected in a variety of conditions.

Speech is assessed by observing articulation, rate, rhythm, and prosody (i.e., the changes in pitch and accentuation of syllables and words).

Language is assessed by observing the content of the patient's verbal and written output, response to spoken commands, and ability to read. A typical testing sequence is to ask the patient to name successively more detailed components of clothing, a watch, or a pen; repeat the phrase "No ifs, ands, or buts"; follow a three-step, verbal command; write a sentence; and read and respond to a written command.

Memory should be analyzed according to three main time scales: (1) immediate memory is assessed by saying a list of three items and having the patient repeat the list immediately; (2) short-term memory is tested by asking the patient to recall the same three items 5 and 15 min later; and (3) long-term memory is evaluated by determining how well the patient is able to provide a coherent chronologic history of his or her illness or personal events.

Fund of information is assessed by asking questions about major historic or current events, with special attention to educational level and life experiences.

Abnormalities of *insight and judgment* are usually detected during the patient interview; a more detailed assessment can be elicited by asking the patient to describe how he or she would respond to situations having a variety of potential outcomes (e.g., "What would you do if you found a wallet on the sidewalk?").

Abstract thought can be tested by asking the patient to describe similarities between various objects or concepts (e.g., apple and orange, desk and chair, poetry and sculpture) or to list items having the same attributes (e.g., a list of four-legged animals).

Calculation ability is assessed by having the patient carry out a computation that is appropriate to the patient's age and education (e.g., serial subtraction of 7 from 100 or 3 from 20; or word problems involving simple arithmetic).

CRANIAL NERVE EXAMINATION

- *The bare minimum: Check the fundi, visual fields, pupil size and reactivity, extraocular movements, and facial movements.*

The CN are best examined in numerical order, except for grouping together CN III, IV, and VI because of their similar function.

CN I (Olfactory) Testing is often omitted unless there is suspicion for inferior frontal lobe disease (e.g., meningioma). With eyes closed, ask the patient to sniff a mild stimulus such as toothpaste or coffee and identify the odorant.

CN II (Optic) Check visual acuity (with eyeglasses or contact lens correction) using a Snellen chart or similar tool. Test the visual fields by confrontation, i.e., by comparing the patient's visual fields to your own. As a screening test, it is usually sufficient to examine the visual fields of both eyes simultaneously; individual eye fields should be tested if there is any reason to suspect a problem of vision by the history or other elements of the examination, or if the screening test reveals an abnormality. Face the patient at a distance of ~0.6–1.0 m (2–3 ft) and place your hands at the periphery of your visual fields in the plane that is equidistant between you and the patient. Instruct the patient to look directly at the center of your face and to indicate when and where he or she sees one of your fingers moving. Beginning with the two inferior quadrants and then the two superior quadrants, move your index finger of the right hand, left hand, or both hands simultaneously and observe whether the patient detects the movements. A single small-amplitude movement of the finger is sufficient for a normal response. Focal perimetry and tangent screen examinations should be used to map out visual field defects fully or to search for subtle abnormalities. Optic fundi should be examined with an ophthalmoscope, and the color, size, and degree of swelling or elevation of the optic disc noted, as well as the color and texture of the retina. The retinal vessels should be checked for size, regularity, arteriovenous nicking at crossing points, hemorrhage, exudates, etc.

CN III, IV, VI (Oculomotor, Trochlear, Abducens) Describe the size and shape of pupils and reaction to light and accommodation (i.e., as the eyes converge while following your finger as it moves toward the bridge of the nose). To check extraocular movements, ask the patient to keep his or her head still while tracking the movement of the tip of your finger. Move the target slowly in the horizontal and vertical planes; observe any paresis, nystagmus, or abnormalities of smooth pursuit (saccades, oculomotor ataxia, etc.). If necessary, the relative position of the two eyes, both in primary and multidirectional gaze, can be assessed by comparing the reflections of a bright light off both pupils. However, in practice it is typically more useful to determine whether the patient describes diplopia in any direction of gaze; true diplopia should almost always resolve with one eye closed. Horizontal nystagmus is best assessed at 45° and not at extreme lateral gaze (which is uncomfortable for the patient); the target must often

be held at the lateral position for at least a few seconds to detect an abnormality.

CN V (Trigeminal) Examine sensation within the three territories of the branches of the trigeminal nerve (ophthalmic, maxillary, and mandibular) on each side of the face. As with other parts of the sensory examination, testing of two sensory modalities derived from different anatomic pathways (e.g., light touch and temperature) is sufficient for a screening examination. Testing of other modalities, the corneal reflex, and the motor component of CN V (jaw clench—masseter muscle) is indicated when suggested by the history.

CN VII (Facial) Look for facial asymmetry at rest and with spontaneous movements. Test eyebrow elevation, forehead wrinkling, eye closure, smiling, and cheek puff. Look in particular for differences in the lower versus upper facial muscles; weakness of the lower two-thirds of the face with preservation of the upper third suggests an upper motor neuron lesion, whereas weakness of an entire side suggests a lower motor neuron lesion.

CN VIII (Vestibulocochlear) Check the patient's ability to hear a finger rub or whispered voice with each ear. Further testing for air versus mastoid bone conduction (Rinne) and lateralization of a 512-Hz tuning fork placed at the center of the forehead (Weber) should be done if an abnormality is detected by history or examination. Any suspected problem should be followed up with formal audiometry. **For further discussion of assessing vestibular nerve function in the setting of dizziness, hearing loss, or coma, see Chaps. 19, 30, and 300, respectively.**

CN IX, X (Glossopharyngeal, Vagus) Observe the position and symmetry of the palate and uvula at rest and with phonation ("aah"). The pharyngeal ("gag") reflex is evaluated by stimulating the posterior pharyngeal wall on each side with a sterile, blunt object (e.g., tongue blade), but the reflex is often absent in normal individuals.

CN XI (Spinal Accessory) Check shoulder shrug (trapezius muscle) and head rotation to each side (sternocleidomastoid) against resistance.

CN XII (Hypoglossal) Inspect the tongue for atrophy or fasciculations, position with protrusion, and strength when extended against the inner surface of the cheeks on each side.

MOTOR EXAMINATION

- *The bare minimum: Look for muscle atrophy and check extremity tone. Assess upper extremity strength by checking for pronator drift and strength of wrist or finger extensors. Assess lower extremity strength by checking strength of the toe extensors and having the patient walk normally and on heels and toes.*

The motor examination includes observations of muscle appearance, tone, and strength. Although gait is in part a test of motor function, it is usually evaluated separately at the end of the examination.

Appearance Inspect and palpate muscle groups under good light and with the patient in a comfortable and symmetric position. Check for muscle fasciculations, tenderness, and atrophy or hypertrophy. Involuntary movements may be present at rest (e.g., tics, myoclonus, choreoathetosis), during maintained posture (pill-rolling tremor of Parkinson's disease), or with voluntary movements (intention tremor of cerebellar disease or familial tremor).

Tone Muscle tone is tested by measuring the resistance to passive movement of a relaxed limb. Patients often have difficulty relaxing during this procedure, so it is useful to distract the patient to minimize active movements. In the upper limbs, tone is assessed by rapid pronation and supination of the forearm and flexion and extension at the wrist. In the lower limbs, while the patient is supine the examiner's hands are placed behind the knees and rapidly raised; with normal tone, the ankles drag along the table surface for a variable distance before rising, whereas increased tone results in an immediate lift of the heel off the surface. Decreased tone is most commonly due to lower motor neuron or peripheral nerve disorders. Increased tone may be evident as spasticity (resistance determined by the angle and velocity of motion; corticospinal tract disease), rigidity (similar resistance in all angles of motion; extrapyramidal disease), or paratonia (fluctuating changes in resistance; frontal lobe pathways or normal difficulty in relaxing). Cogwheel rigidity, in which passive motion elicits jerky interruptions in resistance, is seen in parkinsonism.

Strength Testing for pronator drift is an extremely useful method for screening upper limb weakness. The patient is asked to hold both arms fully extended and parallel to the ground with eyes closed. This position should be maintained for ~10 s; any flexion at the elbow or fingers or pronation of the forearm, especially if asymmetric, is a sign of potential weakness. Muscle strength is further assessed by having the patient exert maximal effort for the particular muscle or muscle group being tested. It is important to isolate the muscles as much as possible, i.e., hold the limb so that only the muscles of interest are active. It is also helpful to palpate accessible muscles as they contract. Grading muscle strength and evaluating the patient's effort is an art that takes time and practice. Muscle strength is traditionally graded using the following scale:

0 = no movement
1 = flicker or trace of contraction but no associated movement at a joint
2 = movement with gravity eliminated
3 = movement against gravity but not against resistance
4– = movement against a mild degree of resistance
4 = movement against moderate resistance
4+ = movement against strong resistance
5 = full power

However, in many cases, it is more practical to use the following terms:

Paralysis = no movement
Severe weakness = movement with gravity eliminated
Moderate weakness = movement against gravity but not against mild resistance
Mild weakness = movement against moderate resistance
Full strength

Noting the pattern of weakness is as important as assessing the magnitude of weakness. Unilateral or bilateral weakness of the upper limb extensors and lower limb flexors ("pyramidal weakness") suggests a lesion of the pyramidal tract, bilateral proximal weakness suggests myopathy, and bilateral distal weakness suggests peripheral neuropathy.

REFLEX EXAMINATION

- *The bare minimum: Check the biceps, patellar, and Achilles reflexes.*

Muscle Stretch Reflexes Those that are typically assessed include the biceps (**C5**, C6), brachioradialis (C5, **C6**), triceps (C6, **C7**), and sometimes finger flexor (**C8**, T1) reflexes in the upper limbs and the patellar or quadriceps (**L3**, L4) and Achilles (**S1**, S2) reflexes in the lower limbs. The patient should be relaxed and the muscle positioned midway between full contraction and extension. Reflexes may be enhanced by asking the patient to voluntarily contract other, distant muscle groups (Jendrassik maneuver). For example, upper limb reflexes may be reinforced by voluntary teeth-clenching, and the Achilles reflex by hooking the flexed fingers of the two hands together and attempting to pull them apart. For each reflex tested, the two sides should be tested sequentially, and it is important to determine the smallest stimulus required to elicit a reflex rather than the maximum response. Reflexes are graded according to the following scale:

0 = absent
1 = present but diminished
2 = normoactive
3 = exaggerated
4 = clonus

Cutaneous Reflexes The plantar reflex is elicited by stroking, with a noxious stimulus such as a tongue blade, the lateral surface of the sole of the foot beginning near the heel and moving across the ball of the foot to the great toe. The normal reflex consists of plantar flexion of the toes. With upper motor neuron lesions above the S1 level of the spinal cord, a paradoxical extension of the toe is observed, associated with fanning and extension of the other toes (termed an *extensor plantar response*, or *Babinski sign*). However, despite its popularity, the reliability and validity of the Babinski sign for identifying upper motor neuron weakness is limited—it is far more useful to rely on tests of tone, strength, stretch reflexes, and coordination. Superficial abdominal reflexes are elicited by gently stroking the abdominal surface near the umbilicus in a diagonal fashion with a sharp object (e.g., the wooden end of a cotton-tipped swab) and observing the movement of the umbilicus. Normally, the umbilicus will pull toward the stimulated quadrant. With upper motor neuron lesions, these reflexes are absent. They are most helpful when there is preservation of the upper (spinal cord level T9) but not lower (T12) abdominal reflexes, indicating a spinal lesion between T9 and T12, or when the response is asymmetric. Other useful cutaneous reflexes include the cremasteric (ipsilateral elevation of the testicle following stroking of the medial thigh; mediated by L1 and L2) and anal (contraction of the anal sphincter when the perianal skin is scratched; mediated by S2, S3, S4) reflexes. It is particularly important to test for these reflexes in any patient with suspected injury to the spinal cord or lumbosacral roots.

Primitive Reflexes With disease of the frontal lobe pathways, several primitive reflexes not normally present in the adult may appear. The suck response is elicited by lightly touching with a tongue blade the center of the lips, and the root response the corner of the lips; the patient will move the lips to suck or root in the direction of the stimulus. The grasp reflex is elicited by touching the palm between the thumb and index finger with the examiner's fingers; a positive response is a forced grasp of the examiner's hand. In many instances, stroking the back of the hand will lead to its release. The palmomental response is contraction of the mentalis muscle (chin) ipsilateral to a scratch stimulus diagonally applied to the palm.

SENSORY EXAMINATION

- *The bare minimum: Ask whether the patient can feel light touch and the temperature of a cool object in each distal extremity. Check double simultaneous stimulation using light touch on the hands. Perform the Romberg maneuver.*

Evaluating sensation is usually the most unreliable part of the examination because it is subjective and is difficult to quantify. In the compliant and discerning patient, the sensory examination can be extremely helpful for the precise localization of a lesion. With patients who are uncooperative or lack an understanding of the tests, it may be useless. The examination should be focused on the suspected lesion. For example, in spinal cord, spinal root, or peripheral nerve abnormalities, all major sensory modalities should be tested while looking for a pattern consistent with a spinal level and dermatomal or nerve distribution. In patients with lesions at or above the brainstem, screening the primary sensory modalities in the distal extremities along with tests of "cortical" sensation is usually sufficient.

The five primary sensory modalities—light touch, pain, temperature, vibration, and joint position—are tested in each limb. Light touch is assessed by stimulating the skin with single, very gentle touches of the examiner's finger or a wisp of cotton. Pain is tested using a new pin, and temperature is assessed using a metal object (e.g., tuning fork) that has been immersed in cold and warm water. Vibration is tested using a 128-Hz tuning fork applied to the distal phalanx of the great toe or index finger just below the nail bed. By placing a finger on the opposite side of the joint being tested, the examiner compares the patient's threshold of vibration perception with his or her own. For joint position testing, the examiner grasps the digit or limb laterally and distal to the joint being assessed; small 1- to 2-mm excursions can usually be sensed. The Romberg maneuver is primarily a test of proprioception. The patient is asked to stand with the feet as close together as necessary to maintain balance while the eyes are open, and the eyes are then closed. A loss of balance with the eyes closed is an abnormal response.

"Cortical" sensation is mediated by the parietal lobes and represents an integration of the primary sensory modalities; testing cortical sensation is only meaningful when primary sensation is intact. Double simultaneous stimulation is especially useful as a screening test for cortical function; with the patient's eyes closed, the examiner lightly touches one or both hands and asks the patient to identify the stimuli. With a parietal lobe lesion, the patient may be unable to identify the stimulus on the contralateral side when both hands are touched. Other modalities relying on the parietal cortex include the discrimination of two closely placed stimuli as separate (two-point discrimination), identification of an object by touch and manipulation alone (stereognosis), and the identification of numbers or letters written on the skin surface (graphesthesia).

COORDINATION EXAMINATION

- *The bare minimum: Observe the patient at rest and during spontaneous movements. Test rapid alternating movements of the hands and feet and finger to nose.*

Coordination refers to the orchestration and fluidity of movements. Even simple acts require cooperation of agonist and antagonist muscles, maintenance of posture, and complex servomechanisms to control the rate and range of movements. Part of this integration relies on normal function of the cerebellar and basal ganglia systems. However, coordination also requires intact muscle strength and kinesthetic and proprioceptive information. Thus, if the examination has disclosed abnormalities of the motor or sensory systems, the patient's coordination should be assessed with these limitations in mind.

Rapid alternating movements in the upper limbs are tested separately on each side by having the patient make a fist, partially extend the index finger, and then tap the index finger on the distal thumb as quickly as possible. In the lower limb, the patient rapidly taps the foot against the floor or the examiner's hand. Finger-to-nose testing is primarily a test of cerebellar function; the patient is asked to touch his or her index finger repetitively to the nose and then to the examiner's outstretched finger, which moves with each repetition. A similar test in the lower extremity is to have the patient raise the leg and touch the examiner's finger with the great toe. Another cerebellar test in the lower limbs is the heel-knee-shin maneuver; in the supine position the patient is asked to slide the heel of each foot from the knee down the shin of the other leg. For all these movements, the accuracy, speed, and rhythm are noted.

GAIT EXAMINATION

- *The bare minimum: Observe the patient while walking normally, on the heels and toes, and along a straight line.*

Watching the patient walk is the most important part of the neurologic examination. Normal gait requires that multiple systems—including strength, sensation, and coordination—function in a highly integrated fashion. Unexpected abnormalities may be detected that prompt the examiner to return in more detail to other aspects of the examination. The patient should be observed while walking and turning normally, walking on the heels, walking on the toes, and walking heel-to-toe along a straight line. The examination may reveal decreased arm swing on one side (corticospinal tract disease), a stooped posture and short-stepped gait (parkinsonism), a broad-based unstable gait (ataxia), scissoring (spasticity), or a high-stepped, slapping gait (posterior column or peripheral nerve disease), or the patient may appear to be stuck in place (apraxia with frontal lobe disease).

NEUROLOGIC DIAGNOSIS

The clinical data obtained from the history and examination are interpreted to arrive at an anatomic localization that best explains the clinical findings (Table 415-2), to narrow the list of diagnostic possibilities, and to select the laboratory tests most likely to be informative. The laboratory assessment may include (1) serum electrolytes; complete blood count; and renal, hepatic, endocrine, and immune studies; (2) cerebrospinal

TABLE 415-2 Findings Helpful for Localizations within the Nervous System

	SIGNS
Cerebrum	Abnormal mental status or cognitive impairment
	Seizures
	Unilateral weakness[a] and sensory abnormalities including head and limbs
	Visual field abnormalities
	Movement abnormalities (e.g., diffuse incoordination, tremor, chorea)
Brainstem	Isolated cranial nerve abnormalities (single or multiple)
	"Crossed" weakness[a] and sensory abnormalities of head and limbs, e.g., weakness of right face and left arm and leg
Spinal cord	Back pain or tenderness
	Weakness[a] and sensory abnormalities sparing the head
	Mixed upper and lower motor neuron findings
	Sensory level
	Sphincter dysfunction
Spinal roots	Radiating limb pain
	Weakness[b] or sensory abnormalities following root distribution **(see Figs. 22-2 and 22-3)**
	Loss of reflexes
Peripheral nerve	Mid or distal limb pain
	Weakness[b] or sensory abnormalities following nerve distribution **(see Figs. 22-2 and 22-3)**
	"Stocking or glove" distribution of sensory loss
	Loss of reflexes
Neuromuscular junction	Bilateral weakness including face (ptosis, diplopia, dysphagia) and proximal limbs
	Increasing weakness with exertion
	Sparing of sensation
Muscle	Bilateral proximal or distal weakness
	Sparing of sensation

[a]Weakness along with other abnormalities having an "upper motor neuron" pattern, i.e., spasticity, weakness of extensors > flexors in the upper extremity and flexors > extensors in the lower extremity, and hyperreflexia. [b]Weakness along with other abnormalities having a "lower motor neuron" pattern, i.e., flaccidity and hyporeflexia.

fluid examination; (3) focused neuroimaging studies **(Chap. 416)**; or (4) electrophysiologic studies. The anatomic localization, mode of onset and course of illness, other medical data, and laboratory findings are then integrated to establish an etiologic diagnosis.

The neurologic examination may be normal even in patients with a serious neurologic disease, such as seizures, chronic meningitis, or a TIA. A comatose patient may arrive with no available history, and in such cases, the approach is as described in **Chap. 300**. In other patients, an inadequate history may be overcome by a succession of examinations from which the course of the illness can be inferred. In perplexing cases it is useful to remember that uncommon presentations of common diseases are more likely than rare etiologies. Thus, even in tertiary care settings, multiple strokes are usually due to emboli and not vasculitis, and dementia with myoclonus is usually Alzheimer's disease and not a prion disorder or a paraneoplastic illness. Finally, the most important task of a primary care physician faced with a patient who has a new neurologic complaint is to assess the urgency of referral to a specialist. Here, the imperative is to rapidly identify patients likely to have nervous system infections, acute strokes, and spinal cord compression or other treatable mass lesions and arrange for immediate care.

FURTHER READING

CAMPBELL WW: *DeJong's Neurological Examination*, 7th ed. Philadelphia, Wolters Kluwer/Lippincott Williams & Wilkins, 2013.

O'BRIEN M: *Aids to the Examination of the Peripheral Nervous System*, 5th ed. Edinburgh, WB Saunders, 2010.

ROPPER AH et al: *Adams and Victor's Principles of Neurology*, 10th ed. New York, McGraw-Hill, 2014.

416 Neuroimaging in Neurologic Disorders

William P. Dillon

Numerous noninvasive imaging options are available to clinicians evaluating patients with neurologic disorders. These include computed tomography (CT) and variations CT angiography (CTA), perfusion CT (pCT), and dual energy CT and magnetic resonance (MR) imaging (MRI) and variations MR angiography (MRA), MR vessel wall imaging, functional MRI (fMRI), MR spectroscopy (MRS), MR neurography (MRN), diffusion and diffusion tensor MR imaging, susceptibility-weighted MR imaging (SWI), arterial spin label imaging (ASL) and perfusion MRI (pMRI). Furthermore, a number of interventional neuroradiologic techniques have matured including catheter embolization, stent retrieval thrombolysis, aneurysm coiling and stenting, as well as numerous techniques for spine disorders, including CT myelography, fluoroscopy and CT-guided transforaminal and translaminar epidural and nerve root injections, radiofrequency ablation and blood patches. Multidetector CTA (MDCTA) and gadolinium-enhanced MRA techniques have reduced the need for catheter-based angiography, which is now reserved for patients in whom small-vessel detail is essential for diagnosis or for whom concurrent interventional therapy is planned **(Table 416-1)**.

In general, MRI is more sensitive than CT for the detection of lesions affecting the peripheral and central nervous system (CNS), particularly those of the spinal cord, cranial nerves, and posterior fossa structures. Diffusion MR, a sequence sensitive to the microscopic motion of water, is the most sensitive technique for detecting acute ischemic stroke of the brain or spinal cord, and it is also useful in the detection and characterization of encephalitis, abscess, Creutzfeldt-Jacob disease, cerebral tumors and acute demyelinating lesions. CT, however, is acquired quickly, making it a pragmatic choice for patients with acute changes in mental status, suspected hemorrhage, and acute intracranial or spinal trauma. CT is also more sensitive than MRI for visualizing fine osseous detail and is indicated in the initial imaging evaluation of conductive hearing loss as well as lesions affecting the skull base and calvarium. MR may, however, add important diagnostic information regarding bone marrow infiltrative processes that are difficult to detect on CT.

COMPUTED TOMOGRAPHY

TECHNIQUE

The CT image is a cross-sectional representation of anatomy created by a computer-generated analysis of the attenuation of x-ray beams passed through a section of the body. As the x-ray beam, collimated to the desired slice width, rotates around the patient, it passes through selected regions in the body. X-rays that are not attenuated by body structures are detected by sensitive x-ray detectors aligned 180° from the x-ray tube. A computer calculates a "back projection" image from the 360° x-ray attenuation profile. Greater x-ray attenuation (e.g., as caused by bone), results in areas of high "density" (whiter) on the scan, whereas soft tissue structures that have poor attenuation of x-rays, such as organs and air-filled cavities, are lower (blacker) in density. The resolution of an image depends on the radiation dose, the detector size, collimation (slice thickness), the field of view, and the matrix size of the display. A modern CT scanner is capable of obtaining sections as thin as 0.5–1 mm with 0.4-mm in-plane resolution at a speed of 0.3 s per rotation; complete studies of the brain can be completed in 1–10 s.

Multidetector CT (MDCT) is now standard. Single or multiple (from 4 to 320) solid-state detectors positioned opposite to the x-ray source result in multiple slices per revolution of the beam around the patient. In helical mode, the table moves continuously through the rotating x-ray beam, generating a continuous "helix" of information that can be reformatted into various slice thicknesses and planes. Advantages of MDCT include shorter scan times and thus reduced patient and

TABLE 416-1 Guidelines for the Use of CT, Ultrasound, and MRI

CONDITION	RECOMMENDED TECHNIQUE
Hemorrhage	
Acute parenchymal	CT, MR
Subacute/chronic	MRI
Subarachnoid hemorrhage	CT, CTA, lumbar puncture → angiography
Aneurysm	Angiography > CTA, MRA
Ischemic infarction	
Hemorrhagic infarction	CT or MRI
Bland infarction	MRI with diffusion > CT, CTA, angiography
Carotid or vertebral dissection	MRI/MRA
Vertebral basilar insufficiency	CTA, MRI/MRA
Carotid stenosis	CTA, MRA > US
Suspected mass lesion	
Neoplasm, primary or metastatic	MRI + contrast
Infection/abscess	MRI + contrast
Immunosuppressed with focal findings	MRI + contrast
Vascular malformation	MRI ± angiography
White matter disorders	MRI
Demyelinating disease	MRI ± contrast
Dementia	MRI > CT
Trauma	
Acute trauma	CT
Shear injury/chronic hemorrhage	MRI + susceptibility-weighted imaging
Headache/migraine	CT/MRI
Seizure	
First time, no focal neurologic deficits	MRI > CT
Partial complex/refractory	MRI
Cranial neuropathy	MRI with contrast
Meningeal disease	MRI with contrast
Spine	
Low back pain	
No neurologic deficits	MRI or CT after >6 weeks
With focal deficits	MRI > CT
Spinal stenosis	MRI or CT
Cervical spondylosis	MRI, CT, CT myelography
Infection	MRI + contrast, CT
Myelopathy	MRI + contrast
Arteriovenous malformation	MRI + contrast, angiography

Abbreviations: CT, computed tomography; CTA, CT angiography; MRA, magnetic resonance angiography; MRI, magnetic resonance imaging.

organ motion, and the ability to acquire images dynamically during the infusion of intravenous contrast, the basis of CTA and CT perfusion (Figs. 416-1*B* and *C*). CTA is displayed in three dimensions to yield angiogram-like images (Figs. 416-1C, 416-2*E* and *F*, and see Fig. 420-3).

Intravenous iodinated contrast is used to identify vascular structures and to detect defects in the blood-brain barrier (BBB) that are caused by tumors, infarcts, and infections. In the normal CNS, only vessels and structures lacking a BBB (e.g., the pituitary gland, choroid plexus, and dura) enhance after contrast administration. While helpful in characterizing mass lesions as well as essential for the acquisition of CTA studies, the decision to use contrast material should always be considered carefully as it carries a small risk of allergic reaction and adds additional expense.

INDICATIONS

CT is the primary study of choice in the evaluation of an acute change in mental status, focal neurologic findings, acute trauma to the brain and spine, suspected subarachnoid hemorrhage, and conductive hearing loss (Table 416-1). CT often is complementary to MR in the evaluation of the skull base, orbit, and osseous structures of the spine. In the spine, CT is useful in evaluating patients with osseous spinal stenosis and spondylosis, but MRI is often preferred in those with neurologic deficits. CT is often acquired following intrathecal contrast injection to evaluate for spinal and intracranial cerebrospinal fluid (CSF) fistula, as well as the spinal subarachnoid space (*CT myelography*) in failed back surgery syndromes.

COMPLICATIONS

CT is safe, fast, and reliable. Radiation exposure depends on the dose used but is normally between 2 and 5 mSv (millisievert) for a routine brain CT study. Care must be taken to reduce exposure when imaging children, who are typically best studied with MR. With the advent of MDCT, CTA, and CT perfusion, the benefit must be weighed against the increased radiation doses associated with these techniques. Advances in postprocessing software now permit acceptable diagnostic CT scans at 30–40% lower radiation doses.

The most frequent complications are those associated with use of intravenous contrast agents. While two broad categories of contrast media, ionic and nonionic, are in use, ionic agents have been largely replaced by safer nonionic compounds.

Contrast nephropathy is rare. It may result from hemodynamic changes, renal tubular obstruction and cell damage, or immunologic reactions to contrast agents. A rise in serum creatinine of at least 44 μmol/L (0.5 mg/dL) within 48 h of contrast administration is often used as a definition of contrast nephropathy, although there is no accepted definition and other causes of acute renal failure must be excluded. The prognosis is usually favorable, with serum creatinine levels returning to baseline

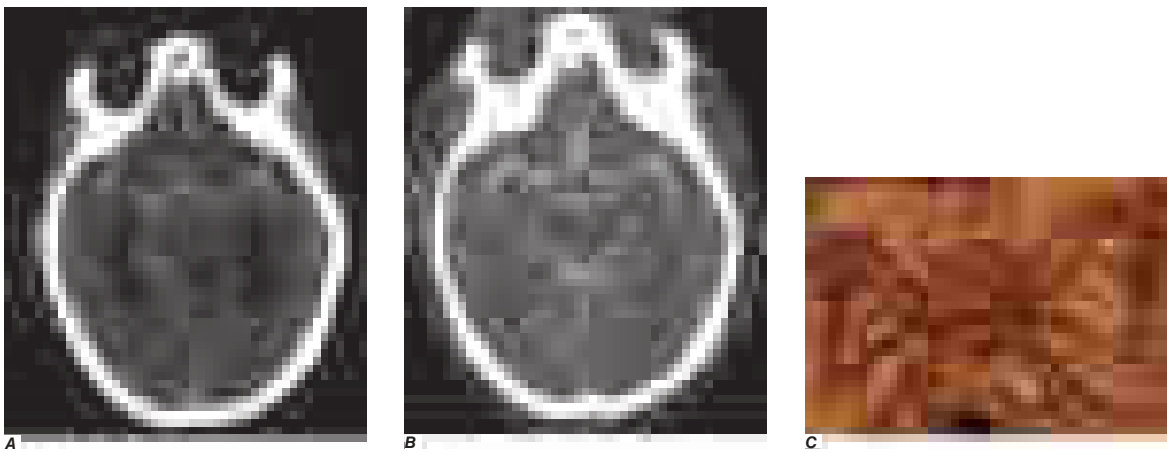


FIGURE 416-1 Computed tomography (CT) angiography (CTA) of ruptured anterior cerebral artery aneurysm in a patient presenting with acute headache. ***A.*** Noncontrast CT demonstrates subarachnoid hemorrhage and mild obstructive hydrocephalus. ***B.*** Axial maximum-intensity projection from CTA demonstrates enlargement of the anterior cerebral artery (*arrow*). ***C.*** Three-dimensional surface reconstruction using a workstation confirms the anterior cerebral aneurysm and demonstrates its orientation and relationship to nearby vessels (*arrow*). CTA image is produced by 0.5- to 1-mm helical CT scans performed during a rapid bolus infusion of intravenous contrast medium.

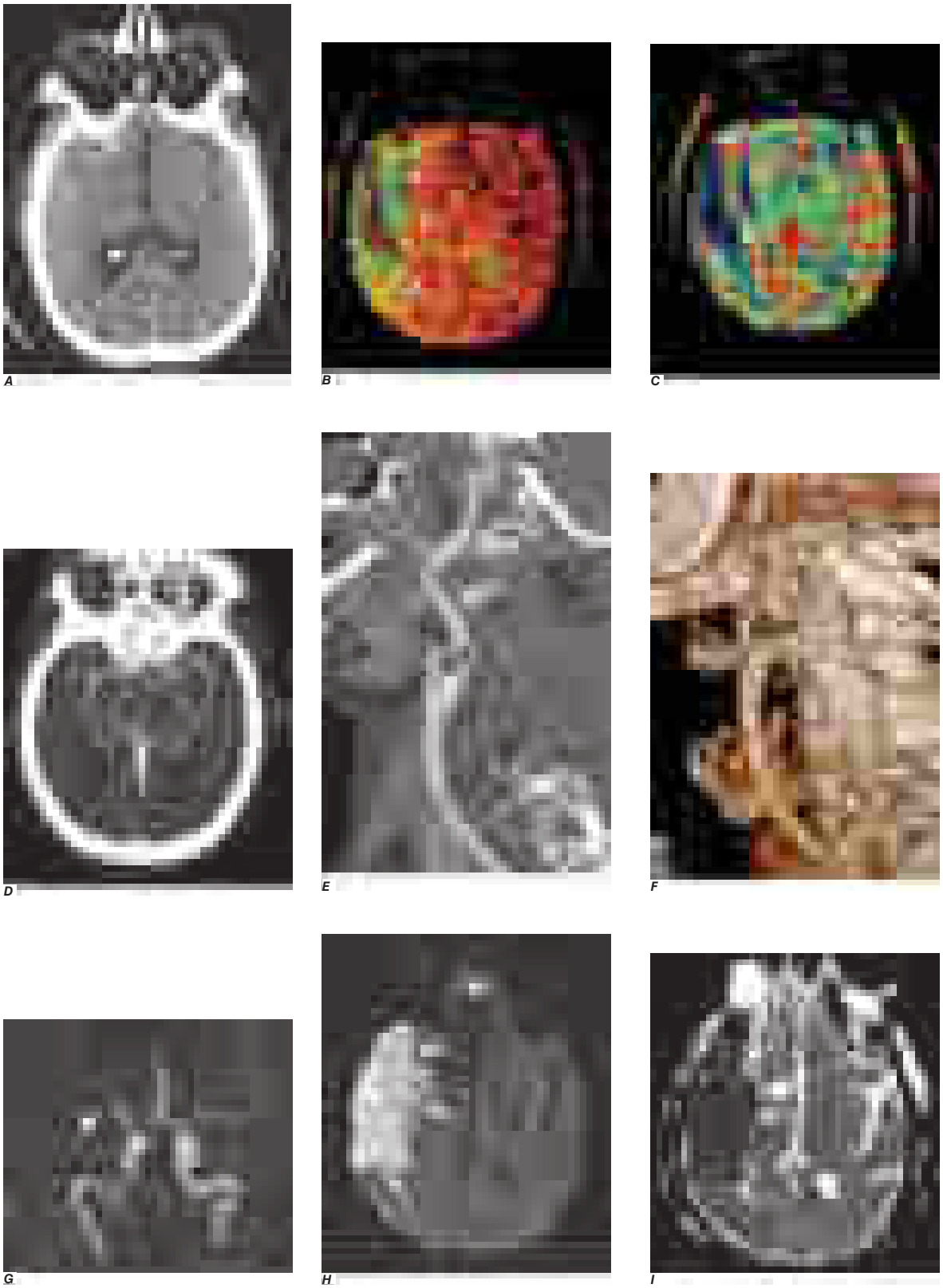


FIGURE 416-2 Acute left hemiparesis due to middle cerebral artery occlusion. ***A.*** Axial noncontrast computed tomography (CT) scan demonstrates high density within the right middle cerebral artery (*arrow*) associated with subtle low density involving the right putamen (*arrowheads*). ***B.*** Mean transit time CT perfusion parametric map indicating prolonged mean transit time involving the right middle cerebral territory (*arrows*). ***C.*** Cerebral blood volume (CBV) map shows reduced CBV involving an area within the defect shown in ***B***, indicating a high likelihood of infarction (*arrows*). ***D.*** Axial maximum-intensity projection from a CT angiography (CTA) study through the circle of Willis demonstrates an abrupt occlusion of the proximal right middle cerebral artery (*arrow*). ***E.*** Sagittal reformation through the right internal carotid artery demonstrates a low-density lipid-laden plaque (*arrowheads*) narrowing the lumen (*black arrow*). ***F.*** Three-dimensional surface-rendered CTA image demonstrates calcification and narrowing of the right internal carotid artery (*arrow*), consistent with atherosclerotic disease. ***G.*** Coronal maximum-intensity projection from magnetic resonance angiography shows right middle cerebral artery (MCA) occlusion (*arrow*). ***H.*** and ***I.*** Axial diffusion-weighted image (***H***) and apparent diffusion coefficient image (***I***) documents the presence of a right middle cerebral artery infarction.

within 1–2 weeks. Risk factors for contrast nephropathy include age (>80 years), preexisting renal disease (serum creatinine exceeding 2 mg/dL), solitary kidney, diabetes mellitus, dehydration, paraproteinemia, concurrent use of nephrotoxic medication or chemotherapeutic agents, and high contrast dose. Patients with diabetes and those with mild renal failure should be well hydrated prior to the administration of contrast agents; careful consideration should be given to alternative imaging techniques such as MRI, noncontrast CT, or ultrasound (US). Nonionic, low-osmolar media produce fewer abnormalities in renal blood flow and less endothelial cell damage but should still be used carefully in patients at risk for allergic reaction. Estimated glomerular filtration rate (eGFR) is a more reliable indicator of renal function compared to creatinine alone because it takes into account age, race, and sex. In one study, 15% of outpatients with a normal serum creatinine had an estimated creatinine clearance of ≤50 mL/min/1.73 m^2 (normal is ≥90 mL/min/1.73 m^2). The exact eGFR threshold, below which withholding intravenous contrast should be considered, is controversial. The risk of contrast nephropathy is minimal in patients with eGFR >30 mL/min/1.73 m^2; however, the majority of these patients will only have a temporary rise in creatinine. The risk of dialysis after receiving contrast significantly increases in patients with eGFR <30 mL/min/1.73 m^2. A creatinine of 1.6 in a 70-year-old, non-African-American male corresponds to an eGFR of ~45 mL/min/1.73 m^2. The American College of Radiology suggests using an eGFR of 30 mL/min/1.73 m^2 as a threshold below which iodinated contrast should not be given without serious consideration of the potential for contrast nephropathy. If contrast must be administered to a patient with an eGFR <30 mL/min/1.73 m^2, the patient should be well hydrated, and a reduction in the dose of contrast should be considered. Use of other agents such as bicarbonate and acetylcysteine may reduce the incidence of contrast nephropathy.

Suggested guidelines for creatinine testing prior to contrast administration:

If serum creatinine is not available, it should be performed IF the patient has ANY of the following risk factors:

- Age >60
- History of "kidney disease" as an adult, including tumor and transplant
- Family history of kidney failure
- Diabetes mellitus treated with insulin or other prescribed medications
- Hypertension
- Paraproteinemia syndromes or diseases (e.g., myeloma)
- Collagen vascular disease (e.g., SLE, scleroderma, rheumatoid arthritis)
- Solid organ transplant recipient

If creatinine testing is required, a creatinine level within the prior 6 weeks is sufficient in most clinical settings.

Allergy Immediate reactions following intravenous contrast media occur through several mechanisms. The most severe reactions are related to allergic hypersensitivity (anaphylaxis) and range from mild hives to bronchospasm and death. The pathogenesis of allergic hypersensitivity reactions is thought to include the release of mediators such as histamine, antibody-antigen reactions, and complement activation. Severe allergic reactions occur in ~0.04% of patients receiving nonionic media, sixfold lower than with ionic media. Risk factors include a history of prior contrast reaction (fivefold increased likelihood), food and or drug allergies, and atopy (asthma and hay fever). The predictive value of specific allergies, such as those to shellfish, once thought important, actually is now recognized to be unreliable. Nonetheless, in patients with a history worrisome for potential allergic reaction, a noncontrast CT or MRI procedure should be considered as an alternative to contrast administration. If iodinated contrast is absolutely required, a nonionic agent should be used in conjunction with pretreatment with glucocorticoids and antihistamines **(Table 416-2)**; however, pretreatment does not guarantee safety. Patients with allergic reactions to iodinated contrast material do not usually react to gadolinium-based MR contrast material, although such reactions can occur. It would be wise to pretreat patients with a prior allergic history to MR contrast administration in a similar fashion. Subacute (>1 h after injection) reactions are frequent and probably related to T cell–mediated immune reactions. These are typically urticarial but can occasionally be more severe. Drug provocation and skin testing may be required to determine the culprit agent involved as well as determine a safe alternative.

TABLE 416-2 Guidelines for Premedication of Patients with Prior Contrast Allergy

12 h prior to examination:
Prednisone, 50 mg PO *or* methylprednisolone, 32 mg PO
2 h prior to examination:
Prednisone, 50 mg PO *or* methylprednisolone, 32 mg PO *and* cimetidine, 300 mg PO *or* ranitidine, 150 mg PO
Immediately prior to examination:
Benadryl, 50 mg IV (alternatively, can be given PO 2 h prior to exam)

Other side effects of CT scanning are rare but include a sensation of warmth throughout the body and a metallic taste during intravenous administration of iodinated contrast media. Extravasation of contrast media, although rare, can be painful and lead to compartment syndrome. When this occurs, consultation with plastic surgery is indicated. Patients with significant cardiac disease may be at increased risk for contrast reactions, and in these patients, limits to the volume and osmolality of the contrast media should be considered. Patients who may undergo systemic radioactive iodine therapy for thyroid disease or cancer should not receive iodinated contrast media if possible, because this will decrease the uptake of the radioisotope into the tumor or thyroid (see the *American College of Radiology Manual on Contrast Media*, Version 10.3, 2017; *https://www.acr.org/-/media/ACR/Files/Clinical-Resources/ContrastMedia.pdf*).

MAGNETIC RESONANCE IMAGING

TECHNIQUE

MRI is a complex interaction between hydrogen protons in biologic tissues, a static magnetic field (the magnet), and energy in the form of radiofrequency (Rf) waves of a specific frequency introduced by coils placed next to the body part of interest. Images are made by computerized processing of resonance information received from protons in the body. Field strength of the magnet is directly related to signal-to-noise ratio. While 1.5 Tesla (T) and 3-T magnets are now widely available and have distinct advantages in the brain and musculoskeletal systems, even igher field magnets (7-T) and positron emission tomography (PET) MR machines promise increased resolution and anatomic-functional information on a variety of disorders. Spatial localization is achieved by magnetic gradients surrounding the main magnet, which impart slight changes in magnetic field throughout the imaging volume. Rf pulses transiently excite the energy state of the hydrogen protons in the body. Rf is administered at a frequency specific for the field strength of the magnet. The subsequent return to equilibrium energy state (*relaxation*) of the hydrogen protons results in a release of Rf energy (the *echo*), which is detected by the coils that delivered the Rf pulses. Fourier analysis is used to transform the echo into the information used to form an MR image. The MR image thus consists of a map of the distribution of hydrogen protons, with signal intensity imparted by both density of hydrogen protons and differences in the relaxation times (see below) of hydrogen protons on different molecules. Although clinical MRI currently makes use of the ubiquitous hydrogen proton, research into sodium and carbon imaging and spectroscopy appears promising.

T1 and T2 Relaxation Times The rate of return to equilibrium of perturbed protons is called the *relaxation rate*. The relaxation rate varies among normal and pathologic tissues. The relaxation rate of a hydrogen proton in a tissue is influenced by local interactions with surrounding molecules and atomic neighbors. Two relaxation rates, T1 and T2, influence the signal intensity of the image. The T1 relaxation time is the time, measured in milliseconds, for 63% of the hydrogen protons to return to their normal equilibrium state, whereas the T2

TABLE 416-3 Some Common Intensities on T1- and T2-Weighted MRI Sequences

IMAGE	TR	TE	SIGNAL INTENSITY			
			CSF	FAT	BRAIN	EDEMA
T1W	Short	Short	Low	High	Low	Low
T2W	Long	Long	High	Low	High	High
FLAIR (T2)	Long	Long	Low	Medium	High	High

Abbreviations: CSF, cerebrospinal fluid; FLAIR, fluid-attenuated inversion recovery; TE, interval between radiofrequency pulse and signal reception; TR, interval between radiofrequency pulses; T1W and T2W, T1- and T2-weighted.

relaxation is the time for 63% of the protons to become dephased owing to interactions among nearby protons. The intensity and image contrast of the signal within various tissues can be modulated by altering acquisition parameters such as the interval between Rf pulses (TR) and the time between the Rf pulse and the signal reception (TE). T1-weighted (T1W) images are produced by keeping the TR and TE relatively short, whereas using longer TR and TE times produces T2-weighted (T2W) images. Fat and subacute hemorrhage have relatively shorter T1 relaxation rates and thus higher signal intensity than brain on T1W images. Structures containing more water, such as CSF and edema, have long T1 and T2 relaxation rates, resulting in relatively lower signal intensity on T1W images and higher signal intensity on T2W images **(Table 416-3)**. Gray matter contains 10–15% more water than white matter, which accounts for much of the intrinsic contrast between the two on MRI **(Fig. 416-4*A*)**. T2W images are more sensitive than T1W images to edema, demyelination, infarction, and chronic hemorrhage, whereas T1W imaging is more sensitive to subacute hemorrhage and fat-containing structures.

Many different MR pulse sequences exist, and each can be obtained in various planes **(Figs. 416-2, 416-3, and 416-4)**. The selection of a proper protocol that will best answer a clinical question depends on an accurate clinical history and indication for the examination. Fluid-attenuated inversion recovery (FLAIR) is a very useful pulse sequence that produces T2W images in which the normally high signal intensity of CSF is suppressed **(Fig. 416-4*B*)**. FLAIR images are more sensitive than standard spin echo images for water-containing lesions or edema, especially those close to CSF filled cisterns and sulci. Diffusion weighted imaging is also routinely obtained in most brain protocols. This sequence interrogates the microscopic motion of water, which is restricted in areas of infarction, abscess, and some tumors. SWI is very sensitive to alterations in local magnetic field generated by blood, calcium, and air. SWI is routinely obtained and helps detects microhemorrhages, such as is typical of amyloid, hemorrhagic metastases, traumatic brain injury, and thrombotic states **(Fig. 416-5*C*)**. MR images can be generated in any plane without changing the patient's position. Each sequence, however, must be obtained separately and takes 1–10 min on average to complete. Three-dimensional volumetric imaging is also possible with MRI, resulting in a three-dimensional volume of data that can be reformatted in any orientation to highlight certain disease processes.

MR Contrast Material The heavy-metal element gadolinium forms the basis of all currently approved intravenous MR contrast agents. Gadolinium is a paramagnetic substance that reduces the T1 and T2 relaxation times of nearby water protons, resulting in a high signal on T1W images and a low signal on T2W images (the latter requires a sufficient local concentration, usually in the form of an intravenous bolus). Unlike iodinated contrast agents, the effect of MR contrast agents depends on the presence of local hydrogen protons on which it must act to achieve the desired effect. There are nine different gadolinium agents approved in the United States for use with MRI. These differ according to the attached chelated moiety, which also affects the strength of chelation of the otherwise toxic gadolinium element. The chelating carrier molecule for gadolinium can be classified by whether it is macrocyclic or has linear geometry and whether it is ionic or nonionic. Macrocyclic ligands (Group 2 agents) are considered more stable as the gadolinium ion is "caged" in the cavity of the ligand, and thus

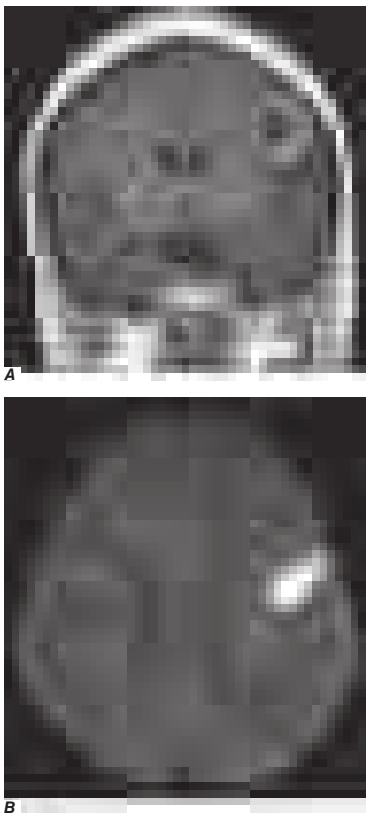


FIGURE 416-3 Cerebral abscess in a patient with fever and a right hemiparesis. ***A.*** Coronal postcontrast T1-weighted image demonstrates a ring-enhancing mass in the left frontal lobe. ***B.*** Axial diffusion-weighted image demonstrates restricted diffusion (high signal intensity) within the lesion, which in this setting is highly suggestive of cerebral abscess.

the rate of dissociation of gadolinium is slower compared to linear ligands (Group 1 agents). Most agents are excreted by the renal system.

BRAIN ACCUMULATION OF GADOLINIUM It recently has become evident that gadolinium accumulates in the dentate nuclei and globus pallidus of the brain after serial administration of some linear Group 1 gadolinium agents. This has not been demonstrated for Group 2 macrocyclic agents. To date there is no clinical effect of this deposition that has been detected.

ALLERGIC HYPERSENSITIVITY Gadolinium-DTPA (diethylenetriaminepentaacetic acid) does not normally cross the intact BBB immediately but will enhance lesions lacking a BBB **(Fig. 416-3*A*)** as well as areas of the brain that normally are devoid of the BBB (pituitary, dura, choroid plexus). However, gadolinium contrast slowly crosses an intact BBB over time and especially in the setting of reduced renal clearance or inflamed meninges. The agents are generally well tolerated; overall adverse events after injection range from 0.07 to 2.4%. True allergic reactions are rare (0.004–0.7%) but have been reported. Severe life-threatening reactions are exceedingly rare; in one report, only 55 reactions out of 20 million doses occurred. However, the adverse reaction rate in patients with a prior history of reaction to gadolinium is eight times higher than normal. Other risk factors include atopy or asthma (3.7%). There is no cross reactivity between different classes of contrast media; a prior reaction to gadolinium-based contrast does not predict a future reaction to iodinated contrast medium, or vice versa, more than any other unrelated allergy. Gadolinium contrast material can be administered safely to children as well as adults, although these agents are generally avoided in those aged <6 months.

NEPHROTOXICITY Contrast-induced renal failure does not occur with gadolinium agents. A rare complication, nephrogenic systemic fibrosis (NSF), has occurred in patients with severe renal insufficiency who have been exposed to linear (Group 1 and 3) gadolinium contrast agents. The onset of NSF has been reported between 5 and 75 days

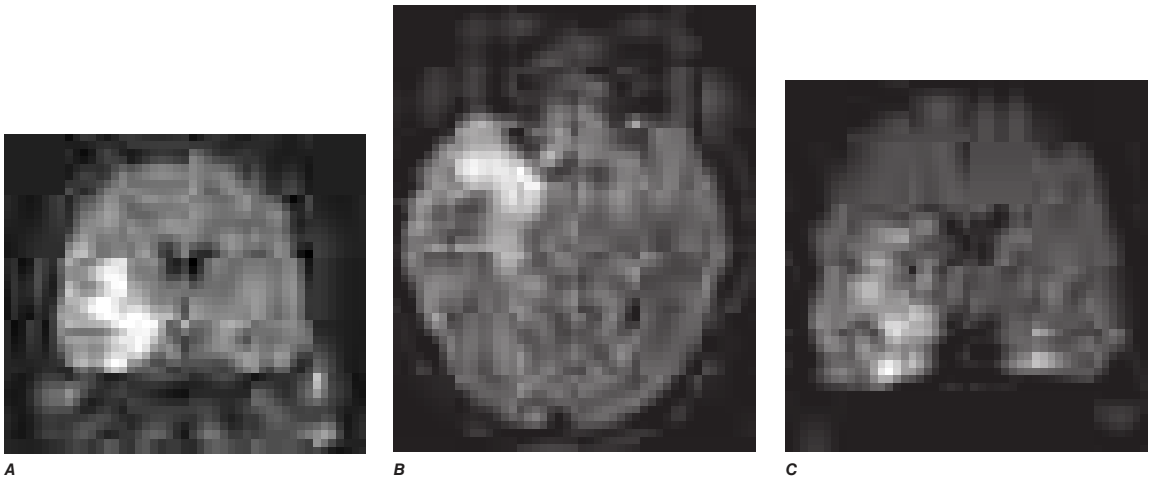


FIGURE 416-4 Herpes simplex encephalitis in a patient presenting with altered mental status and fever. *A.* and ***B.*** Coronal (***A***) and axial (***B***) T2-weighted fluid-attenuated inversion recovery images demonstrate expansion and high signal intensity involving the right medial temporal lobe and insular cortex (*arrows*). ***C.*** Coronal diffusion-weighted image demonstrates high signal intensity indicating restricted diffusion involving the right medial temporal lobe and hippocampus (*arrows*) as well as subtle involvement of the left inferior temporal lobe (*arrowhead*). This is most consistent with neuronal death and can be seen in acute infarction as well as encephalitis and other inflammatory conditions. The suspected diagnosis of herpes simplex encephalitis was confirmed by cerebrospinal fluid polymerase chain reaction analysis.

following exposure; histologic features include thickened collagen bundles with surrounding clefts, mucin deposition, and increased numbers of fibrocytes and elastic fibers in skin. In addition to dermatologic symptoms, other manifestations include widespread fibrosis of the skeletal muscle, bone, lungs, pleura, pericardium, myocardium, kidney, muscle, bone, testes, and dura. The American College of Radiology recommends that a glomerular filtration rate (GFR) assessment be obtained within 6 weeks prior to elective gadolinium-based MR contrast agent administration in patients with:

1. A history of renal disease (including solitary kidney, renal transplant, renal tumor)
2. Age >60 years
3. History of hypertension
4. History of diabetes
5. History of severe hepatic disease, liver transplant, or pending liver transplant; for these patients, it is recommended that the patient's GFR assessment be nearly contemporaneous with the MR examination.

The incidence of NSF in patients with severe renal dysfunction (GFR <30) varies from 0.19 to 4%. Other risk factors for NSF include acute kidney injury, the use of non-macrocyclic agents, and repeated or high-dose exposure to gadolinium. The American College of Radiology Committee on Drugs and Contrast Media considers the risk of NSF among patients exposed to standard or lower doses of Group 2 gadolinium agents (macrocyclic agents) is sufficiently low or possibly non-existent that the assessment of renal function is optimal prior to administration. However, patients receiving any Group 1 (linear) or 3 gadolinium-containing agent should be considered at risk of NSF if they are on dialysis (of any form); have severe or end-stage chronic renal disease (eGFR <30 mL/min/1.73 m^2) without dialysis; eGFR of 30–40 mL/min/1.73 m^2 without dialysis (as the GFR may fluctuate); or have acute renal insufficiency. The use of gadolinium in young children and infants is discouraged due to the unknown risks and their immature renal systems.

COMPLICATIONS AND CONTRAINDICATIONS

From the patient's perspective, an MRI examination can be intimidating, and a higher level of cooperation is required than with CT. The patient lies on a table that is moved into a long, narrow gap within the magnet. Approximately 5% of the population experiences severe claustrophobia in the MR environment. This can be reduced by mild sedation but remains a problem for some. Because it takes between 3 and 10 min per sequence, movement of the patient during an MR examination distorts all of the images; therefore, uncooperative patients should either be sedated for the MR study or scanned with CT. Generally, children aged <8 years usually require conscious sedation in order to complete the MR examination without motion degradation.

MRI is considered safe for patients, even at very high field strengths. Serious injuries have been caused, however, by attraction of

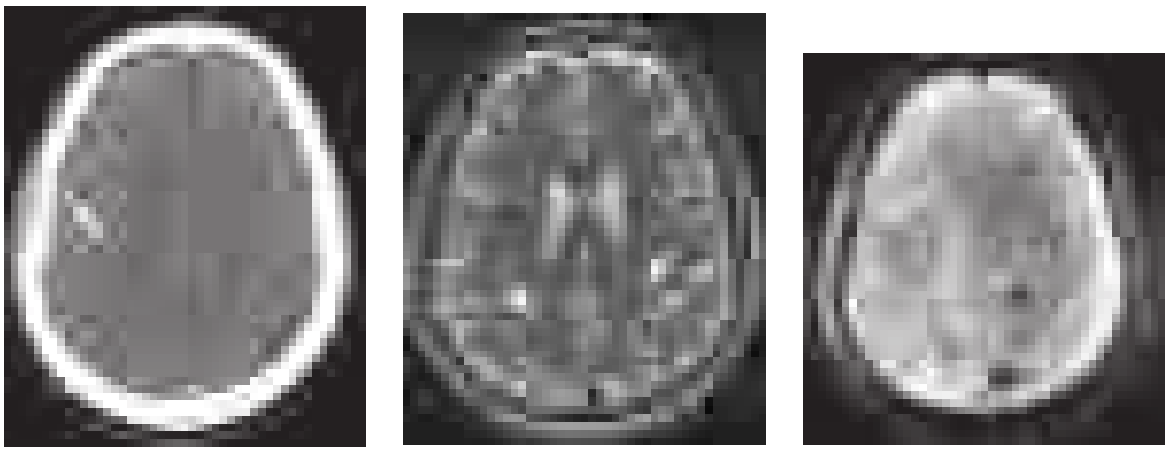


FIGURE 416-5 Susceptibility-weighted imaging in a patient with familial cavernous malformations. *A.* Noncontrast computed tomography scan shows one hyperdense lesion in the right hemisphere (*arrow*). ***B.*** T2-weighted fast spin echo image shows subtle low-intensity lesions (*arrows*). ***C.*** Susceptibility-weighted image shows numerous low-intensity lesions consistent with hemosiderin-laden cavernous malformations (*arrow*).

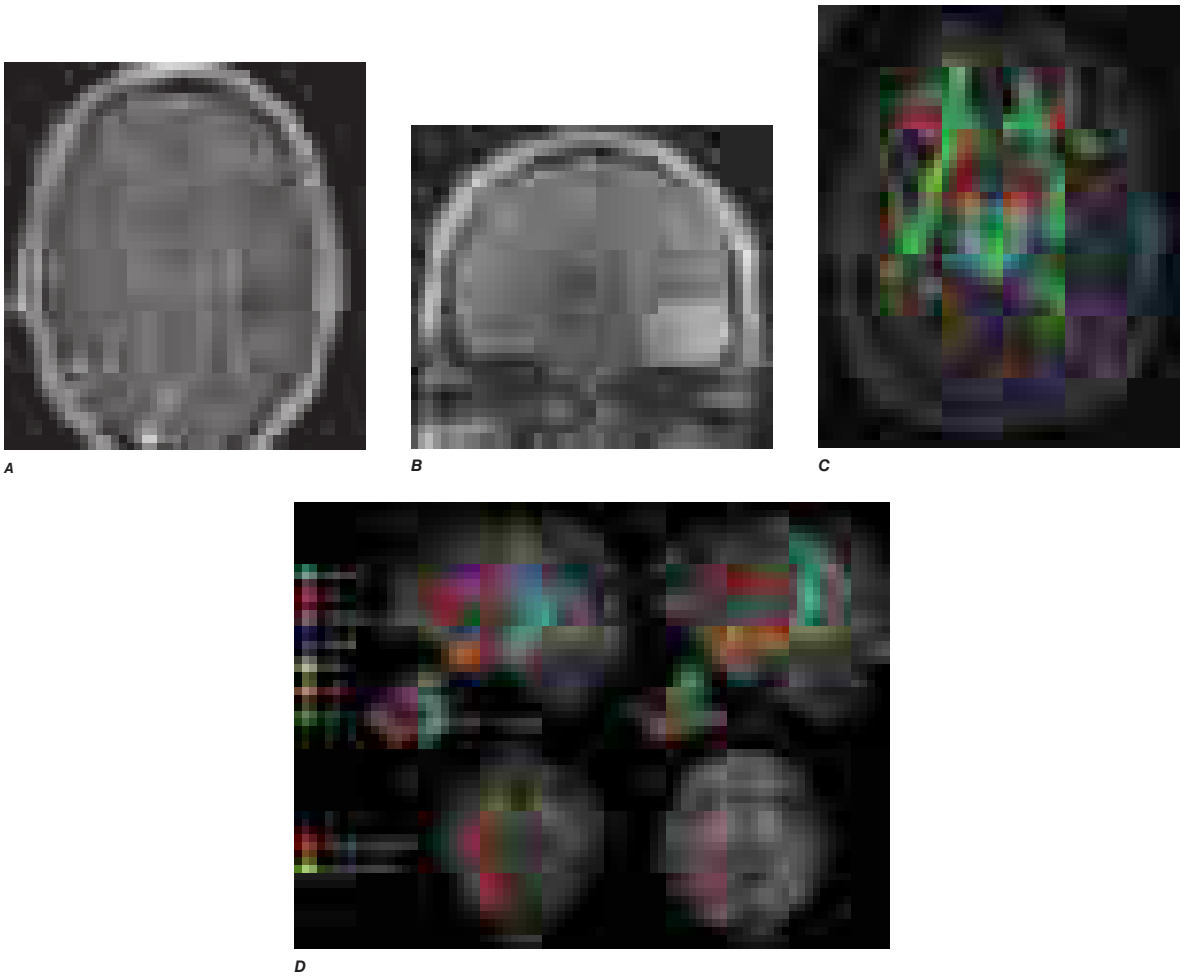


FIGURE 416-6 Diffusion tractography in cerebral glioma. Associative and descending pathways in a healthy subject (***A***) and in a patient with parietal lobe glioblastoma (***B***) presenting with a language deficit: the mass causes a disruption of the arcuate-SLF complex, in particular of its anterior portion (SLF III). Also shown are bilateral optic tract and left optic radiation pathways in a healthy subject (***C***) and in a patient with left occipital grade II oligoastrocytoma (***D***): the mass causes a disruption of the left optic radiation. *Shown in neurologic orientation, i.e., the left brain appears on the left side of the image.* AF, long segment of the arcuate fascicle; CST, corticospinal tract; IFOF: inferior fronto-occipital fascicle; ILF, inferior longitudinal fascicle; SLF III, superior longitudinal fascicle III or anterior segment of the arcuate fascicle; SLF-tp, temporo-parietal portion of the superior longitudinal fascicle or posterior segment of the arcuate fascicle; T, tumor; UF, uncinated fascicle. *(Part D courtesy of Eduardo Caverzasi and Roland Henry.)*

ferromagnetic objects into the magnet, which act as missiles if brought too close to the magnet. Likewise, ferromagnetic implants, such as aneurysm clips, may torque within the magnet, causing damage to vessels and even death. Metallic foreign bodies in the eye have moved and caused intraocular hemorrhage; screening for ocular metallic fragments is indicated in those with a history of metal work or ocular metallic foreign bodies. Implanted cardiac pacemakers are generally a contraindication to MRI owing to the risk of induced arrhythmias; however, some newer pacemakers have been shown to be safe. All health care personnel and patients must be screened and educated thoroughly to prevent such disasters because the magnet is always "on." **Table 416-4** lists common contraindications for MRI.

MAGNETIC RESONANCE ANGIOGRAPHY

MR angiography is a general term describing several MR techniques that result in vascular-weighted images. Non contrast MRA provides a flow map rather than the anatomic map shown by conventional angiography. On routine spin echo MR sequences, moving protons (e.g., flowing blood, CSF) exhibit complex MR signals that range from high- to low-signal intensity relative to background stationary tissue. Fast-flowing blood returns no signal (flow void) on routine T1W or T2W spin echo MR images. Slower-flowing blood, as occurs in veins or distal to arterial stenosis, may appear high in signal. However, using special pulse sequences called *gradient echo sequences*, it is possible to increase the signal intensity of moving protons in contrast to the low signal background intensity of stationary tissue. This creates angiography-like images, which can be manipulated in three dimensions to highlight vascular anatomy and relationships.

So-called time-of-flight (TOF) MRA relies on the suppression of nonmoving tissue to provide a low-intensity background for the high

TABLE 416-4 Common Contraindications to Magnetic Resonance Imaging

Cardiac pacemaker or permanent pacemaker leads
Internal defibrillatory device
Cochlear prostheses
Bone growth stimulators
Spinal cord stimulators
Electronic infusion devices
Intracranial aneurysm clips (some but not all)
Ocular implants (some) or ocular metallic foreign body
McGee stapedectomy piston prosthesis
Duraphase penile implant
Swan-Ganz catheter
Magnetic stoma plugs
Magnetic dental implants
Magnetic sphincters
Ferromagnetic inferior vena cava filters, coils, stents—safe 6 weeks after implantation
Tattooed eyeliner (contains ferromagnetic material and may irritate eyes)

Note: See also *http://www.mrisafety.com*.

signal intensity of flowing blood entering the section; arterial or venous structures may be highlighted. A typical TOF MRA sequence results in a series of contiguous, thin MR sections (0.6–0.9 mm thick), which can be viewed as a stack and manipulated to create an angiographic image data set that can be reformatted and viewed in various planes and angles, much like that seen with conventional angiography **(Fig. 416-2*G*)**.

Phase-contrast MRA has a longer acquisition time than TOF MRA, but in addition to providing anatomic information similar to that of TOF imaging, it can be used to reveal the velocity and direction of blood flow in a given vessel.

MRA is often acquired during infusion of contrast material. Advantages include faster imaging times (1–2 min vs 10 min), fewer flow-related artifacts, and 4D temporal imaging resulting in arterial and venous phases. Recently, contrast-enhanced MRA has become the standard for extracranial vascular MRA. This technique entails rapid imaging using coronal three-dimensional TOF sequences during a bolus infusion of gadolinium contrast agent.

MRA has lower spatial resolution compared with conventional film-based angiography, and therefore the detection of small-vessel abnormalities, such as vasculitis and distal vasospasm, is problematic. MRA is also less sensitive to slowly flowing blood and thus may not reliably differentiate complete from near-complete occlusions. Motion, either by the patient or by anatomic structures, may distort the MRA images, creating artifacts. These limitations notwithstanding, MRA has proved useful in evaluation of the extracranial carotid and vertebral circulation as well as of larger-caliber intracranial arteries and dural sinuses. It has also proved useful in the noninvasive detection of intracranial aneurysms and vascular malformations.

ECHO-PLANAR MRI

Recent improvements in gradients, software, and high-speed computer processors now permit extremely rapid MRI of the brain. With echo-planar MRI (EPI), fast gradients are switched on and off at high speeds to create the information used to form an image. With EPI, all of the information required for processing an image is accumulated in milliseconds, and the information for the entire brain can be obtained in <1–2 min, depending on the degree of resolution required or desired. Fast MRI reduces patient and organ motion and is the basis of perfusion imaging during contrast infusion and kinematic motion studies. EPI is also the sequence used to obtain diffusion imaging and tractography, as well as fMRI and arterial spin-labeled studies **(Figs. 416-2*H*,** 416-3, **416-4*C*,** and 416-6; and **Fig. 419-13)**.

Perfusion and diffusion imaging are EPI techniques that are useful in early detection of ischemic injury of the brain and may be useful together to demonstrate infarcted tissue as well as ischemic but potentially viable tissue at risk of infarction (e.g., the ischemic penumbra). Diffusion-weighted imaging (DWI) assesses microscopic motion of water; abnormal restriction of motion appears as relative high-signal intensity on diffusion-weighted images. Infarcted tissue reduces the water motion within cells and in the interstitial tissues, resulting in high signal on DWI. DWI is the most sensitive technique for detection of acute cerebral infarction of <7 days in duration (Fig. 416-2*H*). It is also quite sensitive for detecting dying or dead brain tissue secondary to encephalitis, as well as abscess formation (Fig. 416-3*B*).

Perfusion MRI can be performed by the acquisition of fast echo planar gradient images during a rapid intravenous bolus of gadolinium contrast material or by arterial spin labeling (ASL) without the use of contrast material. With contrast perfusion imaging, relative cerebral blood volume, mean transit time, and cerebral blood flow maps are derived. Delay in mean transit time and reduction in cerebral blood volume and cerebral blood flow are typical of infarction. In the setting of reduced blood flow, a prolonged mean transit time of contrast but normal or elevated cerebral blood volume may indicate tissue supplied by collateral flow that is at risk of infarction. Perfusion MRI imaging can also be used in the assessment of brain tumors to differentiate intraaxial primary tumors, whose BBB is relatively intact, from extraaxial tumors or metastases, which demonstrate a relatively more permeable BBB.

Diffusion tensor imaging is derived from diffusion MRI imaging sequences. This technique assesses the direction and integrity of white matter architecture. It has proven valuable in preoperative assessment of subcortical white matter tract anatomy prior to brain tumor surgery, as well as determining normal and abnormal white matter architecture in congenital and acquired pathologies (Fig. 416-6).

fMRI of the brain is an EPI technique that localizes regions of activity in the brain following task activation or at rest (so-called resting state fMRI). Neuronal activity elicits a slight increase in the delivery of oxygenated blood flow to a specific region of activated brain. This results in an alteration in the balance of oxyhemoglobin and deoxyhemoglobin, which yields a 2–3% increase in signal intensity within veins and local capillaries. Currently, preoperative somatosensory and auditory cortex localization is possible, and methods to assess motor and language function are in development. This technique has proved useful to neuroscientists interested in interrogating the localization of certain brain functions.

ARTERIAL SPIN LABELING

ASL is a quantitative noninvasive MR technique that measures cerebral blood flow. Blood traversing in the neck is labeled by an MR pulse and then imaged in the brain after a short (2 second) delay. The signal in the brain is reflective of blood flow. ASL is an especially important technique for patients with kidney failure and for pediatric patients in whom the use of radioactive tracers or exogenous contrast agents is contraindicated. Increased cerebral flow is more easily identified than slow flow, which can be sometimes difficult to quantify. This technique has also been useful in detecting arterial venous shunting in arteriovenous malformations and arteriovenous fistulas, as well as increased blood flow in brain tumors, and patients post TIA, post-ictal, or post migrane.

MAGNETIC RESONANCE NEUROGRAPHY

MRN is a T2W MR technique that shows promise in detecting increased signal in irritated, inflamed, or infiltrated peripheral nerves. Images are obtained with fat-suppressed fast spin echo imaging or short inversion recovery sequences. Irritated or infiltrated nerves will demonstrate high signal on T2W imaging. This is indicated in patients with radiculopathy whose conventional MR studies of the spine are normal, or in those suspected of peripheral nerve entrapment or trauma.

POSITRON EMISSION TOMOGRAPHY

PET relies on the detection of positrons emitted during the decay of a radionuclide that has been injected into a patient. The most frequently used moiety is 2-[^{18}F]fluoro-2-deoxy-D-glucose (FDG), which is an analogue of glucose and is taken up by cells competitively with 2-deoxyglucose. Multiple images of glucose uptake activity are formed after 45–60 min. Images reveal differences in regional glucose activity among normal and pathologic brain structures. FDG-PET is used primarily for the detection of extracranial metastatic disease; however, a lower activity of FDG in the parietal lobes is associated with Alzheimer's disease, a finding that may simply reflect atrophy that occurs in the later stages of the disease. Combination PET-CT scanners, in which both CT and PET are obtained at one sitting, have largely replaced PET scans alone for most clinical indications. MR-PET scanners have also been developed and may prove useful for imaging the brain and other organs without the radiation exposure of CT. More recent PET ligand developments include beta-amyloid tracers, such as Pittsburgh compound B (PIB) and 18-F AV-45 (florbetapir), and tau PET tracers, such as 18F-T807 and T808. Studies have shown an increased percentage of amyloid deposition in patients with Alzheimer's disease compared with mild cognitive impairment and healthy controls; however, up to 25% of cognitively "normal" patients show abnormalities on amyloid PET imaging. This may either reflect subclinical disease processes or variation of normal. Tau imaging may be more specific for Alzheimer's disease, and clinical studies are under way.

MYELOGRAPHY

TECHNIQUE

Myelography involves the intrathecal instillation of specially formulated water-soluble iodinated contrast medium into the lumbar or cervical subarachnoid space. CT scanning is typically performed after

myelography (*CT myelography*) to better demonstrate the spinal cord and roots, which appear as filling defects in the opacified subarachnoid space. *Low-dose CT myelography*, in which CT is performed after the subarachnoid injection of a small amount of relatively dilute contrast material, has replaced conventional myelography for many indications, thereby reducing exposure to radiation and contrast media. CT is obtained at a slice thickness ~ 2.5 mm and reconstructed at 0.625-mm thick slices which can quickly be reformated in sagittal and coronal planes, equivalent to traditional myelography projections.

INDICATIONS

CT myelography and MRI have largely replaced conventional myelography for the diagnosis of diseases of the spinal canal and cord (Table 416-1). Remaining indications for conventional plain-film myelography include the evaluation of suspected meningeal or arachnoid cysts and the localization of CSF fistulas. Conventional myelography and CT myelography provide the most precise information in patients with failed back syndrome following spinal fusion procedures.

CONTRAINDICATIONS

Myelography is relatively safe; however, it should be performed with caution in any patient with elevated intracranial pressure, evidence of a spinal block, or a history of allergic reaction to intrathecal contrast media. In patients with a suspected spinal block, MR is the preferred imaging technique. If myelography is necessary, only a small amount of contrast medium should be instilled below the block in order to minimize the risk of neurologic deterioration. Lumbar puncture is to be avoided in patients with bleeding disorders, including patients receiving anticoagulant therapy, as well as in those with infections of the overlying soft tissues.

COMPLICATIONS

Headache is the most frequent complication of myelography and is reported to occur in 5–30% of patients. Nausea and vomiting may also occur rarely. Postural headache (post–lumbar puncture headache) is generally due to continued leakage of CSF from the dural puncture site. A higher incidence is noted among younger women and with the use of larger gauge cutting-type spinal needles. If significant headache persists for >48 h, placement of an epidural blood patch should be considered. **Management of lumbar puncture headache is discussed in Chap. 13.** Vasovagal syncope may occur during lumbar puncture; it is accentuated by the upright position used during lumbar myelography. Adequate hydration before and after myelography will reduce the incidence of this complication.

Hearing loss is a rare complication of myelography. It may result from a direct toxic effect of the contrast medium or from an alteration of the pressure equilibrium between CSF and perilymph in the inner ear. Puncture of the spinal cord is a rare but serious complication of cervical (C1–2) or high lumbar puncture. The risk of cord puncture is greatest in patients with spinal stenosis, Chiari malformations, or conditions that reduce CSF volume. In these settings, a low-dose lumbar injection followed by thin-section CT or MRI is a safer alternative to cervical puncture. Intrathecal contrast reactions are rare, but aseptic meningitis and encephalopathy are reported complications. The latter is usually dose related and associated with contrast entering the intracranial subarachnoid space. Seizures rarely occur following myelography, historically reported in 0.1–0.3% of patients. Risk factors include a preexisting seizure disorder and the use of a total iodine dose of >4500 mg. Other reported complications include hyperthermia, hallucinations, depression, and anxiety states. These side effects have been reduced by the development of nonionic, water-soluble contrast agents as well as by head elevation and generous hydration following myelography.

SPINE INTERVENTIONS

DISKOGRAPHY

The evaluation of back pain and radiculopathy may require diagnostic procedures that attempt either to reproduce the patient's pain or relieve it, indicating its correct source prior to lumbar fusion. Diskography is performed by fluoroscopic placement of a 22- to 25-gauge needle into the intervertebral disk and subsequent injection of 1–3 mL of contrast media. The intradiskal pressure is recorded, as is an assessment of the patient's response to the injection of contrast material. Typically little or no pain is felt during injection of a normal disk, which does not accept much more than 1 mL of contrast material, even at pressures as high as 415–690 kPa (60–100 lb/in^2). CT and plain films are obtained following the procedure. Concerns have been raised that diskography may contribute to an accelerated rate of disk degeneration; furthermore, patients who suffer from depression or anxiety are more likely to find diskography painful and in some cases the procedure-associated pain became persistent, lasting a year or longer.

SELECTIVE NERVE ROOT AND EPIDURAL SPINAL INJECTIONS

Percutaneous selective nerve root and epidural blocks with glucocorticoid and anesthetic mixtures may be both therapeutic and diagnostic, especially if a patient's pain is relieved. Typically, 1–2 mL of an equal mixture of a long-acting glucocorticoid such as betamethasone and a long-acting anesthetic such as bupivacaine 0.75% is instilled under CT or fluoroscopic guidance in the intraspinal epidural space or adjacent to an existing nerve root. This can also be performed in the facet joints, or around the medial nerve branches that supply innervation to the facet joints.

ANGIOGRAPHY

Catheter angiography is indicated for evaluating intracranial small-vessel pathology (such as vasculitis), for assessing vascular malformations and aneurysms, and in endovascular therapeutic procedures (Table 416-1). As noted above, angiography has been replaced for many indications by CT/CTA or MRI/MRA.

Angiography carries the greatest risk of morbidity of all diagnostic imaging procedures, owing to the necessity of inserting a catheter into a blood vessel, directing the catheter to the required location, injecting contrast material to visualize the vessel, and removing the catheter while maintaining hemostasis. Therapeutic transcatheter procedures (see below) have become important options for the treatment of some cerebrovascular diseases. The decision to undertake a diagnostic or therapeutic angiographic procedure requires careful assessment of the goals of the investigation and its attendant risks.

To improve tolerance to contrast agents, patients undergoing angiography should be well hydrated before and after the procedure. Because the femoral route is used most commonly, the femoral artery must be compressed after the procedure to prevent a hematoma from developing. The puncture site and distal pulses should be evaluated carefully after the procedure; complications can include thigh hematoma or lower extremity emboli.

COMPLICATIONS

A common femoral arterial puncture provides retrograde access via the aorta to the aortic arch and great vessels. The most feared complication of cerebral angiography is stroke. Thrombus can form on or inside the tip of the catheter, and atherosclerotic thrombus or plaque can be dislodged by the catheter or guide wire or by the force of injection and can embolize distally in the cerebral circulation. Risk factors for ischemic complications include limited experience on the part of the angiographer, atherosclerosis, vasospasm, low cardiac output, decreased oxygen-carrying capacity, advanced age, and prior history of migraine. The risk of a neurologic complication varies but is ~4% for transient ischemic attack and stroke, 1% for permanent deficit, and <0.1% for death.

Ionic contrast material injected into the cerebral vasculature can be neurotoxic if the BBB is breached, either by an underlying disease or by the injection of hyperosmolar contrast agent. Ionic contrast media are less well tolerated than nonionic media, probably because they can induce changes in cell membrane electrical potentials. Patients with dolichoectasia of the basilar artery can suffer reversible brainstem dysfunction and acute short-term memory loss during angiography, owing to the slow percolation of the contrast material and the consequent prolonged exposure of the brain. Rarely, an intracranial aneurysm ruptures during an angiographic contrast injection, causing subarachnoid hemorrhage, perhaps as a result of injection under high pressure.

SPINAL ANGIOGRAPHY

Spinal angiography may be indicated to evaluate vascular malformations and tumors and to identify the artery of Adamkiewicz (Chap. 434) prior to aortic aneurysm repair. The procedure is lengthy and requires the use of relatively large volumes of contrast; the incidence of serious complications, including paraparesis, subjective visual blurring, and altered speech, is ~2%. Gadolinium-enhanced MRA has been used successfully in this setting, as has iodinated contrast CTA, which has promise for replacing diagnostic spinal angiography for some indications.

INTERVENTIONAL NEURORADIOLOGY

This rapidly developing field is providing new therapeutic options for patients with challenging neurovascular problems. Available procedures include detachable coil therapy for aneurysms, particulate or liquid adhesive embolization of arteriovenous malformations, stent retrieval systems for embolectomy, balloon angioplasty and stenting of arterial stenosis or vasospasm, transarterial or transvenous embolization of dural arteriovenous fistulas, balloon occlusion of carotid-cavernous and vertebral fistulas, endovascular treatment of vein-of-Galen malformations, preoperative embolization of tumors, and thrombolysis of acute arterial or venous thrombosis. Many of these disorders place the patient at high risk of cerebral hemorrhage, stroke, or death.

The highest complication rates are found with the therapies designed to treat the highest risk diseases. The advent of electrolytically detachable coils ushered in a new era in the treatment of cerebral aneurysms. Two randomized trials found reductions of morbidity and mortality at 1 year among those treated for aneurysm with detachable coils compared with neurosurgical clipping. It remains to be determined what the role of coils will be relative to surgical options, but in many centers, coiling has become standard therapy for many aneurysms.

Finally, recent studies of stent retrieval systems used to withdraw emboli have shown improved clinical outcomes in patients presenting with large vessel occlusions and signs of stroke (Chap. 420).

FURTHER READING

Haller S et al: Arterial spin labeling perfusion of the brain: Emerging clinical applications. Radiology 281:337, 2016.

Kim SJ, Kim KA: Safety issues and updates under MR environments. Eur J Radiol 89:7, 2017.

Mandell DM et al: Intracranial vessel wall MRI: Principles and expert consensus recommendations of the American Society of Neuroradiology. AJNR Am J Neuroradiol 38:218, 2017.

Nguyen HS et al: Progressing bevacizumab-induced diffusion restriction is associated with coagulative necrosis surrounded by viable tumor and decreased overall survival in patients with recurrent glioblastoma. AJNR Am J Neuroradiol 37:2201, 2016.

Ramalho J et al: Gadolinium-based contrast agent accumulation and toxicity: An update. AJNR Am J Neuroradiol 37:1192, 2016.

417 Pathobiology of Neurologic Diseases

Stephen L. Hauser, Arnold R. Kriegstein, Stanley B. Prusiner

The human nervous system is the organ of consciousness, cognition, ethics, and behavior; as such, it is the most intricate structure known to exist. More than one-third of the 23,000 genes encoded in the human genome are expressed in the nervous system. Each mature brain is composed of 100 billion neurons, several million miles of axons and dendrites, and $>10^{15}$ synapses. Neurons exist within a dense parenchyma of multifunctional glial cells that synthesize myelin, preserve homeostasis, and regulate immune responses. Measured against this background of complexity, the achievements of molecular neuroscience have been extraordinary. Advances have occurred in parallel with the development of new enabling technologies—in bioengineering and computational sciences, imaging, and cell, molecular and chemical biology—and moving forward it is likely that the pace of new discoveries will only increase. This chapter reviews a number of the most dynamic areas in neuroscience, specifically highlighting advances in immunology and inflammation, neurodegeneration, and stem cell biology. In each of these areas, recent discoveries are providing context for an understanding of the triggers and mechanisms of disease, and offering new hope for prevention, treatment, and repair of nervous system injuries. Discussions of the neurogenetics of behavior, advances in addiction science, and diseases caused by network dysfunction can be found in Chap. 443 (Biology of Psychiatric Disorders); and new approaches to rehabilitation via harnessing of neuroplasticity, neurostimulation, and computer-brain interfaces are presented in Chap. 477 (Emerging Neurotherapeutic Technologies).

NEUROIMMUNOLOGY AND NEUROINFLAMMATION

OLIGODENDROCYTES AND MYELIN

Myelin is the multilayered insulating substance that surrounds axons and speeds impulse conduction by permitting action potentials to jump between naked regions of axons (nodes of Ranvier) and across myelinated segments. Molecular interactions between the myelin membrane and axon are required to maintain the stability, function, and normal life span of both structures. A single oligodendrocyte usually ensheaths multiple axons in the central nervous system (CNS), whereas in the peripheral nervous system (PNS), each Schwann cell typically myelinates a single axon. Myelin is a lipid-rich material formed by a spiraling process of the membrane of the myelinating cell around the axon, creating multiple membrane bilayers that are tightly apposed (compact myelin) by charged protein interactions. Several inhibitors of axon growth are expressed on the innermost (periaxonal) lamellae of the myelin membrane (see below). A number of clinically important neurologic disorders are caused by inherited mutations in myelin proteins of the CNS or PNS (Chap. 438), and constituents of myelin also have a propensity to be targeted as autoantigens in autoimmune demyelinating disorders (Chap. 436).

Premyelinating oligodendrocyte precursor cells (OPCs) are highly motile cells that migrate extensively during development and in the adult brain following injuries to the myelin sheath. OPCs migrate along the inner (or abluminal) surface of endothelial cells, a process regulated by *Wnt* pathway signaling and upregulation of the chemokine receptor Cxcr4 that drives their attachment and retention to the vasculature. Initial specification to OPCs is transcriptionally regulated by the *Olig 2* and *Yin Yang 1* genes, whereas the later stage of myelination mediated by postmitotic oligodendrocytes depends on a different transcription factor, *myelin gene regulatory factor (MRF)*. In the normal adult brain, large numbers of OPCs (expressing PDGFR-α and NG2) are widely distributed but do not myelinate axons, even in demyelinating environments such as in lesions of multiple sclerosis (MS). In addition to *Wnt*, several families of molecules have been identified that regulate oligodendrocyte differentiation and myelination, including LINGO-1, PSA-NCAM, hyaluronan, Nogo-A, the Wnt pathway, notch signaling (and its receptor Jagged), and the M1 muscarinic receptor Chrm1, all of which are inhibitory, and the retinoic acid receptor RXRγ, which is excitatory. All are also potential targets for myelin repair therapies. In an experimental allergic encephalomyelitis (EAE) model, oligodendrocyte-specific knockout of Chrm1 improved remyelination, protected axons and restored function, directly demonstrating that remyelination can be neuroprotective following injury. A recently reported pivotal trial of a monoclonal antibody against LINGO-1 failed to promote remyelination, a disappointing result given that the antibody appeared to have promising clinical effects in an earlier phase 2 trial.

A series of observations has called into question the traditional concept that axon-derived cues are always required for myelination

to occur. Fixed (i.e., dead) axons could be efficiently myelinated by oligodendrocytes in vitro, as could artificial polystyrene nanowires of a similar diameter. This led to development of new high-throughput screening assays based on myelination of polystyrene nanowires to identify compounds that could promote myelination and in a preliminary human trial a molecule that emerged from this assay, the antihistamine clemastine, had clear efficacy as a remyelinating agent in patients with chronic optic neuropathy due to MS. Remarkably, the drug appears to work via binding to the Chrm1 muscarinic receptor.

MACROPHAGES AND MICROGLIA

These represent the major cell types in the nervous system responsible for antigen presentation and innate immunity. Brain microglia migrate from the yolk sac early in embryogenesis before the blood-brain barrier is formed, and are believed to maintain their cell numbers through cell division within the nervous system and not via repopulation from the circulation. Depletion of microglia in adult mice by administration of a selective inhibitor of colony-stimulating factor receptor 1 (CSFR1) was followed by their rapid repopulation, suggesting that a pool of resident microglial precursor cells exist throughout the CNS. Additional roles for brain microglia are known to exist in neurogenesis, through secretion of brain derived neurotrophic factor (BDNF) and other molecules, as well as in the development and regulation of neural circuits through pruning of excitatory synapses and control of dendritic spine densities **(Fig. 417-1)**. Mice depleted of microglia during development exhibit a variety of cognitive, learning and behavioral deficits, including abnormal social behaviors; these processes are dependent on the classical complement pathway molecules and the chemokine receptor CX3CR1. A challenge to the field has been that tools to definitively separate brain microglia from perivascular macrophages do not currently exist. A recent advance that could provide a possible solution utilizes adult skin-derived pluripotent stem cells (iPSCs), and the development of methods to generate microglia-like cells from iPSCs using media containing IL-34 and colony-stimulating factor 1.

Microglia are located throughout the brain parenchyma, whereas brain macrophages occur primarily in perivascular regions, including the meninges and choroid plexus. In contrast to microglia, brain macrophages are derived from monocytes that enter the nervous system at low levels from the bloodstream on a continuing basis and in higher numbers during pathologic states. A possible exception to this rule are the meningeal macrophages, located primarily in the subdural space, that appear to enter the brain at an early developmental stage and remain throughout the life of the individual. In a murine model of autoimmune demyelination, EAE **(Fig. 417-2)**, macrophages derived from bone marrow monocytes, but not microglia, were the critical population that initiated inflammatory demyelination at paraxonal regions near nodes of Ranvier. Brain macrophages have been found to have multiple pro-inflammatory functions, including promoting adhesion, attraction and activation of B and T lymphocytes; providing antigen-specific activation of T cells via antigen presentation of specific immunogenic peptides, including autoantigens, complexed to surface class II major histocompatibility complex (MHC II) molecules; and contributing to cell injury through generation of oxidative stress and cytotoxicity. By contrast, microglia have been traditionally thought to downregulate inflammatory responses and promote tissue repair. This model of M1 (pro-inflammatory) and M2 (regulatory/repair) macrophage/microglial functions, derived primarily from experimental models of autoimmunity, is certainly an oversimplification, and more nuanced functions of these cell types can be revealed depending on the specific context and environmental cues.

Evidence also supports a primary role for brain macrophages and microglia in neurodegenerative diseases, in contrast to earlier views in which their role was seen as largely secondary and involving phagocytosis of cell debris. In different situations, macrophages and microglia can be either protective or pathogenic. In mice, macrophages contribute to spatial memory when activated in the presence of the cytokine interleukin (IL)-4 produced by invading lymphocytes, and microglia through secretion of BDNF support learning and memory through promoting synaptic plasticity. Experimentally, microglia and brain macrophages also participate in clearance of pathogenic β-amyloid aggregates in Alzheimer's disease (AD) mice, and disruption of brain macrophages by knockout of CCR2, a chemokine required for entry of bloodstream monocytes into the CNS, exacerbated AD pathology. On the other hand, data indicate that disease exacerbating effects of microglia and macrophages may predominate in other situations. A direct role for microglia in human AD was suggested by genetic evidence implicating the phagocytosis-associated gene TREM2, and other genes belonging to the complement system, in AD susceptibility. Activation of the classical complement cascade is also assuming an increased role in concepts of pathogenesis, as follows; synapses targeted for elimination express the complement proteins C1q and C3, the levels of which increase in the presence of excess β-amyloid; C3-bearing synapses are then targeted for elimination by microglia that express the complement 3 receptor (CR3); and knockout of C3 can rescue the clinical and pathologic abnormalities associated with neurodegeneration in AD-prone mice. In familial frontotemporal degeneration (FTD) due to mutations of progranulin (pgrn), a prominent immune pathology has also been identified, including the presence of activated microglia expressing high levels of pro-inflammatory cytokines. In pgrn-/- mice, an age-dependent microglial activation is associated with upregulation of genes associated with innate immunity including complement proteins, and with enhanced pruning of inhibitory synapses in key regions of the CNS, leading to behavioral disorders reminiscent of human FTD. Moreover, inhibition of

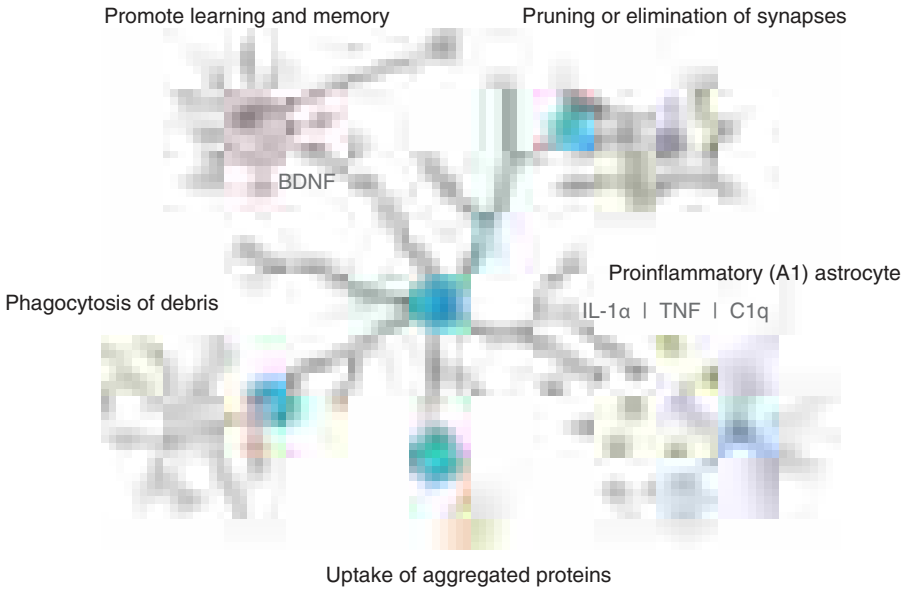


FIGURE 417-1 The multifunctional microglial cell. Microglia have diverse functions that can support healthy development and maintain homeostasis, or contribute to tissue damage in pathologic conditions. Homeostatic functions include promotion of learning and memory through secretion of soluble proteins such as brain derived neurotrophic factor (BDNF); participation in normal synaptic pruning; and clearing cellular debris and protein aggregates via phagocytosis. However, in pathologic states activated microglia also contribute to tissue damage, by targeting normal healthy neurons and synapses; by promoting formation of β-amyloid or other misfolded proteins deposited in neurodegenerative diseases; and secreting cytokines (such as IL-1α, TNF, and the complement component C1q) incriminated in induction of neurotoxic A1 astrocytes. In addition, microglia have diverse functions in adaptive immunity, including roles in antigen presentation and immune regulation (Fig. 417-2). *(Figure adapted from J Herz et al: Myeloid cells in the central nervous system. Immunity 46:943, 2017, Fig. 2.)*

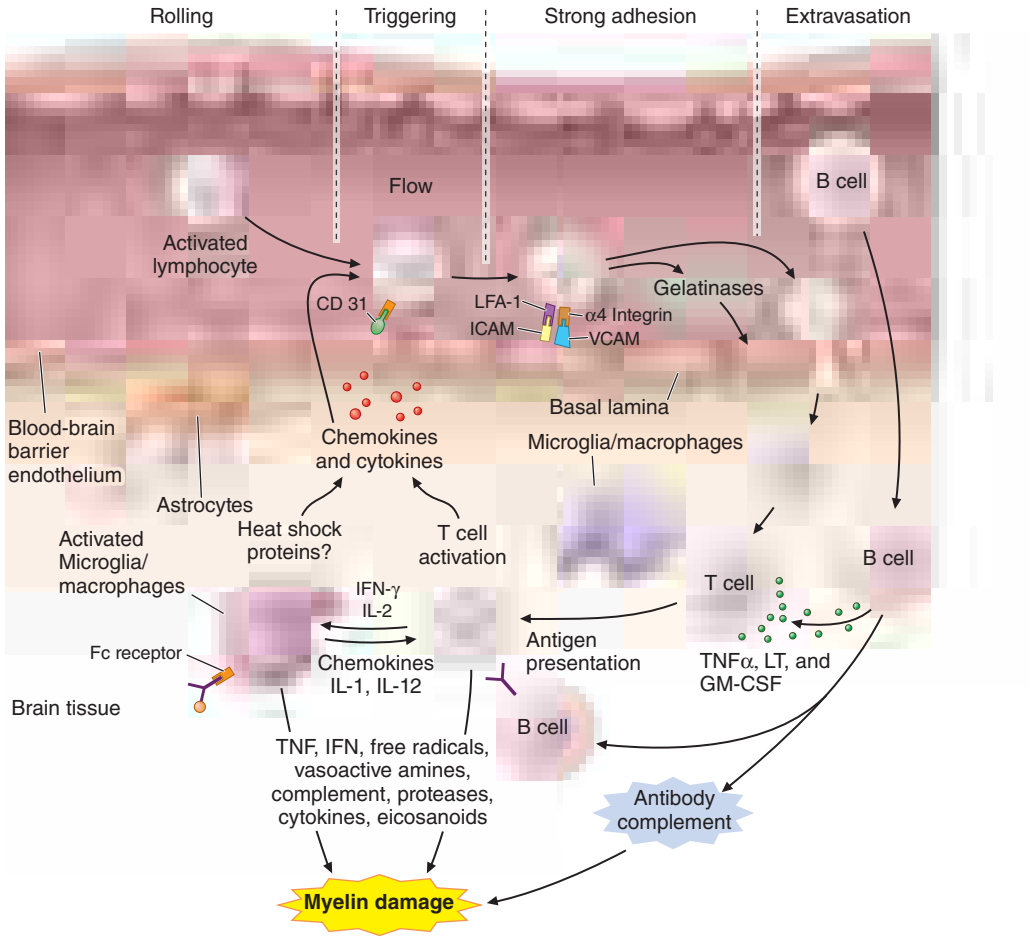


FIGURE 417-2 A model for experimental allergic encephalomyelitis (EAE). Crucial steps for disease initiation and progression include peripheral activation of preexisting autoreactive T cells; homing to the central nervous system (CNS) and extravasation across the blood-brain barrier; reactivation of T cells by exposed autoantigens; secretion of cytokines; activation of microglia and astrocytes and recruitment of a secondary inflammatory wave; and immune-mediated myelin destruction. ICAM, intercellular adhesion molecule; IFN, interferon; IL, interleukin; LFA-1, leukocyte function-associated antigen-1; TNF, tumor necrosis factor; VCAM, vascular cell adhesion molecule.

complement activation rescued all of these deficits. Taken together, these data indicate a primary role for microglial activation in pgrn associated FTD, likely occurring through enhanced lysosomal trafficking and increased production of cleavage products of the C3 complement component, and leading to enhanced and deleterious synaptic pruning in regions of the brain affected in FTD. Although it is likely that the specific mechanisms of complement dependent neurodegeneration will differ in distinct neurodegenerative conditions, these data provide hope that complement pathway interventions could represent a possible approach to control of neurodegenerative pathologies mediated at least in part through the innate immune system.

ASTROCYTES

Astrocytes represent half or more of all cells in the CNS. Traditionally thought to function as simple interstitial supporting cells that provide scaffolds for neuronal migration and contribute to homeostasis, emerging data indicate far more pleiotropic functions for this cell type. Astrocytes exert profound roles in the life of synapses by secreting factors (such as apolipoprotein E, thrombospondins, and glypicans) that regulate development, maintenance, and pruning of presynaptic and postsynaptic structures. Influenced by local neuronal activity, astrocytes actively phagocytose synapses. Astrocytes also participate in dynamic regulation of vascular tone, in part through astrocyte-astrocyte communication mediated through gap junctions and calcium waves modulated by neuronal activity; support blood brain barrier and glymphatic (see below) integrity through extension of foot process to the vascular structures and expression of aquaporin-4 water channels; and carry out additional metabolic functions essential for the maintenance of neuronal health.

One characteristic of the response to many types of brain injury is reactive astrocytosis, or the formation of a glial scar. Recent work has identified two fundamentally different types of reactive astrocytes that appear to have countervailing functions; the terms A1 and A2 astrocytes have been proposed, by analogy to brain macrophage/microglia M1 and M2 designations, described above. A2 astrocytes are induced by ischemia and may serve beneficial functions, including a contribution to tissue repair after injury. By contrast, A1 astrocytes are induced in diverse inflammatory and degenerative states, and appear to actively participate in the injury process. Interestingly, secreted products of activated microglia, specifically IL-1a, TNF, and C1q, induce astrocytes to transform to the A1 type. Functionally, A1 astrocytes lose the capacity to phagocytose synapses and myelin debris, and are strikingly toxic in vitro to various populations of neurons and to mature oligodendrocytes, potentially at least in part via complement mediated damage. Interestingly, OPCs, abundant in active lesions of multiple sclerosis (MS; **Chap. 436**) despite the inflammatory milieu, are resistant to A1 mediated death. The nature of the toxic factor is unknown. There is speculation that products of A1 astrocytes could promote damage in conditions as varied as MS, AD **(Chap. 423)**, Parkinson's disease (PD) **(Chap. 427)**, and amyotrophic lateral sclerosis (ALS) **(Chap. 429)**, despite their distinct etiologies and pathologies.

GLYMPHATICS

Two newly identified lymphatic structures of the CNS are the glymphatic and deep dural lymphoid systems, responsible for clearance of debris in the CNS, and likely also serving a role in immune surveillance. The brain has traditionally been considered to lack a classical lymphatic system, and immune responses against antigens are less

effectively generated in the CNS than in other organ systems, a concept termed "immune privilege." However, there is abundant evidence that the immune privilege status of the brain is only relative and not absolute. Furthermore, given the high metabolic demands of the brain some mechanism for efficient removal of solute and debris must be present. One well-established pathway involves the passive flow of solutes from the brain parenchyma into the cerebrospinal fluid (CSF), and their exit via the arachnoid granulations, as well as along cranial and spinal nerve roots to a series of lymphoid structures located in the cribriform plate and nasal mucosa and elsewhere.

The glymphatic system derives its name from a distinctive architecture involving lymphoid-like structures and astroglial cells. CSF synthesized in the arachnoid villi circulates through the ventricles and subarachnoid space surrounding the convexities of the brain and spinal cord, and exits through conduits surrounding arterioles penetrating into the brain parenchyma. These spaces are lined by endothelial cells internally, and by astrocyte foot processes that form the external walls. Aided by arterial propulsion, CSF moves out of these specialized conduits and into astrocytes via foot processes rich in aquaporin-4 water channels, and then in the interstitium of brain parenchyma picks up solutes and particulate debris that are then carried to perivenous spaces where they passage to exit the brain and drain into the lymphatic system. In mice, knockout of aquaporin-4 markedly reduced the flow of interstitial fluids in the brain, underscoring the critical role of astrocyte uptake of CSF in this process. Interstitial flow in the CNS is also impaired with aging, possibly related to changes in astrocytic aquaporin-4 expression. Another fascinating aspect of the glymphatic system is that the transport of fluids and solutes accelerates with sleep, arguing for a critical role for sleep in promoting clearance of debris needed to meet the high metabolic demands of the nervous system. Furthermore, in disease models, aggregated proteins associated with neurodegenerative disease, such as β-amyloid associated with AD **(Chap. 423)**, were also more efficiently cleared during sleep. Indeed, in mice genetically engineered to produce excess β-amyloid and develop Alzheimer's-like cognitive decline, sleep deprivation increased accumulation of amyloid plaques. Glymphatic pathways are also likely to represent an important egress pathway for lymphocytes in the CNS and a route for lymphocyte encounter with CNS antigens in cervical lymph nodes. In this regard, recent data indicate that deep cervical lymph nodes might be a site for antigen-specific stimulation of B-cells in MS **(Chap. 436)**.

A second recently identified pathway consists of a plexus of small lymphatic-like vessels located on the external surface of meningeal arteries and deep dural sinuses (including the sagittal and transverse sinuses), structures that exit the brain along the surface of veins and arteries and drain to the deep cervical lymph nodes. These conduits are comprised of cells that express a transcriptome indicating that they are components of a lymphoid drainage system distinct from vascular endothelium. These sinus-associated lymphoid structures may be most important in clearing solutes from the CSF, in contrast to the glymphatic system that likely functions to remove waste products from the brain interstitium; however, the exact functions of these two systems and their interrelationships are only beginning to be understood.

MICROBIOTA AND NEUROLOGIC DISEASE

The human microbiome **(Chap. 459)** represents the collective set of genes from the 10^{14} organisms living in our gut, skin, mucosa, and other sites. Different microbial communities are associated with different ethnicities, diets, and environments. In any individual, the predominant gut microbiota can be remarkably stable over decades, but also can be altered by exposure to certain microbial species, for example by ingestion of probiotics.

There is compelling evidence that gut microbes can shape immune responses through the interaction of their metabolism with that of humans. These gut-brain interactions are likely to be important in understanding the pathogenesis of many autoimmune neurologic diseases. For example, mice treated with broad-spectrum antibiotics are resistant to EAE, an effect associated with decreases in production of proinflammatory cytokines, and conversely more production of the immunosuppressive cytokines IL-10 and IL-13 and an increase in regulatory T and B lymphocytes. Oral administration of polysaccharide A (PSA) from *Bacillus fragilis* also protects mice from EAE, via increases in IL-10. Intestinal microbiota from patients with MS were found to promote EAE when transferred to germ free mice, possibly due to imbalances between bacterial species that promote inflammation (such as *Akkermansia muciniphila* and *Acinetobacter calcoaceticus*) and those that induce regulatory immune responses (such as *Parabacteroides distasonis*).

In addition to nonspecific effects on immune homeostasis mediated by cytokines and regulatory cells, some microbial proteins can trigger, in susceptible individuals, a cross-reactive immune response against a homologous protein in the nervous system, a mechanism termed *molecular mimicry*. Examples include cross-reactivity between the astrocyte water channel aquaporin-4 and an ABC transporter permease from *Clostridia perfringens* in neuromyelitis optica **(Chap. 437)**; the neural ganglioside Gm1 and similar sialic acid–containing structures from *Campylobacter jejuni* in Guillain-Barré syndrome **(Chap. 439)**; and the sleep-promoting protein hypocretin and hemagglutinin from H1N1 influenza virus in narcolepsy **(Chap. 27)**.

Recently, a number of tantalizing observations have incriminated the microbial environment in the pathogenesis of a much wider spectrum of neurologic conditions and behaviors, extending well beyond the traditional boundaries of immune-mediated pathologies. It has long been known that gut bacteria can influence brain function, based mostly on classic studies demonstrating that products of gut microbes can worsen hepatic encephalopathy, forming the basis of treatment with antibiotics for this condition.

Mice that developed in a completely germ-free environment displayed less anxiety, lower responses to stressful situations, more exploratory locomotive behaviors, and impaired memory formation compared with non-germ-free counterparts. These behaviors were related to changes in gene expression in pathways related to neural signaling, synaptic function, and modulation of neurotransmitters. Moreover, this behavior could be reversed when the germ-free mice were co-housed with non-germ-free mice. In other experiments, intestinal microbiota were also found to be required for the normal development and function of brain microglia, potentially linking these behavioral effects to specific cellular targets in the CNS.

The enteric autonomic nervous system in humans provides a bidirectional neural connection between the brain and gut. The vagus nerve, which innervates the upper gut and proximal colon, has been implicated in anxiety- and depression-like behaviors in mice. Ingestion of *Lactobacillus rhamnosus* induced changes in expression of the inhibitory neurotransmitter GABA1b in neurons of the limbic cortex, hippocampus, and amygdala, associated with reduced levels of corticosteroids and reduced anxiety- and depression-like behaviors. Remarkably, these changes could be blocked by vagotomy.

Another area of emerging interest is in a possible contribution of the gut microbiome to autism and related disorders. Children with autistic spectrum disorders have long been known to have gastrointestinal disturbances, and it has been claimed that the severity of dysbiosis correlates with the severity of autism. A murine model of autism was recently induced in offspring after injecting the pregnant mother with the viral RNA mimic polyinosinic:polycytidylic acid (poly I:C). Remarkably, oral treatment of offspring with *B. fragilis* corrected a range of autistic behaviors in these mice and also improved gut permeability.

PATHOLOGIC PROTEINS, PRIONS, AND NEURODEGENERATION (FIG. 417-3)

The term "protein aggregation" has become widely used to describe easily recognizable hallmarks of neurodegeneration. While such neuropathologic hallmarks including plaques, neurofibrillary tangles, and inclusion bodies are often thought to cause neurologic dysfunction, numerous new discoveries over the past several decades have rendered this view increasingly unlikely. Instead, protein aggregates represent accumulations of toxic proteins that become less harmful when they are sequestered into plaques, tangles, and inclusion bodies.

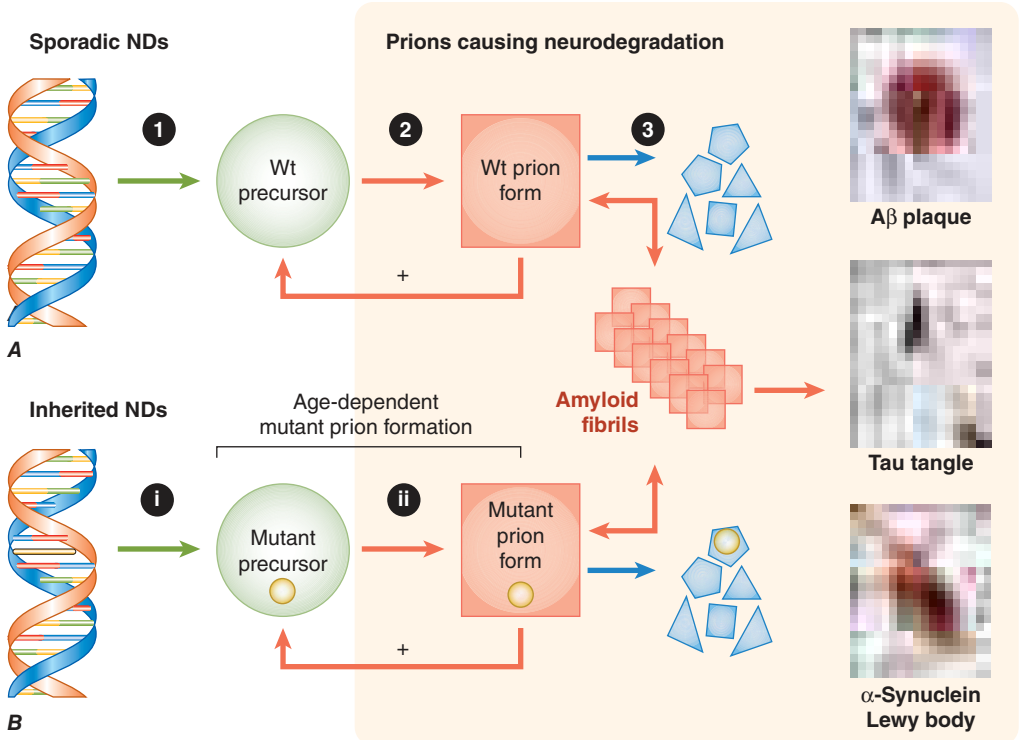


FIGURE 417-3 Neurodegeneration caused by prions. *A.* In sporadic neurodegenerative diseases (NDs), wild-type (Wt) prions multiply through self-propagating cycles of posttranslational modification, during which the precursor protein (*green circle*) is converted into the prion form (*red square*), which generally is high in β-sheet content. Pathogenic prions are most toxic as oligomers and less toxic after polymerization into amyloid fibrils. The small polygons (*blue*) represent proteolytic cleavage products of the prion. Depending on the protein, the fibrils coalesce into Aβ amyloid plaques in AD, neurofibrillary tangles in AD and Pick's disease, or Lewy bodies in PD and Lewy body dementia. Drug targets for the development of therapeutics include: (1) lowering the precursor protein, (2) inhibiting prion formation, and (3) enhancing prion clearance. ***B.*** Late-onset heritable neurodegeneration argues for two discrete events: The (*i*) first event is the synthesis of mutant precursor protein (*green circle*), and the (*ii*) second event is the age-dependent formation of mutant prions (*red square*). The *highlighted yellow bar* in the DNA structure represents mutation of a base pair within an exon, and the *small yellow circles* signify the corresponding mutant amino acid substitution. *Green arrows* represent a normal process; *red arrows*, a pathogenic process; and *blue arrows*, a process that is known to occur but unknown whether it is normal or pathogenic. *(Micrographs prepared by Stephen J. DeArmond. Reprinted with permission from SB Prusiner: Biology and genetics of prions causing neurodegeneration. Annu Rev Genet 47:601, 2013.)*

Deposition of β-amyloid is strongly implicated in the pathogenesis of AD. Genetic mutations in familial AD cause increased production of β-amyloid with 42 amino acids, which has an increased propensity to aggregate, as compared to β-amyloid with 40 amino acids. Furthermore, mutations in the amyloid precursor protein (APP) which reduce the production of β-amyloid protect against the development of AD and are associated with preserved cognition in the elderly. Mutations in genes encoding MAPT lead to altered splicing of tau and the production of neurofibrillary tangles in frontotemporal dementia and progressive supranuclear palsy. Familial PD is associated with mutations in *leucine-rich repeat kinase 2 (LRRK2), α-synuclein, parkin, PINK1,* and *DJ-1*. PINK1 is a mitochondrial kinase (see below), and DJ-1 is a protein involved in protection from oxidative stress. Parkin, which causes autosomal recessive early-onset PD, is a ubiquitin ligase. The characteristic histopathologic feature of PD is the Lewy body, an eosinophilic cytoplasmic inclusion that contains both neurofilaments and α-synuclein. Huntington's disease (HD) and cerebellar degenerations are associated with expansions of polyglutamine repeats in proteins, which aggregate to produce neuronal intranuclear inclusions. Familial ALS is associated with superoxide dismutase mutations and cytoplasmic inclusions containing superoxide dismutase. An important finding was the discovery that ubiquinated inclusions observed in most cases of ALS and the most common form of frontotemporal dementia are composed of TAR DNA binding protein 43 (TDP-43). Subsequently, mutations in the TDP-43 gene, and in the fused in sarcoma gene (FUS), were found in familial ALS. Both of these proteins are involved in transcription regulation as well as RNA metabolism. In autosomal dominant neurohypophyseal diabetes insipidus, mutations in vasopressin result in abnormal protein processing, accumulation in the endoplasmic reticulum, and cell death.

Another key mechanism linked to cell death is mitochondrial dynamics, which refers to the processes involved in movement of mitochondria, as well as in mitochondrial fission and fusion, which play a critical role in mitochondrial turnover and in replenishment of damaged mitochondria. Mitochondrial dysfunction is strongly linked to the pathogenesis of a number of neurodegenerative diseases such as Friedreich's ataxia, which is caused by mutations in an iron-binding protein that plays an important role in transferring iron to iron-sulfur clusters in aconitase and complex I and II of the electron transport chain. Mitochondrial fission is dependent on the dynamin-related proteins (Drp1), which bind to its receptor Fis, whereas mitofuscins 1 and 2 (MF 1/2) and optic atrophy protein 1 (OPA1) are responsible for fusion of the outer and inner mitochondrial membrane, respectively. Mutations in MFN2 cause Charcot-Marie-Tooth neuropathy type 2A, and mutations in OPA1 cause autosomal dominant optic atrophy. Both β-amyloid and mutant huntingtin protein induce mitochondrial fragmentation and neuronal cell death associated with increased activity of Drp1. In addition, mutations in genes causing autosomal recessive PD, *parkin* and *PINK1*, cause abnormal mitochondrial morphology and result in impairment of the ability of the cell to remove damaged mitochondria by autophagy.

One major scientific question is whether protein aggregates directly contribute to neuronal death or whether they are merely secondary bystanders. A current focus in all the neurodegenerative diseases is on small protein aggregates termed *oligomers*. These may be the toxic species of β-amyloid, α-synuclein, and proteins with expanded polyglutamines such as are associated with HD. Protein aggregates are usually ubiquinated, which targets them for degradation by the 26S component of the proteasome. An inability to degrade protein aggregates could lead to cellular dysfunction, impaired axonal transport, and cell death by apoptotic mechanisms.

Autophagy is the degradation of cystolic components in lysosomes. There is increasing evidence that autophagy plays an important role in degradation of protein aggregates in the neurodegenerative diseases,

and it is impaired in AD, PD, and HD. Autophagy is particularly important to the health of neurons, and failure of autophagy contributes to cell death. In HD, a failure of cargo recognition occurs, contributing to protein aggregates and cell death. Rapamycin, which induces autophagy, exerts beneficial therapeutic effects in transgenic mouse models of AD, PD, and HD.

There is other evidence for lysosomal dysfunction and impaired autophagy in PD. Mutations in glucocerebrosidase are associated with 5% of all PD cases as well as 8–9% of patients with dementia with Lewy bodies. Therefore, this is the most important genetic cause of both disorders thus far identified. There appear to be reciprocal interactions between glucocerebrosidase and α-synuclein. It has been shown that glucocerebrosidase concentrations and enzymatic activity are reduced in the substantia nigra of sporadic PD patients. Furthermore, α-synuclein is degraded by chaperone-mediated and macro autophagy. The degradation of α-synuclein has been shown to be impaired in transgenic mice deficient in glucocerebrosidase as well as in mice in which the enzyme has been inhibited. Finally, it is known that α-synuclein inhibits the activity of glucocerebrosidase. Therefore, there is bidirectional feedback between α-synuclein and glucocerebrosidase. An attractive therapeutic intervention could be to use protein chaperones to increase the activity and duration of action of glucocerebrosidase. This would also reduce α-synuclein levels and block the degeneration of dopaminergic neurons.

The retromer complex is a conserved membrane-associated protein complex that functions in the endosome-to-Golgi complex. The retromer complex contains a cargo selective complex consisting of VPS35, VPS26, and VPS29, along with a sorting nexin dimer. Recently, mutations in VPS35 were shown to be a cause of late-onset autosomal dominant PD. The retromer also traffics APP away from endosomes, where it is cleaved to generate β-amyloid. Deficiencies of VPS35 and VPS26 were also identified in hippocampal brain tissue from AD. A new therapeutic approach to these diseases might therefore be to use chaperones to stabilize the retromer and reduce the generation of β-amyloid and α-synuclein.

The LRRK2 mutations were shown to have effects on clearance of Golgi-derived vesicles through the autophagy-lysosome system both in vitro and in vivo. LRRK2 mutations also are linked to elevated protein synthesis mediated by ribosomal protein s15 phosphorylation. Blocking this phosphorylation reduces LRRK2-mediated neurite loss and cell death in human dopamine and cortical neurons.

Interestingly, in experimental models of HD and cerebellar degeneration, protein aggregates are not well correlated with neuronal death and may be protective. A substantial body of evidence suggests that the mutant proteins with polyglutamine expansions in these diseases bind to transcription factors and that this contributes to disease pathogenesis. In HD, there is dysfunction of the transcriptional co-regulator, PGC-1α, a key regulator of mitochondrial biogenesis. There is evidence that impaired function of PGC-1α is also important in both PD and AD, making it an attractive target for treatments. Agents that upregulate gene transcription are neuroprotective in animal models of these diseases. A number of compounds have been developed to block β-amyloid production and/or aggregation, and these agents are being studied in early clinical trials in humans. Another approach under investigation is immunotherapy with antibodies that bind β-amyloid, tau, or α-synuclein. These studies have shown efficacy in preventing the spread of amyloid, tau, and α-synuclein in animal studies, raising hopes that this could lead to effective therapies by blocking neuron-to-neuron propagation. Two large clinical trials of β-amyloid immunotherapy, however, did not show efficacy, although this therapeutic strategy is still being studied.

PRIONS AND NEURODEGENERATIVE DISEASES

As we have learned more about the etiology and pathogenesis of the neurodegenerative diseases, it has become clear that the histologic abnormalities that were once curiosities, in fact, are likely to reflect the etiologies. For example, the amyloid plaques in kuru and Creutzfeldt-Jakob disease (CJD) are filled with the PrP^{Sc} prions that have assembled into fibrils. The past three decades have witnessed an explosion of new knowledge about prions. For many years, kuru, CJD, and scrapie of sheep were thought to be caused by slow-acting viruses, but a large body of experimental evidence argues that the infectious pathogens causing these diseases are devoid of nucleic acid. Such pathogens are called prions, which are composed of host-encoded proteins that adopt alternative conformations that undergo self-propagation **(Chap. 430)**. Prions impose their conformations on the normal, precursor proteins, which in turn become self-templating resulting in faithful copies; most prions are enriched for β-sheet and can assemble into amyloid fibrils.

Similar to the plaques in kuru and CJD that are composed of PrP prions, the amyloid plaques in AD are filled with Aβ prions that have polymerized into fibrils. This relationship between the neuropathologic findings and the etiologic prion was strengthened by the genetic linkage between familial CJD and mutations in the PrP gene, as well as (as noted above) between familial AD and mutations in the *APP* gene. Moreover, a mutation in the *APP* gene that prevents Aβ peptide formation was correlated with a decreased incidence of AD in Iceland.

The heritable neurodegenerative diseases offer an important insight into the pathogenesis of the more common, sporadic ones. Although the mutant proteins that cause these disorders are expressed in the brains of people early in life, the diseases do not occur for many decades. Many explanations for the late onset of familial neurodegenerative diseases have been offered, but none are supported by substantial experimental evidence. The late onset might be due to a second event in which a mutant protein, after its conversion into a prion, begins to accumulate at some rather advanced age. Such a formulation is also consistent with data showing that the protein quality-control mechanisms diminish in efficiency with age. Thus, the prion forms of both wild-type and mutant proteins are likely to be efficiently degraded in younger people but are less well handled in older individuals. This explanation is consistent with the view that neurodegenerative diseases are disorders of the aging nervous system.

A new classification for neurodegenerative diseases can be proposed based on not only the traditional phenotypic presentation and neuropathology, but also the prion etiology **(Table 417-1)**. Over the past decade, an expanding body of experimental data has accumulated implicating prions in each of these illnesses. In addition to kuru and CJD, Gerstmann-Sträussler-Scheinker disease (GSS) and fatal insomnia in humans are caused by PrP^{Sc} prions. In animals, PrP^{Sc} prions cause scrapie of sheep and goats, bovine spongiform encephalopathy (BSE), chronic wasting disease (CWD) of deer and elk, feline spongiform encephalopathy, and transmissible mink encephalopathy (TME). Similar to PrP, Aβ, tau,

TABLE 417-1 A Prion-Based Classification of Neurodegenerative Diseases

NEURODEGENERATIVE DISEASES	CAUSATIVE PRION PROTEINS
Creutzfeldt-Jakob disease (CJD) Kuru Gerstmann-Sträussler-Scheinker (GSS) Fatal insomnia Bovine spongiform encephalopathy (BSE) Scrapie Chronic wasting disease (CWD) Feline spongiform encephalopathy Transmissible mink encephalopathy	PrP^{Sc}
Alzheimer's disease (AD)	Aβ → tau
Parkinson's disease Multiple system atrophy	α-Synuclein
Frontotemporal dementias (FTDs) Posttraumatic FTD, called chronic traumatic encephalopathy	Tau, TDP43, FUS (C9orf72, progranulin)
Amyotrophic lateral sclerosis	SOD1, TDP43, FUS (C9orf72)
Huntington's disease	Huntingtin

α-synuclein, superoxide dismutase 1 (SOD1), and possibly huntingtin all adopt alternative conformations that become self-propagating, and thus, each protein can become a prion and be transferred to synaptically connected neurons. Moreover, each of these prions causes a distinct constellation of neurodegenerative diseases.

Evidence for a prion etiology of AD comes from a series of transmission experiments initially performed in marmosets and more recently in transgenic (Tg) mice inoculated with a synthetic Aβ peptide folded into a prion. Studies with the tau protein have shown that it not only features in the pathogenesis of AD, but also causes such illnesses as the frontotemporal dementias including chronic traumatic encephalopathy, which has been reported in both contact sport athletes and military personnel who have suffered traumatic brain injuries. A series of incisive studies using cultured cells and Tg mice has demonstrated that tau can become a prion and multiply in the brain. In contrast to the Aβ and tau prions, a strain of α-synuclein prions found in the brains of patients who died of multiple system atrophy (MSA) killed the Tg mouse host ~90 days after intracerebral inoculation, whereas mutant α-synuclein (A53T) prions formed spontaneously in Tg mouse brains killed recipient mice in ~200 days.

For many years, the most frequently cited argument against prions was the existence of strains that produced distinct clinical presentations and different patterns of neuropathologic lesions. Some investigators argued that the biologic information carried in different prion strains could only be encoded within a nucleic acid. Subsequently, many studies demonstrated that strain-specified variation is enciphered in the conformation of PrP^{Sc}, but the molecular mechanisms responsible for the storage of biologic information remains enigmatic. The neuroanatomical patterns of prion deposition have been shown to be dependent on the particular strain of prion. Convincing evidence in support of this proposition has been accumulated for PrP, Aβ, tau, and α-synuclein prions.

Although the number of prions identified in mammals and in fungi continues to expand, the existence of prions in other phylogeny remains undetermined. Some mammalian prions perform vital functions and do not cause disease; such nonpathogenic prions include the cytoplasmic polyadenylation element binding (CPEB) protein, the mitochondrial antiviral-signaling (MAVS) protein, and T cell–restricted intracellular antigen 1 (TIA-1).

All mammalian prion proteins adopt a β-sheet-rich conformation and appear to readily oligomerize as this process becomes self-propagating. Control of the self-propagating state of benign mammalian prions is not well understood but is critical for the well-being of the host. In contrast, pathogenic mammalian prions appear to multiply exponentially, but the mechanisms by which they cause disease are poorly defined. We do not know if prions multiply as monomers or as oligomers; notably, the ionizing radiation target size of PrP^{Sc} prions seems to suggest it is a trimer. The oligomeric states of pathogenic mammalian prions are thought to be the toxic forms, and assembly into larger polymers, such as amyloid fibrils, seems to be a mechanism for minimizing toxicity.

To date, there is no medication that halts or even slows a human neurodegenerative disease. The development of drugs designed to inhibit the conversion of the normal precursor proteins into prions or to enhance the degradation of prions focuses on the initial step in prion accumulation. Although a dozen drugs that cross the blood-brain barrier have been identified that prolong the lives of mice infected with scrapie prions, none have been identified that extend the lives of Tg mice that replicate human CJD prions. Despite doubling or tripling the length of incubation times in mice inoculated with scrapie prions, all of the mice eventually succumb to illness. Because all of the treated mice develop neurologic dysfunction at the same time, the mutation rate as judged by drug resistance is likely to approach 100%, which is much higher than mutation rates recorded for bacteria and viruses. Mutations in prions seem likely to represent conformational variants that are selected for in mammals where survival becomes limited by the fastest-replicating prions. The results of these studies make it likely that cocktails of drugs that attack a variety of prion conformers will be required for the development of effective therapeutics.

NEURAL STEM CELL BIOLOGY

Normal and genetically modified ("transgenic") mice are the most widely used model systems to study features of human nervous system diseases. However, modeling genetic diseases in rodents is limited to the relatively small number of monogenic human diseases where the specific gene mutations are known, and is further limited by species differences. The latter can be particularly important in brain regions such as the cerebral cortex that have undergone significant evolutionary expansion in humans. These shortcomings, that likely contribute to the low probability that therapeutic efficacy translates from animal models to humans, can potentially be overcome through stem cell models that enable the use of human cells and tissues to model human diseases. The advent of new stem cell technologies is transforming our understanding of the pathobiology of human neurologic diseases. Stem cell platforms are being used to screen for therapeutic agents, to uncover adverse drug effects, and to discover novel therapeutic targets.

Among the most exciting recent advances in stem cell technology is the ability to convert somatic cells, either skin fibroblasts or blood cells, into pluripotent stem cells known as induced pluripotent stem cells (iPSCs). This technology has introduced an entirely new and powerful approach to study the pathobiology of heritable diseases. Pluripotent stem cells can be easily obtained through minimally invasive procedures such as a skin biopsy or blood sample, and converted to pluripotency through application of a cocktail of reprogramming factors to create iPSCs. Initially, a set of four programming factors, Oct3/4, Klf-4, Sox2, and c-Myc, were delivered to cells using lentiviruses which stably integrated the reprogramming factor genes into the iPSC genome, potentially altering disease phenotypes and also abrogating expression of native genes at the DNA sites where the factors integrated. Newer techniques have been developed that use non-integrating approaches such as through the use of Sendai virus, messenger RNA (mRNA), or episomal vectors that circumvent these problems. Once created, iPSC lines can be expanded indefinitely to produce a limitless supply of stem cells. These cells are the starting material for the derivation of specific cell types based on protocols that use small molecules, proteins, or direct gene induction to recapitulate developmental programs. Most current protocols derive neuronal progenitors through "dual-SMAD inhibition," a step that involves the use of small molecule inhibitors to block endoderm and mesodermal cell fates, thereby creating neural cells by default. Multiple protocols have been developed over the last decade for creating large numbers of human neuron progenitor cell types and directing them toward specific nervous system cell fates, including neuron subtypes from multiple regions of brain and spinal cord as well as retinal cells, glial cells including astrocytes and oligodendrocytes, immune cells, and PNS cells.

The primary medical benefit of iPSC technology is that it enables the creation of patient-specific cells or tissues that are genetically matched to individual patients. This approach not only enables the study of monogenetic disorders, but also sporadic forms of disease, and complex polygenic disorders including those with unidentified risk loci. Furthermore, by deriving iPSC cell lines from multiple patients it would be possible to explore how disease phenotypes may vary according to genetic background. Another approach that has been used to generate specific neuron and glial cell types from somatic cells such as fibroblasts is through direct reprogramming. This approach relies on a cocktail of specific transcription factors to directly convert somatic cells into the alternate desired cell type. This approach bypasses the epigenetic reset that accompanies cells as they are reprogrammed to a pluripotent state. The advantage of this approach is that age-related epigenetic signatures are not erased, so that derived neurons may more readily reflect diseases that manifest in older cells.

Despite the advantages of using *in vitro* models of nervous system diseases derived from patient-specific iPSCs, several potential roadblocks remain. There are no standard reprogramming or derivation protocols, and the different methods can result in considerable variability in the disease phenotypes reported by different laboratories. Confidence in the specificity of a particular phenotype is therefore increased if it has been validated across multiple laboratories. There is also the problem of inherent variability between patient lines that may

result from their different genetic backgrounds. One solution, available only in the case of monogenic disorders, is to use isogenic controls generated using gene editing, such as with CRISPR-Cas9 technology, to create disease and control lines on an identical genetic background. However, because differences in genetic background can influence the penetrance of a particular trait, it will still be necessary to compare disease lines from multiple patients to discern a true disease phenotype. For polygenic disorders where the causative mutations are unknown it will not be possible to create isogenic controls, and in these situations the best strategy for improving reliability and sensitivity is to compare lines from multiple patients.

Organoids Most nervous system disorders, including autism spectrum disorder, schizophrenia, PD, AD, and ALS are complex disorders, resulting from an unknown combination of gene mutations and manifest not only in specific cell types, but also in alterations of the local tissue environment. These disorders are difficult to model in animals, but they are approachable using three-dimensional human iPSC stem cell models, often referred to as "organoids." Organoids are derived from pluripotent stem cells that are directed along a tissue-specific lineage through the timed application of growth factors, genes, or small molecule activators or inhibitors, and allowed to aggregate into three-dimensional structures. With time, cell intrinsic programs are spontaneously engaged and the cellular aggregates begin to self-organize and develop into structures that recapitulate the complex topographical and cellular diversity of normal organ development. In this way it has been possible to create, at least in part, in vitro brain-like organoids that resemble the human forebrain at early stages of development. These structures, when allowed to develop from an anterior neural tube stage, can become heterogeneous containing regions with forebrain, midbrain, and/or hindbrain identity and can often include retina-like structures. The high degree of variability in such "cerebral organoids" can be a liability for controlled studies, and can be reduced by the use of more directed protocols that restrict outcomes to more defined brain regions, such as forebrain, cortex, or ganglionic eminence. A variety of protocols have now been developed to generate organoids with specific regional identity, and fusing organoids of different regional identity with each other has been used to reproduce cellular interactions such as neuronal migration across regions. Many protocols are focused on modeling cortical development, and they can reproduce developmental features including a diversity of progenitor and neuronal cell types topographically distributed within ventricular and subventricular progenitor regions and rudimentary cortical layers. However, the organoids follow a human developmental timetable and still remain at stages roughly comparable to late fetal development after 6–9 months. Moreover they lack key cell types such as endothelial cells, pericytes, microglia, and have few if any astrocytes or oligodendrocytes. Nonetheless, while still only reflecting rudimentary organizational and compositional features, organoids have become attractive models to study human brain development and the pathophysiology of human nervous system diseases in the context of an organized brain-like structure.

Brain Development and Developmental Disorders: Microcephaly and Lissencephaly Transcriptional analysis has suggested that the neurons produced by most stem cell protocols resemble early to mid-gestational stages of human brain development. The immaturity of stem cell-derived human neurons may limit their utility for modeling adult diseases, but makes them ideally suited for the study of brain development and the pathophysiology of neurodevelopmental disorders.

Primary autosomal recessive microcephaly (MCPH) is a rare neurodevelopmental disorder producing severe microcephaly with simplified cortical gyration and intellectual disability. MCPH was one of the first disorders to be studied using cerebral organoids. Mutations in genes encoding microtubule spindle components and spindle-associated proteins are the most frequent causes of congenital microcephaly. Among them is cyclin-dependent kinase 5 related activator protein 2 (CDK5RAP2). Skin fibroblasts derived from a single microcephalic patient carrying a mutation in CDK5RAP2 were used to generate four iPSC lines. Cerebral organoids grown from these cell lines contained fewer proliferating progenitor cells and showed premature neural differentiation compared to wild type controls. Introducing functional CDK5RAP2 by electroporation partially rescued the disease phenotype, supporting the notion that failure of the founder population of neural progenitors to properly expand underlies the smaller brain. This study demonstrated that brain organoids derived from patients with microcephaly can be used to reproduce features of the disease, but did not reveal new insights or disease features of CDK5RAP2 microcephaly that had not already been described in mouse models.

In a study using cortical organoids to model Miller-Dieker syndrome (MDS), a severe congenital form of lissencephaly or "smooth-brain," features of the human disease were observed that had not been noted in murine models. Classical lissencephaly is a genetic neurological disorder associated with mental retardation and intractable epilepsy, and MDS is a severe form of the disorder. Cortical folding in humans begins toward the end of the second trimester, a stage of development that has not yet been modeled in organoids, but gyrencephaly depends upon earlier events such as neural progenitor cell proliferation and neuronal migration that can be modeled in organoids. The human organoid model of MDS exhibited several neural progenitor cell phenotypes that had already been reported in mouse models, including altered mitotic spindle orientation and neuronal migration defects. But the organoids also displayed a mitotic defect in a specific neural stem cell subtype, the outer radial glia cell (oRG) that had not been observed in mice. oRG cells are enriched in the outer subventricular zone, a proliferative region that is large in primates and not present in rodents. These cells are particularly numerous in the developing human cortex and are thought to underlie the developmental and evolutionary expansion of the human cortex. oRG cells from MDS patients behaved abnormally and had arrested or delayed mitoses. MDS organoids also identified non-cell autonomous defects in WNT signaling as an underlying mechanism. These insights into mechanistic and cell type specific features of human disease highlight how organoid technology can provide new and valuable perspectives on the pathophysiology of disorders of in utero development.

Acquired Neurodevelopmental Disorders: Zika The recent outbreak of Zika virus (ZIKV) and associated microcephaly cases in the Americas provided a test case for the utility of brain organoids to model acquired human microcephaly. Despite a correlation between Zika infection rates and the incidence of congenital microcephaly, compelling evidence that ZIKV caused microcephaly was lacking in the early phases of the epidemic. The causal link between ZIKV and congenital microcephaly was buttressed by two studies in 2016 that used human iPSC-derived neural progenitor cells and organoids to demonstrate ZIKV tropism for human neural progenitor cells. Neural progenitor cells (radial glia) were readily infected in vitro with subsequent progenitor cell death and involution of organoid size. Forebrain organoids were further used to highlight the role of the flavivirus entry factor, AXL, in determining viral tropism, and were also used to explore the disease mechanism by demonstrating upregulation of the innate immune receptor toll-like receptor 3 (TLR) in response to ZIKV infection. Stem cell-derived models of human brain development have also demonstrated centrosomal abnormalities in radial glia and alteration in the cleavage plane of mitotic radial glia associated with premature neural differentiation. Mouse models are also being used to study the pathophysiology of congenital ZIKV syndrome, but the availability of unlimited numbers of human neural cells produced using stem cell technology has enabled high-throughput screening assays to test libraries of clinically approved compounds for potential therapeutic agents. This strategy has already highlighted several compounds that could potentially help protect against ZIKV microcephaly.

Neurodevelopmental Disorders: Autism and Schizophrenia Autism spectrum disorders (ASD) are complex and heterogeneous neurodevelopmental disorders usually manifesting in childhood with difficulties in social interaction, verbal and nonverbal communication and repetitive behaviors. The cellular and molecular mechanisms underlying ASD are thought to arise at stages of fetal brain

development, making them well-suited for exploration using human iPSC-derived disease models. The pathophysiology of disorders associated with ASD that are caused by monogenic mutations have been studied using iPSC-derived neurons; these include Fragile X, Rett, and Timothy syndromes.

Fragile X is the most common heritable cause of intellectual disability, affecting 1 in 4000 males and 1 in 8000 females, and is a leading genetic cause of ASD. Patients also have speech delay, growth and motor abnormalities, hyperactivity, and anxiety. The causative mutation lies in the *FMR1* gene and produces a CGG triplet repeat expansion from a normal number of 5–20 to >200, leading to epigenetic silencing of the *FMR1* gene and loss of the Fragile X mental retardation protein. The epigenetic mechanism means that unlike a simple gene deletion that would lead to ubiquitous loss of expression, the FMR1 locus becomes hypermethylated and epigenetically silenced during differentiation, thus FMR1 protein is expressed by the early embryo and becomes absent only around the beginning of the second trimester. Interestingly, this expression pattern is recapitulated during cellular differentiation in stem cell models. Pluripotent Fragile X stem cell lines have been derived from embryos identified through pre-implantation genetic diagnosis and by reprogramming skin fibroblasts from Fragile X patients to create iPSC lines. In both cases, *FMR1* was expressed by the pluripotent stem cells, but underwent transcriptional silencing following differentiation. Fragile X stem cell lines can therefore be used to study the mechanism of FMR1 silencing, an effort that is ongoing. Neurons generated from Fragile X iPSC cells reproduce features observed in neurons from transgenic *FMR1* mouse models and patients, including stunted neurites with decreased branching, increasing confidence in the iPSC model. In addition to providing a model that can be used to study disease pathogenesis, Fragile X iPSC-derived neurons could be used to screen for potential therapeutic agents or gene editing strategies that could be able to remove the repressive epigenetic marks induced by the mutation and rescue the phenotype.

Rett syndrome is an X-linked neurodevelopmental disorder with dominant inheritance caused by a mutation in the MECP2 gene. Because males carrying one copy of the defect gene usually die in infancy, most patients are girls. Random inactivation of the X chromosome in girls results in mosaic cellular expression of the mutation that circumvents fatality and produces a variable phenotype. The symptoms are present in early childhood and include microcephaly associated with developmental delay, autistic-like behaviors and cognitive dysfunction, seizures, and repetitive motor actions; these then progress to include difficulties with gait, swallowing, and breathing before usually stabilizing with patients surviving to adulthood. The pathophysiology of RETT syndrome is presumed to involve abnormal epigenetic regulation leading to decreased transcriptional repression of genes whose overexpression produces the disease phenotype, although this concept has been contested. In one of the first studies to use iPSC modeling to study RETT syndrome, it was discovered that when fibroblasts from patients were reprogrammed to pluripotent stem cells, X inactivation was erased. In apparent recapitulation of endogenous events, X chromosome inactivation re-occurred during neuronal differentiation, producing a mosaic of cells carrying the mutant gene intermingled with normal cells. RETT neurons had fewer dendritic spines and synapses, smaller cell bodies, and reduced network activity. Another iPSC model of RETT syndrome highlighted the potential role of altered inhibitory function. RETT neurons were found to have a deficit of a potassium/chloride cotransporter (KCC2) that is developmentally regulated and normally leads to a switch in GABA signaling from excitatory at embryonic ages to inhibitory by birth. In RETT neurons KCC2 expression level was low, and the functional switch in GABA effects was delayed, contributing to some of the disease features and possibly accounting for the developmental onset of the disease. One curious feature of some iPSC RETT lines was that despite the mosaic expression of the mutation, disease phenotypes were observed in all cells. Possibly, this could reflect a non-cell autonomous effect, but as in all iPSC disease models, confidence in disease-specific features will be increased when similar phenotypes are seen across multiple independent studies.

Timothy syndrome, another severe neurodevelopmental disease associated with ASD has been modeled using iPSC-derived organoids. Timothy syndrome is caused by a mutation in the *CACNA1C* gene coding for a voltage-gated calcium channel, and neuron defects in Timothy syndrome organoids were rescued by selectively altering calcium channel activity. In one study two separate organoids were produced with different regional identity, one represented neocortex and one a more ventral structure known as the medial ganglionic eminence, which is the source of most cortical interneurons. The two organoids were then fused together to allow the interneurons to migrate into the cortex, mimicking their endogenous behavior. The ability to model interneuron migration led to the discovery of a cell-autonomous migration defect in the disease-carrying neurons.

The majority of nervous system diseases, including ASD, are multigenic and cannot be modeled in animals but can be modeled using patient-derived iPSCs. For example, a subset of patients with ASD have large head size, and a cohort of patients with this phenotype were used to generate iPSCs which were converted to neural progenitor cells and forebrain neurons. The progenitors had an accelerated cell cycle and produced an excess of inhibitory interneurons and had exuberant cellular overgrowth of neurites and synapses. This last feature is in contrast to the decrease in spines and synapses observed in other iPSC models of ASD such as Fragile X and Rett syndrome and underscores the need for replication and validation of purported disease phenotypes given the high variability based on differences between stem cell lines, protocols, patient genetic background, and other factors. Moreover, the clinical features of most neuropsychiatric diseases reflect disorders in processes such as circuit formation and refinement that occur after birth and may be difficult to capture at the fetal stage of development reflected in stem cell models.

Patient stem cells have also been used by multiple groups to study the pathophysiology of schizophrenia, producing a variety of diverse and sometimes contradictory results. Reports claim obvious phenotypes such as disruptions in the adherens junctions of forebrain radial glia or aberrant neuronal migration, although such gross abnormalities observed at the equivalent of in utero stages of development seem very unlikely to underlie a disease that usually manifests at adolescence or young adulthood. Other studies report abnormalities related to abnormal microRNA expression, disordered cyclic AMP and Wnt signaling, abnormal stress responses, diminished neuronal connectivity, fewer neuronal processes, problems with neuronal differentiation, and mitochondrial abnormalities among others. While the pathophysiology of as complex a neurodevelopmental disorder as schizophrenia may be multidimensional, it is unclear which, if any, of the reported findings in iPSC models reflect the true pathology of schizophrenia. Progress will likely depend on the adoption of more standard and reproducible protocols, more rigorous identification of cell types, markers of regional identity, and indicators of maturity.

Alzheimer's Disease As noted above, the leading concept of AD pathogenesis, the amyloid hypothesis, suggests that an imbalance between production and clearance of β-amyloid leads to excessive accumulation of β-amyloid peptide and the formation of neurofibrillary tangles within neurons, composed of aggregated hyperphosphorylated tau proteins. Additionally, aggregates of amyloid fibrils are deposited outside neurons in the form of neuritic plaques. Recent failures of anti-β-amyloid therapies, which were highly effective in mouse models, have led to a search for alternative models that might be more predictive of therapeutic effectiveness in humans. Among the causes of familial AD are mutations in genes involved in β-amyloid production, including APP and presenilin 1 and 2. Shortly after the introduction of iPSC technology, human stem cell-derived neurons were generated from patients carrying mutations in AD causative genes as well as from sporadic AD cases. The disease neurons developed hallmarks of AD including intracellular accumulation of β-amyloid and phosphorylated tau, as well as secretion of APP cleavage products, features that could be reduced by adding β- or γ-secretase inhibitors or β-amyloid specific antibodies. The neurons also demonstrated other disease features observed in postmortem AD tissues. However, extracellular β-amyloid

aggregation and neurofibrillary tangles were not robustly modeled in these two-dimensional systems, presumably because secreted factors were able to readily diffuse away. The use of three-dimensional organoids to model AD overcame this limitation, presumably by recreating a more faithful extracellular matrix. Organoid models promoted the aggregation of β-amyloid, and more readily recapitulated the pathologic features of AD, including the formation of neurofibrillary tangles and neuritic plaques.

It is hoped that the new stem cell models, particularly organoid models, will accelerate our understanding of AD by enabling the study of human disease-carrying cells in a quasi in situ setting. These new models may lead to discovery of novel druggable targets and new diagnostic and prognostic biomarkers. One concern is that the pathogenic features of AD usually appear in the sixth or seventh decade of life and progress slowly over years, while most protocols for the derivation of human cortical neurons generate cells over weeks or months and most remain comparable to immature neurons at fetal stages of development. Nonetheless, these young cells have been used to model neurodegenerative diseases such as AD and HD that strike patients in mid to late adulthood. Possibly the onset of disease phenotype is accelerated in stem cell models due to increased cellular stress, or disease features may actually have a subtle onset at earlier stages than generally suspected. Indeed, 3-year-old children at genetic risk of developing early-onset AD appear to have smaller hippocampal size and lower scores on memory tests than children in a non-risk group. The phenotypes of adult neurodegenerative diseases that are visible at fetal stages may or may not correspond to those manifest at later, adult stages, but they may offer the possibility of devising preventative strategies effective at very early stages of the disease.

Cell Type Disorders: ALS and HD In diseases such as ALS, PD, and HD, that mostly target specific neuron subtypes, stem cells provide an ideal means to study the vulnerable human cell populations. By enabling the production of unlimited numbers of normal and diseased human midbrain dopaminergic neurons for the study of PD, medium spiny striatal neurons for HD, and spinal and cortical motor neurons for ALS, iPSC approaches have the potential to transform our understanding and management of these diseases. Stem cell-derived neurons serve as platforms to explore mechanisms of cell vulnerability, to screen drugs for neural protection, and potentially to derive neurons for replacement therapy.

Amyotrophic Lateral Sclerosis One of the first protocols for producing neurons of a specific subtype from embryonic stem cells recapitulated normal developmental programs to generate mouse spinal motor neurons. Pluripotent mouse stem cells underwent neural induction and adopted a caudal identity through the application of retinoic acid, and subsequently adopted motor neuron fate through the action of sonic hedgehog, a ventralizing factor. Generating human motor neurons proved more complex, requiring additional steps, such as early exposure to the growth factor, FGF2. The first application of stem cell-derived motor neurons to study ALS involved the use of mouse motor neurons generated from transgenic mice expressing a mutation in the SOD1 gene, the most common mutation responsible for familial ALS. Only 5–10% of ALS cases are familial, but the known mutations provide a useful entry point to tease apart the causative pathophysiology. Mutations in SOD1 produce ALS through a toxic gain of function for which the mechanism remains unclear, despite the use of multiple transgenic animal and iPSC models. The use of mouse ESC-derived motor neurons, however, demonstrated that toxic factors secreted by SOD1 astrocytes contribute to the death of motor neurons. Interestingly, stem cell-derived interneurons were spared, indicating a specific vulnerability of motor neurons. These findings helped establish the notion that a non-cell autonomous toxic mechanism contributes to ALS pathogenesis and may ultimately lead to novel treatment strategies. These findings also highlight that modeling the full pathophysiology of ALS may require the reproduction of a complex environment including motor neurons, astrocytes, and possibly additional cell types such as microglia. A variety of approaches including co-culture of specific cell types, three-dimensional spinal cord organoids, and microfluidic organ-on-chip models are being explored to achieve a more complete facscimile of spinal cord organization. Similar to other neurologic disorders where a clearly defined phenotype has been observed in human stem cell-derived models, there is hope that drug screening using human disease-expressing cells will identify a potential therapeutic compound.

Huntington's Disease HD is caused by an expansion in CAG triplet repeats in the huntingtin gene which leads to an expanded polyglutamine tract in the huntingtin protein. HD is dominantly inherited, with symptoms of cognitive decline and uncontrollable gait and limb motions beginning in the third to fifth decade of life with progression to dementia and death ~20 years later. Mutant huntingtin causes a toxic gain-of-function, with the degree of effect related to the CAG repeat length. For example, a CAG length of 40–60 repeats produces adult onset HD, while repeats of ≥60 produce juvenile onset disease. Although it has been 25 years since the discovery of this causative mutation, the disease mechanism remains poorly understood. Excess huntingtin protein and protein fragments accumulate in specific subtypes of neurons where they misfold and form aggregates that are visible as cellular inclusions. Affected cells eventually die, possibly as a result of metabolic toxicity. The medium spiny neurons of the striatum are the most vulnerable neurons, spurring ongoing attempts to produce replacement cells derived from stem cells, but neuron loss is widespread including in the cortex, complicating a cell replacement approach for this disease. HD iPSCs have been generated from patients with various CAG repeat lengths, but those from juvenile onset disease with the longest repeat lengths have been favored as being most likely to express robust disease phenotypes at an early stage. This is particularly important given the immature stage of maturation of stem cell-derived human neurons. This approach has been able to produce disease phenotypes observed in patients including huntingtin protein aggregation, decreased metabolic capacity, increased oxidative stress with mitochondrial fragmentation, and apoptosis enhanced by withdrawal of growth factor support. However, many of these phenotypes were observed in pluripotent cells prior to neural differentiation and in neural progenitors and a broad array of CNS neurons in contrast to the cell type specific features of the disease. Nonetheless, neurons that assumed striatal fate appear to be more vulnerable to stress and apoptosis than other cell types. As with other iPSC models of nervous system diseases, there have so far been few efforts to validate results in multiple iPSC lines having different genetic backgrounds but with similar CAG repeat lengths. An HD consortium has been formed to address this problem by generating a series of iPSC lines from multiple patients. An alternative strategy to validate disease phenotypes has been to use gene editing to create isogenic iPSC lines that are corrected to produce wild type control and HD iPSC lines against the same genetic background.

Future Perspectives Despite early successes, it may prove difficult to reconstitute neurodegenerative disease conditions in human cells in vitro over a short time-course because the pathogenic changes of degenerative diseases progress slowly and commence in the later stages of life. The differentiation and maturation of human neurons from stem cell lines occurs over a span of months, which may not be long enough to establish the aged brain conditions under which patients develop robust neurodegenerative pathology. Possible manipulation through gene editing or by application of aging-associated stresses, such as DNA damaging agents or proteasome inhibitors, may accelerate the expression of degenerative phenotypes in human iPSC-derived cellular models. Stem cell-derived organoid models are also ideal platforms to apply methods for cellular level visualization such as clarity and multi-electrode recording techniques to better evaluate three-dimensional organoid structures and explore early-forming circuits. These applications are only just beginning.

Two-dimensional cell cultures are ideal for production and evaluation of large numbers of specific cells of a particular identity, but may not provide the complex extracellular environment necessary to model

certain disease processes, such as extracellular protein aggregation. These features can be best modeled using three-dimensional organoids, but current methods do not reproduce all the relevant features of brain tissue. Optimization will be needed to better reproduce the cellular composition of brain, including endothelial cells, astrocytes, microglia, and oligodendrocytes. It may also be necessary to combine different brain regions generated separately, possibly by fusion of tissues such as dorsal cortex, subpallium, thalamus, retina, and others. However, currently there is a limited ability to recreate tissues or neurons with regional brain identity, such as hippocampus, thalamus, or cerebellum. More faithful organoid models could also emerge through the application of bioengineered scaffolds, matrices, or perfusion systems that might allow the growth of larger structures. Of course, not all aspects of mature brain architecture and function will be modeled by these tissue structures, particularly as they represent fetal stages of development, but perhaps the most precocious events in disease etiology can be captured and investigated and these may share mechanistic pathways with disease features that manifest at later stages.

The current excitement surrounding human stem cells has more to do with their promise to improve on animal models of disease than their potential as a source for cell based therapies. Even without new insights into disease pathogenesis, there is promise that iPSC models such as brain organoids will act as drug screening platforms for discovery of novel therapeutics and for detection of off-target and toxic effects. The failure of many neurotherapeutic approaches to translate from animal models to clinical practice underscores the need for better predictive models, and stem cell models and brain organoids based on human cells may be ideally suited to bridge this divide.

A Current Perspective on Neural Stem Cells in the Clinic The prospect of stem cell therapies to treat diseases or injuries of the nervous system has captured the attention of researchers, clinicians, and the public. The pace of research is usually slow and deliberate, but in the stem cell arena there has been enormous pressure to accelerate the pace of progress in order to bring cell-based therapies to the clinic. Expectations have been raised, and clinics have already begun offering unproven or dangerous treatments to a public that is ill-informed and vulnerable to exploitation. Nonetheless, there is cautious optimism that stem cells will eventually realize the promise of regenerative therapy for at least some currently untreatable or incurable nervous system diseases.

Pursuit of a cell-based therapy for PD has been ongoing for many decades. Following anecdotal success in a handful of patients who appeared to improve following striatal grafts of fetal midbrain dopaminergic cells, two NIH-funded double-blind control studies were launched in the 1990s. However, only a small number of younger patients showed some benefit, and several patients developed spontaneous dyskinetic movements related to the therapy. These efforts constituted a failed trial as the treated patients who did not experience side effects failed to improve significantly. The dyskinesias that curtailed the trials were eventually ascribed to an abundance of serotonergic neurons that were inadvertently included in some of the cell grafts. Protocols for deriving dopaminergic neurons from stem cells could potentially avoid this complication by providing a more purified cell population, and several groups in the United States and Europe have been aggressively pursuing a stem cell-based approach and are nearing clinical trials. Meanwhile, techniques to extract dopaminergic cells from fetal tissue have been improved, and on the basis of encouraging results in individual transplanted patients, some of whom have managed to go off their Parkinson's medication, a new trial of fetal cell transplantation for PD has started in Europe. This is a very consequential trial, as a poor clinical outcome could dampen enthusiasm for the planned follow-on stem cell trials in PD and possibly in other disorders as well.

One of the first cell-based clinical trials for a neurological disease targeted patients suffering from an untreatable childhood disorder, Batten disease. Batten disease is an autosomal recessive metabolic disorder resulting from an inability to synthesize a lysosomal enzyme critical to brain function. The Phase 1 trial involved six patients with infantile and late infantile forms of the disease who received neural stem cells rather than any specific postmitotic cell type. Neural stem cells derived from donated fetal tissue were expanded in vitro prior to surgical grafting into the brain. This approach was not without risk, as the neural stem cells were proliferating and could potentially form an abnormal growth. The rationale was that the cells would be capable of synthesizing and secreting the missing lysosomal enzyme and would therefore serve as a delivery device. Animal studies using a transgenic mouse model of Batten disease demonstrated rescue, and this promising result led to a small Phase 1 trial. The Phase 1 study was considered a success as no adverse events were reported and the cells appeared to be safe, though there was no clinical improvement and no clear evidence of whether the cells had dispersed, transformed into neurons or glia, or indeed survived at all. Despite clearing the Phase 1 trial, the company did not pursue further trials for Batten disease, but instead initiated clinical trials using the same cell product for several other indications, including an inherited fatal dysmyelination syndrome known as Pelizaeus-Merzbacher disease (PMD). The human neural stem cells have both neurogenic and gliogenic potential, and when delivered to white matter regions in experimental animals most persisting cells had become oligodendrocytes. This supported use of the cells to promote myelin formation in conditions such as PMD. The company also initiated trials in spinal cord injury. However, the spinal cord trial failed to achieve sufficient benefit in Phase II and the company ceased its work on stem cell therapies.

Spinal cord injury is an attractive target for novel therapies since there are no effective treatment options currently. A series of stem cell trials designed to treat subacute spinal cord injury are underway in the United States and Europe. The first to enter clinical trials in the United States was based on a protocol designed to generate oligodendrocytes from pluripotent embryonic stem cells. Evidence of efficacy was obtained in animal models following surgical grafting of cells to sites of spinal cord injury. However, evidence of myelination of host axons was minimal, and other mechanisms were invoked for improvement in gait, including trophic support and immune modulation. Regulatory permission for a Phase 1 trial for subacute mid-thoracic injury was initially stalled by concern over abnormal growths at sites of cell deposit in some animals, but this was satisfactorily addressed and patient trials commenced. However, following a change in leadership, the stem cell program was terminated. The program was acquired by another company that has resumed the spinal cord injury trial and received regulatory approval to advance to include cervical level injuries.

The possibility of treating ALS by replacing dying motor neurons with stem cell–derived substitutes has excited interest but this prospect seems very remote. Even if new neurons are able to integrate into spinal cord circuits and become properly innervated, they would have to grow long axons that would take many months to years to project to appropriate targets and attract myelinating Schwann cells. Furthermore, cells would need to be grafted at multiple spinal cord and brainstem levels, and the upper motor neuron deficit would need to be treated by replacing projecting neurons in the motor cortex. An additional complication is the recent finding that spinal motor neurons have unique segmental identity, and replacement cells might need to be generated with a range of molecular identities in order to integrate at multiple spinal levels. This would still leave unaddressed the toxic effects recently shown to be produced in ALS by diseased astrocytes and microglia that could attack the replacement cells. A more tractable near-term solution would be to graft support cells that could rescue or protect endogenous motor neurons from damage. This approach was tried in a mouse model of ALS. Human stem cell–derived neural progenitor cells engineered to express GDNF, a growth factor known to provide trophic support for neurons, were grafted to the spinal cord of young ALS mice. The cells dispersed and were able to rescue motor neurons, a very promising result, but disappointingly, the animals became weak and died at the same rate as untreated control animals. However, ALS is a deadly disease with no known treatment. In the hope that patients will respond differently than mice, a clinical trial based on this approach has been approved by the U.S. Food and Drug Administration (FDA) and will begin soon.

Following Shinya Yamanaka's discovery of iPSCs, the Japanese government has invested in bringing iPSC-derived cell therapy to the clinic. Banks of iPSC lines selected to capture the diversity of HLA haplotypes found in the Japanese population are being produced in the hope that these will allow cell therapies to be matched to individual patient haplotypes in order to avoid immune rejection. While these stem cell banks were still being produced, the first Japanese study to use stem cells was approved in August, 2013, and involved patients who were to receive customized therapy using cells derived from their own skin fibroblasts. The targeted disease was age-related macular degeneration, a common cause of blindness in the elderly that results from loss of retinal pigment epithelial (RPE) cells. RPE cells are relatively easy to generate from pluripotent stem cells, making replacement therapy an attractive target in this condition. A challenge is to coax the replacement cells to recreate an epithelium in the subretinal space. The Japanese approach involves surgical insertion of a biofilm seeded with RPE cells into the retina. One patient was treated with his/her own stem cell-derived RPE cells, but prior to treating a second patient, the genome of the RPE cell line was sequenced, and a mutation was discovered in a known oncogene. The trial was halted and a decision made to discontinue the effort for customized cell therapy in favor of using RPE cells derived from the national repository of banked iPSC lines which undergo extensive gene sequencing and quality controls. This outcome serves as a caution for the challenges involved in bringing a customized cell therapy to the clinic.

By far the largest number of human trials have been performed using mesenchymal stem cells (MSCs) sourced from a variety of sites including bone marrow, peripheral blood, adipose tissue, umbilical cord, etc. Interest in the potential utility of MSCs for regenerative therapy began with the optimistic report that bone marrow stem cells were pluripotent and capable of generating nerve and heart muscle as well as blood cells. The possibility that easily obtainable MSCs could be used to regenerate injured or diseased cells or organs to treat diseases ranging from stroke, neurodegenerative disease, myocardial infarct, and even diabetes, generated enormous enthusiasm. The enthusiasm proved irresistible to many, and even after the initial reports were discredited—MSCs turned out not to be pluripotent stem cells as initially thought—a veritable flood of papers began to appear claiming disease-modifying activity of MSCs in mouse models of almost every degenerative disease and injury model. But when it became clear that the MSCs were not transforming into or generating new neurons or cardiac myocytes, alternative mechanisms of action were invoked, including the release of trophic factors, cytokines, or inflammatory modulators that were credited with producing their remarkable restorative effects. The relative ease with which blood or adipose tissue can be harvested from patients or donors and MSCs extracted has led to a rapidly expanding number of clinical trials for conditions ranging from stroke and MS to AD and PD. Furthermore, a loophole in the regulatory framework of the FDA allows autologous cell therapy to escape regulation provided that the cells have not been significantly processed. This lax regulation has spawned a veritable industry of stem cell clinics making unsubstantiated claims of success in treating nervous system diseases. Patients have died from treatments in unregulated clinics operating in countries around the world and three patients became blind in a well-publicized incident following stem cell treatments delivered by a Florida clinic. The "stem cells" were derived from the patients' own fat tissue and blood. These activities represent the dark side of the stem cell revolution perpetrated by practitioners who exploit the desperation of patients and their families. Legitimate and effective stem cell therapies will emerge over time, but given the prevalence and abundance of misleading information available on the internet and elsewhere, a trusted and well-informed physician can play a key role in helping patients navigate the current cell therapy minefield.

FURTHER READING

Chung WS et al: Do glia drive synaptic and cognitive impairment in disease? Nat Neurosci 18:1539, 2015.

Di Lullo E, Kriegstein AR: The use of brain organoids to investigate neural development and disease. Nat Rev Neurosci 18:573, 2017.

Fung TC et al: Interactions between the microbiota, immune and nervous systems in health and disease. Nat Neurosci 20:145, 2017.

Herz J et al: Myeloid cells in the central nervous system. Immunity 46:943, 2017.

Hong S et al: Complement and microglia mediate early synapse loss in Alzheimer mouse models. Science 352:712, 2016.

Liddelow SA et al: Neurotoxic reactive astrocytes are induced by activated microglia. Nature 481:541, 2017.

Prusiner SB et al: Evidence for α-synuclein prions causing multiple system atrophy in humans with parkinsonism. Proc Natl Acad Sci USA 112:E5308, 2015.

Turner L, Knoepfler P: Selling stem cells in the USA: Assessing the direct-to-consumer industry. Cell Stem Cell 19:154, 2016.

Yoo BB, Masmanian SK: The Enteric Network: Interactions between the immune and nervous systems of the gut. Immunity 46:910, 2016.

Section 2 Diseases of the Central Nervous System

418 Seizures and Epilepsy

Daniel H. Lowenstein

A *seizure* (from the Latin *sacire*, "to take possession of") is a transient occurrence of signs or symptoms due to abnormal excessive or synchronous neuronal activity in the brain. Depending on the distribution of discharges, this abnormal brain activity can have various manifestations, ranging from dramatic convulsive activity to experiential phenomena not readily discernible by an observer. Although a variety of factors influence the incidence and prevalence of seizures, ~5–10% of the population will have at least one seizure, with the highest incidence occurring in early childhood and late adulthood.

The meaning of the term *seizure* needs to be carefully distinguished from that of epilepsy. *Epilepsy* describes a condition in which a person has a risk of *recurrent* seizures due to a chronic, underlying process. This definition implies that a person with a single seizure, or recurrent seizures due to correctable or avoidable circumstances, does not necessarily have epilepsy (although a single seizure associated with particular clinical or electroencephalographic features may establish the diagnosis of epilepsy). Epilepsy refers to a clinical phenomenon rather than a single disease entity, because there are many forms and causes of epilepsy. However, among the many causes of epilepsy there are various *epilepsy syndromes* in which the clinical and pathologic characteristics are distinctive and suggest a specific underlying etiology.

Using the definition of epilepsy as two or more unprovoked seizures, the incidence of epilepsy is ~0.3–0.5% in different populations throughout the world, and the prevalence of epilepsy has been estimated at 5–30 persons per 1000.

CLASSIFICATION OF SEIZURES

Determining the type of seizure that has occurred is essential for focusing the diagnostic approach on particular etiologies, selecting the appropriate therapy, and providing potentially vital information regarding prognosis. The International League Against Epilepsy (ILAE) Commission on Classification and Terminology provided an updated approach to classification of seizures in 2017 **(Table 418-1)**. This system is based on the clinical features of seizures and associated electroencephalographic findings. Other potentially distinctive features such as etiology or cellular substrate are not considered in this classification system, although this will undoubtedly change in the future as more is learned about the pathophysiologic mechanisms that underlie specific seizure types.

A fundamental principle is that seizures may be either focal or generalized. *Focal seizures* originate within networks limited to one brain region (note that the term *partial seizures* is no longer used). *Generalized*

TABLE 418-1 Classification of Seizures*

1. **Focal Onset**
   (Can be further described as having intact or impaired awareness, motor or nonmotor onset, or evolve from focal to bilateral tonic clonic)
2. **Generalized Onset**
   a. Motor
      Tonic-clonic
      Other motor (e.g., atonic, myoclonic)
   b. Nonmotor (absence)
3. **Unknown Onset**
   (Can be further described as motor or nonmotor, or unclassified)

*Based on the new 2017 International League Against Epilepsy classification of seizure types (RS Fisher et al: Epilepsia 58: 522, 2017).

seizures arise within and rapidly engage networks distributed across both cerebral hemispheres. Focal seizures are usually associated with structural abnormalities of the brain. In contrast, generalized seizures may result from cellular, biochemical, or structural abnormalities that have a more widespread distribution. There are clear exceptions in both cases, however.

FOCAL ONSET SEIZURES

Focal seizures arise from a neuronal network either discretely localized within one brain region or more broadly distributed but still within a cerebral hemisphere. With the new classification system, the subcategories of "simple focal seizures" and "complex focal seizures" have been eliminated. Instead, the classification emphasizes the effect on awareness (intact or impaired) and nature of the onset (motor or nonmotor). Focal seizures can also evolve into generalized seizures. In the past this was referred to as *focal seizures with secondary generalization*, but the new system relies on descriptions of the type of generalized seizures that evolve from the focal seizure.

The routine interictal (i.e., between seizures) electroencephalogram (EEG) in patients with focal seizures is often normal or may show brief discharges termed *epileptiform spikes*, or *sharp waves*. Because focal seizures can arise from the medial temporal lobe or inferior frontal lobe (i.e., regions distant from the scalp), the EEG recorded during the seizure may be nonlocalizing. However, the region of seizure onset may be detected using surgically placed intracranial electrodes.

Focal Seizures with Intact Awareness Focal seizures can have motor manifestations (such as tonic, clonic, or myoclonic movements) or nonmotor manifestations (such as sensory, autonomic, or emotional symptoms) without impairment of awareness. For example, a patient having a focal motor seizure arising from the right primary motor cortex near the area controlling hand movement will note the onset of involuntary movements of the contralateral, left hand. Since the cortical region controlling hand movement is immediately adjacent to the region for facial expression, the seizure may also cause abnormal movements of the face synchronous with the movements of the hand. The EEG recorded with scalp electrodes during the seizure (i.e., an ictal EEG) may show abnormal discharges in a very limited region over the appropriate area of cerebral cortex if the seizure focus involves the cerebral convexity.

Three additional features of focal motor seizures are worth noting. First, in some patients, the abnormal motor movements may begin in a very restricted region such as the fingers and gradually progress (over seconds to minutes) to include a larger portion of the extremity. This phenomenon, described by Hughlings Jackson and known as a "Jacksonian march," represents the spread of seizure activity over a progressively larger region of motor cortex. Second, patients may experience a localized paresis (Todd's paralysis) for minutes to many hours in the involved region following the seizure. Third, in rare instances, the seizure may continue for hours or days. This condition, termed *epilepsia partialis continua*, is often refractory to medical therapy.

Focal seizures may also manifest as changes in somatic sensation (e.g., paresthesias), vision (flashing lights or formed hallucinations), equilibrium (sensation of falling or vertigo), or autonomic function (flushing, sweating, piloerection). Focal seizures arising from the temporal or frontal cortex may also cause alterations in hearing, olfaction, or emotional state. This includes the sensation of unusual, intense odors (e.g., burning rubber or kerosene) or sounds (crude or highly complex sounds), or an epigastric sensation that rises from the stomach or chest to the head. Some patients describe odd, internal feelings such as fear, a sense of impending change, detachment, depersonalization, déjá vu, or illusions that objects are growing smaller (micropsia) or larger (macropsia). These subjective, "internal" events that are not directly observable by someone else are referred to as *auras*.

Focal Seizures with Impaired Awareness Focal seizures may also be accompanied by a transient impairment of the patient's ability to maintain normal contact with the environment. The patient is unable to respond appropriately to visual or verbal commands during the seizure and has impaired recollection or awareness of the ictal phase. The seizures frequently begin with an aura (i.e., a focal seizure without cognitive disturbance) that is stereotypic for the patient. The start of the ictal phase is often a motionless stare, which marks the onset of the period of impaired awareness. The impaired awareness is usually accompanied by *automatisms*, which are involuntary, automatic behaviors that have a wide range of manifestations. Automatisms may consist of very basic behaviors such as chewing, lip smacking, swallowing, or "picking" movements of the hands, or more elaborate behaviors such as a display of emotion or running. The patient is typically confused following the seizure, and the transition to full recovery of consciousness may range from seconds up to an hour or longer. Examination immediately following the seizure may show an anterograde amnesia or transient neurological deficits (such as aphasia, hemi-neglect, or visual loss) caused by postictal inhibition of the cortical regions most involved in the seizure itself.

The range of potential clinical behaviors linked to focal seizures is so broad that extreme caution is advised before concluding that stereotypic episodes of bizarre or atypical behavior are not due to seizure activity. In such cases additional, detailed EEG studies may be helpful.

EVOLUTION OF FOCAL SEIZURES TO GENERALIZED SEIZURES

Focal seizures can spread to involve both cerebral hemispheres and produce a generalized seizure, usually of the tonic-clonic variety (discussed below). This evolution is observed frequently following focal seizures arising from a region in the frontal lobe, but may also be associated with focal seizures occurring elsewhere in the brain. A focal seizure that evolves into a generalized seizure is often difficult to distinguish from a primary generalized onset tonic-clonic seizure, because bystanders tend to emphasize the more dramatic, generalized convulsive phase of the seizure and overlook the more subtle, focal symptoms present at onset. In some cases, the focal onset of the seizure becomes apparent only when a careful history identifies a preceding aura. Often, however, the focal onset is not clinically evident and may be established only through careful EEG analysis. Nonetheless, distinguishing between these two entities is extremely important, because there may be substantial differences in the evaluation and treatment of epilepsies characterized by focal versus generalized onset seizures.

GENERALIZED ONSET SEIZURES

Generalized seizures arise at some point in the brain but immediately and rapidly engage neuronal networks in both cerebral hemispheres. Several types of generalized seizures have features that place them in distinctive categories and facilitate clinical diagnosis.

Typical Absence Seizures Typical absence seizures are characterized by sudden, brief lapses of consciousness without loss of postural control. The seizure usually lasts for only seconds, consciousness returns as suddenly as it was lost, and there is no postictal confusion. Although the brief loss of consciousness may be clinically inapparent or the sole manifestation of the seizure discharge, absence seizures are usually accompanied by subtle, bilateral motor signs such as rapid blinking of the eyelids, chewing movements, or small-amplitude, clonic movements of the hands.

Typical absence seizures are associated with a group of genetically determined epilepsies with onset usually in childhood (ages 4–10 years) or early adolescence and are the main seizure type in 15–20% of children with epilepsy. The seizures can occur hundreds of times per day, but the child may be unaware of or unable to convey their existence. Because the clinical signs of the seizures are subtle, especially to parents who may not have had previous experience with seizures, it is not surprising that the first clue to absence epilepsy is often unexplained "daydreaming" and a decline in school performance recognized by a teacher.

The electrophysiologic hallmark of typical absence seizures is a generalized, symmetric, 3-Hz spike-and-slow-wave discharges that begins and ends suddenly, superimposed on a normal EEG background. Periods of spike-and-slow-wave discharges lasting more than a few seconds usually correlate with clinical signs, but the EEG often shows many more brief bursts of abnormal cortical activity than were suspected clinically. Hyperventilation tends to provoke these electrographic discharges and even the seizures themselves and is routinely used when recording the EEG.

Atypical Absence Seizures Atypical absence seizures have features that deviate both clinically and electrophysiologically from typical absence seizures. For example, the lapse of consciousness is usually of longer duration and less abrupt in onset and cessation, and the seizure is accompanied by more obvious motor signs that may include focal or lateralizing features. The EEG shows a generalized, slow spike-and-slow-wave pattern with a frequency of ≤2.5 per second, as well as other abnormal activity. Atypical absence seizures are usually associated with diffuse or multifocal structural abnormalities of the brain and therefore may accompany other signs of neurologic dysfunction such as mental retardation. Furthermore, the seizures are less responsive to anticonvulsants compared to typical absence seizures.

Generalized, Tonic-Clonic Seizures Generalized onset tonic-clonic seizures are the main seizure type in ~10% of all persons with epilepsy. They are also the most common seizure type resulting from metabolic derangements and are therefore frequently encountered in many different clinical settings. The seizure usually begins abruptly without warning, although some patients describe vague premonitory symptoms in the hours leading up to the seizure. This prodrome is distinct from the stereotypic auras associated with focal seizures that generalize. The initial phase of the seizure is usually tonic contraction of muscles throughout the body, accounting for a number of the classic features of the event. Tonic contraction of the muscles of expiration and the larynx at the onset will produce a loud moan or "ictal cry." Respirations are impaired, secretions pool in the oropharynx, and cyanosis develops. Contraction of the jaw muscles may cause biting of the tongue. A marked enhancement of sympathetic tone leads to increases in heart rate, blood pressure, and pupillary size. After 10–20 s, the tonic phase of the seizure typically evolves into the clonic phase, produced by the superimposition of periods of muscle relaxation on the tonic muscle contraction. The periods of relaxation progressively increase until the end of the ictal phase, which usually lasts no more than 1 min. The postictal phase is characterized by unresponsiveness, muscular flaccidity, and excessive salivation that can cause stridorous breathing and partial airway obstruction. Bladder or bowel incontinence may occur at this point. Patients gradually regain consciousness over minutes to hours, and during this transition, there is typically a period of postictal confusion. Patients subsequently complain of headache, fatigue, and muscle ache that can last for many hours. The duration of impaired consciousness in the postictal phase can be extremely long (i.e., many hours) in patients with prolonged seizures or underlying central nervous system (CNS) diseases such as alcoholic cerebral atrophy.

The EEG during the tonic phase of the seizure shows a progressive increase in generalized low-voltage fast activity, followed by generalized high-amplitude, polyspike discharges. In the clonic phase, the high-amplitude activity is typically interrupted by slow waves to create a spike-and-slow-wave pattern. The postictal EEG shows diffuse suppression of all cerebral activity, then slowing that gradually recovers as the patient awakens.

There are a number of variants of generalized motor seizures, including pure tonic and pure clonic seizures. Brief tonic seizures lasting only a few seconds are especially noteworthy since they are usually associated with specific epilepsy syndromes having mixed seizure phenotypes, such as the Lennox-Gastaut syndrome (discussed below).

Atonic Seizures Atonic seizures are characterized by sudden loss of postural muscle tone lasting 1–2 s. Consciousness is briefly impaired, but there is usually no postictal confusion. A very brief seizure may cause only a quick head drop or nodding movement, whereas a longer seizure will cause the patient to collapse. This can be extremely dangerous, because there is a substantial risk of direct head injury with the fall. The EEG shows brief, generalized spike-and-wave discharges followed immediately by diffuse slow waves that correlate with the loss of muscle tone. Similar to pure tonic seizures, atonic seizures are usually seen in association with known epilepsy syndromes.

Myoclonic Seizures Myoclonus is a sudden and brief muscle contraction that may involve one part of the body or the entire body. A normal, common physiologic form of myoclonus is the sudden jerking movement observed while falling asleep. Pathologic myoclonus is most commonly seen in association with metabolic disorders, degenerative CNS diseases, or anoxic brain injury (Chap. 301). Although the distinction from other forms of myoclonus is imprecise, myoclonic seizures are considered to be true epileptic events because they are caused by cortical (versus subcortical or spinal) dysfunction. The EEG shows bilaterally synchronous spike-and-slow-wave discharges immediately prior to the movement and muscle artifact associated with the myoclonus. Myoclonic seizures usually coexist with other forms of generalized seizures but are the predominant feature of juvenile myoclonic epilepsy (JME) (discussed below).

Epileptic Spasms Epileptic spasms are characterized by a briefly sustained flexion or extension of predominantly proximal muscles, including truncal muscles. The EEG usually shows hypsarrhythmia, which consist of diffuse, giant slow waves with a chaotic background of irregular, multifocal spikes and sharp waves. During the clinical spasm, there is a marked suppression of the EEG background (the "electrodecremental response"). The electromyogram (EMG) also reveals a characteristic rhomboid pattern that may help distinguish spasms from brief tonic and myoclonic seizures. Epileptic spasms occur predominantly in infants and likely result from differences in neuronal function and connectivity in the immature versus mature CNS.

EPILEPSY SYNDROMES

Epilepsy syndromes are disorders in which epilepsy is a predominant feature, and there is sufficient evidence (e.g., through clinical, EEG, radiologic, or genetic observations) to suggest a common underlying mechanism. Three important epilepsy syndromes are listed below; additional examples with a known genetic basis are shown in Table 418-2.

JUVENILE MYOCLONIC EPILEPSY

JME is a generalized seizure disorder of unknown cause that appears in early adolescence and is usually characterized by bilateral myoclonic jerks that may be single or repetitive. The myoclonic seizures are most frequent in the morning after awakening and can be provoked by sleep deprivation. Consciousness is preserved unless the myoclonus is especially severe. Many patients also experience generalized tonic-clonic seizures, and up to one-third have absence seizures. Although complete remission is relatively uncommon, the seizures usually respond well to appropriate anticonvulsant medication. There is often a family history of epilepsy, and genetic linkage studies suggest a polygenic cause.

LENNOX-GASTAUT SYNDROME

Lennox-Gastaut syndrome occurs in children and is defined by the following triad: (1) multiple seizure types (usually including generalized tonic-clonic, atonic, and atypical absence seizures); (2) an EEG showing slow (<3 Hz) spike-and-wave discharges and a variety of other abnormalities; and (3) impaired cognitive function in most but not all cases.

TABLE 418-2 Examples of Genes Associated with Epilepsy Syndromes[a]

GENE (LOCUS)	FUNCTION OF GENE	CLINICAL SYNDROME	COMMENTS
CHRNA4 (20q13.2)	Nicotinic acetylcholine receptor subunit; mutations cause alterations in Ca^{2+} flux through the receptor; this may reduce the amount of GABA release in presynaptic terminals	Autosomal dominant nocturnal frontal lobe epilepsy (ADNFLE); childhood onset; brief, nighttime seizures with prominent motor movements; often misdiagnosed as primary sleep disorder	Rare; first identified in a large Australian family; other families found to have mutations in *CHRNA2* or *CHRNB2*, and some families appear to have mutations at other loci
KCNQ2 (20q13.3)	Voltage-gated potassium channel subunits; mutation in pore regions may cause a 20–40% reduction of potassium currents, which will lead to impaired repolarization	Benign familial neonatal seizures (BFNSs); autosomal dominant inheritance; onset in first week of life in infants who are otherwise normal; remission usually within weeks to months; long-term epilepsy in 10–15%	Rare; other families found to have mutations in *KCNQ3* or an inversion in chromosomal 5; sequence and functional homology to *KCNQ1*, mutations of which cause long QT syndrome and a cardiac-auditory syndrome
SCN1A (2q24.3)	α-Subunit of a voltage-gated sodium channel; numerous mutations affecting sodium currents that cause either gain or loss of function; network effects appear related to expression in excitatory or inhibitory cells	Generalized epilepsy with febrile seizures plus (GEFS+); autosomal dominant inheritance; presents with febrile seizures at median 1 year, which may persist >6 years, then variable seizure types not associated with fever; numerous other syndromes, including almost 80% of patients with Dravet's syndrome (severe myoclonic epilepsy of infancy) and some cases of Lennox-Gastaut syndrome	Incidence uncertain; GEFS+ identified in other families with mutations in other sodium channel subunits (*SCN2B* and *SCN2A*) and $GABA_A$ receptor subunit (*GABRG2* and *GABRA1*); significant phenotypic heterogeneity within same family, including members with febrile seizures only
LGI1 (10q24)	Leucine-rich glioma-inactivated 1 gene; previous evidence for role in glial tumor progression; recent studies suggest an influence in the postnatal development of glutamatergic circuits in the hippocampus	Autosomal dominant partial epilepsy with auditory features (ADPEAF); a form of idiopathic lateral temporal lobe epilepsy with auditory symptoms or aphasia as a major focal seizure manifestation; age of onset usually between 10 and 25 years	Mutations found in up to 50% of families containing two or more subjects with idiopathic localization-related epilepsy with ictal auditory symptoms, suggesting that at least one other gene may underlie this syndrome
DEPDC5 (22q12.2)	Disheveled, Egl-10 and pleckstrin domain containing protein 5; exerts an inhibitory effect on mammalian target of rapamycin (mTOR)-mediated processes, such as cell growth and proliferation	Autosomal dominant familial focal epilepsy with variable foci (FFEVF); family members have seizures originating from different cortical regions; neuroimaging usually normal but may harbor subtle malformations; recent studies also suggest association with benign epilepsy with centrotemporal spikes	Study of families with the limited number of affected members revealed mutations in ~12% of families; thus may be a relatively common cause of lesion-negative focal epilepsies with suspected genetic basis
CSTB (21q22.3)	Cystatin B, a noncaspase cysteine protease inhibitor; normal protein may block neuronal apoptosis by inhibiting caspases directly or indirectly (via cathepsins), or controlling proteolysis	Progressive myoclonus epilepsy (PME) (Unverricht-Lundborg disease); autosomal recessive inheritance; age of onset between 6 and 15 years, myoclonic seizures, ataxia, and progressive cognitive decline; brain shows neuronal degeneration	Overall rare, but relatively common in Finland and Western Mediterranean (>1 in 20,000); precise role of cystatin B in human disease unknown, although mice with null mutations of cystatin B have similar syndrome
EPM2A (6q24)	Laforin, a protein tyrosine phosphatase (PTP); involved in glycogen metabolism and may have antiapoptotic activity	Progressive myoclonus epilepsy (Lafora's disease); autosomal recessive inheritance; age of onset 6–19 years, death within 10 years; brain degeneration associated with polyglucosan intracellular inclusion bodies in numerous organs	Most common PME in Southern Europe, Middle East, Northern Africa, and Indian subcontinent; genetic heterogeneity; unknown whether seizure phenotype due to degeneration or direct effects of abnormal laforin expression
Doublecortin (Xq21-24)	Doublecortin, expressed primarily in frontal lobes; directly regulates microtubule polymerization and bundling	Classic lissencephaly associated with severe mental retardation and seizures in males; subcortical band heterotopia with more subtle findings in females (presumably due to random X-inactivation); X-linked dominant	Relatively rare but of uncertain incidence; recent increased ascertainment due to improved imaging techniques; relationship between migration defect and seizure phenotype unknown

[a]The first five syndromes listed in the table (ADNFLE, BFNC, GEFS+, ADPEAF, and FFEVF) are examples of idiopathic epilepsies associated with identified gene mutations. The last three syndromes are examples of the numerous Mendelian disorders in which seizures are one part of the phenotype.

Abbreviations: GABA, γ-aminobutyric acid; PME, progressive myoclonus epilepsy.

Lennox-Gastaut syndrome is associated with CNS disease or dysfunction from a variety of causes, including *de novo* mutations, developmental abnormalities, perinatal hypoxia/ischemia, trauma, infection, and other acquired lesions. The multifactorial nature of this syndrome suggests that it is a nonspecific response of the brain to diffuse neuronal dysfunction. Unfortunately, many patients have a poor prognosis due to the underlying CNS disease and the physical and psychosocial consequences of severe, poorly controlled epilepsy.

MESIAL TEMPORAL LOBE EPILEPSY SYNDROME

Mesial temporal lobe epilepsy (MTLE) is the most common syndrome associated with focal seizures with impairment of consciousness and is an example of an epilepsy syndrome with distinctive clinical, electroencephalographic, and pathologic features **(Table 418-3)**. High-resolution magnetic resonance imaging (MRI) can detect the characteristic hippocampal sclerosis that appears to be essential in the pathophysiology of MTLE for many patients **(Fig. 418-1)**. Recognition of this syndrome is especially important because it tends to be refractory to treatment with anticonvulsants but responds well to surgical intervention. Advances in the understanding of basic mechanisms of epilepsy have come through studies of experimental models of MTLE, discussed below.

THE CAUSES OF SEIZURES AND EPILEPSY

Seizures are a result of a shift in the normal balance of excitation and inhibition within the CNS. Given the numerous properties that control neuronal excitability, it is not surprising that there are many different ways to perturb this normal balance, and therefore many different causes of both seizures and epilepsy. Three clinical observations emphasize how a variety of factors determine why certain conditions may cause seizures or epilepsy in a given patient.

1. *The normal brain is capable of having a seizure under the appropriate circumstances, and there are differences between individuals in the susceptibility or threshold for seizures*. For example, seizures may be induced by high fevers in children who are otherwise normal and who never develop other neurologic problems, including epilepsy. However, febrile seizures occur only in a relatively small proportion of children. This implies there are various underlying *endogenous*

TABLE 418-3 Characteristics of the Mesial Temporal Lobe Epilepsy Syndrome

History	
History of febrile seizures	Rare generalized seizures
Family history of epilepsy	Seizures may remit and reappear
Early onset	Seizures often intractable
Clinical Observations	
Aura common	Postictal disorientation
Behavioral arrest/stare	Memory loss
Complex automatisms	Dysphasia (with focus in dominant hemisphere)
Unilateral posturing	
Laboratory Studies	
Unilateral or bilateral anterior temporal spikes on EEG	
Hypometabolism on interictal PET	
Hypoperfusion on interictal SPECT	
Material-specific memory deficits on intracranial amobarbital (Wada) test	
MRI Findings	
Small hippocampus with increased signal on T2-weighted sequences	
Small temporal lobe	
Enlarged temporal horn	
Pathologic Findings	
Highly selective loss of specific cell populations within hippocampus in most cases	

Abbreviations: EEG, electroencephalogram; MRI, magnetic resonance imaging; PET, positron emission tomography; SPECT, single-photon emission computed tomography.

factors that influence the threshold for having a seizure. Some of these factors are genetic, as a family history of epilepsy has a clear influence on the likelihood of seizures occurring in otherwise normal individuals. Normal development also plays an important role, because the brain appears to have different seizure thresholds at different maturational stages.

2. *There are a variety of conditions that have an extremely high likelihood of resulting in a chronic seizure disorder*. One of the best examples of this is severe, penetrating head trauma, which is associated with up to a 45% risk of subsequent epilepsy. The high propensity for severe traumatic brain injury to lead to epilepsy suggests that the injury results in a long-lasting pathologic change in the CNS that transforms a presumably normal neural network into one that is abnormally hyperexcitable. This process is known as *epileptogenesis*, and the specific changes that result in a lowered seizure threshold can be considered *epileptogenic factors*. Other processes associated with epileptogenesis include stroke, infections, and abnormalities of CNS development. Likewise, the genetic abnormalities associated with epilepsy likely involve processes that trigger the appearance of specific sets of epileptogenic factors.
3. *Seizures are episodic*. Patients with epilepsy have seizures intermittently and, depending on the underlying cause, many patients are completely normal for months or even years between seizures. This implies there are important provocative or *precipitating factors* that induce seizures in patients with epilepsy. Similarly, precipitating factors are responsible for causing the single seizure in someone without epilepsy. Precipitants include those due to intrinsic physiologic processes such as psychological or physical stress, sleep deprivation, or hormonal changes. They also include exogenous factors such as exposure to toxic substances and certain medications.

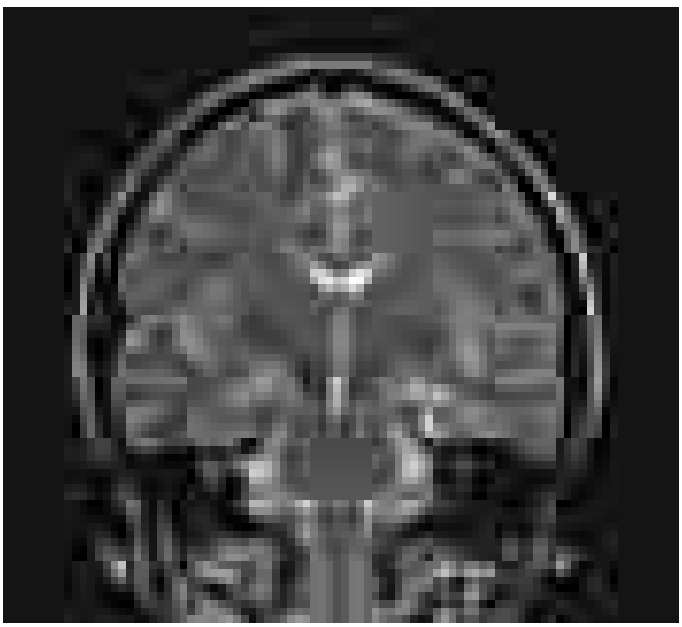


FIGURE 418-1 Mesial temporal lobe epilepsy. The electroencephalogram and seizure semiology were consistent with a left temporal lobe focus. This coronal high-resolution T2-weighted fast spin echo magnetic resonance image obtained at 3 tesla is at the level of the hippocampal bodies, and shows abnormal high signal intensity, blurring of internal laminar architecture, and reduced size of the left hippocampus (*arrow*) relative to the right. This triad of imaging findings is consistent with hippocampal sclerosis.

These observations emphasize the concept that the many causes of seizures and epilepsy result from a dynamic interplay between endogenous factors, epileptogenic factors, and precipitating factors. The potential role of each needs to be carefully considered when determining the appropriate management of a patient with seizures. For example, the identification of predisposing factors (e.g., family history of epilepsy) in a patient with febrile seizures may increase the necessity for closer follow-up and a more aggressive diagnostic evaluation. Finding an epileptogenic lesion may help in the estimation of seizure recurrence and duration of therapy. Finally, removal or modification of a precipitating factor may be an effective and safer method for preventing further seizures than the prophylactic use of anticonvulsant drugs.

CAUSES ACCORDING TO AGE

In practice, it is useful to consider the etiologies of seizures based on the age of the patient, because age is one of the most important factors determining both the incidence and the likely causes of seizures or epilepsy **(Table 418-4)**. During the *neonatal period and early infancy*, potential causes include hypoxic-ischemic encephalopathy, trauma, CNS infection, congenital CNS abnormalities, and metabolic disorders. Babies born to mothers using neurotoxic drugs such as cocaine, heroin, or ethanol are susceptible to drug-withdrawal seizures in the first few days after delivery. Hypoglycemia and hypocalcemia, which can occur as secondary complications of perinatal injury, are also causes of seizures early after delivery. Seizures due to inborn errors of metabolism usually present once regular feeding begins, typically 2–3 days after birth. Pyridoxine (vitamin B_6) deficiency, an important cause of neonatal seizures, can be effectively treated with pyridoxine replacement. The idiopathic or inherited forms of benign neonatal seizures are also seen during this time period.

The most common seizures arising in *late infancy and early childhood* are febrile seizures, which are seizures associated with fevers but without evidence of CNS infection or other defined causes. The overall prevalence is 3–5% and even higher in some parts of the world such as Asia. Patients often have a family history of febrile seizures or epilepsy. Febrile seizures usually occur between 3 months and 5 years of age and have a peak incidence between 18 and 24 months. The typical scenario is a child who has a generalized, tonic-clonic seizure during a febrile illness in the setting of a common childhood infection such as otitis media, respiratory infection, or gastroenteritis. The seizure is likely to occur during the rising phase of the temperature curve (i.e., during the first day) rather than well into the course of the illness. A *simple* febrile seizure is a single, isolated event, brief, and symmetric in appearance. *Complex* febrile seizures are characterized by repeated seizure activity, duration >15 minutes, or by focal features. Approximately one-third of patients with febrile seizures will have a recurrence, but <10% have three or more episodes. Recurrences are much more likely when the

TABLE 418-4 Causes of Seizures

Neonates (<1 month)	Perinatal hypoxia and ischemia
	Intracranial hemorrhage and trauma
	CNS infection
	Metabolic disturbances (hypoglycemia, hypocalcemia, hypomagnesemia, pyridoxine deficiency)
	Drug withdrawal
	Developmental disorders
	Genetic disorders
Infants and children (>1 month and <12 years)	Febrile seizures
	Genetic disorders (metabolic, degenerative, primary epilepsy syndromes)
	CNS infection
	Developmental disorders
	Trauma
Adolescents (12–18 years)	Trauma
	Genetic disorders
	Infection
	Illicit drug use
	Brain tumor
Young adults (18–35 years)	Trauma
	Alcohol withdrawal
	Illicit drug use
	Brain tumor
	Autoantibodies
Older adults (>35 years)	Cerebrovascular disease
	Brain tumor
	Alcohol withdrawal
	Metabolic disorders (uremia, hepatic failure, electrolyte abnormalities, hypoglycemia, hyperglycemia)
	Alzheimer's disease and other degenerative CNS diseases
	Autoantibodies

Abbreviation: CNS, central nervous system.

febrile seizure occurs in the first year of life. Simple febrile seizures are not associated with an increase in the risk of developing epilepsy, while complex febrile seizures have a risk of 2–5%; other risk factors include the presence of preexisting neurologic deficits and a family history of nonfebrile seizures.

Childhood marks the age at which many of the well-defined epilepsy syndromes present. Some children who are otherwise normal develop idiopathic, generalized tonic-clonic seizures without other features that fit into specific syndromes. Temporal lobe epilepsy usually presents in childhood and may be related to mesial temporal lobe sclerosis (as part of the MTLE syndrome) or other focal abnormalities such as cortical dysgenesis. Other types of focal seizures, including those that evolve into generalized seizures, may be the relatively late manifestation of a developmental disorder, an acquired lesion such as head trauma, CNS infection (especially viral encephalitis), or very rarely a CNS tumor.

The period of *adolescence and early adulthood* is one of transition during which the idiopathic or genetically based epilepsy syndromes, including JME and juvenile absence epilepsy, become less common, while epilepsies secondary to acquired CNS lesions begin to predominate. Seizures that arise in patients in this age range may be associated with head trauma, CNS infections (including parasitic infections such as cysticercosis), brain tumors, congenital CNS abnormalities, illicit drug use, or alcohol withdrawal. Autoantibodies directed against CNS antigens such as potassium channels or glutamate receptors are a newly recognized cause of epilepsy that also begins to appear in this age group (although cases of autoimmunity are being increasingly described in the pediatric population), including patients without an identifiable cancer. This etiology should be suspected when a previously normal individual presents with a particularly aggressive seizure pattern developing over weeks to months and characterized by increasingly frequent and prolonged seizures, especially when combined with psychiatric symptoms and changes in cognitive function **(Chap. 90)**.

Head trauma is a common cause of epilepsy in adolescents and adults. The head injury can be caused by a variety of mechanisms, and the likelihood of developing epilepsy is strongly correlated with the severity of the injury. A patient with a penetrating head wound, depressed skull fracture, intracranial hemorrhage, or prolonged posttraumatic coma or amnesia has a 30–50% risk of developing epilepsy, whereas a patient with a closed head injury and cerebral contusion has a 5–25% risk. Recurrent seizures usually develop within 1 year after head trauma, although intervals of >10 years are well known. In controlled studies, mild head injury, defined as a concussion with amnesia or loss of consciousness of <30 min, was found to be associated with only a slightly increased likelihood of epilepsy. Nonetheless, most epileptologists know of patients who have focal seizures within hours or days of a mild head injury and subsequently develop chronic seizures of the same type; such cases may represent rare examples of chronic epilepsy resulting from mild head injury.

The causes of seizures in *older adults* include cerebrovascular disease, trauma (including subdural hematoma), CNS tumors, and degenerative diseases. Cerebrovascular disease may account for ~50% of new cases of epilepsy in patients >65 years. Acute seizures (i.e., occurring at the time of the stroke) are seen more often with embolic rather than hemorrhagic or thrombotic stroke. Chronic seizures typically appear months to years after the initial event and are associated with all forms of stroke.

Metabolic disturbances such as electrolyte imbalance, hypo- or hyperglycemia, renal failure, and hepatic failure may cause seizures at any age. Similarly, endocrine disorders, hematologic disorders, vasculitides, and many other systemic diseases may cause seizures over a broad age range. A wide variety of medications and abused substances are known to precipitate seizures as well **(Table 418-5)**.

BASIC MECHANISMS

MECHANISMS OF SEIZURE INITIATION AND PROPAGATION

Focal seizure activity can begin in a very discrete region of cortex and then slowly invade the surrounding regions. The hallmark of an established seizure is typically an electrographic "spike" due to intense near-simultaneous firing of a large number of local excitatory neurons, resulting in an apparent hypersynchronization of the excitatory bursts across a relatively large cortical region. The bursting activity in individual neurons (the "paroxysmal depolarization shift") is caused by a relatively long-lasting depolarization of the neuronal membrane due to influx of extracellular calcium (Ca^{2+}), which leads to the opening of voltage-dependent sodium (Na^+) channels, influx of Na^+, and generation of repetitive action potentials. This is followed by a hyperpolarizing afterpotential mediated by γ-aminobutyric acid (GABA) receptors or potassium (K^+) channels, depending on the cell type. The synchronized bursts from a sufficient number of neurons result in a so-called spike discharge on the EEG.

The spreading seizure wavefront is thought to slow and ultimately halt by intact hyperpolarization and a "surround" inhibition created by feedforward activation of inhibitory neurons. With sufficient activation, there is a recruitment of surrounding neurons via a number of synaptic and nonsynaptic mechanisms, including: (1) an increase in extracellular K^+, which blunts hyperpolarization and depolarizes neighboring neurons; (2) accumulation of Ca^{2+} in presynaptic terminals, leading to enhanced neurotransmitter release; (3) depolarization-induced activation of the *N*-methyl-D-aspartate (NMDA) subtype of the excitatory amino acid receptor, which causes additional Ca^{2+} influx and neuronal activation; and (4) ephaptic interactions related to changes in tissue osmolarity and cell swelling. The recruitment of a sufficient number of neurons leads to the propagation of excitatory currents into contiguous areas via local cortical connections and to more distant areas via long commissural pathways such as the corpus callosum.

TABLE 418-5 Drugs and Other Substances That Can Cause Seizures

Alkylating agents (e.g., busulfan, chlorambucil)
Antimalarials (chloroquine, mefloquine)
Antimicrobials/antivirals
β-lactam and related compounds
Quinolones
Acyclovir
Isoniazid
Ganciclovir
Anesthetics and analgesics
Meperidine
Fentanyl
Tramadol
Local anesthetics
Dietary supplements
Ephedra (ma huang)
Gingko
Immunomodulatory drugs
Cyclosporine
OKT3 (monoclonal antibodies to T cells)
Tacrolimus
Interferons
Psychotropics
Antidepressants (e.g., bupropion)
Antipsychotics (e.g., clozapine)
Lithium
Radiographic contrast agents
Drug withdrawal
Alcohol
Baclofen
Barbiturates (short-acting)
Benzodiazepines (short-acting)
Zolpidem
Drugs of abuse
Amphetamine
Cocaine
Phencyclidine
Methylphenidate
Flumazenil[a]

[a]In benzodiazepine-dependent patients.

Many factors control neuronal excitability, and thus there are many potential mechanisms for altering a neuron's propensity to have bursting activity. Mechanisms *intrinsic* to the neuron include changes in the conductance of ion channels, response characteristics of membrane receptors, cytoplasmic buffering, second-messenger systems, and protein expression as determined by gene transcription, translation, and posttranslational modification. Mechanisms *extrinsic* to the neuron include changes in the amount or type of neurotransmitters present at the synapse, modulation of receptors by extracellular ions and other molecules, and temporal and spatial properties of synaptic and nonsynaptic input. Nonneural cells, such as astrocytes and oligodendrocytes, have an important role in many of these mechanisms as well.

Certain recognized causes of seizures are explained by these mechanisms. For example, accidental ingestion of domoic acid, which is an analogue of glutamate (the principal excitatory neurotransmitter in the brain), causes profound seizures via direct activation of excitatory amino acid receptors throughout the CNS. Penicillin, which can lower the seizure threshold in humans and is a potent convulsant in experimental models, reduces inhibition by antagonizing the effects of GABA at its receptor. The basic mechanisms of other precipitating factors of seizures, such as sleep deprivation, fever, alcohol withdrawal, hypoxia, and infection, are not as well understood but presumably involve analogous perturbations in neuronal excitability. Similarly, the endogenous factors that determine an individual's seizure threshold may relate to these properties as well.

Knowledge of the mechanisms responsible for initiation and propagation of most generalized seizures (including tonic-clonic, myoclonic, and atonic types) remains rudimentary and reflects the limited understanding of the connectivity of the brain at a systems level. Much more is understood about the origin of generalized spike-and-wave discharges in absence seizures. These appear to be related to oscillatory rhythms normally generated during sleep by circuits connecting the thalamus and cortex. This oscillatory behavior involves an interaction between $GABA_B$ receptors, T-type Ca^{2+} channels, and K^+ channels located within the thalamus. Pharmacologic studies indicate that modulation of these receptors and channels can induce absence seizures, and there is good evidence that the genetic forms of absence epilepsy may be associated with mutations of components of this system.

MECHANISMS OF EPILEPTOGENESIS

Epileptogenesis refers to the transformation of a normal neuronal network into one that is chronically hyperexcitable. There is often a delay of months to years between an initial CNS injury such as trauma, stroke, or infection and the first clinically evident seizure. The injury appears to initiate a process that gradually lowers the seizure threshold in the affected region until a spontaneous seizure occurs. In many genetic and idiopathic forms of epilepsy, epileptogenesis is presumably determined by developmentally regulated events.

Pathologic studies of the hippocampus from patients with temporal lobe epilepsy have led to the suggestion that some forms of epileptogenesis are related to *structural changes in neuronal networks*. For example, many patients with MTLE have a highly selective loss of neurons that normally contribute to inhibition of the main excitatory neurons within the dentate gyrus. There is also evidence that, in response to the loss of neurons, there is reorganization of surviving neurons in a way that affects the excitability of the network. Some of these changes can be seen in experimental models of prolonged electrical seizures or traumatic brain injury. Thus, an initial injury such as head injury may lead to a very focal, confined region of structural change that causes local hyperexcitability. The local hyperexcitability leads to further structural changes that evolve over time until the focal lesion produces clinically evident seizures. Similar models have provided strong evidence for long-term alterations in *intrinsic, biochemical properties of cells* within the network such as chronic changes in glutamate or GABA receptor function. Induction of inflammatory cascades may be a critical factor in these processes as well.

GENETIC CAUSES OF EPILEPSY

The most important recent progress in epilepsy research has been the identification of genetic mutations associated with a variety of epilepsy syndromes (Table 418-2). Although most of the mutations identified to date cause rare forms of epilepsy, their discovery has led to extremely important conceptual advances. For example, it appears that many of the inherited epilepsies are due to mutations affecting ion channel function. These syndromes are therefore part of the larger group of channelopathies causing paroxysmal disorders such as cardiac arrhythmias, episodic ataxia, periodic weakness, and familial hemiplegic migraine. Other gene mutations are proving to be associated with pathways influencing CNS development or neuronal homeostasis. *De novo* mutations may explain a significant proportion of these syndromes, especially those with onset in early childhood. A current challenge is to identify the multiple susceptibility genes that underlie the more common forms of idiopathic epilepsies. Recent studies suggest that ion channel mutations and copy number variants may contribute to causation in a subset of these patients.

MECHANISMS OF ACTION OF ANTIEPILEPTIC DRUGS

Antiepileptic drugs appear to act primarily by blocking the initiation or spread of seizures. This occurs through a variety of mechanisms

that modify the activity of ion channels or neurotransmitters, and in most cases, the drugs have pleiotropic effects. The mechanisms include inhibition of Na^+-dependent action potentials in a frequency-dependent manner (e.g., phenytoin, carbamazepine, lamotrigine, topiramate, zonisamide, lacosamide, rufinamide), inhibition of voltage-gated Ca^{2+} channels (phenytoin, gabapentin, pregabalin), facilitating the opening of potassium channels (ezogabine), attenuation of glutamate activity (lamotrigine, topiramate, felbamate), potentiation of GABA receptor function (benzodiazepines and barbiturates), increase in the availability of GABA (valproic acid, gabapentin, tiagabine), and modulation of release of synaptic vesicles (levetiracetam, brivaracetam). Two of the effective drugs for absence seizures, ethosuximide and valproic acid, probably act by inhibiting T-type Ca^{2+} channels in thalamic neurons.

In contrast to the relatively large number of antiepileptic drugs that can attenuate seizure activity, there are currently no drugs known to prevent the formation of a seizure focus following CNS injury. The eventual development of such "antiepileptogenic" drugs will provide an important means of preventing the emergence of epilepsy following injuries such as head trauma, stroke, and CNS infection.

APPROACH TO THE PATIENT

Seizure

When a patient presents shortly after a seizure, the first priorities are attention to vital signs, respiratory and cardiovascular support, and treatment of seizures if they resume (see "Treatment: Seizures and Epilepsy"). Life-threatening conditions such as CNS infection, metabolic derangement, or drug toxicity must be recognized and managed appropriately.

When the patient is not acutely ill, the evaluation will initially focus on whether there is a history of earlier seizures **(Fig. 418-2)**. If this is the first seizure, then the emphasis will be to: (1) establish whether the reported episode was a seizure rather than another paroxysmal event, (2) determine the cause of the seizure by identifying risk factors and precipitating events, and (3) decide whether anticonvulsant therapy is required in addition to treatment for any underlying illness.

In the patient with prior seizures or a known history of epilepsy, the evaluation is directed toward: (1) identification of the underlying cause and precipitating factors, and (2) determination of the adequacy of the patient's current therapy.

HISTORY AND EXAMINATION

The first goal is to determine whether the event was truly a seizure. An in-depth history is essential, because *in many cases the diagnosis of a seizure is based solely on clinical grounds—the examination and laboratory studies are often normal*. Questions should focus on the symptoms before, during, and after the episode in order to differentiate a seizure from other paroxysmal events (see "Differential Diagnosis of Seizures" below). Seizures frequently occur out-of-hospital, and the patient may be unaware of the ictal and immediate postictal phases; thus, witnesses to the event should be interviewed carefully.

The history should also focus on risk factors and predisposing events. Clues for a predisposition to seizures include a history of febrile seizures, a family history of seizures, and, of particular importance, earlier auras or brief seizures not recognized as such. Epileptogenic factors such as prior head trauma, stroke, tumor, or CNS infection should be identified. In children, a careful assessment of developmental milestones may provide evidence for underlying CNS disease. Precipitating factors such as sleep deprivation, systemic diseases, electrolyte or metabolic derangements, acute infection, drugs that lower the seizure threshold (Table 418-5), or alcohol or illicit drug use should also be identified.

The general physical examination includes a search for signs of infection or systemic illness. Careful examination of the skin may reveal signs of neurocutaneous disorders such as tuberous sclerosis or neurofibromatosis, or chronic liver or renal disease. A finding of organomegaly may indicate a metabolic storage disease, and limb asymmetry may provide a clue to brain injury early in development. Signs of head trauma and use of alcohol or illicit drugs should be sought. Auscultation of the heart and carotid arteries may identify an abnormality that predisposes to cerebrovascular disease.

All patients require a complete neurologic examination, with particular emphasis on eliciting signs of cerebral hemispheric disease **(Chap. 415)**. Careful assessment of mental status (including memory, language function, and abstract thinking) may suggest lesions in the anterior frontal, parietal, or temporal lobes. Testing of visual fields will help screen for lesions in the optic pathways and occipital lobes. Screening tests of motor function such as pronator drift, deep tendon reflexes, gait, and coordination may suggest lesions in motor (frontal) cortex, and cortical sensory testing (e.g., double simultaneous stimulation) may detect lesions in the parietal cortex.

LABORATORY STUDIES

Routine blood studies are indicated to identify the more common metabolic causes of seizures such as abnormalities in electrolytes, glucose, calcium, or magnesium, and hepatic or renal disease. A screen for toxins in blood and urine should also be obtained from all patients in appropriate risk groups, especially when no clear precipitating factor has been identified. A lumbar puncture is indicated if there is any suspicion of meningitis or encephalitis, and it is mandatory in all patients infected with HIV, even in the absence of symptoms or signs suggesting infection. Testing for autoantibodies in the serum and cerebrospinal fluid (CSF) should be considered in patients presenting with a seemingly aggressive form of epilepsy associated with other abnormalities such as psychiatric symptoms or cognitive disturbances.

ELECTROPHYSIOLOGIC STUDIES

The electrical activity of the brain (the EEG) is easily recorded from electrodes placed on the scalp. The potential difference between pairs of electrodes on the scalp (bipolar derivation) or between individual scalp electrodes and a relatively inactive common reference point (referential derivation) is amplified and displayed on a computer monitor, oscilloscope, or paper. Digital systems allow the EEG to be reconstructed and displayed with any desired format and to be manipulated for more detailed analysis and also permit computerized techniques to be used to detect certain abnormalities. The characteristics of the normal EEG depend on the patient's age and level of arousal. The rhythmic activity normally recorded represents the postsynaptic potentials of vertically oriented pyramidal cells of the cerebral cortex and is characterized by its frequency. In normal awake adults lying quietly with the eyes closed, an 8- to 13-Hz alpha rhythm is seen posteriorly in the EEG, intermixed with a variable amount of generalized faster (beta) activity (>13 Hz); the alpha rhythm is attenuated when the eyes are opened **(Fig. 418-3)**. During drowsiness, the alpha rhythm is also attenuated; with light sleep, slower activity in the theta (4–7 Hz) and delta (<4 Hz) ranges becomes more conspicuous.

All patients who have a possible seizure disorder should be evaluated with an EEG as soon as possible. In the evaluation of a patient with suspected epilepsy, the presence of *electrographic seizure activity* during the clinically evident event (i.e., abnormal, repetitive, rhythmic activity having a discrete onset and termination) clearly establishes the diagnosis. The absence of electrographic seizure activity does not exclude a seizure disorder, however, because focal seizures may originate from a region of the cortex that cannot be detected by standard scalp electrodes. The EEG is always abnormal during generalized tonic-clonic seizures. Because seizures are typically infrequent and unpredictable, it is often not possible to obtain the EEG during a clinical event. In such situations, activating procedures are generally undertaken while the EEG is recorded in an attempt to provoke abnormalities. These procedures commonly include hyperventilation (for 3 or 4 min), photic stimulation, sleep, and sleep deprivation on the night prior to the recording. Continuous monitoring for prolonged periods in video-EEG telemetry units for hospitalized patients or the use of portable equipment to

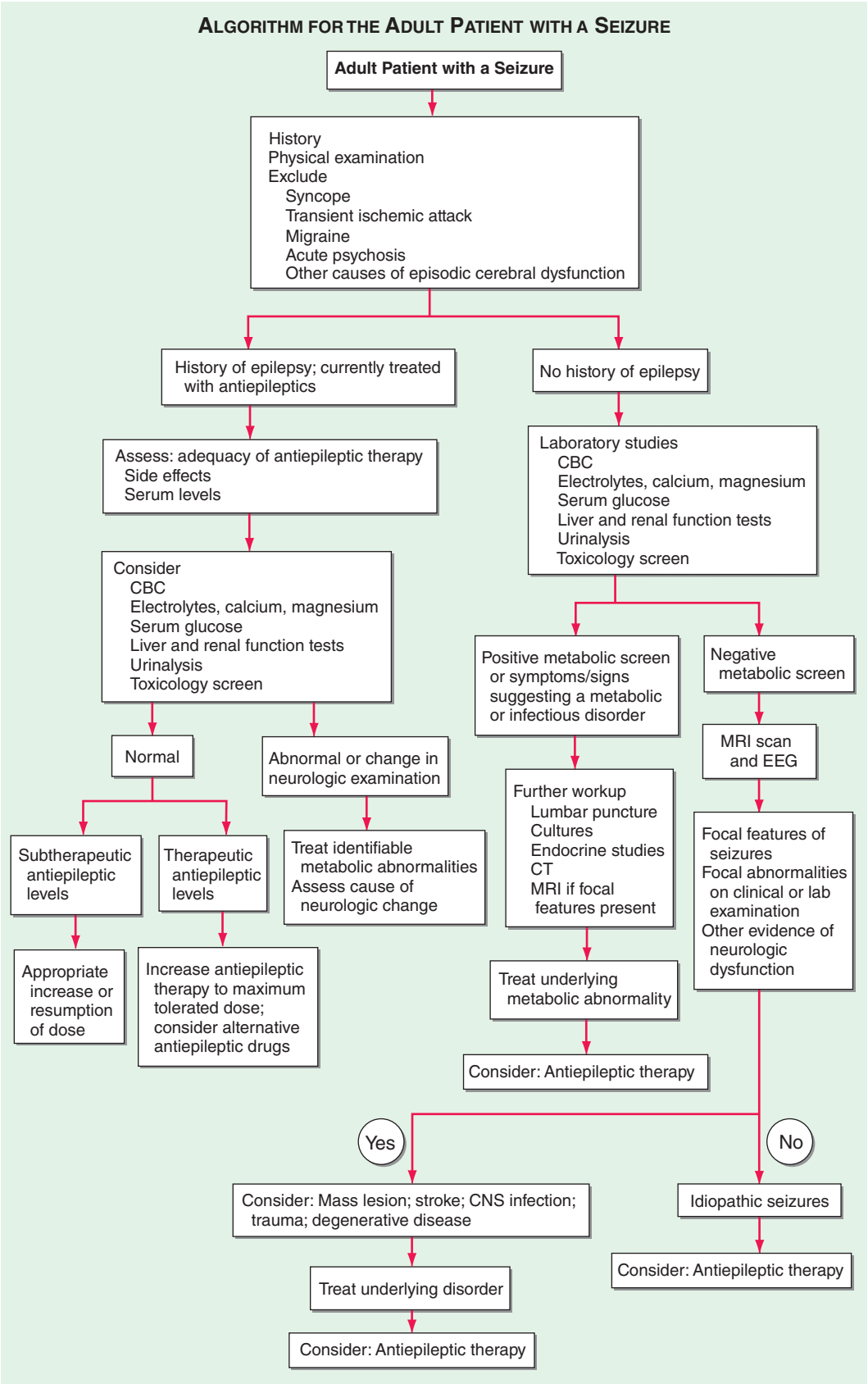


FIGURE 418-2 Evaluation of the adult patient with a seizure. CBC, complete blood count; CNS, central nervous system; CT, computed tomography; EEG, electroencephalogram; MRI, magnetic resonance imaging.

record the EEG continuously for ≥24 h in ambulatory patients has made it easier to capture the electrophysiologic accompaniments of clinical events. In particular, video-EEG telemetry is now a routine approach for the accurate diagnosis of epilepsy in patients with poorly characterized events or seizures that are difficult to control.

The EEG may also be helpful in the interictal period by showing certain abnormalities that are highly supportive of the diagnosis of epilepsy. Such *epileptiform activity* consists of bursts of abnormal discharges containing spikes or sharp waves. The presence of epileptiform activity is not entirely specific for epilepsy, but it has a much greater

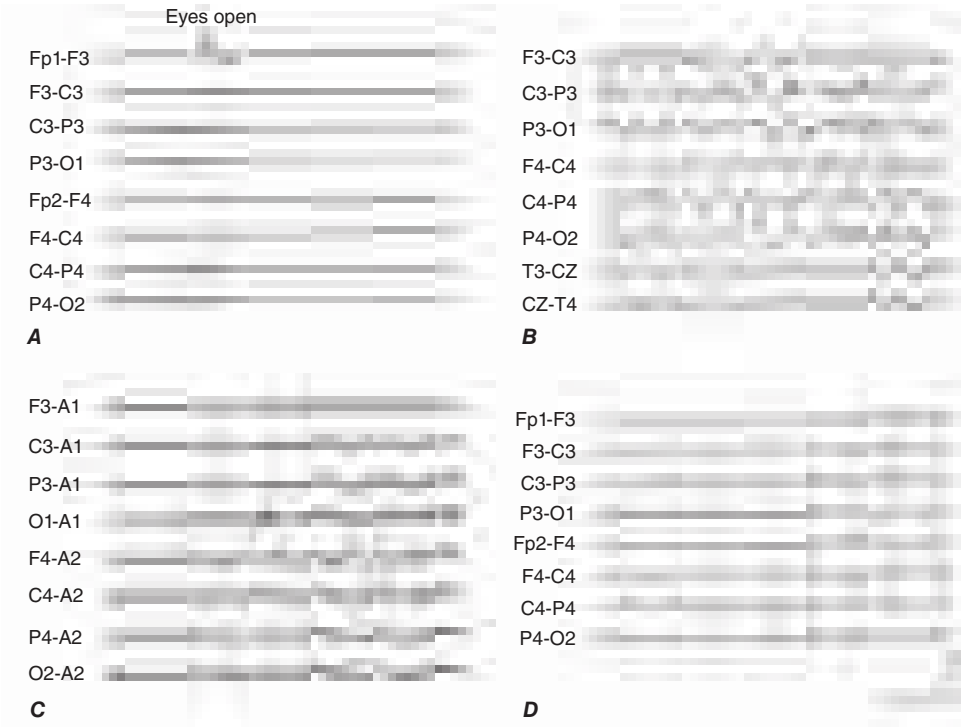


FIGURE 418-3 ***A.*** Normal electroencephalogram (EEG) showing a posteriorly situated 9-Hz alpha rhythm that attenuates with eye opening. ***B.*** Abnormal EEG showing irregular diffuse slow activity in an obtunded patient with encephalitis. ***C.*** Irregular slow activity in the right central region, on a diffusely slowed background, in a patient with a right parietal glioma. ***D.*** Periodic complexes occurring once every second in a patient with Creutzfeldt-Jakob disease. Horizontal calibration: 1 s; vertical calibration: 200 μV in A, 300 μV in other panels. In this and the following figure, electrode placements are indicated at the left of each panel and accord with the international 10:20 system. A, earlobe; C, central; F, frontal; Fp, frontal polar; P, parietal; T, temporal; O, occipital. Right-sided placements are indicated by even numbers, left-sided placements by odd numbers, and midline placements by Z. (*From MJ Aminoff [ed]: Aminoff's Electrodiagnosis in Clinical Neurology, 6th ed. Oxford, Elsevier Saunders, 2012.*)

prevalence in patients with epilepsy than in normal individuals. However, even in an individual who is known to have epilepsy, the initial routine interictal EEG may be normal up to 60% of the time. Thus, the EEG cannot establish the diagnosis of epilepsy in many cases.

The EEG is also used for classifying seizure disorders and aiding in the selection of anticonvulsant medications **(Fig. 418-4)**. For example, episodic generalized spike-wave activity is usually seen in patients with typical absence epilepsy and may be seen with other generalized epilepsy syndromes. Focal interictal epileptiform discharges would support the diagnosis of a focal seizure disorder such as temporal lobe epilepsy or frontal lobe seizures, depending on the location of the discharges.

The routine scalp-recorded EEG may also be used to assess the prognosis of seizure disorders; in general, a normal EEG implies a better prognosis, whereas an abnormal background or profuse epileptiform activity suggests a worse outcome. Unfortunately, the EEG has not proved to be useful in predicting which patients with predisposing conditions such as head injury or brain tumor will go on to develop epilepsy, because in such circumstances epileptiform activity is commonly encountered regardless of whether seizures occur.

Magnetoencephalography (MEG) provides another way of looking noninvasively at cortical activity. Instead of measuring electrical activity of the brain, it measures the small magnetic fields that are generated by this activity. The source of epileptiform activity seen on MEG can be analyzed, and its source in the brain can be estimated using a variety of mathematical techniques. These source estimates can then be plotted on an anatomic image of the brain such as an MRI (discussed below) to generate a magnetic source image (MSI). MSI can be useful to localize potential seizure foci.

BRAIN IMAGING

Almost all patients with new-onset seizures should have a brain imaging study to determine whether there is an underlying structural abnormality that is responsible. The only potential exception to this rule is children who have an unambiguous history and examination suggestive of a benign, generalized seizure disorder such as absence epilepsy. MRI has been shown to be superior to computed tomography (CT) for the detection of cerebral lesions associated with epilepsy. In some cases, MRI will identify lesions such as tumors, vascular malformations, or other pathologies that need urgent therapy. The availability of newer MRI methods such as 3-tesla scanners, parallel imaging with multichannel head coils, three-dimensional structural imaging at submillimeter resolution, and widespread use of pulse sequences such as fluid-attenuated inversion recovery (FLAIR), has increased the sensitivity for detection of abnormalities of cortical architecture, including hippocampal atrophy associated with mesial temporal sclerosis, as well as abnormalities of cortical neuronal migration. In such cases, the findings may not lead to immediate therapy, but they do provide an explanation for the patient's seizures and point to the need for chronic antiepileptic drug therapy or possible surgical resection.

In the patient with a suspected CNS infection or mass lesion, CT scanning should be performed emergently when MRI is not immediately available. Otherwise, it is usually appropriate to obtain an MRI study within a few days of the initial evaluation. Functional imaging procedures such as positron emission tomography (PET) and single-photon emission computed tomography (SPECT) are also used to evaluate certain patients with medically refractory seizures (discussed below).

GENETIC TESTING

With the increasing recognition of specific gene mutations causing epilepsy, genetic testing is beginning to emerge as part of the diagnostic evaluation of patients with epilepsy. In addition to providing a definitive diagnosis (which may be of great benefit to the patient and family members, and curtail the pursuit of additional, unrevealing laboratory testing), genetic testing may offer a guide for therapeutic options (see section "Selection of Antiepileptic Drugs" below). Presently, genetic testing is being done mainly in infants and children with epilepsy syndromes thought to have a genetic cause. However, genetic testing should also be considered in older patients with a history suggesting an undiagnosed genetic epilepsy syndrome that began early in life.

DIFFERENTIAL DIAGNOSIS OF SEIZURES

Disorders that may mimic seizures are listed in **Table 418-6**. In most cases, seizures can be distinguished from other conditions by meticulous attention to the history and relevant laboratory studies. On occasion, additional studies such as video-EEG monitoring, sleep studies, tilt-table analysis, or cardiac electrophysiology may be required to reach a correct diagnosis. Two of the more common nonepileptic syndromes in the differential diagnosis are detailed below.

SYNCOPE

(See also Chap. 18) The diagnostic dilemma encountered most frequently is the distinction between a generalized seizure and syncope. Observations by the patient and bystanders that can help differentiate between the two are listed in **Table 418-7**. Characteristics of a seizure include the presence of an aura, cyanosis, unconsciousness, motor

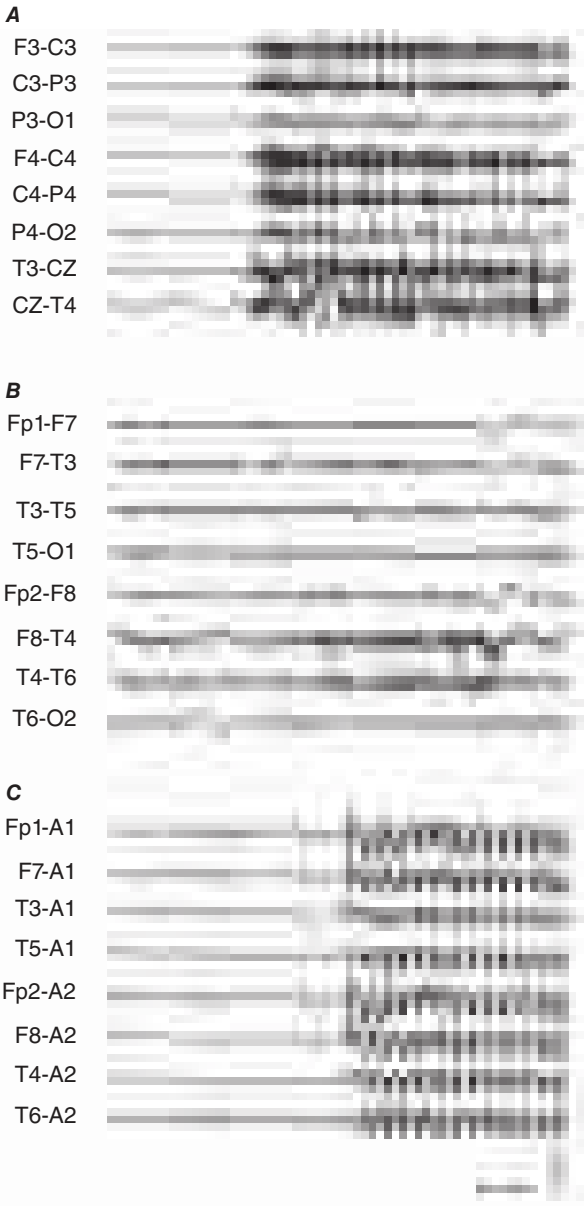


FIGURE 418-4 Electrographic seizures. ***A.*** Onset of a tonic seizure showing generalized repetitive sharp activity with synchronous onset over both hemispheres. ***B.*** Burst of repetitive spikes occurring with sudden onset in the right temporal region during a clinical spell characterized by transient impairment of external awareness. ***C.*** Generalized 3-Hz spike-wave activity occurring synchronously over both hemispheres during an absence (petit mal) attack. Horizontal calibration: 1 s; vertical calibration: 400 μV in *A*, 200 μV in *B*, and 750 μV in *C*. *(From MJ Aminoff [ed]: Aminoff's Electrodiagnosis in Clinical Neurology, 6th ed. Oxford, Elsevier Saunders, 2012.)*

manifestations lasting >15 s, postictal disorientation, muscle soreness, and sleepiness. In contrast, a syncopal episode is more likely if the event was provoked by acute pain or emotional stress or occurred immediately after arising from the lying or sitting position. Patients with syncope often describe a stereotyped transition from consciousness to unconsciousness that includes tiredness, sweating, nausea, and tunneling of vision, and they experience a relatively brief loss of consciousness. Headache or incontinence usually suggests a seizure but may on occasion also occur with syncope. A brief period (i.e., 1–10 s) of convulsive motor activity is frequently seen immediately at the onset of a syncopal episode, especially if the patient remains in an upright posture after fainting (e.g., in a dentist's chair) and therefore has a sustained decrease in cerebral perfusion. Rarely, a syncopal episode can induce a full tonic-clonic seizure. In such cases, the evaluation must focus on both the cause of the syncopal event as well as the possibility that the patient has a propensity for recurrent seizures.

TABLE 418-6 Differential Diagnosis of Seizures

Syncope	**Transient ischemic attack (TIA)**
Vasovagal syncope	Basilar artery TIA
Cardiac arrhythmia	**Sleep disorders**
Valvular heart disease	Narcolepsy/cataplexy
Cardiac failure	Benign sleep myoclonus
Orthostatic hypotension	**Movement disorders**
Psychological disorders	Tics
Psychogenic seizure	Nonepileptic myoclonus
Hyperventilation	Paroxysmal choreoathetosis
Panic attack	**Special considerations in children**
Metabolic disturbances	Breath-holding spells
Alcoholic blackouts	Migraine with recurrent abdominal pain and cyclic vomiting
Delirium tremens	Benign paroxysmal vertigo
Hypoglycemia	Apnea
Hypoxia	Night terrors
Psychoactive drugs (e.g., hallucinogens)	Sleepwalking
Migraine	
Confusional migraine	
Basilar migraine	

PSYCHOGENIC SEIZURES

Psychogenic seizures are nonepileptic behaviors that resemble seizures. They are often part of a conversion reaction precipitated by underlying psychological distress. Certain behaviors such as side-to-side turning of the head, asymmetric and large-amplitude shaking movements of the limbs, twitching of all four extremities without loss of consciousness, and pelvic thrusting are more commonly associated with psychogenic rather than epileptic seizures. Psychogenic seizures often last longer than epileptic seizures and may wax and wane over minutes to hours. However, the distinction is sometimes difficult on clinical grounds alone, and there are many examples of diagnostic errors made by experienced epileptologists. This is especially true for psychogenic seizures that resemble focal seizures, because the behavioral manifestations of focal seizures (especially of frontal lobe origin) can be extremely unusual, and in both cases, the routine surface EEG may be normal. Video-EEG monitoring is very useful when historic features are nondiagnostic. Generalized tonic-clonic seizures always

TABLE 418-7 Features That Distinguish Generalized Tonic-Clonic Seizure from Syncope

FEATURES	SEIZURE	SYNCOPE
Immediate precipitating factors	Usually none	Emotional stress, Valsalva, orthostatic hypotension, cardiac etiologies
Premonitory symptoms	None or aura (e.g., odd odor)	Tiredness, nausea, diaphoresis, tunneling of vision
Posture at onset	Variable	Usually erect
Transition to unconsciousness	Often immediate	Gradual over seconds[a]
Duration of unconsciousness	Minutes	Seconds
Duration of tonic or clonic movements	30–60 s	Never >15 s
Facial appearance during event	Cyanosis, frothing at mouth	Pallor
Disorientation and sleepiness after event	Many minutes to hours	<5 min
Aching of muscles after event	Often	Sometimes
Biting of tongue	Sometimes	Rarely
Incontinence	Sometimes	Sometimes
Headache	Sometimes	Rarely

[a]May be sudden with certain cardiac arrhythmias.

produce marked EEG abnormalities during and after the seizure. For suspected focal seizures of temporal lobe origin, the use of additional electrodes may help to localize a seizure focus. Measurement of serum prolactin levels may also help to distinguish between epileptic and psychogenic seizures, because most generalized seizures and some focal seizures are accompanied by rises in serum prolactin (during the immediate 30-min postictal period), whereas psychogenic seizures are not. The diagnosis of psychogenic seizures does not exclude a concurrent diagnosis of epilepsy, because the two may coexist.

TREATMENT

Seizures and Epilepsy

Therapy for a patient with a seizure disorder is almost always multimodal and includes treatment of underlying conditions that cause or contribute to the seizures, avoidance of precipitating factors, suppression of recurrent seizures by prophylactic therapy with antiepileptic medications or surgery, and addressing a variety of psychological and social issues. Treatment plans must be individualized, given the many different types and causes of seizures as well as the differences in efficacy and toxicity of antiepileptic medications for each patient. In almost all cases, a neurologist with experience in the treatment of epilepsy should design and oversee implementation of the treatment strategy. Furthermore, patients with refractory epilepsy or those who require polypharmacy with antiepileptic drugs should remain under the regular care of a neurologist.

TREATMENT OF UNDERLYING CONDITIONS

If the sole cause of a seizure is a metabolic disturbance such as an abnormality of serum electrolytes or glucose, then treatment is aimed at reversing the metabolic problem and preventing its recurrence. Therapy with antiepileptic drugs is usually unnecessary unless the metabolic disorder cannot be corrected promptly and the patient is at risk of having further seizures. If the apparent cause of a seizure was a medication (e.g., theophylline) or illicit drug use (e.g., cocaine), then appropriate therapy is avoidance of the drug; there is usually no need for antiepileptic medications unless subsequent seizures occur in the absence of these precipitants.

Seizures caused by a structural CNS lesion such as a brain tumor, vascular malformation, or brain abscess may not recur after appropriate treatment of the underlying lesion. However, despite removal of the structural lesion, there is a risk that the seizure focus will remain in the surrounding tissue or develop *de novo* as a result of gliosis and other processes induced by surgery, radiation, or other therapies. Most patients are therefore maintained on an antiepileptic medication for at least 1 year, and an attempt is made to withdraw medications only if the patient has been completely seizure free. If seizures are refractory to medication, the patient may benefit from surgical removal of the epileptic brain region (see below).

AVOIDANCE OF PRECIPITATING FACTORS

Unfortunately, little is known about the specific factors that determine precisely when a seizure will occur in a patient with epilepsy. An almost universal precipitating factor for seizures is sleep deprivation, so patients should do everything possible to optimize their sleep quality. Many patients can identify other particular situations that appear to lower their seizure threshold; these situations should be avoided. For example, patients may note an association between alcohol intake and seizures, and they should be encouraged to modify their drinking habits accordingly. There are also relatively rare cases of patients with seizures that are induced by highly specific stimuli such as a video game monitor, music, or an individual's voice ("reflex epilepsy"). Because there is often an association between stress and seizures, stress reduction techniques such as physical exercise, meditation, or counseling may be helpful.

ANTIEPILEPTIC DRUG THERAPY

Antiepileptic drug therapy is the mainstay of treatment for most patients with epilepsy. The overall goal is to completely prevent seizures without causing any untoward side effects, preferably with a single medication and a dosing schedule that is easy for the patient to follow. Seizure classification is an important element in designing the treatment plan, because some antiepileptic drugs have different activities against various seizure types. However, there is considerable overlap between many antiepileptic drugs such that the choice of therapy is often determined more by the patient's specific needs, especially his or her assessment of side effects.

When to Initiate Antiepileptic Drug Therapy Antiepileptic drug therapy should be started in any patient with recurrent seizures of unknown etiology or a known cause that cannot be reversed. Whether to initiate therapy in a patient with a single seizure is controversial. Patients with a single seizure due to an identified lesion such as a CNS tumor, infection, or trauma, in which there is strong evidence that the lesion is epileptogenic, should be treated. The risk of seizure recurrence in a patient with an apparently unprovoked or idiopathic seizure is uncertain, with estimates ranging from 31 to 71% in the first 12 months after the initial seizure. This uncertainty arises from differences in the underlying seizure types and etiologies in various published epidemiologic studies. Generally accepted risk factors associated with recurrent seizures include the following: (1) an abnormal neurologic examination, (2) seizures presenting as status epilepticus, (3) postictal Todd's paralysis, (4) a strong family history of seizures, or (5) an abnormal EEG. Most patients with one or more of these risk factors should be treated. Issues such as employment or driving may influence the decision whether to start medications as well. For example, a patient with a single, idiopathic seizure whose job depends on driving may prefer taking antiepileptic drugs rather than risk a seizure recurrence and the potential loss of driving privileges.

Selection of Antiepileptic Drugs Antiepileptic drugs available in the United States are shown in Table 418-8, and the main pharmacologic characteristics of commonly used drugs are listed in Table 418-9. Worldwide, older medications such as phenytoin, valproic acid, carbamazepine, phenobarbital, and ethosuximide are generally used as first-line therapy for most seizure disorders because, overall, they are as effective as recently marketed drugs and significantly less expensive overall. Most of the new drugs that have become available in the past decade are used as add-on or alternative therapy, although many are now being used as first-line monotherapy.

TABLE 418-8 Selection of Antiepileptic Drugs

GENERALIZED-ONSET TONIC-CLONIC	FOCAL	TYPICAL ABSENCE	ATYPICAL ABSENCE, MYOCLONIC, ATONIC
First-Line			
Lamotrigine Valproic acid	Lamotrigine Carbamazepine Oxcarbazepine Phenytoin Levetiracetam	Valproic acid Ethosuximide Lamotrigine	Valproic acid Lamotrigine Topiramate
Alternatives			
Zonisamide[a] Phenytoin Levetiracetam Carbamazepine Oxcarbazepine Topiramate Phenobarbital Primidone Felbamate	Zonisamide[a] Brivaracetam Topiramate Valproic acid Tiagabine[a] Gabapentin[a] Lacosamide[a] Phenobarbital Primidone Felbamate	Lamotrigine Clonazepam	Clonazepam Felbamate Clobazam Rufinamide

[a]As adjunctive therapy.

TABLE 418-9 Dosage and Adverse Effects of Commonly Used Antiepileptic Drugs

GENERIC NAME	TRADE NAME	PRINCIPAL USES	TYPICAL DOSE; DOSE INTERVAL	HALF-LIFE	THERAPEUTIC RANGE	ADVERSE EFFECTS		DRUG INTERACTIONS[a]
						NEUROLOGIC	SYSTEMIC	
Brivaracetam	Briviact	Focal-onset	100–200 mg/d; bid	7–10 h	Not established	Fatigue Dizziness Weakness Ataxia Mood changes	Gastrointestinal irritation	May increase carbamazepine-epoxide causing decreased tolerability May increase phenytoin
Carbamazepine	Tegretol[c]	Tonic-clonic Focal-onset	600–1800 mg/d (15–35 mg/kg, child); bid (capsules or tablets), tid-qid (oral suspension)	10–17 h (variable due to autoinduction: complete 3–5 wk after initiation)	4–12 μg/mL	Ataxia Dizziness Diplopia Vertigo	Aplastic anemia Leukopenia Gastrointestinal irritation Hepatotoxicity Hyponatremia	Level decreased by enzyme-inducing drugs[b] Level increased by erythromycin, propoxyphene, isoniazid, cimetidine, fluoxetine
Clobazam	Onfi	Lennox-Gastaut syndrome	10–40 mg/d (5–20 mg/d for patients <30 kg body weight); bid	36–42 h (71–82 h for less active metabolite)	Not established	Fatigue Sedation Ataxia Aggression Insomnia	Constipation Anorexia Skin rash	Level increased by CYP2C19 inhibitors
Clonazepam	Klonopin	Absence Atypical absence Myoclonic	1–12 mg/d; qd-tid	24–48 h	10–70 ng/mL	Ataxia Sedation Lethargy	Anorexia	Level decreased by enzyme-inducing drugs[b]
Ethosuximide	Zarontin	Absence	750–1250 mg/d (20–40 mg/kg); qd-bid	60 h, adult 30 h, child	40–100 μg/mL	Ataxia Lethargy Headache	Gastrointestinal irritation Skin rash Bone marrow suppression	Level decreased by enzyme-inducing drugs[b] Level increased by valproic acid
Felbamate	Felbatol	Focal-onset Lennox-Gastaut syndrome Tonic-clonic	2400–3600 mg/d, tid-qid	16–22 h	30–60 μg/mL	Insomnia Dizziness Sedation Headache	Aplastic anemia Hepatic failure Weight loss Gastrointestinal irritation	Increases phenytoin, valproic acid, active carbamazepine metabolite
Gabapentin	Neurontin	Focal-onset	900–2400 mg/d; tid-qid	5–9 h	2–20 μg/mL	Sedation Dizziness Ataxia Fatigue	Gastrointestinal irritation Weight gain Edema	No known significant interactions
Lacosamide	Vimpat	Focal-onset	200–400 mg/d; bid	13 h	Not established	Dizziness Ataxia Diplopia Vertigo	Gastrointestinal irritation Cardiac conduction (PR interval prolongation)	Level decreased by enzyme-inducing drugs[b]
Lamotrigine	Lamictal[c]	Focal-onset Tonic-clonic Atypical absence Myoclonic Lennox-Gastaut syndrome	150–500 mg/d; bid (immediate release), daily (extended release) (lower daily dose for regimens with valproic acid; higher daily dose for regimens with an enzyme inducer)	25 h 14 h (with enzyme-inducers), 59 h (with valproic acid)	2.5–20 μg/mL	Dizziness Diplopia Sedation Ataxia Headache	Skin rash Stevens-Johnson syndrome	Level decreased by enzyme-inducing drugs[b] and oral contraceptives Level increased by valproic acid
Levetiracetam	Keppra[c]	Focal-onset	1000–3000 mg/d; bid (immediate release), daily (extended release)	6–8 h	5–45 μg/mL	Sedation Fatigue Incoordination Mood changes	Anemia Leukopenia	No known significant interactions
Oxcarbazepine[c]	Trileptal	Focal-onset Tonic-clonic	900–2400 mg/d (30–45 mg/kg, child); bid	10–17 h (for active metabolite)	10–35 μg/mL	Fatigue Ataxia Dizziness Diplopia Vertigo Headache	See carbamazepine	Level decreased by enzyme-inducing drugs[b] May increase phenytoin

(*Continued*)

TABLE 418-9 Dosage and Adverse Effects of Commonly Used Antiepileptic Drugs (Continued)

GENERIC NAME	TRADE NAME	PRINCIPAL USES	TYPICAL DOSE; DOSE INTERVAL	HALF-LIFE	THERAPEUTIC RANGE	ADVERSE EFFECTS		DRUG INTERACTIONS[a]
						NEUROLOGIC	SYSTEMIC	
Phenobarbital	Luminal	Tonic-clonic Focal-onset	60–180 mg/d; qd-tid	90 h	10–40 µg/mL	Sedation Ataxia Confusion Dizziness Decreased libido Depression	Skin rash	Level increased by valproic acid, phenytoin
Phenytoin (diphenylhydantoin)	Dilantin	Tonic-clonic Focal-onset	300–400 mg/d (3–6 mg/kg, adult; 4–8 mg/kg, child); qd-tid	24 h (wide variation, dose-dependent)	10–20 µg/mL	Dizziness Diplopia Ataxia Incoordination Confusion	Gingival hyperplasia Lymphadenopathy Hirsutism Osteomalacia Facial coarsening Skin rash	Level increased by isoniazid, sulfonamides, fluoxetine Level decreased by enzyme-inducing drugs[b] Altered folate metabolism
Primidone	Mysoline	Tonic-clonic Focal-onset	750–1000 mg/d; bid-tid	Primidone, 8–15 h Phenobarbital, 90 h	Primidone, 4–12 µg/mL Phenobarbital, 10–40 µg/mL	Same as phenobarbital		Level increased by valproic acid Level decreased by phenytoin (increased conversion to phenobarbital)
Rufinamide	Banzel	Lennox-Gastaut syndrome	3200 mg/d (45 mg/kg, child); bid	6–10 h	Not established	Sedation Fatigue Dizziness Ataxia Headache Diplopia	Gastrointestinal irritation Leukopenia Cardiac conduction (QT interval shortening)	Level decreased by enzyme-inducing drugs[b] Level increased by valproic acid May increase phenytoin
Tiagabine	Gabitril	Focal-onset	32–56 mg/d; bid-qid (as adjunct to enzyme-inducing antiepileptic drug regimen)	2–5 h (with enzyme inducer), 7–9 h (without enzyme inducer)	Not established	Confusion Sedation Depression Dizziness Speech or language problems Paresthesias Psychosis	Gastrointestinal irritation	Level decreased by enzyme-inducing drugs[b]
Topiramate	Topamax	Focal-onset Tonic-clonic Lennox-Gastaut syndrome	200–400 mg/d; bid (immediate release), daily (extended release)	20 h (immediate release), 30 h (extended release)	2–20 µg/mL	Psychomotor slowing Sedation Speech or language problems Fatigue Paresthesias	Renal stones (avoid use with other carbonic anhydrase inhibitors) Glaucoma Weight loss Hypohidrosis	Level decreased by enzyme-inducing drugs[b]
Valproic acid (valproate sodium, divalproex sodium)	Depakene Depakote[c]	Tonic-clonic Absence Atypical absence Myoclonic Focal-onset Atonic	750–2000 mg/d (20–60 mg/kg); bid-qid (immediate and delayed release), daily (extended release)	15 h	50–125 µg/mL	Ataxia Sedation Tremor	Hepatotoxicity Thrombocytopenia Gastrointestinal irritation Weight gain Transient alopecia Hyperammonemia	Level decreased by enzyme-inducing drugs[b]
Zonisamide	Zonegran	Focal-onset Tonic-clonic	200–400 mg/d; qd-bid	50–68 h	10–40 µg/mL	Sedation Dizziness Confusion Headache Psychosis	Anorexia Renal stones Hypohidrosis	Level decreased by enzyme-inducing drugs[b]

[a]Examples only; please refer to other sources for comprehensive listings of all potential drug–drug interactions. [b]Phenytoin, carbamazepine, phenobarbital. [c]Extended-release product available.

In addition to efficacy, factors influencing the choice of an initial medication include the convenience of dosing (e.g., once daily versus three or four times daily) and potential side effects. In this regard, a number of the newer drugs have the advantage of reduced drug–drug interactions and easier dosing. Almost all of the commonly used antiepileptic drugs can cause similar, dose-related side effects such as sedation, ataxia, and diplopia. Long-term use of some agents in adults, especially the elderly, can lead to osteoporosis. Close follow-up is required to ensure these side effects are promptly recognized and reversed. Most of the older drugs and some of the newer ones can also cause idiosyncratic toxicity such as rash, bone marrow suppression, or hepatotoxicity. Although rare, these side effects should be considered during drug selection, and patients must be instructed about symptoms or signs that should signal the need to alert their health care provider. For some drugs, laboratory tests (e.g., complete blood count and liver function tests) are recommended prior to the institution of therapy (to establish baseline values) and during initial dosing and titration of the agent.

An important recent advance in the care of patients with epilepsy has been the application of genetic testing to help guide the choice of therapy (as well as establishing the underlying cause of a patient's syndrome). For example, the identification of a mutation in the *SLC2A1* gene, which encodes the glucose type 1 transporter (GLUT-1) and is a cause of GLUT-1 deficiency, should immediately prompt treatment with the ketogenic diet. Mutations of the *ALDH7A1* gene, which encodes antiquitin, can cause alterations in pyridoxine metabolism that are reversed by treatment with pyridoxine. There is also mounting evidence that certain gene mutations may indicate better or worse response to specific antiepileptic drugs. For example, patients with mutations in the sodium channel subunit *SCN1A* should generally avoid taking phenytoin or lamotrigine, whereas patients with mutations in the *SCN2A* or *SCN8A* sodium channel subunits appear to respond favorably to high-dose phenytoin. Genetic testing may also help predict antiepileptic drug toxicity. Studies have shown that Asian individuals carrying the human leukocyte antigen allele, HLA-B*1502, are at particularly high risk of developing serious skin reactions from carbamazepine, phenytoin, oxcarbazepine, and lamotrigine. HLA-A*31:01 has also been found to be associated with carbamazepine-induced hypersensitivity reactions in patients of European or Japanese ancestry. As a result, racial background and genotype are additional factors to consider in drug selection.

Antiepileptic Drug Selection for Focal Seizures Carbamazepine (or a related drug, oxcarbazepine), lamotrigine, phenytoin, and levetiracetam are currently the drugs of choice approved for the initial treatment of focal seizures, including those that evolve into generalized seizures. Overall they have very similar efficacy, but differences in pharmacokinetics and toxicity are the main determinants for use in a given patient. For example, an advantage of carbamazepine (which is also available in an extended-release form) is that its metabolism follows first-order pharmacokinetics, which allows for a linear relationship between drug dose, serum levels, and toxicity. Carbamazepine can cause leukopenia, aplastic anemia, or hepatotoxicity and would therefore be contraindicated in patients with predispositions to these problems. Oxcarbazepine has the advantage of being metabolized in a way that avoids an intermediate metabolite associated with some of the side effects of carbamazepine. Oxcarbazepine also has fewer drug interactions than carbamazepine. Lamotrigine tends to be well tolerated in terms of side effects. However, patients need to be particularly vigilant about the possibility of a skin rash during the initiation of therapy. This can be extremely severe and lead to Stevens-Johnson syndrome if unrecognized and if the medication is not discontinued immediately. This risk can be reduced by the use of low initial doses and slow titration. Lamotrigine must be started at lower initial doses when used as add-on therapy with valproic acid, because valproic acid inhibits lamotrigine metabolism and results in a substantially prolonged half-life. Phenytoin has a relatively long half-life and offers the advantage of once or twice daily dosing compared to two or three times daily dosing for many of the other drugs. However, phenytoin shows properties of nonlinear kinetics, such that small increases in phenytoin doses above a standard maintenance dose can precipitate marked side effects. This is one of the main causes of acute phenytoin toxicity. Long-term use of phenytoin is associated with untoward cosmetic effects (e.g., hirsutism, coarsening of facial features, gingival hypertrophy) and effects on bone metabolism. Due to these side effects, phenytoin is often avoided in young patients who are likely to require the drug for many years. Levetiracetam has the advantage of having no known drug–drug interactions, making it especially useful in the elderly and patients on other medications. However, a significant number of patients taking levetiracetam complain of irritability, anxiety, and other psychiatric symptoms. Topiramate can be used for both focal and generalized seizures. Similar to some of the other antiepileptic drugs, topiramate can cause significant psychomotor slowing and other cognitive problems. Additionally, it should not be used in patients at risk for the development of glaucoma or renal stones.

Valproic acid is an effective alternative for some patients with focal seizures, especially when the seizures generalize. Gastrointestinal side effects are fewer when using the delayed-release formulation (Depakote). Laboratory testing is required to monitor toxicity because valproic acid can rarely cause reversible bone marrow suppression and hepatotoxicity. This drug should generally be avoided in patients with preexisting bone marrow or liver disease. Valproic acid also has relatively high risks of unacceptable adverse effects for women of childbearing age, including hyperandrogenism that may affect fertility and teratogenesis (e.g., neural tube defects) in offspring. Irreversible, fatal hepatic failure appearing as an idiosyncratic rather than dose-related side effect is a relatively rare complication; its risk is highest in children <2 years old, especially those taking other antiepileptic drugs or with inborn errors of metabolism.

Zonisamide, brivaracetam, tiagabine, gabapentin, and lacosamide are additional drugs currently used for the treatment of focal seizures with or without evolution into generalized seizures. Phenobarbital and other barbiturate compounds were commonly used in the past as first-line therapy for many forms of epilepsy. However, the barbiturates frequently cause sedation in adults, hyperactivity in children, and other more subtle cognitive changes; thus, their use should be limited to situations in which no other suitable treatment alternatives exist.

Antiepileptic Drug Selection for Generalized Seizures Lamotrigine, valproic acid and levetiracetam are currently considered the best initial choice for the treatment of primary generalized, tonic-clonic seizures. Topiramate, zonisamide, phenytoin, carbamazepine, and oxcarbazepine are suitable alternatives, although carbamazepine, oxcarbazepine, and phenytoin can worsen certain types of generalized seizures. Valproic acid is particularly effective in absence, myoclonic, and atonic seizures. It is therefore commonly used in patients with generalized epilepsy syndromes having mixed seizure types. However, levetiracetam, rather than valproic acid, is increasingly considered the initial drug of choice for women with epilepsies having mixed seizure types given the adverse effects of valproic acid for women of childbearing age. Lamotrigine is also an alternative to valproate, especially for absence epilepsies. Ethosuximide is a particularly effective drug for the treatment of uncomplicated absence seizures, but it is not useful for tonic-clonic or focal seizures. Periodic monitoring of blood cell counts is required since ethosuximide rarely causes bone marrow suppression.

INITIATION AND MONITORING OF THERAPY

Because the response to any antiepileptic drug is unpredictable, patients should be carefully educated about the approach to therapy. The goal is to prevent seizures and minimize the side effects of treatment; determination of the optimal dose is often a matter of trial and error. This process may take months or longer if the baseline seizure frequency is low. Most antiepileptic drugs need to be introduced

relatively slow to minimize side effects. Patients should expect that minor side effects such as mild sedation, slight changes in cognition, or imbalance will typically resolve within a few days. Starting doses are usually the lowest value listed under the dosage column in **Table 418-9**. Subsequent increases should be made only after achieving a steady state with the previous dose (i.e., after an interval of five or more half-lives).

Monitoring of serum antiepileptic drug levels can be very useful for establishing the initial dosing schedule. However, the published therapeutic ranges of serum drug concentrations are only an approximate guide for determining the proper dose for a given patient. The key determinants are the clinical measures of seizure frequency and presence of side effects, not the laboratory values. Conventional assays of serum drug levels measure the total drug (i.e., both free and protein bound). However, it is the concentration of free drug that reflects extracellular levels in the brain and correlates best with efficacy. Thus, patients with decreased levels of serum proteins (e.g., decreased serum albumin due to impaired liver or renal function) may have an increased ratio of free to bound drug, yet the concentration of free drug may be adequate for seizure control. These patients may have a "subtherapeutic" drug level, but the dose should be changed only if seizures remain uncontrolled, not just to achieve a "therapeutic" level. It is also useful to monitor free drug levels in such patients. In practice, other than during the initiation or modification of therapy, monitoring of antiepileptic drug levels is most useful for documenting adherence or assessing clinical suspicion of toxicity.

If seizures continue despite gradual increases to the maximum tolerated dose and documented compliance, then it becomes necessary to switch to another antiepileptic drug. This is usually done by maintaining the patient on the first drug while a second drug is added. The dose of the second drug should be adjusted to decrease seizure frequency without causing toxicity. Once this is achieved, the first drug can be gradually withdrawn (usually over weeks unless there is significant toxicity). The dose of the second drug is then further optimized based on seizure response and side effects. Monotherapy should be the goal whenever possible.

WHEN TO DISCONTINUE THERAPY

Overall, about 50–60% of patients who have their seizures completely controlled with antiepileptic drugs can eventually discontinue therapy. The following patient profile yields the greatest chance of remaining seizure free after drug withdrawal: (1) complete medical control of seizures for 1–5 years; (2) single seizure type, with generalized seizures having a better prognosis than focal seizures; (3) normal neurologic examination, including intelligence; (4) no family history of epilepsy; and (5) normal EEG. The appropriate seizure-free interval is unknown and undoubtedly varies for different forms of epilepsy. However, it seems reasonable to attempt withdrawal of therapy after 2 years in a patient who meets all of the above criteria, is motivated to discontinue the medication, and clearly understands the potential risks and benefits. In most cases, it is preferable to reduce the dose of the drug gradually over 2–3 months. Most recurrences occur in the first 3 months after discontinuing therapy, and patients should be advised to avoid potentially dangerous situations such as driving or swimming during this period.

TREATMENT OF REFRACTORY EPILEPSY

Approximately one-third of patients with epilepsy do not respond to treatment with a single antiepileptic drug, and it becomes necessary to try a combination of drugs to control seizures. Patients who have focal epilepsy related to an underlying structural lesion or those with multiple seizure types and developmental delay are particularly likely to require multiple drugs. There are currently no clear guidelines for rational polypharmacy, although in theory a combination of drugs with different mechanisms of action may be most useful. In most cases, the initial combination therapy combines first-line drugs (i.e., carbamazepine, oxcarbazepine, lamotrigine, valproic acid, levetiracetam, and phenytoin). If these drugs are unsuccessful, then the addition of other drugs such as zonisamide, brivaracetam, topiramate, lacosamide, or tiagabine is indicated. Patients with myoclonic seizures resistant to valproic acid may benefit from the addition of clonazepam or clobazam, and those with absence seizures may respond to a combination of valproic acid and ethosuximide. The same principles concerning the monitoring of therapeutic response, toxicity, and serum levels for monotherapy apply to polypharmacy, and potential drug interactions need to be recognized. If there is no improvement, a third drug can be added while the first two are maintained. If there is a response, the less effective or less well tolerated of the first two drugs should be gradually withdrawn.

SURGICAL TREATMENT OF REFRACTORY EPILEPSY

Approximately 20–30% of patients with epilepsy continue to have seizures despite efforts to find an effective combination of antiepileptic drugs. For some, surgery can be extremely effective in substantially reducing seizure frequency and even providing complete seizure control. Understanding the potential value of surgery is especially important when a patient's seizures are not controlled with initial treatment, as such patients often do not respond to subsequent medication trials. Rather than submitting the patient to years of unsuccessful medical therapy and the psychosocial trauma and increased mortality associated with ongoing seizures, the patient should have an efficient but relatively brief attempt at medical therapy and then be referred for surgical evaluation.

The most common surgical procedure for patients with temporal lobe epilepsy involves resection of the anteromedial temporal lobe (temporal lobectomy) or a more limited removal of the underlying hippocampus and amygdala (amygdalohippocampectomy). Focal seizures arising from extratemporal regions may be abolished by a focal neocortical resection with precise removal of an identified lesion (lesionectomy). Localized neocortical resection without a clear lesion identified on MRI is also possible when other tests (e.g., MEG, PET, SPECT) implicate a focal cortical region as a seizure onset zone. When the cortical region cannot be removed, multiple subpial transection, which disrupts intracortical connections, is sometimes used to prevent seizure spread. Hemispherectomy or multilobar resection is useful for some patients with severe seizures due to hemispheric abnormalities such as hemimegalencephaly or other dysplastic abnormalities, and corpus callosotomy has been shown to be effective for disabling tonic or atonic seizures, usually when they are part of a mixed-seizure syndrome (e.g., Lennox-Gastaut syndrome).

Presurgical evaluation is designed to identify the functional and structural basis of the patient's seizure disorder. Inpatient video-EEG monitoring is used to define the anatomic location of the seizure focus and to correlate the abnormal electrophysiologic activity with behavioral manifestations of the seizure. Routine scalp or scalp-sphenoidal recordings and a high-resolution MRI scan are usually sufficient for localization of the epileptogenic focus, especially when the findings are concordant. Functional imaging studies such as SPECT, PET, and MEG are adjunctive tests that may help to reveal or verify the localization of an apparent epileptogenic region. Once the presumed location of the seizure onset is identified, additional studies, including neuropsychological testing, the intracarotid amobarbital test (Wada test), and functional MRI may be used to assess language and memory localization and to determine the possible functional consequences of surgical removal of the epileptogenic region. In some cases, standard noninvasive evaluation is not sufficient to localize the seizure onset zone, and invasive electrophysiologic monitoring, such as implanted depth or subdural electrodes, is required for more definitive localization. The exact extent of the resection to be undertaken can also be determined by performing cortical mapping at the time of the surgical procedure, allowing for a tailored resection. This involves electrocorticographic recordings made with electrodes on the surface of the brain to identify the extent of epileptiform disturbances. If the region to be resected is within or near brain regions suspected of

having sensorimotor or language function, electrical cortical stimulation mapping is performed on the awake patient to determine the function of cortical regions in question in order to avoid resection of so-called eloquent cortex and thereby minimize postsurgical deficits.

Advances in presurgical evaluation and microsurgical techniques have led to a steady increase in the success of epilepsy surgery. Clinically significant complications of surgery are <5%, and the use of functional mapping procedures has markedly reduced the neurologic sequelae due to removal or sectioning of brain tissue. For example, about 70% of patients treated with temporal lobectomy will become seizure free, and another 15–25% will have at least a 90% reduction in seizure frequency. Marked improvement is also usually seen in patients treated with hemispherectomy for catastrophic seizure disorders due to large hemispheric abnormalities. Postoperatively, patients generally need to remain on antiepileptic drug therapy, but the marked reduction of seizures following resective surgery can have a very beneficial effect on quality of life.

Not all medically refractory patients are suitable candidates for resective surgery. For example, some patients have seizures arising from more than one location, making the risk of ongoing seizures or potential harm from the surgery unacceptably high. Vagus nerve stimulation (VNS) has been used in some of these cases, although the results are limited and it is difficult to predict who will benefit. An implantable device that can detect the onset of a seizure (in some instances before the seizure becomes clinically apparent) and deliver an electrical stimulation to abort the seizure (Responsive NeuroStimulation) has proved to be of benefit in selected patients. Studies are currently evaluating the efficacy of stereotactic radiosurgery, laser thermoablation, and deep brain stimulation (DBS) as other options for surgical treatment of refractory epilepsy.

STATUS EPILEPTICUS

Status epilepticus refers to continuous seizures or repetitive, discrete seizures with impaired consciousness in the interictal period. Status epilepticus has numerous subtypes, including generalized convulsive status epilepticus (GCSE) (e.g., persistent, generalized electrographic seizures, coma, and tonic-clonic movements) and nonconvulsive status epilepticus (e.g., persistent absence seizures or focal seizures with confusion or partially impaired consciousness, and minimal motor abnormalities). The duration of seizure activity sufficient to meet the definition of status epilepticus has traditionally been specified as 15–30 min. However, a more practical definition is to consider status epilepticus as a situation in which the duration of seizures prompts the acute use of anticonvulsant therapy. For GCSE, this is typically when seizures last beyond 5 min.

GCSE is an emergency and must be treated immediately, because cardiorespiratory dysfunction, hyperthermia, and metabolic derangements can develop as a consequence of prolonged seizures, and these can lead to irreversible neuronal injury. Furthermore, CNS injury can occur even when the patient is paralyzed with neuromuscular blockade but continues to have electrographic seizures. The most common causes of GCSE are anticonvulsant withdrawal or noncompliance, metabolic disturbances, drug toxicity, CNS infection, CNS tumors, refractory epilepsy, and head trauma.

GCSE is obvious when the patient is having overt seizures. However, after 30–45 min of uninterrupted seizures, the signs may become increasingly subtle. Patients may have mild clonic movements of only the fingers or fine, rapid movements of the eyes. There may be paroxysmal episodes of tachycardia, hypertension, and pupillary dilation. In such cases, the EEG may be the only method of establishing the diagnosis. Thus, if the patient stops having overt seizures, yet remains comatose, an EEG should be performed to rule out ongoing status epilepticus. This is obviously also essential when a patient with GCSE has been paralyzed with neuromuscular blockade in the process of protecting the airway.

The first steps in the management of a patient in GCSE are to attend to any acute cardiorespiratory problems or hyperthermia, perform a brief medical and neurologic examination, establish venous access, and send samples for laboratory studies to identify metabolic abnormalities. Anticonvulsant therapy should then begin without delay; a treatment approach is shown in **Fig. 418-5**.

The treatment of nonconvulsive status epilepticus is thought to be less urgent than GCSE, because the ongoing seizures are not accompanied by the severe metabolic disturbances seen with GCSE. However, evidence suggests that nonconvulsive status epilepticus, especially that caused by ongoing, focal seizure activity, is associated with cellular injury in the region of the seizure focus; therefore, this condition should be treated as promptly as possible using the general approach described for GCSE.

BEYOND SEIZURES: OTHER MANAGEMENT ISSUES

EPILEPSY COMORBIDITIES

The adverse effects of epilepsy often go beyond clinical seizures. Many epilepsy patients are completely normal between seizures and live highly successful and productive lives. However, a significant proportion of patients suffer from varying degrees of cognitive dysfunction, including psychiatric disease, and it has become increasingly clear that the network dysfunction underlying epilepsy can have effects well beyond the occurrence of seizures. For example, patients with seizures secondary to developmental abnormalities or acquired brain injury may have impaired cognitive function and other neurologic deficits due to abnormal brain structure. Frequent interictal EEG abnormalities are associated with subtle dysfunction of memory and attention. Patients with many seizures, especially those emanating from the temporal lobe, often note an impairment of short-term memory that may progress over time.

The psychiatric problems associated with epilepsy include depression, anxiety, and psychosis. This risk varies considerably depending on many factors, including the etiology, frequency, and severity of seizures and the patient's age and previous personal or family history of psychiatric disorder. Depression occurs in ~20% of patients, and the incidence of suicide is higher in patients with epilepsy than in the general population. Depression should be treated through counseling or medication. The selective serotonin reuptake inhibitors (SSRIs) typically have minimal effect on seizures, whereas tricyclic antidepressants may lower the seizure threshold. Anxiety can be a seizure symptom, and anxious or psychotic behavior can occur during a postictal delirium. Postictal psychosis is a rare phenomenon that typically occurs after a period of increased seizure frequency. There is usually a brief lucid interval lasting up to a week, followed by days to weeks of agitated, psychotic behavior. The psychosis usually resolves spontaneously but frequently will require short-term treatment with antipsychotic or anxiolytic medications.

MORTALITY OF EPILEPSY

Patients with epilepsy have a risk of death that is roughly two to three times greater than expected in a matched population without epilepsy. Most of the increased mortality is due to the underlying etiology of epilepsy (e.g., tumors or strokes in older adults). However, a significant number of patients die from accidents, status epilepticus, and a syndrome known as *sudden unexpected death in epilepsy* (SUDEP), which usually affects young people with convulsive seizures and tends to occur at night. The cause of SUDEP is unknown; it may result from brainstem-mediated effects of seizures on pulmonary, cardiac, and arousal functions. Recent studies suggest that, in some cases, a genetic mutation may be the cause of both epilepsy and a cardiac conduction defect that gives rise to sudden death.

PSYCHOSOCIAL ISSUES

There continues to be a cultural stigma about epilepsy, although it is slowly declining in societies with effective health education programs. Many patients with epilepsy harbor fears such as the fear of becoming mentally retarded or dying during a seizure. These issues need to be carefully addressed by educating the patient about epilepsy and by ensuring that family members, teachers, fellow employees, and other associates are equally well informed. A useful source of educational material is the website *www.epilepsy.com*.

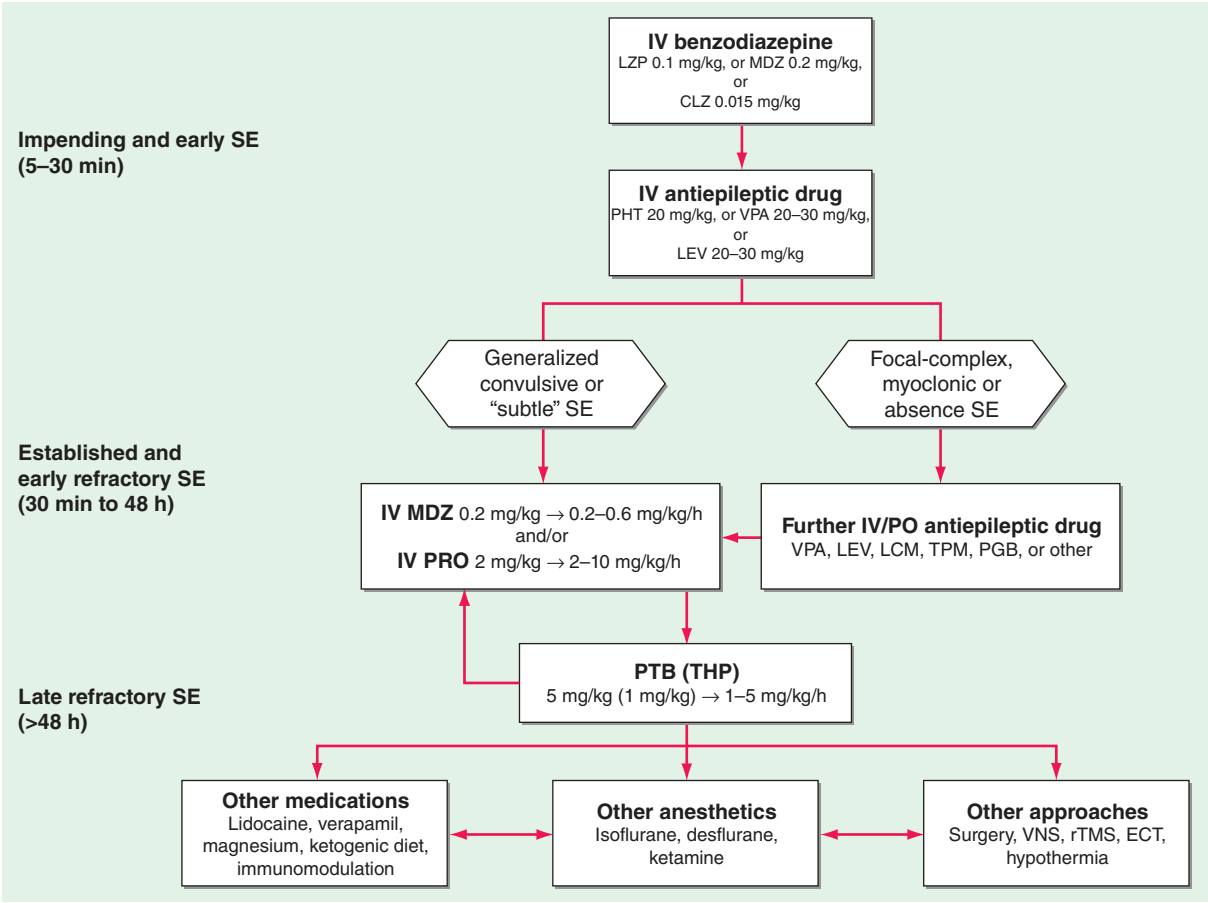


FIGURE 418-5 Pharmacologic treatment of generalized tonic-clonic status epilepticus (SE) in adults. CLZ, clonazepam; ECT, electroconvulsive therapy; LCM, lacosamide; LEV, levetiracetam; LZP, lorazepam; MDZ, midazolam; PGB, pregabalin; PHT, phenytoin or fos-phenytoin; PRO, propofol; PTB, pentobarbital; rTMS, repetitive transcranial magnetic stimulation; THP, thiopental; TPM, topiramate; VNS, vagus nerve stimulation; VPA, valproic acid. *(From AO Rossetti, DH Lowenstein: Lancet Neurol 10:922, 2011.)*

EMPLOYMENT, DRIVING, AND OTHER ACTIVITIES

Many patients with epilepsy face difficulty in obtaining or maintaining employment, even when their seizures are well controlled. Federal and state legislation is designed to prevent employers from discriminating against patients with epilepsy, and patients should be encouraged to understand and claim their legal rights. Patients in these circumstances also benefit greatly from the assistance of health providers who act as strong patient advocates.

Loss of driving privileges is one of the most disruptive social consequences of epilepsy. Physicians should be very clear about local regulations concerning driving and epilepsy, because the laws vary considerably among states and countries. In all cases, it is the physician's responsibility to warn patients of the danger imposed on themselves and others while driving if their seizures are uncontrolled (unless the seizures are not associated with impairment of consciousness or motor control). In general, most states allow patients to drive after a seizure-free interval (on or off medications) of between 3 months and 2 years.

Patients with incompletely controlled seizures must also contend with the risk of being in other situations where an impairment of consciousness or loss of motor control could lead to major injury or death. Thus, depending on the type and frequency of seizures, many patients need to be instructed to avoid working at heights or with machinery or to have someone close by for activities such as bathing and swimming.

SPECIAL ISSUES RELATED TO WOMEN AND EPILEPSY

CATAMENIAL EPILEPSY

Some women experience a marked increase in seizure frequency around the time of menses. This is believed to be mediated by either the effects of estrogen and progesterone on neuronal excitability or changes in antiepileptic drug levels due to altered protein binding or metabolism. Some patients may benefit from increases in antiepileptic drug dosages during menses. Natural progestins or intramuscular medroxyprogesterone may be of benefit to a subset of women.

PREGNANCY

Most women with epilepsy who become pregnant will have an uncomplicated gestation and deliver a normal baby. However, epilepsy poses some important risks to a pregnancy. Seizure frequency during pregnancy will remain unchanged in ~50% of women, increase in ~30%, and decrease in ~20%. Changes in seizure frequency are attributed to endocrine effects on the CNS, variations in antiepileptic drug pharmacokinetics (such as acceleration of hepatic drug metabolism or effects on plasma protein binding), and changes in medication compliance. It is useful to see patients at frequent intervals during pregnancy and monitor serum antiepileptic drug levels. Measurement of the unbound drug concentrations may be useful if there is an increase in seizure frequency or worsening of side effects of antiepileptic drugs.

The overall incidence of fetal abnormalities in children born to mothers with epilepsy is 5–6%, compared to 2–3% in healthy women. Part of the higher incidence is due to teratogenic effects of antiepileptic drugs, and the risk increases with the number of medications used (e.g., 10–20% risk of malformations with three drugs) and possibly with higher doses. A meta-analysis of published pregnancy registries and cohorts found that the most common malformations were defects in the cardiovascular and musculoskeletal system (1.4–1.8%). Valproic acid is strongly associated with an increased risk of adverse fetal outcomes (7–20%). Findings from a large pregnancy registry suggest that, other than topiramate, the newer antiepileptic drugs are far safer than valproic acid.

Because the potential harm of uncontrolled convulsive seizures on the mother and fetus is considered greater than the teratogenic

effects of antiepileptic drugs, it is currently recommended that pregnant women be maintained on effective drug therapy. When possible, it seems prudent to have the patient on monotherapy at the lowest effective dose, especially during the first trimester. For some women, however, the type and frequency of their seizures may allow for them to safely wean off antiepileptic drugs prior to conception. Patients should also take folate (1–4 mg/d), because the antifolate effects of anticonvulsants are thought to play a role in the development of neural tube defects, although the benefits of this treatment remain unproved in this setting.

Enzyme-inducing drugs such as phenytoin, carbamazepine, oxcarbazepine, topiramate, phenobarbital, and primidone cause a transient and reversible deficiency of vitamin K–dependent clotting factors in ~50% of newborn infants. Although neonatal hemorrhage is uncommon, the mother should be treated with oral vitamin K (20 mg/d, phylloquinone) in the last 2 weeks of pregnancy, and the infant should receive intramuscular vitamin K (1 mg) at birth.

CONTRACEPTION

Special care should be taken when prescribing antiepileptic medications for women who are taking oral contraceptive agents. Drugs such as carbamazepine, phenytoin, phenobarbital, and topiramate can significantly decrease the efficacy of oral contraceptives via enzyme induction and other mechanisms. Patients should be advised to consider alternative forms of contraception, or their contraceptive medications should be modified to offset the effects of the antiepileptic medications.

BREAST-FEEDING

Antiepileptic medications are excreted into breast milk to a variable degree. The ratio of drug concentration in breast milk relative to serum ranges from ~5% (valproic acid) to 300% (levetiracetam). Given the overall benefits of breast-feeding and the lack of evidence for long-term harm to the infant by being exposed to antiepileptic drugs, mothers with epilepsy can be encouraged to breast-feed. This should be reconsidered, however, if there is any evidence of drug effects on the infant such as lethargy or poor feeding.

Acknowledgment

Dr. Michael J. Aminoff contributed to the section on EEG interpretation in earlier editions.

FURTHER READING

Chen DK et al: Psychogenic non-epileptic seizures. Curr Neurol Neurosci Rep 17:71, 2017.

Crepeau AZ, Sirven JI: Management of adult onset seizures. Mayo Clin Proc 92:306, 2017.

Dalkilic EB: Neurostimulation devices used in treatment of epilepsy. Curr Treat Options Neurol 19:7, 2017.

Epi PM Consortium: A roadmap for precision medicine in the epilepsies. Lancet Neurol 14:1219, 2015.

Gavvala JR, Schuele SU: New-onset seizure in adults and adolescents: A review. JAMA 316:2657, 2016.

Golyala A, Kwan P: Drug development for refractory epilepsy: The past 25 years and beyond. Seizure 44:147, 2017.

Jetté N et al: Surgical treatment for epilepsy: The potential gap between evidence and practice. Lancet Neurol 15:982, 2016.

Keezer MR et al: Comorbidities of epilepsy: Current concepts and future perspectives. Lancet Neurol 15:106, 2016.

Lamberink HJ et al: Individualised prediction model of seizure recurrence and long-term outcomes after withdrawal of antiepileptic drugs in seizure-free patients: A systematic review and individual participant data meta-analysis. Lancet Neurol 16:523, 2017.

McGovern RA et al: New techniques and progress in epilepsy surgery. Curr Neurol Neurosci Rep 16:65, 2016.

Pitkänen A et al: Advances in the development of biomarkers for epilepsy. Lancet Neurol 15:843, 2016.

419 Cerebrovascular Diseases

Wade S. Smith, S. Claiborne Johnston, J. Claude Hemphill, III

Cerebrovascular diseases include some of the most common and devastating disorders: ischemic stroke and hemorrhagic stroke. Stroke is the second leading cause of death worldwide, with 6.2 million dying from stroke in 2015, an increase of 830,000 since the year 2000. While stroke has grown in incidence worldwide, it is declining among the affluent and rising among those with less access to medical care. In the United States, the incidence of stroke has declined steadily since at least 1958, and stroke is currently the fifth leading cause of death with 133,000 dying in 2014. Despite this progress, however, stroke remains the most common disabling disease in the United States and in many forms is preventable.

A stroke, or cerebrovascular accident, is defined as an abrupt onset of a neurologic deficit that is attributable to a focal vascular cause. Thus, the definition of stroke is clinical, and laboratory studies including brain imaging are used to support the diagnosis. The clinical manifestations of stroke are highly variable because of the complex anatomy of the brain and its vasculature. *Cerebral ischemia* is caused by a reduction in blood flow that lasts longer than several seconds. Neurologic symptoms are manifest within seconds because neurons lack glycogen, so energy failure is rapid. If the cessation of flow lasts for more than a few minutes, *infarction* or death of brain tissue results. When blood flow is quickly restored, brain tissue can recover fully and the patient's symptoms are only transient: this is called a *transient ischemic attack* (TIA). The definition of TIA requires that all neurologic signs and symptoms resolve within 24 h without evidence of brain infarction on brain imaging. Stroke has occurred if the neurologic signs and symptoms last for >24 h or brain infarction is demonstrated. A generalized reduction in cerebral blood flow due to systemic hypotension (e.g., cardiac arrhythmia, myocardial infarction, or hemorrhagic shock) usually produces syncope **(Chap. 18)**. If low cerebral blood flow persists for a longer duration, then infarction in the border zones between the major cerebral artery distributions may develop. In more severe instances, *global hypoxia-ischemia* causes widespread brain injury; the constellation of cognitive sequelae that ensues is called *hypoxic-ischemic encephalopathy* **(Chap. 301)**. *Focal ischemia* or infarction, conversely, is usually caused by thrombosis of the cerebral vessels themselves or by emboli from a proximal arterial source or the heart **(Chap. 420)**. *Intracranial hemorrhage* is caused by bleeding directly into or around the brain; it produces neurologic symptoms by producing a mass effect on neural structures, from the toxic effects of blood itself, or by increasing intracranial pressure **(Chap. 421)**.

APPROACH TO THE PATIENT

Cerebrovascular Disease

Rapid evaluation is essential for use of acute treatments such as thrombolysis or thrombectomy. However, patients with acute stroke often do not seek medical assistance on their own because they may lose the appreciation that something is wrong (anosognosia) or lack the knowledge that acute treatment is beneficial; it is often a family member or a bystander who calls for help. Therefore, patients and their family members should be counseled to call emergency medical services immediately if they experience or witness the sudden onset of any of the following: loss of sensory and/or motor function on one side of the body (nearly 85% of ischemic stroke patients have hemiparesis); change in vision, gait, or ability to speak or understand; or a sudden, severe headache. The acronym FAST (Facial weakness, Arm weakness, Speech abnormality and Time) is simple and helpful to teach to the lay public about the common physical

symptoms of stroke and to underscore that treatments are highly time-sensitive.

Other causes of sudden-onset neurologic symptoms that may mimic stroke include seizure, intracranial tumor, migraine, and metabolic encephalopathy. An adequate history from an observer that no convulsive activity occurred at the onset usually excludes seizure, although ongoing complex partial seizures without tonic-clonic activity can on occasion mimic stroke. Tumors may present with acute neurologic symptoms due to hemorrhage, seizure, or hydrocephalus. Surprisingly, migraine **(Chap. 422)** can mimic stroke, even in patients without a significant migraine history. When migraine develops without head pain (*acephalgic migraine*), the diagnosis can be especially difficult. Patients without any prior history of migraine may develop acephalgic migraine even after age 65. A sensory disturbance is often prominent, and the sensory deficit, as well as any motor deficits, tends to migrate slowly across a limb, over minutes rather than seconds as with stroke. The diagnosis of migraine becomes more secure as the cortical disturbance begins to cross vascular boundaries or if classic visual symptoms are present such as scintillating scotomata. At times, it may be impossible to make the diagnosis of migraine until there have been multiple episodes with no residual symptoms or signs and no changes on brain magnetic resonance imaging (MRI). Metabolic encephalopathies typically produce fluctuating mental status changes without focal neurologic findings. However, in the setting of prior stroke or brain injury, a patient with fever or sepsis may manifest a recurrent hemiparesis, which clears rapidly when the infection is treated. The metabolic process serves to "unmask" a prior deficit.

Once the diagnosis of stroke is made, a brain imaging study is necessary to determine if the cause of stroke is ischemia or hemorrhage **(Fig. 419-1)**. Computed tomography (CT) imaging of the brain is the standard imaging modality to detect the presence or absence of intracranial hemorrhage (see "Imaging Studies," below). If the stroke is ischemic, administration of recombinant tissue plasminogen activator (rtPA) or endovascular mechanical thrombectomy may be beneficial in restoring cerebral perfusion **(Chap. 420)**. Medical management to reduce the risk of complications becomes the next priority, followed by plans for secondary prevention. For ischemic stroke, several strategies can reduce the risk of subsequent stroke in all patients, while other strategies are effective for patients with specific causes of stroke such as cardiac embolus and carotid atherosclerosis. For hemorrhagic stroke, aneurysmal subarachnoid hemorrhage (SAH) and hypertensive intracerebral hemorrhage are two important causes. **The treatment and prevention of hypertensive intracerebral hemorrhage are discussed in Chap. 421. SAH is discussed in Chap. 302.**

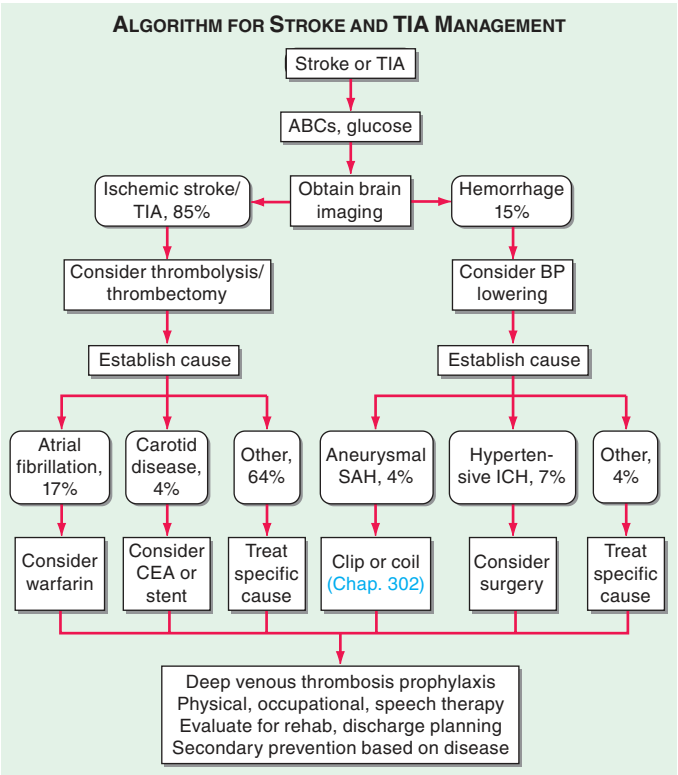


FIGURE 419-1 Medical management of stroke and TIA. *Rounded boxes* are diagnoses; *rectangles* are interventions. Numbers are percentages of stroke overall. ABCs, airway, breathing, circulation; BP, blood pressure; CEA, carotid endarterectomy; ICH, intracerebral hemorrhage; SAH, subarachnoid hemorrhage; TIA, transient ischemic attack.

STROKE SYNDROMES

A careful history and neurologic examination can often localize the region of brain dysfunction; if this region corresponds to an arterial distribution, the possible causes responsible for the syndrome can be narrowed. This is of particular importance when the patient presents with a TIA and a normal examination. For example, if a patient develops language loss and a right homonymous hemianopia, a search for causes of left middle cerebral emboli should be performed. A finding of an isolated stenosis of the right internal carotid artery in that patient, for example, suggests an asymptomatic carotid stenosis, and the search for other causes of stroke should continue. The following sections describe the clinical findings of cerebral ischemia associated with cerebral vascular territories depicted in **Figs. 419-2 through 419-11**. Stroke syndromes are divided into: (1) large-vessel stroke within the anterior circulation, (2) large-vessel stroke within the posterior circulation, and (3) small-vessel disease of either vascular bed.

Stroke within the Anterior Circulation The internal carotid artery and its branches comprise the anterior circulation of the brain. These vessels can be occluded by intrinsic disease of the vessel (e.g., atherosclerosis or dissection) or by embolic occlusion from a proximal source as discussed above. Occlusion of each major intracranial vessel has distinct clinical manifestations.

MIDDLE CEREBRAL ARTERY Occlusion of the proximal middle cerebral artery (MCA) or one of its major branches is most often due to an embolus (artery-to-artery, cardiac, or of unknown source) rather than intracranial atherothrombosis. Atherosclerosis of the proximal MCA may cause distal emboli to the middle cerebral territory or, less commonly, may produce low-flow TIAs. Collateral formation via leptomeningeal vessels often prevents MCA stenosis from becoming symptomatic.

The cortical branches of the MCA supply the lateral surface of the hemisphere except for (1) the frontal pole and a strip along the superomedial border of the frontal and parietal lobes supplied by the anterior cerebral artery (ACA) and (2) the lower temporal and occipital pole convolutions supplied by the posterior cerebral artery (PCA) (Figs. 419-2–419-5).

The proximal MCA (M1 segment) gives rise to penetrating branches (termed *lenticulostriate arteries*) that supply the putamen, outer globus pallidus, posterior limb of the internal capsule, adjacent corona radiata, and most of the caudate nucleus **(**Fig. 419-2**)**. In the sylvian fissure, the MCA in most patients divides into *superior* and *inferior* divisions (M2 branches). Branches of the inferior division supply the inferior parietal and temporal cortex, and those from the superior division supply the frontal and superior parietal cortex (Fig. 419-3).

If the entire MCA is occluded at its origin (blocking both its penetrating and cortical branches) and the distal collaterals are limited, the clinical findings are contralateral hemiplegia, hemianesthesia, homonymous hemianopia, and a day or two of gaze preference to the ipsilateral side. Dysarthria is common because of facial weakness. When the dominant hemisphere is involved, global aphasia is present

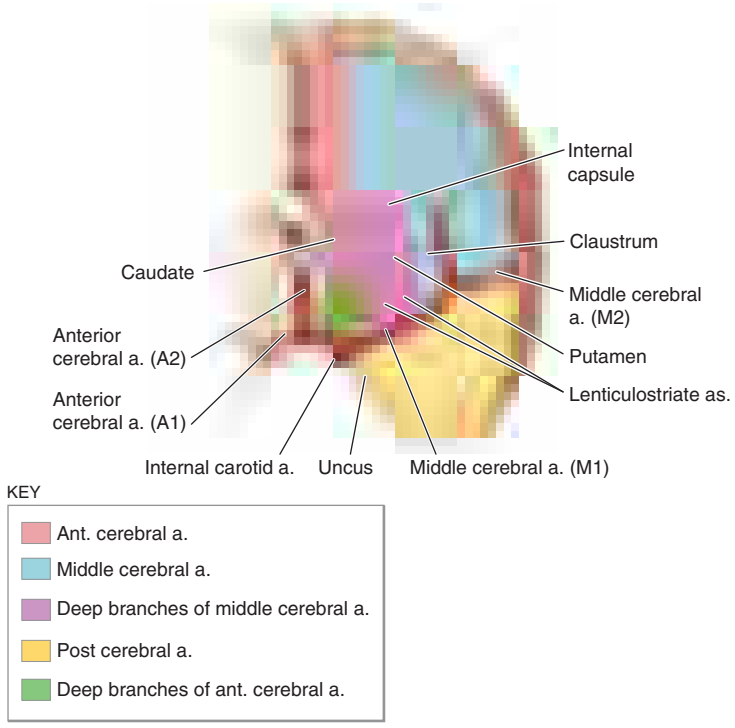


FIGURE 419-2 Diagram of a cerebral hemisphere in coronal section showing the territories of the major cerebral vessels that branch from the internal carotid arteries.

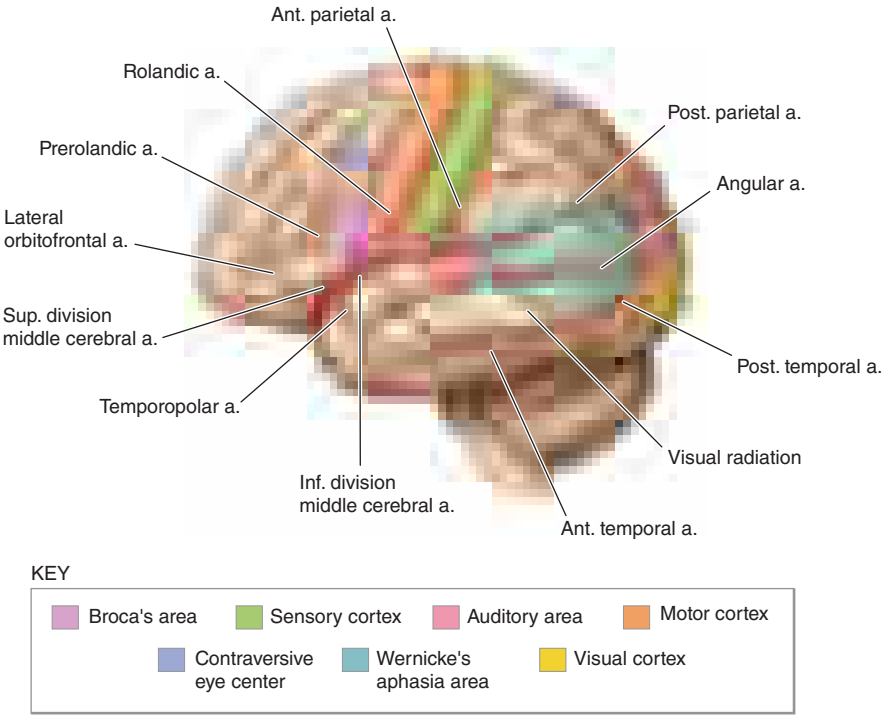


FIGURE 419-3 Diagram of a cerebral hemisphere, lateral aspect, showing the branches and distribution of the middle cerebral artery (MCA) and the principal regions of cerebral localization. Note the bifurcation of the MCA into a superior and inferior division.

Signs and symptoms: *Structures involved*

Paralysis of the contralateral face, arm, and leg; sensory impairment over the same area (pinprick, cotton touch, vibration, position, two-point discrimination, stereognosis, tactile localization, barognosis, cutaneographia): *Somatic motor area for face and arm and the fibers descending from the leg area to enter the corona radiata and corresponding somatic sensory system*

Motor aphasia: Motor speech area of the dominant hemisphere

Central aphasia, word deafness, anomia, jargon speech, sensory agraphia, acalculia, alexia, finger agnosia, right-left confusion (the last four comprise the Gerstmann syndrome): Central, suprasylvian speech area and parietooccipital cortex of the dominant hemisphere

Conduction aphasia: Central speech area (parietal operculum)

Apractagnosia of the nondominant hemisphere, anosognosia, hemiasomatognosia, unilateral neglect, agnosia for the left half of external space, dressing "apraxia," constructional "apraxia," distortion of visual coordinates, inaccurate localization in the half field, impaired ability to judge distance, upside-down reading, visual illusions (e.g., it may appear that another person walks through a table): Nondominant parietal lobe (area corresponding to speech *area in dominant hemisphere); loss of topographic memory is usually due to a nondominant lesion, occasionally to a dominant one*

Homonymous hemianopia (often homonymous inferior quadrantanopia): *Optic radiation deep to second temporal convolution*

Paralysis of conjugate gaze to the opposite side: *Frontal contraversive eye field or projecting fibers*

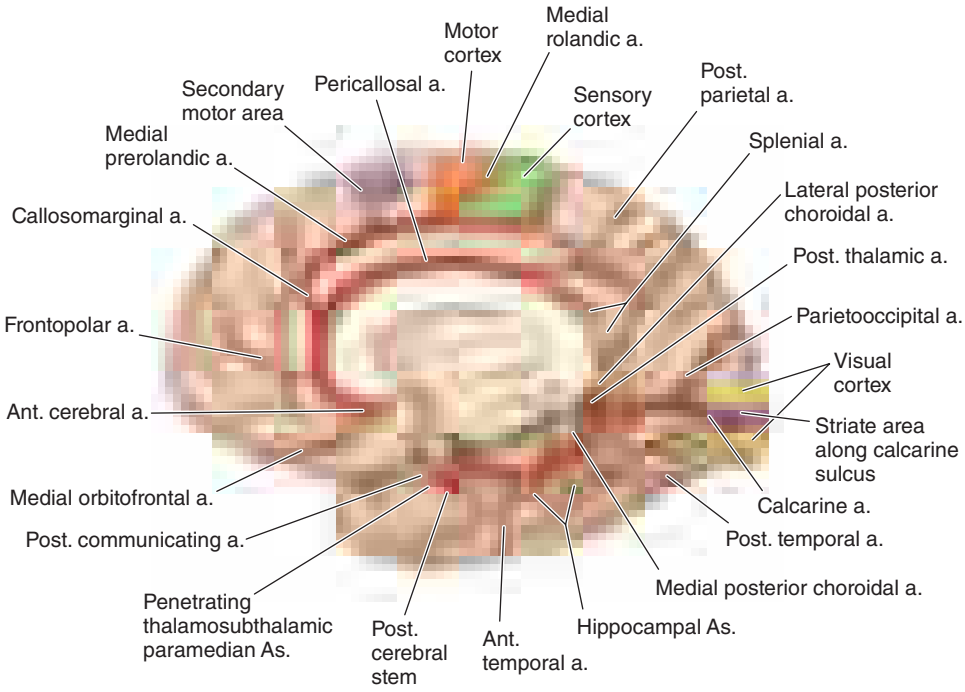


FIGURE 419-4 Diagram of a cerebral hemisphere, medial aspect, showing the branches and distribution of the anterior cerebral artery and the principal regions of cerebral localization.

Signs and symptoms: *Structures involved*

Paralysis of opposite foot and leg: *Motor leg area*

A lesser degree of paresis of opposite arm: *Arm area of cortex or fibers descending to corona radiata*

Cortical sensory loss over toes, foot, and leg: *Sensory area for foot and leg*

Urinary incontinence: *Sensorimotor area in paracentral lobule*

Contralateral grasp reflex, sucking reflex, gegenhalten (paratonic rigidity): *Medial surface of the posterior frontal lobe; likely supplemental motor area*

Abulia (akinetic mutism), slowness, delay, intermittent interruption, lack of spontaneity, whispering, reflex distraction to sights and sounds: *Uncertain localization—probably cingulate gyrus and medial inferior portion of frontal, parietal, and temporal lobes*

Impairment of gait and stance (gait apraxia): *Frontal cortex near leg motor area*

Dyspraxia of left limbs, tactile aphasia in left limbs: Corpus callosum

also, and when the nondominant hemisphere is affected, anosognosia, constructional apraxia, and neglect are found **(Chap. 26)**.

Complete MCA syndromes occur most often when an embolus occludes the stem of the artery. Cortical collateral blood flow and differing arterial configurations are probably responsible for the development of many partial syndromes. Partial syndromes may also be due to emboli that enter the proximal MCA without complete occlusion, occlude distal MCA branches, or fragment and move distally.

Partial syndromes due to embolic occlusion of a single branch include hand, or arm and hand, weakness alone (brachial syndrome) or facial weakness with nonfluent (Broca) aphasia **(Chap. 26)**, with or without arm weakness (frontal opercular syndrome). A combination of sensory disturbance, motor weakness, and nonfluent aphasia suggests that an embolus has occluded the proximal superior division and infarcted large portions of the frontal and parietal cortices (Fig. 419-3). If a fluent (Wernicke's) aphasia occurs without weakness, the inferior division of the MCA supplying the posterior part (temporal cortex) of the dominant hemisphere is probably involved. Jargon speech and an inability to comprehend written and spoken language are prominent features, often accompanied by a contralateral, homonymous superior quadrantanopia. Hemineglect or spatial agnosia without weakness indicates that the inferior division of the MCA in the nondominant hemisphere is involved.

Occlusion of a lenticulostriate vessel produces small-vessel (lacunar) stroke within the internal capsule (Fig. 419-2). This produces pure motor stroke or sensory-motor stroke contralateral to the lesion. Ischemia within the genu of the internal capsule causes primarily facial weakness followed by arm and then leg weakness as the ischemia moves posterior within the capsule. Alternatively, the contralateral hand may become ataxic, and dysarthria will be prominent (clumsy hand, dysarthria lacunar syndrome). Lacunar infarction affecting the globus pallidus and putamen often has few clinical signs, but parkinsonism and hemiballismus have been reported.

ANTERIOR CEREBRAL ARTERY The ACA is divided into two segments: the precommunal (A1) circle of Willis, or stem, which connects the internal carotid artery to the anterior communicating artery, and the postcommunal (A2) segment distal to the anterior communicating artery (Figs. 419-2 and 419-4). The A1 segment gives rise to several deep penetrating branches that supply the anterior limb of the internal capsule, the anterior perforate substance, amygdala, anterior hypothalamus, and the inferior part of the head of the caudate nucleus.

Occlusion of the proximal ACA is usually well tolerated because of collateral flow through the anterior communicating artery and collaterals through the MCA and PCA. Occlusion of a single A2 segment results in the contralateral symptoms noted in Fig. 419-4. If both A2 segments arise from a single anterior cerebral stem (contralateral A1 segment atresia), the occlusion may affect both hemispheres. Profound abulia (a delay in verbal and motor response) and bilateral pyramidal signs with paraparesis or quadriparesis and urinary incontinence result.

ANTERIOR CHOROIDAL ARTERY This artery arises from the internal carotid artery and supplies the posterior limb of the internal capsule and the white matter posterolateral to it, through which pass some of the geniculocalcarine fibers (Fig. 419-5). The complete syndrome of anterior choroidal artery occlusion consists of contralateral hemiplegia, hemianesthesia (hypesthesia), and homonymous hemianopia. However, because this territory is also supplied by penetrating vessels of the proximal MCA and the posterior communicating and posterior choroidal arteries, minimal deficits may occur, and patients frequently recover substantially. Anterior choroidal strokes are usually the result of in situ thrombosis of the vessel, and the vessel is particularly

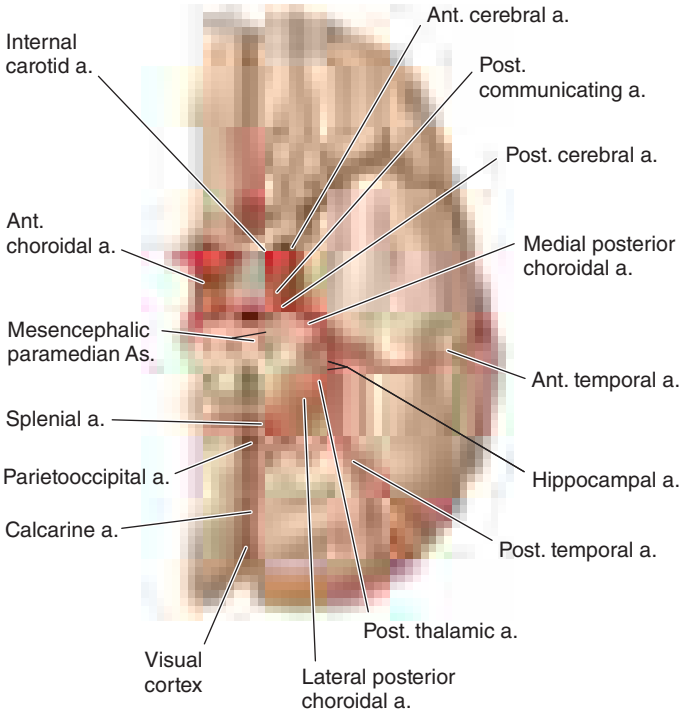


FIGURE 419-5 Inferior aspect of the brain with the branches and distribution of the posterior cerebral artery and the principal anatomic structures shown.

Signs and symptoms: *Structures involved*

Peripheral territory (see also Fig. 419-9). Homonymous hemianopia (often upper quadrantic): *Calcarine cortex or optic radiation nearby*. Bilateral homonymous hemianopia, cortical blindness, awareness or denial of blindness; tactile naming, achromatopia (color blindness), failure to see to-and-fro movements, inability to perceive objects not centrally located, apraxia of ocular movements, inability to count or enumerate objects, tendency to run into things that the patient sees and tries to avoid: *Bilateral occipital lobe with possibly the parietal lobe involved*. Verbal dyslexia without agraphia, color anomia: *Dominant calcarine lesion and posterior part of corpus callosum*. Memory defect: *Hippocampal lesion bilaterally or on the dominant side only*. Topographic disorientation and prosopagnosia: *Usually with lesions of nondominant, calcarine, and lingual gyrus*. Simultanagnosia, hemivisual neglect: *Dominant visual cortex, contralateral hemisphere*. Unformed visual hallucinations, peduncular hallucinosis, metamorphopsia, teleopsia, illusory visual spread, palinopsia, distortion of outlines, central photophobia: *Calcarine cortex*. Complex hallucinations: *Usually nondominant hemisphere*.

Central territory. Thalamic syndrome: sensory loss (all modalities), spontaneous pain and dysesthesias, choreoathetosis, intention tremor, spasms of hand, mild hemiparesis: *Posteroventral nucleus of thalamus; involvement of the adjacent subthalamus body or its afferent tracts*. Thalamoperforate syndrome: crossed cerebellar ataxia with ipsilateral third nerve palsy (Claude's syndrome): *Dentatothalamic tract and issuing third nerve*. Weber's syndrome: third nerve palsy and contralateral hemiplegia: *Third nerve and cerebral peduncle*. Contralateral hemiplegia: *Cerebral peduncle*. Paralysis or paresis of vertical eye movement, skew deviation, sluggish pupillary responses to light, slight miosis and ptosis (retraction nystagmus and "tucking" of the eyelids may be associated): *Supranuclear fibers to third nerve, interstitial nucleus of Cajal, nucleus of Darkschewitsch, and posterior commissure*. Contralateral rhythmic, ataxic action tremor; rhythmic postural or "holding" tremor (rubral tremor): *Dentatothalamic tract*.

vulnerable to iatrogenic occlusion during surgical clipping of aneurysms arising from the internal carotid artery.

INTERNAL CAROTID ARTERY The clinical picture of internal carotid occlusion varies depending on whether the cause of ischemia is propagated thrombus, embolism, or low flow. The cortex supplied by the MCA territory is affected most often. With a competent circle of Willis, occlusion may go unnoticed. If the thrombus propagates up the internal carotid artery into the MCA or embolizes it, symptoms are identical to proximal MCA occlusion (see above). Sometimes there is massive infarction of the entire deep white matter and cortical surface. When the origins of both the ACA and MCA are occluded at the top of the carotid artery, abulia or stupor occurs with hemiplegia, hemianesthesia, and aphasia or anosognosia. When the PCA arises from the internal carotid artery (a configuration called a *fetal PCA*), it may also become occluded and give rise to symptoms referable to its peripheral territory (Figs. 419-4 and 419-5).

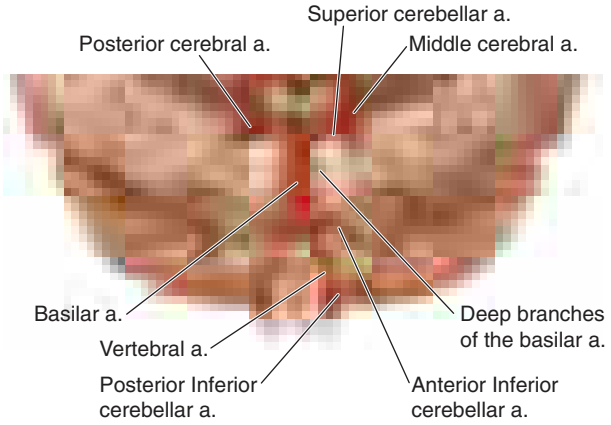


FIGURE 419-6 Diagram of the posterior circulation, showing the intracranial vertebral arteries forming the basilar artery that gives off the anterior inferior cerebellar, superior cerebellar, and posterior cerebral arteries. The posterior inferior cerebellar artery arises from each of the vertebral segments. The majority of brainstem blood flow arises from numerous deep branches of the basilar artery that penetrate directly into the brainstem.

In addition to supplying the ipsilateral brain, the internal carotid artery perfuses the optic nerve and retina via the ophthalmic artery. In ~25% of symptomatic internal carotid disease, recurrent transient monocular blindness (amaurosis fugax) warns of the lesion. Patients typically describe a horizontal shade that sweeps down or up across the field of vision. They may also complain that their vision was blurred in that eye or that the upper or lower half of vision disappeared. In most cases, these symptoms last only a few minutes. Rarely, ischemia or infarction of the ophthalmic artery or central retinal arteries occurs at the time of cerebral TIA or infarction.

A high-pitched prolonged carotid bruit fading into diastole is often associated with tightly stenotic lesions. As the stenosis grows tighter and flow distal to the stenosis becomes reduced, the bruit becomes fainter and may disappear when occlusion is imminent.

COMMON CAROTID ARTERY All symptoms and signs of internal carotid occlusion may also be present with occlusion of the common carotid artery. Jaw claudication may result from low flow in the external carotid branches. Bilateral common carotid artery occlusions at their origin may occur in Takayasu's arteritis **(Chap. 356)**.

Stroke within the Posterior Circulation

The posterior circulation is composed of the paired vertebral arteries, the basilar artery, and the paired PCAs. The vertebral arteries join to form the basilar artery at the pontomedullary junction. The basilar artery divides into two PCAs in the interpeduncular fossa (Figs. 419-4–419-6). These major arteries give rise to long and short circumferential branches and to smaller deep penetrating branches that supply the cerebellum, medulla, pons, midbrain, subthalamus, thalamus, hippocampus, and medial temporal and occipital lobes. Occlusion of each vessel produces its own distinctive syndrome.

POSTERIOR CEREBRAL ARTERY In 75% of cases, both PCAs arise from the bifurcation of the basilar artery; in 20%, one has its origin from the ipsilateral internal carotid artery via the posterior communicating artery; in 5%, both originate from the respective ipsilateral internal carotid arteries (Figs. 419-4–419-6). The precommunal, or P1, segment of the true PCA is atretic in such cases.

PCA syndromes usually result from atheroma formation or emboli that lodge at the top of the basilar artery; posterior circulation disease may also be caused by dissection of either vertebral artery or fibromuscular dysplasia.

Two clinical syndromes are commonly observed with occlusion of the PCA: (1) *P1 syndrome*: midbrain, subthalamic, and thalamic signs, which are due to disease of the proximal P1 segment of the PCA or its penetrating branches (thalamogeniculate, Percheron, and posterior

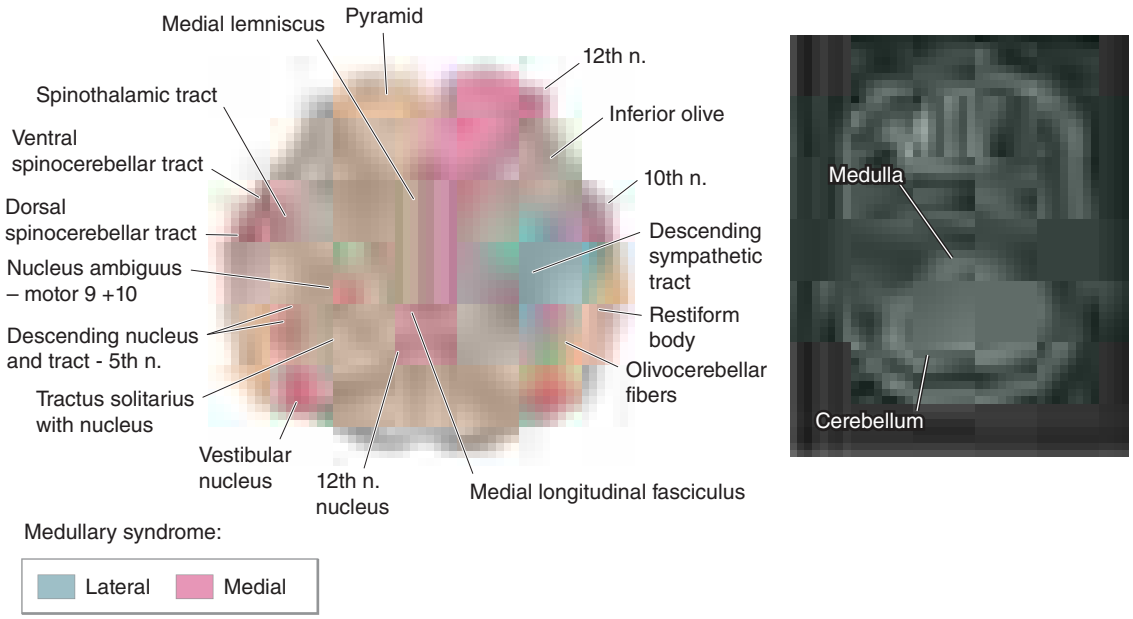


FIGURE 419-7 Axial section at the level of the medulla, depicted schematically on the left, with a corresponding magnetic resonance image on the right. Note that in Figs. 419-7 through 419-11, all drawings are oriented with the dorsal surface at the bottom, matching the orientation of the brainstem that is commonly seen in all modern neuroimaging studies. Approximate regions involved in medial and lateral medullary stroke syndromes are shown.

Signs and symptoms: *Structures involved*

1. Medial medullary syndrome (occlusion of vertebral artery or of branch of vertebral or lower basilar artery)
   - On side of lesion
     - Paralysis with atrophy of one-half half the tongue: *Ipsilateral twelfth nerve*
   - On side opposite lesion
     - Paralysis of arm and leg, sparing face; impaired tactile and proprioceptive sense over one-half the body: *Contralateral pyramidal tract and medial lemniscus*
2. Lateral medullary syndrome (occlusion of any of five vessels may be responsible—vertebral, posterior inferior cerebellar, superior, middle, or inferior lateral medullary arteries)
   - On side of lesion
     - Pain, numbness, impaired sensation over one-half the face: *Descending tract and nucleus fifth nerve*
     - Ataxia of limbs, falling to side of lesion: *Uncertain—restiform body, cerebellar hemisphere, cerebellar fibers, spinocerebellar tract (?)*
     - Nystagmus, diplopia, oscillopsia, vertigo, nausea, vomiting: *Vestibular nucleus*
     - Horner's syndrome (miosis, ptosis, decreased sweating): *Descending sympathetic tract*
     - Dysphagia, hoarseness, paralysis of palate, paralysis of vocal cord, diminished gag reflex: *Issuing fibers ninth and tenth nerves*
     - Loss of taste: *Nucleus and tractus solitarius*
     - Numbness of ipsilateral arm, trunk, or leg: *Cuneate and gracile nuclei*
     - Weakness of lower face: *Genuflected upper motor neuron fibers to ipsilateral facial nucleus*
   - On side opposite lesion
     - Impaired pain and thermal sense over half the body, sometimes face: *Spinothalamic tract*
3. Total unilateral medullary syndrome (occlusion of vertebral artery): Combination of medial and lateral syndromes
4. Lateral pontomedullary syndrome (occlusion of vertebral artery): Combination of lateral medullary and lateral inferior pontine syndrome
5. Basilar artery syndrome (the syndrome of the lone vertebral artery is equivalent): A combination of the various brainstem syndromes plus those arising in the posterior cerebral artery distribution.
   - Bilateral long tract signs (sensory and motor; cerebellar and peripheral cranial nerve abnormalities): *Bilateral long tract; cerebellar and peripheral cranial nerves*
   - Paralysis or weakness of all extremities, plus all bulbar musculature: *Corticobulbar and corticospinal tracts bilaterally*

choroidal arteries); and (2) *P2 syndrome*: cortical temporal and occipital lobe signs, due to occlusion of the P2 segment distal to the junction of the PCA with the posterior communicating artery.

P1 SYNDROMES Infarction usually occurs in the ipsilateral subthalamus and medial thalamus and in the ipsilateral cerebral peduncle and midbrain (Figs. 419-5 and 419-11). A third nerve palsy with contralateral ataxia (Claude's syndrome) or with contralateral hemiplegia (Weber's syndrome) may result. The ataxia indicates involvement of the red nucleus or dentatorubrothalamic tract; the hemiplegia is localized to the cerebral peduncle (Fig. 419-11). If the subthalamic nucleus is involved, contralateral hemiballismus may occur. Occlusion of the artery of Percheron produces paresis of upward gaze and drowsiness and often abulia. Extensive infarction in the midbrain and subthalamus occurring with bilateral proximal PCA occlusion presents as coma, unreactive pupils, bilateral pyramidal signs, and decerebrate rigidity.

Occlusion of the penetrating branches of thalamic and thalamogeniculate arteries produces less extensive thalamic and thalamocapsular lacunar syndromes. The *thalamic Déjérine-Roussy syndrome* consists of contralateral hemisensory loss followed later by an agonizing, searing, or burning pain in the affected areas. It is persistent and responds poorly to analgesics. Anticonvulsants (carbamazepine or gabapentin) or tricyclic antidepressants may be beneficial.

P2 SYNDROMES (Figs. 419-4 and 419-5) Occlusion of the distal PCA causes infarction of the medial temporal and occipital lobes. Contralateral homonymous hemianopia without macula sparing is the usual manifestation. (MCA strokes often produce hemianopia but typically spare the macula as calcarine cortex is perfused by the P2 segment). Occasionally, only the upper quadrant of visual field is involved or the macula vision is spared. If the visual association areas are spared and only the calcarine cortex is involved, the patient may be aware of visual defects. Medial temporal lobe and hippocampal involvement may cause an acute disturbance in memory, particularly if it occurs in the dominant hemisphere. The defect usually clears because memory has bilateral representation. If the dominant hemisphere is affected and the infarct extends to involve the splenium of the corpus callosum, the patient may demonstrate alexia without agraphia. Visual agnosia for

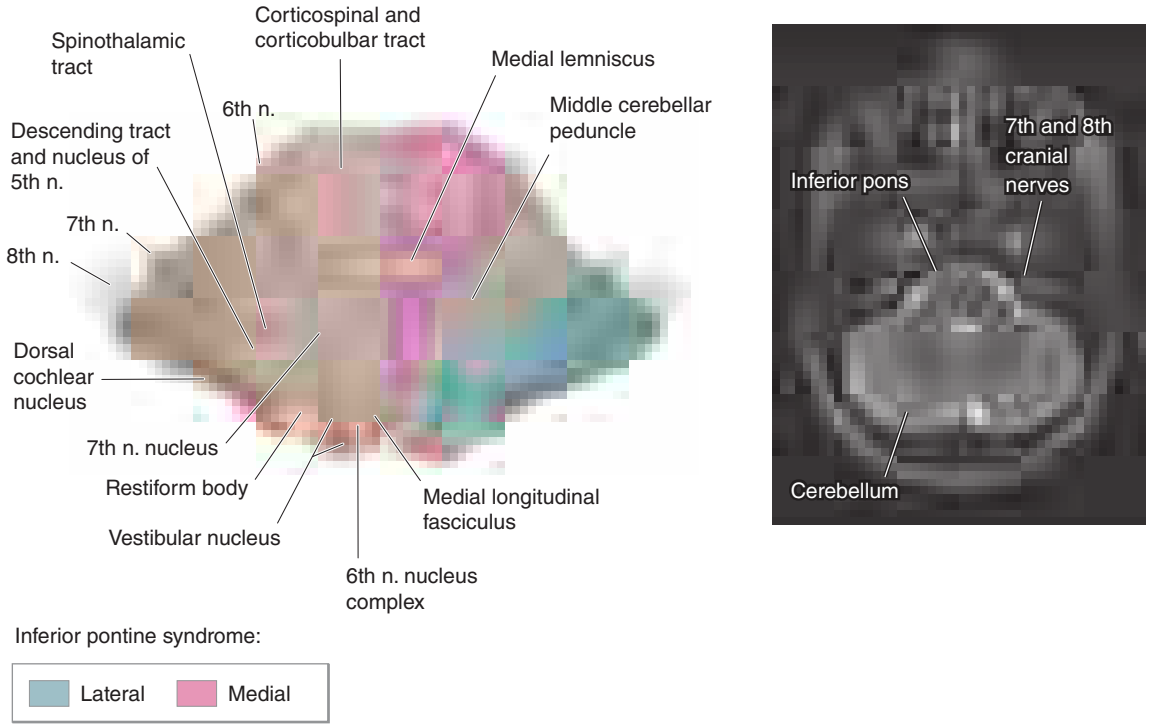


FIGURE 419-8 Axial section at the level of the inferior pons, depicted schematically on the left, with a corresponding magnetic resonance image on the right. Approximate regions involved in medial and lateral inferior pontine stroke syndromes are shown.

Signs and symptoms: *Structures involved*

1. Medial inferior pontine syndrome (occlusion of paramedian branch of basilar artery)
   - On side of lesion
     - Paralysis of conjugate gaze to side of lesion (preservation of convergence): *Center for conjugate lateral gaze*
     - Nystagmus: *Vestibular nucleus*
     - Ataxia of limbs and gait: Likely *middle cerebellar peduncle*
     - Diplopia on lateral gaze: *Abducens nerve*
   - On side opposite lesion
     - Paralysis of face, arm, and leg: *Corticobulbar and corticospinal tract in lower pons*
     - Impaired tactile and proprioceptive sense over one-half of the body: *Medial lemniscus*
2. Lateral inferior pontine syndrome (occlusion of anterior inferior cerebellar artery)
   - On side of lesion
     - Horizontal and vertical nystagmus, vertigo, nausea, vomiting, oscillopsia: *Vestibular nerve or nucleus*
     - Facial paralysis: *Seventh nerve*
     - Paralysis of conjugate gaze to side of lesion: *Center for conjugate lateral gaze*
     - Deafness, tinnitus: *Auditory nerve or cochlear nucleus*
     - Ataxia: *Middle cerebellar peduncle and cerebellar hemisphere*
     - Impaired sensation over face: *Descending tract and nucleus fifth nerve*
   - On side opposite lesion
     - Impaired pain and thermal sense over one-half the body (may include face): *Spinothalamic tract*

faces, objects, mathematical symbols, and colors and anomia with paraphasic errors (amnestic aphasia) may also occur, even without callosal involvement. Occlusion of the PCA can produce *peduncular hallucinosis* (visual hallucinations of brightly colored scenes and objects).

Bilateral infarction in the distal PCAs produces cortical blindness (blindness with preserved pupillary light reaction). The patient is often unaware of the blindness or may even deny it (*Anton's syndrome*). Tiny islands of vision may persist, and the patient may report that vision fluctuates as images are captured in the preserved portions. Rarely, only peripheral vision is lost and central vision is spared, resulting in "gun-barrel" vision. Bilateral visual association area lesions may result in *Balint's syndrome*, a disorder of the orderly visual scanning of the environment **(Chap. 26)**, usually resulting from infarctions secondary to low flow in the "watershed" between the distal PCA and MCA territories, as occurs after cardiac arrest. Patients may experience persistence of a visual image for several minutes despite gazing at another scene (*palinopsia*) or an inability to synthesize the whole of an image (*asimultanagnosia*). Embolic occlusion of the top of the basilar artery can produce any or all the central or peripheral territory symptoms. The hallmark is the sudden onset of bilateral signs, including ptosis, pupillary asymmetry or lack of reaction to light, and somnolence. Patients will often have posturing and myoclonic jerking that simulates seizure. Interrogation of the noncontrast CT scan for a hyperdense basilar artery sign (indicating thrombus in the basilar artery), or CT angiography (CTA) establishes this diagnosis. Physicians should be suspicious of this rare, but potentially treatable stroke syndrome in the setting of presumed new onset seizure and cranial nerve deficits.

VERTEBRAL AND POSTERIOR INFERIOR CEREBELLAR ARTERIES The vertebral artery, which arises from the innominate artery on the right and the subclavian artery on the left, consists of four segments. The first (V1) extends from its origin to its entrance into the sixth or fifth transverse vertebral foramen. The second segment (V2) traverses the vertebral foramina from C6 to C2. The third (V3) passes through the transverse foramen and circles around the arch of the atlas to pierce the dura at the foramen magnum. The fourth (V4) segment courses upward to join the other vertebral artery to form the basilar artery (Fig. 419-6); only the fourth segment gives rise to branches that supply the brainstem and cerebellum. The posterior inferior cerebellar artery (PICA) in its proximal segment supplies the lateral medulla and, in its distal branches, the inferior surface of the cerebellum.

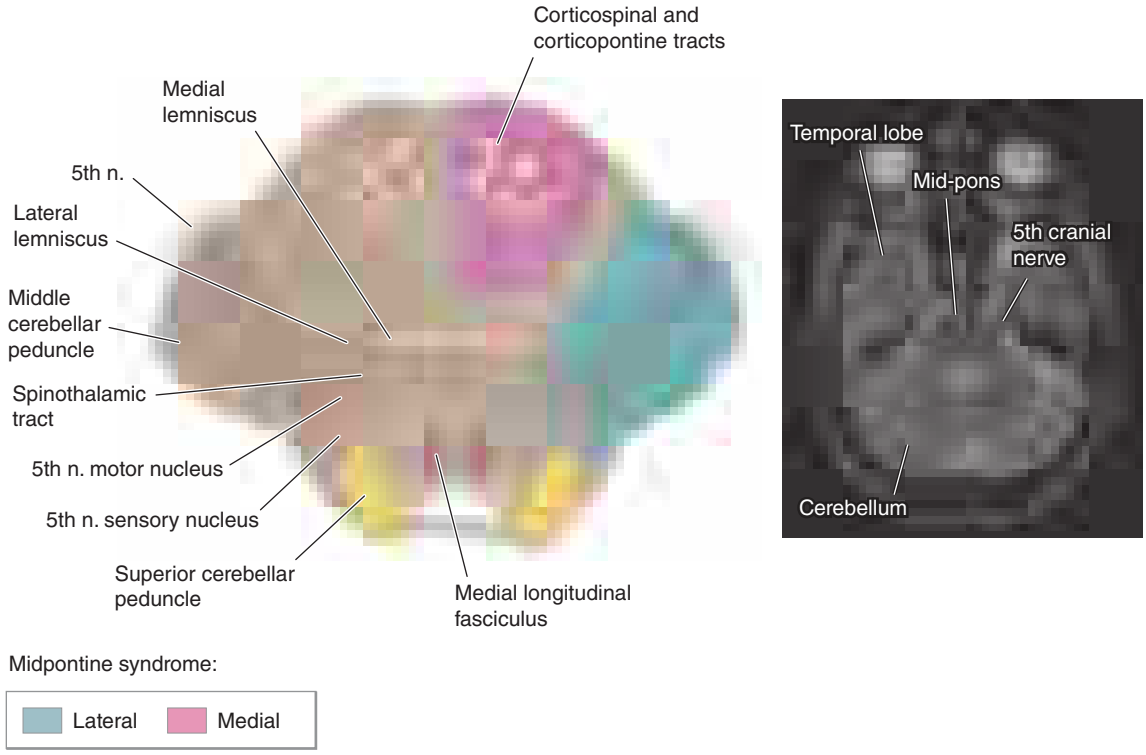


FIGURE 419-9 Axial section at the level of the midpons, depicted schematically on the left, with a corresponding magnetic resonance image on the right. Approximate regions involved in medial and lateral midpontine stroke syndromes are shown.

Signs and symptoms: *Structures involved*

1. Medial midpontine syndrome (paramedian branch of midbasilar artery)
   - On side of lesion
     - Ataxia of limbs and gait (more prominent in bilateral involvement): *Pontine nuclei*
   - On side opposite lesion
     - Paralysis of face, arm, and leg: *Corticobulbar and corticospinal tract*
     - Variable impaired touch and proprioception when lesion extends posteriorly: *Medial lemniscus*
2. Lateral midpontine syndrome (short circumferential artery)
   - On side of lesion
     - Ataxia of limbs: *Middle cerebellar peduncle*
     - Paralysis of muscles of mastication: *Motor fibers or nucleus of fifth nerve*
     - Impaired sensation over side of face: *Sensory fibers or nucleus of fifth nerve*
   - On side opposite lesion
     - Impaired pain and thermal sense on limbs and trunk: *Spinothalamic tract*

Atherothrombotic lesions have a predilection for V1 and V4 segments of the vertebral artery. The first segment may become diseased at the origin of the vessel and may produce posterior circulation emboli; collateral flow from the contralateral vertebral artery or the ascending cervical, thyrocervical, or occipital arteries is usually sufficient to prevent low-flow TIAs or stroke. When one vertebral artery is atretic and an atherothrombotic lesion threatens the origin of the other, the collateral circulation, which may also include retrograde flow down the basilar artery, is often insufficient (Figs. 419-5 and 419-6). In this setting, low-flow TIAs may occur, consisting of syncope, vertigo, and alternating hemiplegia; this state also sets the stage for thrombosis. Disease of the distal fourth segment of the vertebral artery can promote thrombus formation manifest as embolism or with propagation as basilar artery thrombosis. Stenosis proximal to the origin of the PICA can threaten the lateral medulla and posterior inferior surface of the cerebellum.

If the subclavian artery is occluded proximal to the origin of the vertebral artery, there is a reversal in the direction of blood flow in the ipsilateral vertebral artery. Exercise of the ipsilateral arm may increase demand on vertebral flow, producing posterior circulation TIAs, or "subclavian steal."

Although atheromatous disease rarely narrows the second and third segments of the vertebral artery, this region is subject to dissection, fibromuscular dysplasia, and, rarely, encroachment by osteophytic spurs within the vertebral foramina.

Embolic occlusion or thrombosis of a V4 segment causes ischemia of the lateral medulla. The constellation of vertigo, numbness of the ipsilateral face and contralateral limbs, diplopia, hoarseness, dysarthria, dysphagia, and ipsilateral Horner's syndrome is called the *lateral medullary (or Wallenberg's) syndrome* (Fig. 419-7). Ipsilateral upper motor neuron facial weakness can also occur. Most cases result from ipsilateral vertebral artery occlusion; in the remainder, PICA occlusion is responsible. Occlusion of the medullary penetrating branches of the vertebral artery or PICA results in partial syndromes. *Hemiparesis is not a typical feature of vertebral artery occlusion; however, quadriparesis may result from occlusion of the anterior spinal artery.*

Rarely, a *medial medullary syndrome* occurs with infarction of the pyramid and contralateral hemiparesis of the arm and leg, sparing the face. If the medial lemniscus and emerging hypoglossal nerve fibers are involved, contralateral loss of joint position sense and ipsilateral tongue weakness occur.

Cerebellar infarction can lead to *respiratory arrest* due to raised intracranial pressure from cerebellar swelling, closure of the aqueduct of Silvius or fourth ventricle, followed by hydrocephalus and central herniation. Displacement of the brainstem from cerebellar edema will also cause respiratory and hemodynamic instability. Drowsiness, Babinski signs, dysarthria, and bifacial weakness may be absent, or present only briefly, before respiratory arrest ensues. Gait unsteadiness, headache, dizziness, nausea, and vomiting may be the only early

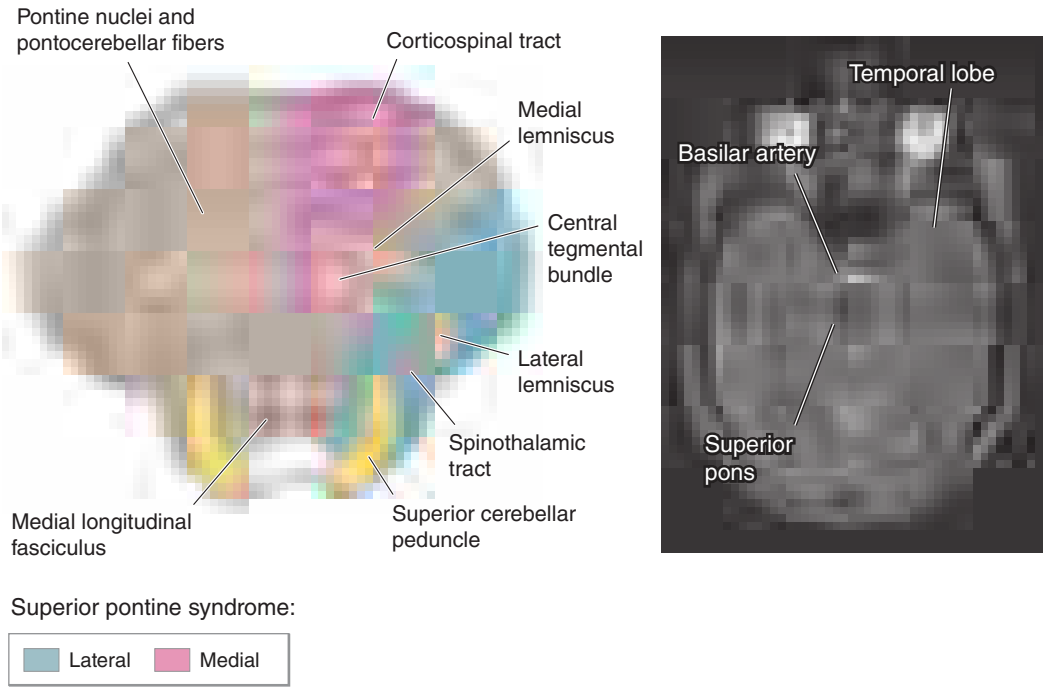


FIGURE 419-10 Axial section at the level of the superior pons, depicted schematically on the left, with a corresponding magnetic resonance image on the right. Approximate regions involved in medial and lateral superior pontine stroke syndromes are shown.

Signs and symptoms: *Structures involved*

1. Medial superior pontine syndrome (paramedian branches of upper basilar artery)
   - On side of lesion
     - Cerebellar ataxia (probably): *Superior and/or middle cerebellar peduncle*
     - Internuclear ophthalmoplegia: *Medial longitudinal fasciculus*
     - Myoclonic syndrome, palate, pharynx, vocal cords, respiratory apparatus, face, oculomotor apparatus, etc.: *Localization uncertain—central tegmental bundle, dentate projection, inferior olivary nucleus*
   - On side opposite lesion
     - Paralysis of face, arm, and leg: *Corticobulbar and corticospinal tract*
     - Rarely touch, vibration, and position are affected: *Medial lemniscus*
2. Lateral superior pontine syndrome (syndrome of superior cerebellar artery)
   - On side of lesion
     - Ataxia of limbs and gait, falling to side of lesion: *Middle and superior cerebellar peduncles, superior surface of cerebellum, dentate nucleus*
     - Dizziness, nausea, vomiting; horizontal nystagmus: *Vestibular nucleus*
     - Paresis of conjugate gaze (ipsilateral): *Pontine contralateral gaze*
     - Skew deviation: *Uncertain*
     - Miosis, ptosis, decreased sweating over face (Horner's syndrome): *Descending sympathetic fibers*
     - Tremor: Localization unclear—*Dentate nucleus, superior cerebellar peduncle*
   - On side opposite lesion
     - Impaired pain and thermal sense on face, limbs, and trunk: *Spinothalamic tract*
     - Impaired touch, vibration, and position sense, more in leg than arm (there is a tendency to incongruity of pain and touch deficits): *Medial lemniscus (lateral portion)*

symptoms and signs and should arouse suspicion of this impending complication, which may require neurosurgical decompression, often with an excellent outcome. Separating these symptoms from those of viral labyrinthitis can be a challenge, but headache, neck stiffness, and unilateral dysmetria favor stroke.

BASILAR ARTERY Branches of the basilar artery (Fig. 419-6) supply the base of the pons and superior cerebellum and fall into three groups: (1) paramedian, 7–10 in number, which supply a wedge of pons on either side of the midline; (2) short circumferential, 5–7 in number, that supply the lateral two-thirds of the pons and middle and superior cerebellar peduncles; and (3) bilateral long circumferential (superior cerebellar and anterior inferior cerebellar arteries), which course around the pons to supply the cerebellar hemispheres.

Atheromatous lesions can occur anywhere along the basilar trunk but are most frequent in the proximal basilar and distal vertebral segments. Typically, lesions occlude either the proximal basilar and one or both vertebral arteries. The clinical picture varies depending on the availability of retrograde collateral flow from the posterior communicating arteries. Rarely, dissection of a vertebral artery may involve the basilar artery and, depending on the location of true and false lumen, may produce multiple penetrating artery strokes.

Although atherothrombosis occasionally occludes the distal portion of the basilar artery, emboli from the heart or proximal vertebral or basilar segments are more commonly responsible for "top of the basilar" syndromes.

Because the brainstem contains many structures in close apposition, a diversity of clinical syndromes may emerge with ischemia, reflecting involvement of the corticospinal and corticobulbar tracts, ascending sensory tracts, and cranial nerve nuclei (Figs. 419-7–419-11).

The symptoms of transient ischemia or infarction in the territory of the basilar artery often do not indicate whether the basilar artery itself or one of its branches is diseased, yet this distinction has important implications for therapy. *The picture of complete basilar occlusion, however, is easy to recognize as a constellation of bilateral long tract signs (sensory and motor) with signs of cranial nerve and cerebellar dysfunction.* Patients may have spontaneous posturing movements that are myoclonic in nature and simulate seizure activity. A "locked-in" state of preserved consciousness with quadriplegia and cranial nerve signs suggests complete pontine and lower midbrain infarction. The

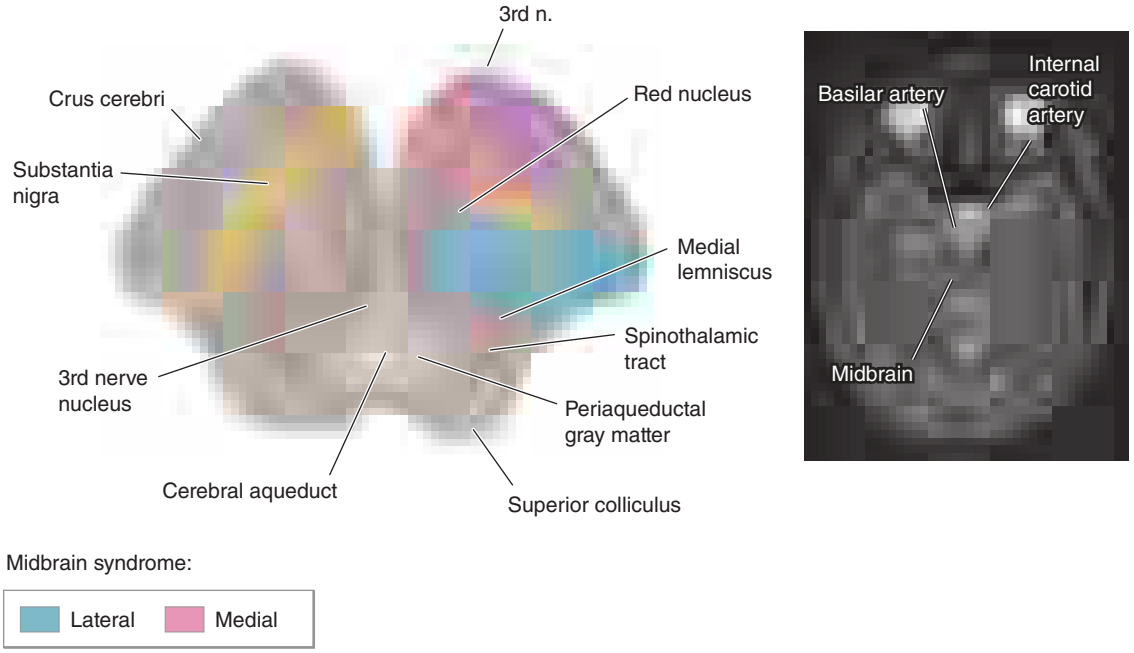


FIGURE 419-11 Axial section at the level of the midbrain, depicted schematically on the left, with a corresponding magnetic resonance image on the right. Approximate regions involved in medial and lateral midbrain stroke syndromes are shown.

Signs and symptoms: *Structures involved*

1. Medial midbrain syndrome (paramedian branches of upper basilar and proximal posterior cerebral arteries)
   - On side of lesion
     - Eye "down and out" secondary to unopposed action of fourth and sixth cranial nerves, with dilated and unresponsive pupil: *Third nerve fibers*
   - On side opposite lesion
     - Paralysis of face, arm, and leg: *Corticobulbar and corticospinal tract descending in crus cerebri*
2. Lateral midbrain syndrome (syndrome of small penetrating arteries arising from posterior cerebral artery)
   - On side of lesion
     - Eye "down and out" secondary to unopposed action of fourth and sixth cranial nerves, with dilated and unresponsive pupil: *Third nerve fibers and/or third nerve nucleus*
   - On side opposite lesion
     - Hemiataxia, hyperkinesias, tremor: *Red nucleus, dentatorubrothalamic pathway*

therapeutic goal is to identify *impending* basilar occlusion before devastating infarction occurs. A series of TIAs and a slowly progressive, fluctuating stroke are extremely significant, because they often herald an atherothrombotic occlusion of the distal vertebral or proximal basilar artery.

TIAs in the proximal basilar distribution may produce vertigo (often described by patients as "swimming," "swaying," "moving," "unsteadiness," or "light-headedness"). Other symptoms that warn of basilar thrombosis include diplopia, dysarthria, facial or circumoral numbness, and hemisensory symptoms. In general, symptoms of basilar branch TIAs affect one side of the brainstem, whereas symptoms of basilar artery TIAs usually affect both sides, although a "herald" hemiparesis has been emphasized as an initial symptom of basilar occlusion. Most often, TIAs, whether due to impending occlusion of the basilar artery or a basilar branch, are short lived (5–30 min) and repetitive, occurring several times a day. The pattern suggests intermittent reduction of flow. Although treatment with intravenous heparin or various combinations of antiplatelet agents have been used to prevent clot propagation there is no specific evidence to support any one approach, and endovascular intervention is also an option.

Atherothrombotic occlusion of the basilar artery with infarction usually causes *bilateral* brainstem signs. A gaze paresis or internuclear ophthalmoplegia associated with ipsilateral hemiparesis may be the only manifestation of bilateral brainstem ischemia. More often, unequivocal signs of bilateral pontine disease are present. Complete basilar thrombosis carries a high mortality.

Occlusion of a branch of the basilar artery usually causes *unilateral* symptoms and signs involving motor, sensory, and cranial nerves. If symptoms remain unilateral, concern over pending basilar occlusion should be reduced.

Occlusion of the superior cerebellar artery results in severe ipsilateral cerebellar ataxia, nausea and vomiting, dysarthria, and contralateral loss of pain and temperature sensation over the extremities, body, and face (spino- and trigeminothalamic tract). Partial deafness, ataxic tremor of the ipsilateral upper extremity, Horner's syndrome, and palatal myoclonus may occur rarely. Partial syndromes occur frequently (Fig. 419-10). With large strokes, swelling and mass effects may compress the midbrain or produce hydrocephalus; these symptoms may evolve rapidly. Neurosurgical intervention may be lifesaving in such cases.

Occlusion of the anterior inferior cerebellar artery produces variable degrees of infarction because the size of this artery and the territory it supplies vary inversely with those of the PICA. The principal symptoms include: (1) ipsilateral deafness, facial weakness, vertigo, nausea and vomiting, nystagmus, tinnitus, cerebellar ataxia, Horner's syndrome, and paresis of conjugate lateral gaze; and (2) contralateral loss of pain and temperature sensation. An occlusion close to the origin of the artery may cause corticospinal tract signs (Fig. 419-8).

Occlusion of one of the short circumferential branches of the basilar artery affects the lateral two-thirds of the pons and middle or superior cerebellar peduncle, whereas occlusion of one of the paramedian branches affects a wedge-shaped area on either side of the medial pons (Figs. 419-8–419-10).

IMAGING STUDIES

See also Chap. 416.

CT Scans CT radiographic images identify or exclude hemorrhage as the cause of stroke, and they identify extraparenchymal hemorrhages, neoplasms, abscesses, and other conditions masquerading

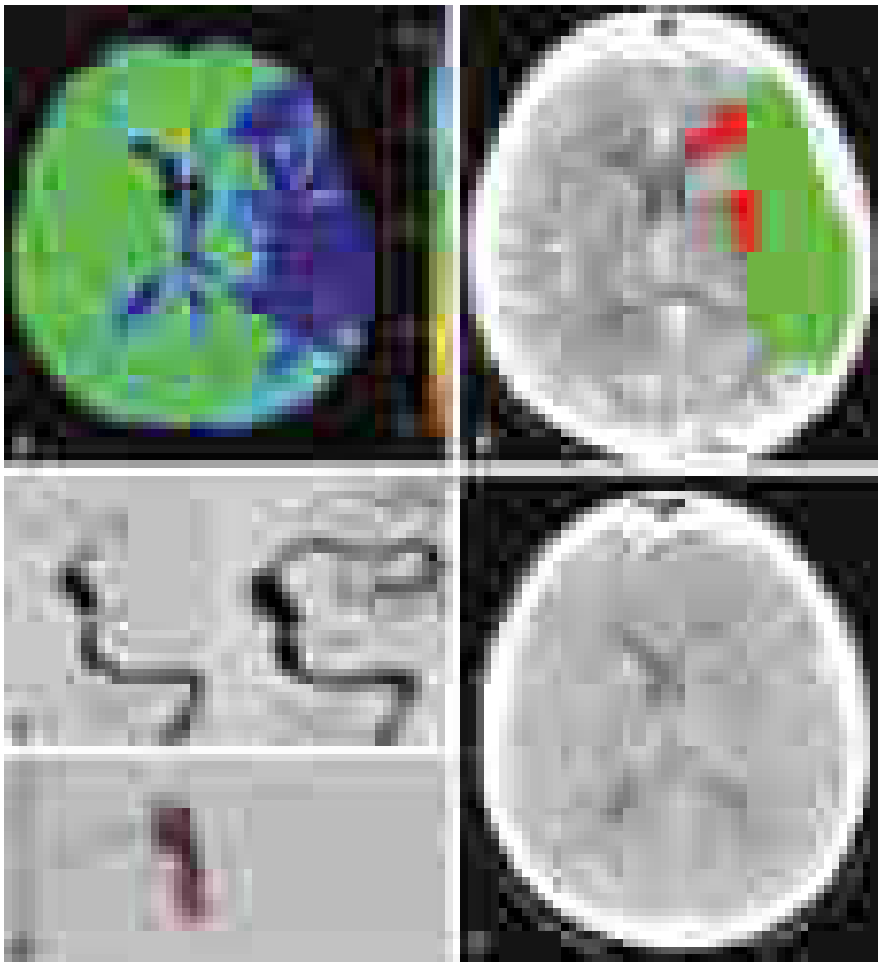


FIGURE 419-12 Acute left middle cerebral artery (MCA) stroke with right hemiplegia but preserved language. ***A.*** Computed tomography (CT) perfusion mean-transit time map showing delayed perfusion of the left MCA distribution (*blue*). ***B.*** Predicted region of infarct (*red*) and penumbra (*green*) based on CT perfusion data. ***C.*** Conventional angiogram showing occlusion of the left internal carotid–MCA bifurcation (*left panel*), and revascularization of the vessels following successful thrombectomy 8 h after stroke symptom onset (*right panel*). ***D.*** The clot removed with a thrombectomy device (L5, Concentric Medical, Inc.). ***E.*** CT scan of the brain 2 days later; note infarction in the region predicted in ***B*** but preservation of the penumbral region by successful revascularization.

as stroke. Brain CT scans obtained in the first several hours after an infarction generally show no abnormality, and the infarct may not be seen reliably for 24–48 h. CT may fail to show small ischemic strokes in the posterior fossa because of bone artifact; small infarcts on the cortical surface may also be missed.

Contrast-enhanced CT scans add specificity by showing contrast enhancement of subacute infarcts and allow visualization of venous structures. Coupled with multidetector scanners, CT angiography can be performed with administration of IV iodinated contrast allowing visualization of the cervical and intracranial arteries, intracranial veins, aortic arch, and even the coronary arteries in one imaging session. Carotid disease and intracranial vascular occlusions are readily identified with this method **(see Fig. 420-2)**. After an IV bolus of contrast, deficits in brain perfusion produced by vascular occlusion can also be demonstrated **(Fig. 419-12)** and used to predict the region of infarcted brain and the brain at risk of further infarction (i.e., the ischemic penumbra, see "Pathophysiology of Ischemic Stroke" in **Chap. 420**). CT imaging is also sensitive for detecting SAH (although by itself does not rule it out), and CTA can readily identify intracranial aneurysms **(Chap. 301)**. Because of its speed and wide availability, noncontrast head CT is the imaging modality of choice in patients with acute stroke (Fig. 419-1), and CTA and CT perfusion imaging may also be useful and convenient adjuncts.

MRI

MRI reliably documents the extent and location of infarction in all areas of the brain, including the posterior fossa and cortical surface. It also identifies intracranial hemorrhage and other abnormalities and, using special sequences, can be as sensitive as CT for detecting acute intracerebral hemorrhage. MRI scanners with magnets of higher field strength produce more reliable and precise images. Diffusion-weighted imaging is more sensitive for early brain infarction than standard MR sequences or CT **(Fig. 419-13)**, as is fluid-attenuated inversion recovery (FLAIR) imaging **(Chap. 416)**. Using IV administration of gadolinium contrast, MR perfusion studies can be performed. Brain regions showing poor perfusion but no abnormality on diffusion provide, compared to CT, an equivalent measure of the ischemic penumbra. MR angiography is highly sensitive for stenosis of extracranial internal carotid arteries and of large intracranial vessels. With higher degrees of stenosis, MR angiography tends to overestimate the degree of stenosis when compared to conventional x-ray angiography. MRI with fat saturation is an imaging sequence used to visualize extra or intracranial arterial dissection. This sensitive technique images clotted blood within the dissected vessel wall. Iron-sensitive imaging (ISI) is helpful to detect cerebral microbleeds that may be present in cerebral amyloid angiopathy and other hemorrhagic disorders.

MRI is more expensive and time consuming than CT and less readily available. Claustrophobia and the logistics of imaging acutely critically ill patients also limit its application. Most acute stroke protocols use CT because of these limitations. However, MRI is useful outside the acute period by more clearly defining the extent of tissue injury and discriminating new from old regions of brain infarction. MRI may have utility in patients with TIA, because it is also more likely to identify new infarction, which is a strong predictor of subsequent stroke.

Cerebral Angiography Conventional x-ray cerebral angiography is the gold standard for identifying and quantifying atherosclerotic

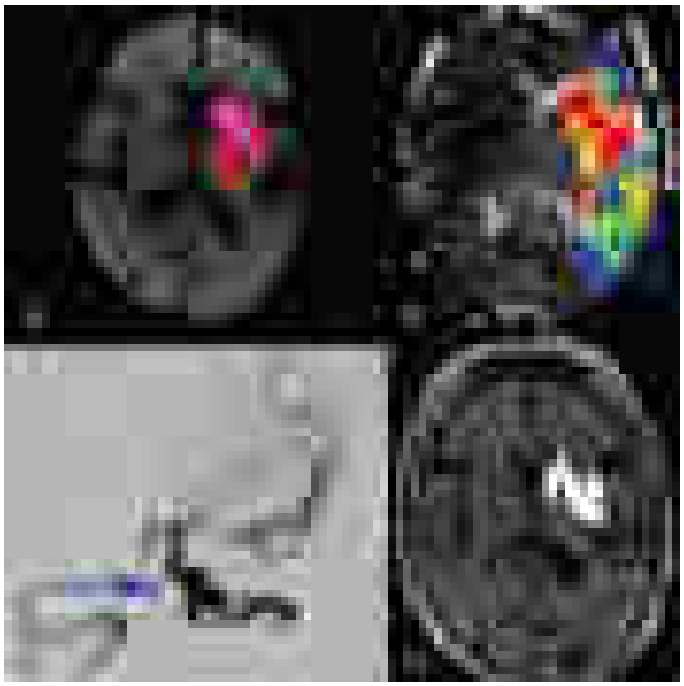


FIGURE 419-13 Magnetic resonance imaging (MRI) of acute stroke. ***A.*** MRI diffusion-weighted image (DWI) of an 82-year-old woman 2.5 h after onset of right-sided weakness and aphasia reveals restricted diffusion within the left basal ganglia and internal capsule (*colored regions*). ***B.*** Perfusion defect within the left hemisphere (*colored signal*) imaged after administration of an IV bolus of gadolinium contrast. The discrepancy between the region of poor perfusion shown in ***B*** and the diffusion deficit shown in ***A*** is called *diffusion-perfusion mismatch* and provides an estimate of the ischemic penumbra. Without specific therapy, the region of infarction will expand into much or all the perfusion deficit. ***C.*** Cerebral angiogram of the left internal carotid artery in this patient before (*left*) and after (*right*) successful endovascular embolectomy. The occlusion is within the carotid terminus. ***D.*** Fluid-attenuated inversion recovery image obtained 3 days later showing a region of infarction (*coded as white*) that corresponds to the initial DWI image in ***A***, but not the entire area at risk shown in ***B***, suggesting that successful embolectomy saved a large region of brain tissue from infarction. *(Courtesy of Gregory Albers, MD, Stanford University; with permission.)*

stenoses of the cerebral arteries and for identifying and characterizing other pathologies, including aneurysms, vasospasm, intraluminal thrombi, fibromuscular dysplasia, arteriovenous fistulae, vasculitis, and collateral channels of blood flow. Conventional angiography carries risks of arterial damage, groin hemorrhage, embolic stroke, and renal failure from contrast nephropathy, so it should be reserved for situations where less invasive means are inadequate. Acute stroke treatment with endovascular thrombectomy has proven effective in ischemic strokes caused by internal carotid terminus or MCA occlusions and has now part of routine clinical practice at centers that have this capability **(see Chap. 420)**.

Ultrasound Techniques Stenosis at the origin of the internal carotid artery can be identified and quantified reliably by ultrasonography that combines a B-mode ultrasound image with a Doppler ultrasound assessment of flow velocity ("duplex" ultrasound). Transcranial Doppler (TCD) assessment of MCA, ACA, and PCA flow and of vertebrobasilar flow is also useful. This latter technique can detect stenotic lesions in the large intracranial arteries because such lesions increase systolic flow velocity. TCD can also detect microemboli from otherwise asymptomatic carotid plaques. In many cases, MR angiography combined with carotid and transcranial ultrasound studies eliminates the need for conventional x-ray angiography in evaluating vascular stenosis. Alternatively, CTA of the entire head and neck can be performed during the initial imaging of acute stroke. Because this images the entire arterial system relevant to stroke, with the exception of the heart, much of the clinician's stroke workup can be completed with this single imaging study.

Perfusion Techniques Both xenon techniques (principally xenon-CT) and positron emission tomography (PET) can quantify cerebral blood flow. These tools are generally used for research **(Chap. 416)** but can be useful for determining the significance of arterial stenosis and planning for revascularization surgery. Single-photon emission computed tomography (SPECT) and MR perfusion techniques report relative cerebral blood flow. As noted above, CT imaging is used as the initial imaging modality for acute stroke, and some centers combine both CTA and CT perfusion imaging together with the noncontrast CT scan. CT perfusion imaging increases the sensitivity for detecting ischemia and can measure the ischemic penumbra (Fig. 419-12). Alternatively, MR perfusion can be combined with MR diffusion imaging to identify the ischemic penumbra as the mismatch between these two imaging sequences (Fig. 419-13).

FURTHER READING

CAPLAN LR: *Caplan's Stroke: A Clinical Approach*, 5th ed. Cambridge, UK, Cambridge University Press, 2016.

TAMUTZER AA et al: ED misdiagnosis of cerebrovascular events in the era of modern neuroimaging: A meta-analysis. Neurology 88:1468, 2017.

420 Ischemic Stroke

Wade S. Smith, S. Claiborne Johnston, J. Claude Hemphill, III

The clinical diagnosis of stroke is discussed in **Chap. 419**. Once this diagnosis is made, and either a non-contrast CT scan or MRI has been performed, rapid reversal of ischemia is paramount. This chapter will focus on the stroke treatment timeline and subsequent secondary stroke prevention.

PATHOPHYSIOLOGY OF ISCHEMIC STROKE

Acute occlusion of an intracranial vessel causes reduction in blood flow to the brain region it supplies. The magnitude of flow reduction is a function of collateral blood flow, and this depends on individual vascular anatomy (which may be altered by disease), the site of occlusion, and systemic blood pressure. A decrease in cerebral blood flow to zero causes death of brain tissue within 4–10 min; values <16–18 mL/100 g tissue per minute cause infarction within an hour; and values <20 mL/100 g tissue per minute cause ischemia without infarction unless prolonged for several hours or days. If blood flow is restored to ischemic tissue before significant infarction develops, the patient may experience only transient symptoms, and the clinical syndrome is called a transient ischemic attack (TIA). Another important concept is the *ischemic penumbra*, defined as the ischemic but reversibly dysfunctional tissue surrounding a core area of infarction. The penumbra can be imaged by perfusion imaging using MRI or CT (see below and **Figs. 419-12** and **419-13**). The ischemic penumbra will eventually progress to infarction if no change in flow occurs, and hence saving the ischemic penumbra is the goal of revascularization therapies.

Focal cerebral infarction occurs via two distinct pathways **(Fig. 420-1)**: (1) a necrotic pathway in which cellular cytoskeletal breakdown is rapid, due principally to energy failure of the cell; and (2) an apoptotic pathway in which cells become programmed to die. Ischemia produces necrosis by starving neurons of glucose and oxygen, which in turn results in failure of mitochondria to produce ATP. Without ATP, membrane ion pumps stop functioning and neurons depolarize, allowing intracellular calcium to rise. Cellular depolarization also causes glutamate release from synaptic terminals; excess extracellular glutamate produces neurotoxicity by activating postsynaptic glutamate receptors that increase neuronal calcium influx. Free radicals are produced by degradation of membrane lipids and mitochondrial dysfunction. Free radicals cause catalytic destruction of membranes and likely damage other

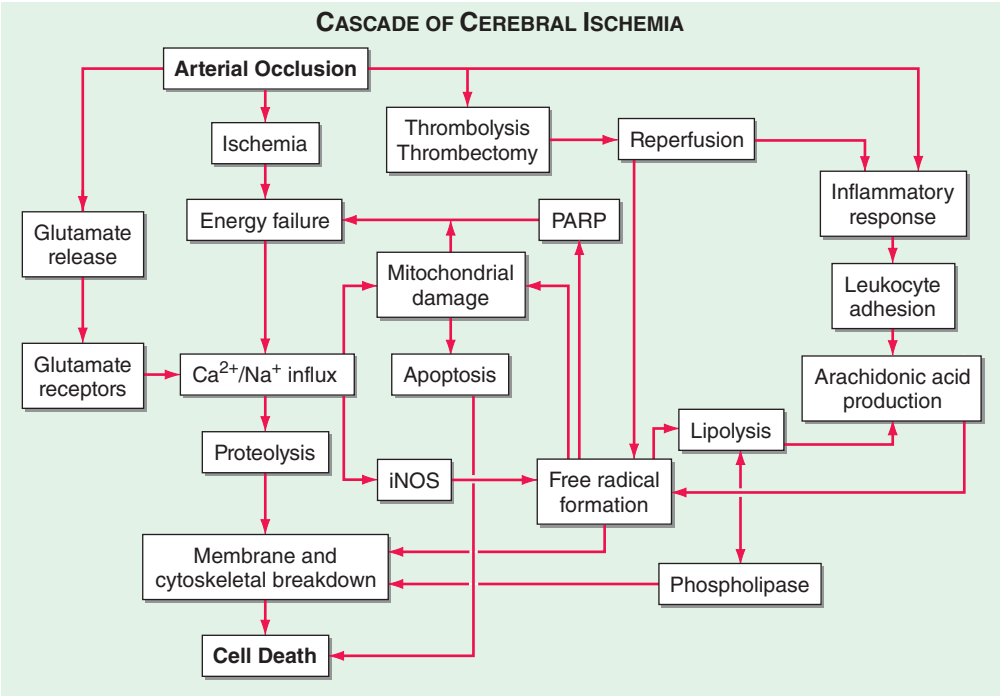


FIGURE 420-1 Major steps in the cascade of cerebral ischemia. See text for details. iNOS, inducible nitric oxide synthase; PARP, poly-A ribose polymerase.

vital functions of cells. Lesser degrees of ischemia, as are seen within the ischemic penumbra, favor apoptotic cellular death causing cells to die days to weeks later. Fever dramatically worsens brain injury during ischemia, as does hyperglycemia (glucose >11.1 mmol/L [200 mg/dL]), so it is reasonable to suppress fever and prevent hyperglycemia as much as possible. The value of induced mild hypothermia to improve stroke outcomes is the subject of continuing clinical research.

TREATMENT

Acute Ischemic Stroke

After the clinical diagnosis of stroke is made, an orderly process of evaluation and treatment should follow. The first goal is to prevent or reverse brain injury. Attend to the patient's airway, breathing, and circulation (ABCs), and treat hypoglycemia or hyperglycemia if identified by finger stick testing. Perform an emergency noncontrast head CT scan to differentiate between ischemic stroke and hemorrhagic stroke; there are no reliable clinical findings that conclusively separate ischemia from hemorrhage, although a more depressed level of consciousness, higher initial blood pressure, or worsening of symptoms after onset favor hemorrhage, and a deficit that is maximal at onset, or remits, suggests ischemia. Treatments designed to reverse or lessen the amount of tissue infarction and improve clinical outcome fall within six categories: (1) medical support, (2) IV thrombolysis, (3) endovascular revascularization, (4) antithrombotic treatment, (5) neuroprotection, and (6) stroke centers and rehabilitation.

MEDICAL SUPPORT

When ischemic stroke occurs, the immediate goal is to optimize cerebral perfusion in the surrounding ischemic penumbra. Attention is also directed toward preventing the common complications of bedridden patients—infections (pneumonia, urinary, and skin) and deep-venous thrombosis (DVT) with pulmonary embolism. Subcutaneous heparin (unfractionated and low-molecular-weight) is safe and can be used concomitantly. Use of pneumatic compression stockings is of proven benefit in reducing risk of DVT and is a safe alternative to heparin.

Because collateral blood flow within the ischemic brain may be blood pressure dependent, there is controversy about whether blood pressure should be lowered acutely. Blood pressure should be reduced if it exceeds 220/120 mmHg, if there is malignant hypertension (Chap. 271), concomitant myocardial ischemia, or if blood pressure is >185/110 mmHg and thrombolytic therapy is anticipated. When faced with the competing demands of myocardium and brain, lowering the heart rate with a β_1-adrenergic blocker (such as esmolol) can be a first step to decrease cardiac work and maintain blood pressure. Routine lowering of blood pressure below the limits listed above has the potential to worsen outcomes. Fever is detrimental and should be treated with antipyretics and surface cooling. Serum glucose should be monitored and kept <10.0 mmol/L (180 mg/dL) using an insulin infusion if necessary, and above at least 3.3 mmol/L (60 mg/dL).

Between 5 and 10% of patients develop enough cerebral edema to cause obtundation or brain herniation. Edema peaks on the second or third day but can cause mass effect for ~10 days. The larger the infarct, the greater the likelihood that clinically significant edema will develop. Water restriction and IV mannitol may be used to raise the serum osmolarity, but hypovolemia should be avoided because this may contribute to hypotension and worsening infarction. Combined analysis of three randomized European trials of hemicraniectomy (craniotomy and temporary removal of part of the skull) shows that hemicraniectomy reduces mortality by 50%, and the clinical outcomes of survivors are significantly improved. Older patients (age >60 years) benefit less, but still significantly. The size of the diffusion-weighted imaging volume of brain infarction during the acute stroke is a predictor of deterioration requiring hemicraniectomy.

Special vigilance is warranted for patients with cerebellar infarction. These strokes may mimic labyrinthitis because of prominent vertigo and vomiting; the presence of head or neck pain should alert the physician to consider cerebellar stroke due to vertebral artery dissection. Even small amounts of cerebellar edema can acutely increase intracranial pressure (ICP) by obstructing cerebrospinal fluid (CSF) flow leading to hydrocephalus or by directly compressing the brainstem. The resulting brainstem compression can manifest as coma and respiratory arrest and require emergency surgical decompression. Suboccipital decompression is recommended in patients with cerebellar infarcts who demonstrate neurological deterioration and should be performed before significant brainstem compression occurs.

INTRAVENOUS THROMBOLYSIS

The National Institute of Neurological Disorders and Stroke (NINDS) rtPA Stroke Study showed a clear benefit for IV rtPA in selected patients with acute stroke. The NINDS study used IV rtPA (0.9 mg/kg to a 90-mg maximum; 10% as a bolus, then the remainder over 60 min) versus placebo in ischemic stroke within 3 h of onset. One-half of the patients were treated within 90 min. Symptomatic intracranial hemorrhage occurred in 6.4% of patients on rtPA and 0.6% on placebo. In the rTPA group, there was a significant 12% absolute increase in the number of patients with only minimal disability (32% on placebo and 44% on rtPA) and a nonsignificant 4% reduction in mortality (21% on placebo and 17% on rtPA). Thus, despite an increased incidence of symptomatic intracranial hemorrhage, treatment with IV rtPA within 3 h of the onset of ischemic stroke improved clinical outcome.

Three subsequent trials of IV rtPA did not confirm this benefit, perhaps because of the dose of rtPA used, the timing of its delivery, and small sample size. When data from all randomized IV rtPA trails were combined, however, efficacy was confirmed in the <3-h time window, and efficacy likely extended to 4.5 h and possibly to 6 h. Based on these combined results, the European Cooperative Acute Stroke Study (ECASS) III explored the safety and efficacy of rtPA in the 3- to 4.5-h time window. Unlike the NINDS study, patients aged >80 years and diabetic patients with a previous stroke were excluded. In this 821-patient randomized study, efficacy was again confirmed, although the treatment effect was less robust than in the 0- to 3-h time window. In the rtPA group, 52.4% of patients achieved a good outcome at 90 days, compared to 45.2% of the placebo group (odds ratio [OR] 1.34, p = .04). The symptomatic intracranial hemorrhage rate was 2.4% in the rtPA group and 0.2% in the placebo group (p = .008).

Based on these data, rtPA is approved in the 3- to 4.5-h window in Europe and Canada, but is still only approved for 0–3 h in the United States. A dose of 0.6 mg/kg is typically used in Japan and other Asian countries based on observation of >600 patients given this lower dose, and observing similar outcomes to historical controls and a lower rate of intracranial hemorrhage. This dose also mitigates concerns that patients of Asian descent have a higher propensity to bleed from most antithrombotic and thrombolytic medications.. Use of IV tPA is now considered a central component of primary stroke centers (see below). It represents the first treatment proven to improve clinical outcomes in ischemic stroke and is cost-effective and cost-saving. Advanced neuroimaging techniques **(see Chap. 419)** may help to select patients beyond the 4.5-h window who will benefit from thrombolysis. The time of stroke onset is defined as the time the patient's symptoms were witnessed to begin or the time the patient was last seen as normal. Patients who awaken with stroke have the onset defined as when they went to bed. **Table 420-1** summarizes eligibility criteria and instructions for administration of IV rtPA.

TABLE 420-1 Administration of Intravenous Recombinant Tissue Plasminogen Activator (rtPA) for Acute Ischemic Stroke (AIS)[a]

INDICATION	CONTRAINDICATION
Clinical diagnosis of stroke	Sustained BP >185/110 mmHg despite treatment
Onset of symptoms to time of drug administration ≤4.5 h[b]	Bleeding diathesis
CT scan showing no hemorrhage or edema of >1/3 of the MCA territory	Recent head injury or intracerebral hemorrhage
Age 18 ≥ years	Major surgery in preceding 14 days
	Gastrointestinal bleeding in preceding 21 days
	Recent myocardial infarction

Administration of rtPA
IV access with two peripheral IV lines (avoid arterial or central line placement)
Review eligibility for rtPA
Administer 0.9 mg/kg IV (maximum 90 mg) IV as 10% of total dose by bolus, followed by remainder of total dose over 1 h[c]
Frequent cuff blood pressure monitoring
No other antithrombotic treatment for 24 h
For decline in neurologic status or uncontrolled blood pressure, stop infusion, give cryoprecipitate, and reimage brain emergently
Avoid urethral catheterization for ≥2 h

[a]See Activase (tissue plasminogen activator) package insert for complete list of contraindications and dosing. [b]Depending on the country, IV rtPA may be approved for up to 4.5 h with additional restrictions. [c]A dose of 0.6 mg/kg is commonly used in Asia (Japan and China) based on randomized data indicating less hemorrhage and similar efficacy using this lower-dose.

Abbreviations: BP, blood pressure; CT, computed tomography; HCT, hematocrit; INR, international normalized ratio; MCA, middle cerebral artery; PTT, partial thromboplastin time.

ENDOVASCULAR REVASCULARIZATION

Ischemic stroke from large-vessel intracranial occlusion results in high rates of mortality and morbidity. Occlusions in such large vessels (middle cerebral artery [MCA], intracranial internal carotid artery, and the basilar artery) generally involve a large clot volume and often fail to open with IV rtPA alone. As proof of concept, thrombolytics were tested via an intraarterial route to increase the concentration of drug at the clot and minimize systemic bleeding complications. The Prolyse in Acute Cerebral Thromboembolism (PROACT) II trial found benefit for intraarterial prourokinase in acute MCA occlusions up to the sixth hour following onset of stroke. Intraarterial treatment of basilar artery occlusions may also be beneficial for selected patients but has not been tested in a randomized trial. Intraarterial administration of a thrombolytic agent for acute ischemic stroke (AIS) is not approved by the U.S. Food and Drug Administration (FDA); however, based on these data many stroke centers consider this treatment if more advanced techniques of mechanical thrombectomy fail.

Endovascular mechanical thrombectomy has been studied as an alternative or adjunctive treatment of acute stroke in patients who are ineligible for, or have contraindications to, thrombolytics or in those who failed to achieve vascular recanalization with IV thrombolytics (**see Fig. 419-12**). First generation thrombectomy devices produced promising results with recanalization in observational studies, leading to FDA approval. Three randomized stroke trials published in 2013 concluded that endovascular therapy did not improve outcomes, but results may have been influenced by methodologic issues: angiography was not required for study entry and less effective mechanical devices were employed. In 2015, the results of six randomized trials were published, all demonstrating that endovascular therapy improved clinical outcomes for internal carotid and MCA occlusions proven by CTA, under 6 h from stroke onset, with or without pretreatment with IV t-PA. One study concluded that patients were home nearly 2 months earlier if they received endovascular therapy. A combined meta-analysis of all 1287 patients in these trials confirmed a large benefit with endovascular therapy (OR 2.49, 95% CI 1.76–3.53; p<0.001). The percentage of patients who achieved modified Rankin scores of 0–2 (normal or symptomatic but independent) was 46% in the endovascular group and 26.5% in the medical arm. Mortality was unchanged. As with IV t-PA treatment, clinical outcome is dependent on time to effective therapy. The odds of a good outcome exceed 3 if groin puncture occurs within 2 h of symptom onset, but is only 2 if 8 h elapse. Over 80% of patients who had vessel opening within 1 h of arrival to the Emergency Department had a good outcome, while only one-third had a good outcome if 6 h elapsed.

Extending the time window beyond 6 h appears to be effective if the patient has specific imaging findings demonstrating good vascular collaterals (CT perfusion or MR perfusion techniques, **see Chap. 419**) and can be treated within 24 h; the Clinical Mismatch in the Triage of Wake Up and Late Presenting Strokes Undergoing Neurointervention With Trevo (DAWN) trial reported good outcomes more frequently with endovascular therapy than with medical care alone (47 vs 13%, p<0.0001). The Endovascular Therapy Following Imaging Evaluation for Ischemic Stroke 3 (DEFUSE-3) trial confirmed these results (45 vs 17%, p<0.001) if treated up to 16 hours from stroke onset.

Now that endovascular stroke therapy is proven to be effective, the creation of comprehensive stroke centers designed to rapidly identify and treat patients with large vessel cerebral ischemia are a major focus internationally. Creating geographical systems of care whereby stroke patients are first evaluated at primary stroke centers (which can administer IV t-PA) then transferred to comprehensive centers if needed, or directly triaged to comprehensive centers based on field assessment, appears to be an effective strategy to improve patient outcomes.

ANTITHROMBOTIC TREATMENT

Platelet Inhibition Aspirin is the only antiplatelet agent that has been proven to be effective for the acute treatment of ischemic stroke; there are several antiplatelet agents proven for the secondary prevention of stroke (see below). Two large trials, the International Stroke Trial (IST) and the Chinese Acute Stroke Trial (CAST), found that the use of aspirin within 48 h of stroke onset reduced both stroke recurrence risk and mortality minimally. Among 19,435 patients in IST, those allocated to aspirin, 300 mg/d, had slightly fewer deaths within 14 days (9.0 vs 9.4%), significantly fewer recurrent ischemic strokes (2.8 vs 3.9%), no excess of hemorrhagic strokes (0.9 vs 0.8%), and a trend toward a reduction in death or dependence at 6 months (61.2 vs 63.5%). In CAST, 21,106 patients with ischemic stroke received 160 mg/d of aspirin or a placebo for up to 4 weeks. There were very small reductions in the aspirin group in early mortality (3.3 vs 3.9%), recurrent ischemic strokes (1.6 vs 2.1%), and dependency at discharge or death (30.5 vs 31.6%). These trials demonstrate that the use of aspirin in the treatment of AIS is safe and produces a small net benefit. For every 1000 acute strokes treated with aspirin, about 9 deaths or nonfatal stroke recurrences will be prevented in the first few weeks and ~13 fewer patients will be dead or dependent at 6 months.

Anticoagulation Numerous clinical trials have failed to demonstrate any benefit of routine anticoagulation in the primary treatment of atherothrombotic cerebral ischemia, and have also shown an increase in the risk of brain and systemic hemorrhage. Therefore the routine use of heparin or other anticoagulants for patients with atherothrombotic stroke is not warranted. Heparin and oral anticoagulation are likely no more effective than aspirin for stroke associated with arterial dissection. However, there may be benefit of anticoagulation for halting progression of dural sinus thrombosis.

NEUROPROTECTION

Neuroprotection is the concept of providing a treatment that prolongs the brain's tolerance to ischemia. Drugs that block the excitatory amino acid pathways have been shown to protect neurons and glia in animals, but despite multiple human trials, they have not yet been proven to be beneficial. Hypothermia is a powerful neuroprotective treatment in patients with cardiac arrest **(Chap. 301)** and is neuroprotective in animal models of stroke, but it has not been adequately studied in patients with ischemic stroke and is associated with an increase in pneumonia rates that could adversely impact stroke outcomes.

STROKE CENTERS AND REHABILITATION

Patient care in stroke units followed by rehabilitation services improves neurologic outcomes and reduces mortality. Use of clinical pathways and staff dedicated to the stroke patient can improve care. This includes use of standardized stroke order sets. Stroke teams that provide emergency 24-h evaluation of acute stroke patients for acute medical management and consideration of thrombolysis or endovascular treatments are essential components of primary and comprehensive stroke centers, respectively.

Proper rehabilitation of the stroke patient includes early physical, occupational, and speech therapy. It is directed toward educating the patient and family about the patient's neurologic deficit, preventing the complications of immobility (e.g., pneumonia, DVT and pulmonary embolism, pressure sores of the skin, and muscle contractures), and providing encouragement and instruction in overcoming the deficit. Use of pneumatic compression stockings is of proven benefit in reducing risk of DVT and is a safe alternative to heparin. The goal of rehabilitation is to return the patient home and to maximize recovery by providing a safe, progressive regimen suited to the individual patient. Additionally, the use of constrained movement therapy (immobilizing the unaffected side) has been shown to improve hemiparesis following stroke, even years after the stroke, suggesting that physical therapy can recruit unused neural pathways. Newer robotic therapies appear promising as well. The human nervous system is more adaptable than previously thought, and developing physical and pharmacologic strategies to enhance long-term neural recovery is an active area of research.

ETIOLOGY OF ISCHEMIC STROKE

(Fig. 420-2 and Table 420-2) Although the initial management of AIS often does not depend on the etiology, establishing a cause is essential to reduce the risk of recurrence. Focus should be on atrial fibrillation and carotid atherosclerosis, because these etiologies have proven secondary prevention strategies. The clinical presentation and examination findings often establish the cause of stroke or narrow the possibilities to a few. Judicious use of laboratory testing and imaging studies completes the initial evaluation. Nevertheless, nearly 30% of strokes remain unexplained despite extensive evaluation.

Clinical examination should focus on the peripheral and cervical vascular system (carotid auscultation for bruits and blood pressure), the heart (dysrhythmia, murmurs), extremities (peripheral emboli), and retina (effects of hypertension and cholesterol emboli [Hollenhorst plaques]). A complete neurologic examination is performed to localize the anatomic site of stroke. An imaging study of the brain is nearly always indicated and is required for patients being considered for thrombolysis; it may be combined with CT- or MRI-based angiography to visualize the vasculature of the neck and intracranial vessels **(see "Imaging Studies," Chap. 419)**. A chest x-ray, electrocardiogram (ECG), urinalysis, complete blood count, erythrocyte sedimentation rate (ESR), serum electrolytes, blood urea nitrogen (BUN), creatinine, blood glucose, serum lipid profile, prothrombin time (PT), and partial thromboplastin time (PTT) are often useful and should be considered in all patients. An ECG, and subsequent cardiac telemetry, may demonstrate arrhythmias or reveal evidence of recent myocardial infarction (MI). Of all these studies, only brain imaging and capillary blood glucose are necessary prior to IV rtPA; the results of other studies should not delay the rapid administration of IV rtPA if the patient is eligible.

Cardioembolic Stroke Cardioembolism is responsible for ~20% of all ischemic strokes. Stroke caused by heart disease is primarily due to embolism of thrombotic material forming on the atrial or ventricular wall or the left heart valves. These thrombi then detach and embolize into the arterial circulation. The thrombus may fragment or lyse quickly, producing only a TIA. Alternatively, the arterial occlusion may last longer, producing stroke. Embolic strokes tend to occur suddenly with maximum neurologic deficit present at onset. With reperfusion following more prolonged ischemia, petechial hemorrhages can occur within the ischemic territory. These are usually of no clinical significance and should be distinguished from frank intracranial hemorrhage into a region of ischemic stroke where the mass effect from the hemorrhage can cause a significant decline in neurologic function.

Emboli from the heart most often lodge in the intracranial internal carotid artery, the MCA, the posterior cerebral artery (PCA), or one of their branches; infrequently, the anterior cerebral artery (ACA) is involved. Emboli large enough to occlude the stem of the MCA (3–4 mm) or internal carotid terminus lead to large infarcts that involve both deep gray and white matter and some portions of the cortical surface and its underlying white matter. A smaller embolus may occlude a small cortical or penetrating arterial branch. The location and size of an infarct within a vascular territory depend on the extent of the collateral circulation.

The most significant cause of cardioembolic stroke in most of the world is nonrheumatic (often called nonvalvular) atrial fibrillation. MI, prosthetic valves, rheumatic heart disease, and ischemic

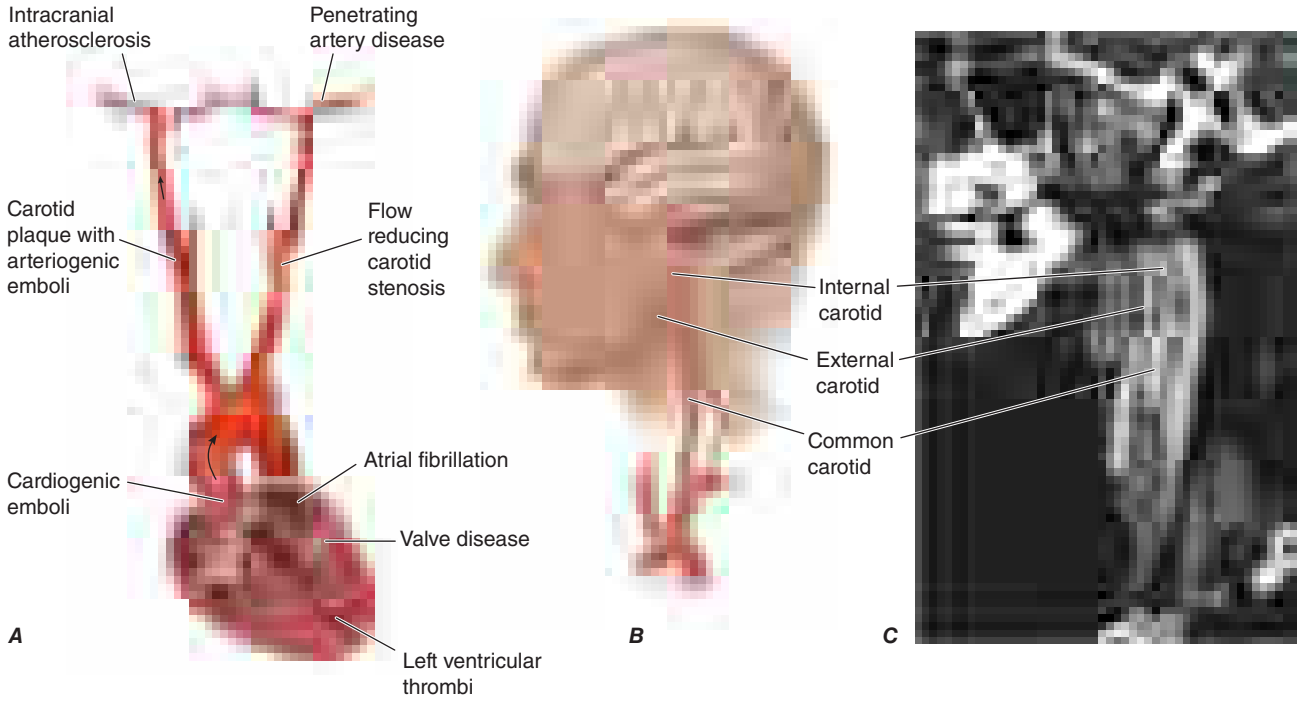


FIGURE 420-2 Pathophysiology of ischemic stroke. ***A.*** Diagram illustrating the three major mechanisms that underlie ischemic stroke: (1) occlusion of an intracranial vessel by an embolus that arises at a distant site (e.g., cardiogenic sources such as atrial fibrillation or artery-to-artery emboli from carotid atherosclerotic plaque), often affecting the large intracranial vessels; (2) in situ thrombosis of an intracranial vessel, typically affecting the small penetrating arteries that arise from the major intracranial arteries; (3) hypoperfusion caused by flow-limiting stenosis of a major extracranial (e.g., internal carotid) or intracranial vessel, often producing "watershed" ischemia. ***B.*** and ***C.*** Diagram and reformatted computed tomography angiogram of the common, internal, and external carotid arteries. High-grade stenosis of the internal carotid artery, which may be associated with either cerebral emboli or flow-limiting ischemia, was identified in this patient.

cardiomyopathy are other considerations (Table 420-2). Cardiac disorders causing brain embolism are discussed in the chapters on heart diseases, but a few pertinent aspects are highlighted here.

Nonrheumatic atrial fibrillation is the most common cause of cerebral embolism overall. The presumed stroke mechanism is thrombus formation in the fibrillating atrium or atrial appendage, with subsequent embolization. Patients with atrial fibrillation have an average annual risk of stroke of ~5%. The risk of stroke can be estimated by calculating the CHA_2DS_2-VASc score **(Table 420-3)**. Left atrial enlargement is an additional risk factor for formation of atrial thrombi. Rheumatic heart disease usually causes ischemic stroke when there is prominent mitral stenosis or atrial fibrillation. Recent MI may be a source of emboli, especially when transmural and involving the anteroapical ventricular wall, and prophylactic anticoagulation following MI has been shown to reduce stroke risk. Mitral valve prolapse is not usually a source of emboli unless the prolapse is severe.

Paradoxical embolization occurs when venous thrombi migrate to the arterial circulation, usually via a patent foramen ovale (PFO) or atrial septal defect. Bubble-contrast echocardiography (IV injection of agitated saline coupled with either transthoracic or transesophageal echocardiography) can demonstrate a right-to-left cardiac shunt, revealing the conduit for paradoxical embolization. Alternatively, a right-to-left shunt is implied if immediately following IV injection of agitated saline, the ultrasound signature of bubbles is observed during transcranial Doppler insonation of the MCA; pulmonary arteriovenous malformations should be considered if this test is positive yet an echocardiogram fails to reveal an intracardiac shunt. Both techniques are highly sensitive for detection of right-to-left shunts. Besides venous clot, fat and tumor emboli, bacterial endocarditis, IV air, and amniotic fluid emboli at childbirth may occasionally be responsible for paradoxical embolization. The importance of a PFO as a cause of stroke is debated, particularly because they are present in ~15% of the general population. Some studies have suggested that the risk is only elevated in the presence of a coexisting atrial septal aneurysm. The presence of a venous source of embolus, most commonly a deep-venous thrombus, may provide confirmation of the importance of a PFO with an accompanying right-to-left shunt in a particular case. Two recent trials found about a 1% per year absolute reduction in stroke risk using percutaneous occlusion devices in patients with no other explanation for their stroke.

Bacterial endocarditis can be a source of valvular vegetations that give rise to septic emboli. The appearance of multifocal symptoms and signs in a patient with stroke makes bacterial endocarditis more likely. Infarcts of microscopic size occur, and large septic infarcts may evolve into brain abscesses or cause hemorrhage into the infarct, which generally precludes use of anticoagulation or thrombolytics. Mycotic aneurysms caused by septic emboli may also present as subarachnoid hemorrhage (SAH) or intracerebral hemorrhage.

Artery-to-Artery Embolic Stroke Thrombus formation on atherosclerotic plaques may embolize to intracranial arteries producing an artery-to-artery embolic stroke. Less commonly, a diseased vessel may acutely thrombose. Unlike the myocardial vessels, artery-to-artery embolism, rather than local thrombosis, appears to be the dominant vascular mechanism causing large-vessel brain ischemia. Any diseased vessel may be an embolic source, including the aortic arch, common carotid, internal carotid, vertebral, and basilar arteries.

CAROTID ATHEROSCLEROSIS Atherosclerosis within the carotid artery occurs most frequently within the common carotid bifurcation and proximal internal carotid artery; the carotid siphon (portion within the cavernous sinus) is also vulnerable to atherosclerosis. Male gender, older age, smoking, hypertension, diabetes, and hypercholesterolemia are risk factors for carotid disease, as they are for stroke in general **(Table 420-4)**. Carotid atherosclerosis produces an estimated 10% of ischemic stroke. **For further discussion of the pathogenesis of atherosclerosis, see Chap. 232.**

Carotid disease can be classified by whether the stenosis is symptomatic or asymptomatic and by the degree of stenosis (percent narrowing of the narrowest segment compared to a nondiseased segment). Symptomatic carotid disease implies that the patient has experienced a stroke or TIA within the vascular distribution of the artery, and it is associated with a greater risk of subsequent stroke than asymptomatic

TABLE 420-2 Causes of Ischemic Stroke

COMMON CAUSES	UNCOMMON CAUSES
Thrombosis	Hypercoagulable disorders
Lacunar stroke (small vessel)	Protein C deficiency[a]
Large-vessel thrombosis	Protein S deficiency[a]
Dehydration	Antithrombin III deficiency[a]
Embolic occlusion	Antiphospholipid syndrome
Artery-to-artery	Factor V Leiden mutation[a]
Carotid bifurcation	Prothrombin G20210 mutation[a]
Aortic arch	Systemic malignancy
Arterial dissection	Sickle cell anemia
Cardioembolic	β Thalassemia
Atrial fibrillation	Polycythemia vera
Mural thrombus	Systemic lupus erythematosus
Myocardial infarction	Homocysteinemia
Dilated cardiomyopathy	Thrombotic thrombocytopenic purpura
Valvular lesions	
Mitral stenosis	Disseminated intravascular coagulation
Mechanical valve	Dysproteinemias[a]
Bacterial endocarditis	Nephrotic syndrome[a]
Paradoxical embolus	Inflammatory bowel disease[a]
Atrial septal defect	Oral contraceptives
Patent foramen ovale	Venous sinus thrombosis[b]
Atrial septal aneurysm	Fibromuscular dysplasia
Spontaneous echo contrast	Vasculitis
Stimulant drugs: cocaine, amphetamine	Systemic vasculitis (PAN, granulomatosis with polyangiitis [Wegener's], Takayasu's, giant cell arteritis)
	Primary CNS vasculitis
	Meningitis (syphilis, tuberculosis, fungal, bacterial, zoster)
	Noninflammatory vasculopathy
	Reversible vasoconstriction syndrome
	Fabry's disease
	Angiocentric lymphoma
	Cardiogenic
	Mitral valve calcification
	Atrial myxoma
	Intracardiac tumor
	Marantic endocarditis
	Libman-Sacks endocarditis
	Subarachnoid hemorrhage vasospasm
	Moyamoya disease
	Eclampsia

[a]Chiefly cause venous sinus thrombosis. [b]May be associated with any hypercoagulable disorder.

Abbreviations: CNS, central nervous system; PAN, polyarteritis nodosa.

TABLE 420-3 Recommendations on Chronic Use of Antithrombotics for Various Cardiac Conditions

CONDITION	RECOMMENDATION
Nonvalvular atrial fibrillation	Calculate CHA_2DS_2-VASc score[a]
• CHA_2DS_2-VASc score of 0	Aspirin or no antithrombotic
• CHA_2DS_2-VASc score of 1	Aspirin or OAC
• CHA_2DS_2-VASc score of 2 or greater	OAC
Rheumatic mitral valve disease	
• With atrial fibrillation, previous embolization, or atrial appendage thrombus, or left atrial diameter >55 mm	OAC
• Embolization or appendage clot despite OAC	OAC plus aspirin
Mitral valve prolapse	
• Asymptomatic	No therapy
• With otherwise cryptogenic stroke or TIA	Aspirin
• Atrial fibrillation	OAC
Mitral annular calcification	
• Without atrial fibrillation but systemic embolization, or otherwise cryptogenic stroke or TIA	Aspirin
• Recurrent embolization despite aspirin	OAC
• With atrial fibrillation	OAC
Aortic valve calcification	
• Asymptomatic	No therapy
• Otherwise cryptogenic stroke or TIA	Aspirin
Aortic arch mobile atheroma	
• Otherwise cryptogenic stroke or TIA	Aspirin or OAC
Patent foramen ovale	
• Otherwise cryptogenic ischemic stroke or TIA	Aspirin or closure with device
• Indication for OAC (deep-venous thrombosis or hypercoagulable state)	OAC
Mechanical heart value	
• Aortic position, bileaflet or Medtronic Hall tilting disk with normal left atrial size and sinus rhythm	VKA INR 2.5, range 2–3
• Mitral position tilting disk or bileaflet valve	VKA INR 3.0, range 2.5–3.5
• Mitral or aortic position, anterior-apical myocardial infarct or left atrial enlargement	VKA INR 3.0, range 2.5–3.5
• Mitral or aortic position, with atrial fibrillation, or hypercoagulable state, or low ejection fraction, or atherosclerotic vascular disease	Aspirin plus VKA INR 3.0, range 2.5–3.5
• Systemic embolization despite target INR	Add aspirin and/or increase INR: prior target was 2.5 increase to 3.0, range 2.5–3.5; prior target was 3.0 increase to 3.5, range 3–4
Bioprosthetic valve	
• No other indication for VKA therapy	Aspirin
Infective endocarditis	Avoid antithrombotic agents
Nonbacterial thrombotic endocarditis	
• With systemic embolization	Full-dose unfractionated heparin or SC LMWH

[a]CHA_2DS_2-VASc score is calculated as follows: 1 point for Congestive heart failure, 1 point for Hypertension, 2 points for Age ≥ 75 y, 1 point for Diabetes mellitus, 2 points for Stroke or TIA, 1 point for Vascular disease (prior MI, peripheral vascular disease or aortic plaque), 1 point for Age 65–74 y, 1 point for female Sex category; sum of point is the total [b]CHA_2DS_2-VASc score.

Note: Dose of aspirin is 50–325 mg/d; target INR for OAC is between 2 and 3 unless otherwise specified.

Abbreviations: INR, international normalized ratio; LMWH, low-molecular-weight heparin; OAC, oral anticoagulant (VKA, thrombin inhibitor, or oral factor Xa inhibitors); TIA, transient ischemic attack; VKA, vitamin K antagonist.

Sources: Modified from DE Singer et al: Chest 133:546S, 2008; DN Salem et al: Chest 133:593S, 2008; CT January et al: JACC 64:2246, 2014.

stenosis, in which the patient is symptom free and the stenosis is detected through screening. Greater degrees of arterial narrowing are generally associated with a higher risk of stroke, except that those with near occlusions are at lower risk of stroke.

OTHER CAUSES OF ARTERY-TO-ARTERY EMBOLIC STROKE *Intracranial atherosclerosis* produces stroke either by an embolic mechanism or by in situ thrombosis of a diseased vessel. It is more common in patients of Asian and African-American descent. Recurrent stroke risk is ~15% per year, similar to untreated symptomatic carotid atherosclerosis.

Dissection of the internal carotid or vertebral arteries or even vessels beyond the circle of Willis is a common source of embolic stroke in young (age <60 years) patients. The dissection is usually painful and precedes the stroke by several hours or days. Extracranial dissections do not cause hemorrhage, presumably because of the tough adventitia

TABLE 420-4 Risk Factors for Stroke

RISK FACTOR	RELATIVE RISK	RELATIVE RISK REDUCTION WITH TREATMENT	NUMBER NEEDED TO TREAT[a]	
			PRIMARY PREVENTION	SECONDARY PREVENTION
Hypertension	2–5	38%	100–300	50–100
Atrial fibrillation	1.8–2.9	68% warfarin, 21% aspirin	20–83	13
Diabetes	1.8–6	No proven effect		
Smoking	1.8	50% at 1 year, baseline risk at 5 years postcessation		
Hyperlipidemia	1.8–2.6	16–30%	560	230
Asymptomatic carotid stenosis	2.0	53%	85	N/A
Symptomatic carotid stenosis (70–99%)		65% at 2 years	N/A	12
Symptomatic carotid stenosis (50–69%)		29% at 5 years	N/A	77

[a]Number needed to treat to prevent one stroke annually. Prevention of other cardiovascular outcomes is not considered here.

Abbreviation: N/A, not applicable.

of these vessels. Intracranial dissections, conversely, may produce SAH because the adventitia of intracranial vessels is thin and pseudoaneurysms may form, requiring urgent treatment to prevent rerupture. Treating asymptomatic pseudoaneurysms following dissection is likely not necessary. The cause of dissection is usually unknown, and recurrence is rare. Ehlers-Danlos type IV, Marfan's disease, cystic medial necrosis, and fibromuscular dysplasia are associated with dissections. Trauma (usually a motor vehicle accident or a sports injury) can cause carotid and vertebral artery dissections. Spinal manipulative therapy is associated with vertebral artery dissection and stroke. Most dissections heal spontaneously, and stroke or TIA is uncommon beyond 2 weeks. A recent trial showed no difference in stroke prevention with aspirin compared to anticoagulation, with a low recurrent stroke rate of 2%.

SMALL-VESSEL STROKE

The term *lacunar infarction* refers to infarction following atherothrombotic or lipohyalinotic occlusion of a small artery in the brain. The term *small-vessel stroke* denotes occlusion of such a small penetrating artery and is now the preferred term. Small-vessel strokes account for ~20% of all strokes.

Pathophysiology The MCA stem, the arteries comprising the circle of Willis (A1 segment, anterior and posterior communicating arteries, and P1 segment), and the basilar and vertebral arteries all give rise to 30- to 300-μm branches that penetrate the deep gray and white matter of the cerebrum or brainstem **(Fig. 420-3)**. Each of these small branches can occlude either by atherothrombotic disease at its origin or by the development of lipohyalinotic thickening. Thrombosis of these vessels causes small infarcts that are referred to as *lacunes* (Latin for "lake" of fluid noted at autopsy). These infarcts range in size from 3 mm to 2 cm in diameter. Hypertension and age are the principal risk factors.

Clinical Manifestations The most common small-vessel stroke syndromes are the following: (1) *Pure motor hemiparesis* from an infarct in the posterior limb of the internal capsule or the pons; the face, arm, and leg are almost always involved; (2) *pure sensory stroke* from an infarct in the ventral thalamus; (3) *ataxic hemiparesis* from an infarct in the ventral pons or internal capsule; (4) and *dysarthria and a clumsy hand* or arm due to infarction in the ventral pons or in the genu of the internal capsule.

Transient symptoms (small-vessel TIAs) may herald a small-vessel infarct; they may occur several times a day and last only a few minutes. Recovery from small-vessel strokes tends to be more rapid and complete than recovery from large-vessel strokes; in some cases, however, there is severe permanent disability.

A large-vessel source (either thrombosis or embolism) may manifest initially as a small-vessel infarction. Therefore, the search for embolic sources (carotid and heart) should not be completely abandoned in the evaluation of these patients. Secondary prevention of small-vessel stroke involves risk factor modification, specifically reduction in blood pressure (see "Treatment: Primary and Secondary Prevention of Stroke and TIA," below).

LESS COMMON CAUSES OF STROKE

(Table 420-2) *Hypercoagulable disorders* **(Chap. 61)** primarily increase the risk of cortical vein or cerebral venous sinus thrombosis. Systemic lupus erythematosus with Libman-Sacks endocarditis can be a cause of embolic stroke. These conditions overlap with the antiphospholipid syndrome, which probably requires long-term anticoagulation to prevent further stroke. Homocysteinemia may cause arterial thromboses as well; this disorder is caused by various mutations in the homocysteine pathways and responds to different forms of cobalamin depending on the mutation.

Venous sinus thrombosis of the lateral or sagittal sinus or of small cortical veins (cortical vein thrombosis) occurs as a complication of oral contraceptive use, pregnancy and the postpartum period, inflammatory bowel disease, intracranial infections (meningitis), and dehydration. It is also seen in patients with laboratory-confirmed thrombophilia including antiphospholipid syndrome, polycythemia, sickle cell anemia, deficiencies of proteins C and S, factor V Leiden mutation (resistance to activated protein C), antithrombin III deficiency, homocysteinemia, and the prothrombin G20210 mutation. Women who take oral contraceptives and have the prothrombin G20210 mutation may be at particularly high risk for sinus thrombosis. Patients present with headache and may also have focal neurologic signs (especially paraparesis) and seizures. Often, CT imaging is normal unless an intracranial venous hemorrhage has occurred, but the venous sinus occlusion is readily visualized using magnetic resonance (MR) or CT venography or conventional x-ray angiography. With greater degrees of sinus thrombosis, the patient may develop signs of increased ICP and coma. Intravenous heparin, regardless of the presence of intracranial hemorrhage, reduces morbidity and mortality, and the long-term outcome is generally good. Heparin prevents further thrombosis and reduces venous hypertension and ischemia. If an underlying hypercoagulable state is not found, many physicians treat with vitamin K antagonists (VKAs) for 3–6 months and then convert to aspirin, depending on the degree of resolution of the venous sinus thrombus. Anticoagulation is often continued indefinitely if thrombophilia is diagnosed.

Sickle cell anemia (SS disease) is a common cause of stroke in children. A subset of homozygous carriers of this hemoglobin mutation develop stroke in childhood, and this may be predicted by documenting high-velocity blood flow within the MCAs using transcranial Doppler ultrasonography. In children who are identified to have high velocities, treatment with aggressive exchange transfusion dramatically reduces risk of stroke, and if exchange transfusion is ceased, their stroke rate increases again along with MCA velocities.

Fibromuscular dysplasia affects the cervical arteries and occurs mainly in women. The carotid or vertebral arteries show multiple rings of segmental narrowing alternating with dilatation. Vascular occlusion is usually incomplete. The process is often asymptomatic but occasionally

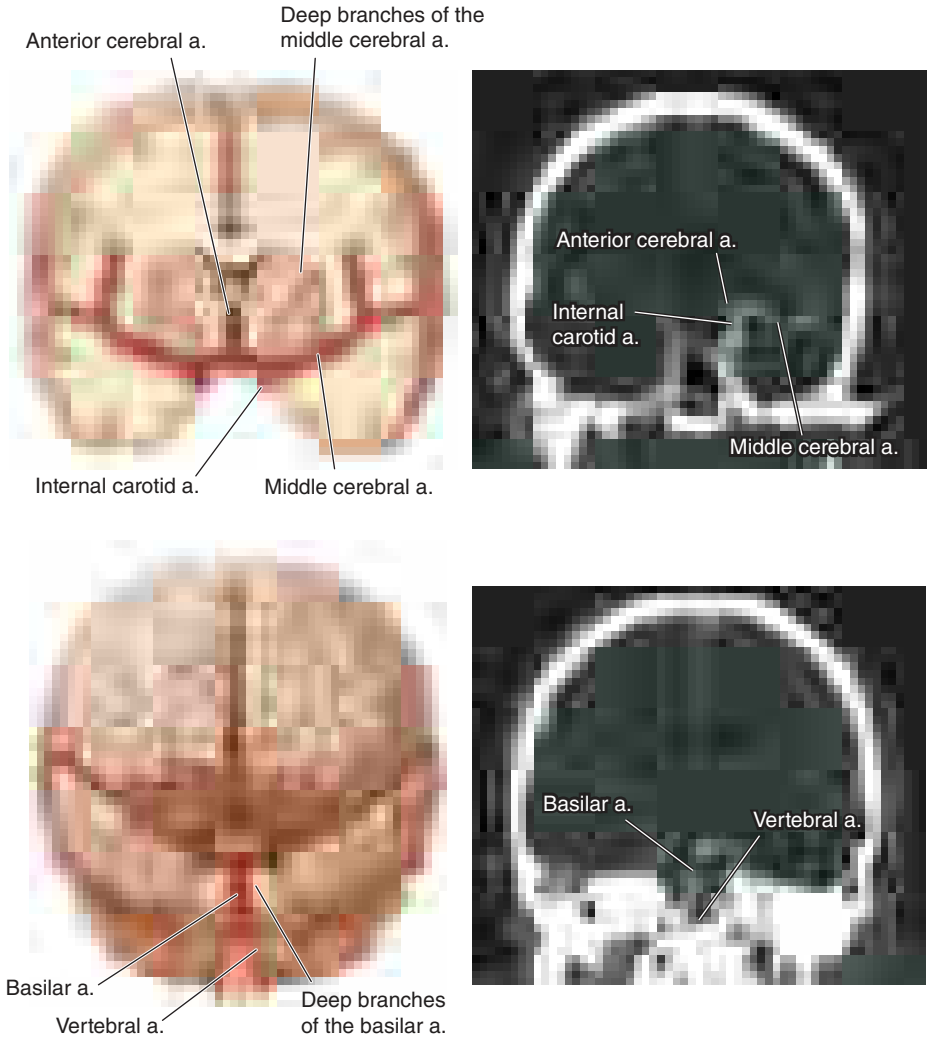


FIGURE 420-3 Diagrams and reformatted computed tomography (CT) angiograms in the coronal section illustrating the deep penetrating arteries involved in small-vessel strokes. In the anterior circulation, small penetrating arteries called *lenticulostriates* arise from the proximal portion of the anterior and middle cerebral arteries and supply deep subcortical structures (***upper panels***). In the posterior circulation, similar arteries arise directly from the vertebral and basilar arteries to supply the brainstem (***lower panels***). Occlusion of a single penetrating artery gives rise to a discrete area of infarct (pathologically termed a "lacune," or lake). Note that these vessels are too small to be visualized on CT angiography.

is associated with an audible bruit, TIAs, or stroke. Involvement of the renal arteries is common and may cause hypertension. The cause and natural history of fibromuscular dysplasia are unknown **(Chap. 275)**. TIA or stroke generally occurs only when the artery is severely narrowed or dissects. Anticoagulation or antiplatelet therapy may be helpful.

Temporal (giant cell) arteritis **(Chap. 356)** is a relatively common affliction of elderly individuals in which the external carotid system, particularly the temporal arteries, undergo subacute granulomatous inflammation with giant cells. Occlusion of posterior ciliary arteries derived from the ophthalmic artery results in blindness in one or both eyes and can be prevented with glucocorticoids. It rarely causes stroke because the internal carotid artery is usually not inflamed. Idiopathic giant cell arteritis involving the great vessels arising from the aortic arch (*Takayasu's arteritis*) may cause carotid or vertebral thrombosis; it is rare in the Western Hemisphere.

Necrotizing (or granulomatous) arteritis, occurring alone or in association with generalized polyarteritis nodosa or granulomatosis with polyangiitis (Wegener's), involves the distal small branches (<2 mm diameter) of the main intracranial arteries and produces small ischemic infarcts in the brain, optic nerve, and spinal cord. The CSF often shows pleocytosis, and the protein level is elevated. *Primary central nervous system vasculitis* is rare; small or medium-sized vessels are usually affected, without apparent systemic vasculitis. The differential diagnosis includes other inflammatory vasculopathies including infection (tuberculous, fungal), sarcoidosis, angiocentric lymphoma, carcinomatous meningitis, and noninflammatory causes such as atherosclerosis, emboli, connective tissue disease, vasospasm, migraine-associated vasculopathy, and drug-associated causes. Some cases develop in the postpartum period and are self-limited.

Patients with any form of vasculopathy may present with insidious progression of combined white and gray matter infarctions, prominent headache, and cognitive decline. Brain biopsy or high-resolution conventional x-ray angiography is usually required to make the diagnosis **(Fig. 420-4)**. An inflammatory profile (elevated WBCs, elevated IgG index, bands on electrophoresis) found on lumbar puncture favors an inflammatory cause. In cases where inflammation is confirmed, aggressive immunosuppression with glucocorticoids, and often cyclophosphamide, is usually necessary to prevent progression; a diligent investigation for infectious causes such as tuberculosis is essential prior to immunosuppression. With prompt recognition and treatment, many patients can make an excellent recovery.

Drugs, in particular amphetamines and perhaps cocaine, may cause stroke on the basis of acute hypertension or drug-induced vasculopathy. This vasculopathy is commonly due to vasospasm or atherosclerosis but cases of inflammatory vasculitis have also been reported. No data exist on the value of any treatment, but cessation of stimulants is prudent. Phenylpropanolamine has been linked with intracranial hemorrhage, as has cocaine and methamphetamine, perhaps related to a drug-induced vasculopathy. *Moyamoya disease* is a poorly understood occlusive disease involving large intracranial arteries, especially the distal internal carotid artery and the stem of the MCA and

FIGURE 420-4 Cerebral angiogram from a 32-year-old male with central nervous system vasculopathy. Dramatic beading (*arrows*) typical of vasculopathy is seen.

ACA. Vascular inflammation is absent. The lenticulostriate arteries develop a rich collateral circulation around the occlusive lesion, which gives the impression of a "puff of smoke" (*moyamoya* in Japanese) on conventional x-ray angiography. Other collaterals include transdural anastomoses between the cortical surface branches of the meningeal and scalp arteries. The disease occurs mainly in Asian children or young adults, but the appearance may be identical in adults who have atherosclerosis, particularly in association with diabetes. Intracranial hemorrhage may result from rupture of the moyamoya collaterals; thus, anticoagulation is risky. Progressive occlusion of large surface arteries can occur, producing large-artery distribution strokes. Surgical bypass of extracranial carotid arteries to the dura or MCAs may prevent stroke and hemorrhage.

Posterior reversible encephalopathy syndrome (PRES) can occur with head injury, seizure, migraine, sympathomimetic drug use, eclampsia, and in the postpartum period. The pathophysiology is uncertain but likely involves a hyperperfusion state where blood pressure exceeds the upper limit of cerebral autoregulation resulting in cerebral edema **(Chap. 301)**. Patients complain of headache and manifest fluctuating neurologic symptoms and signs, especially visual symptoms. Sometimes cerebral infarction ensues, but typically the clinical and imaging findings reverse completely. MRI findings are characteristic with the edema present within the occipital lobes but can be generalized and do not respect any single vascular territory. A closely related *reversible cerebral vasoconstriction syndrome* (RCVS) typically presents with sudden, severe headache closely mimicking SAH. Patients may experience ischemic infarction and intracerebral hemorrhage and typically have new-onset, severe hypertension. Conventional x-ray angiography reveals changes in the vascular caliber throughout the hemispheres resembling vasculitis, but the process is noninflammatory. Oral calcium channel blockers may be effective in producing remission, and recurrence is rare.

Leukoaraiosis, or *periventricular white matter disease*, is the result of multiple small-vessel infarcts within the subcortical white matter. It is readily seen on CT or MRI scans as areas of white matter injury surrounding the ventricles and within the corona radiata. The pathophysiologic basis of the disease is lipohyalinosis of small penetrating arteries within the white matter, likely produced by chronic hypertension. Patients with periventricular white matter disease may develop a subcortical dementia syndrome, and it is likely that this common form of dementia may be delayed or prevented with antihypertensive medications **(Chap. 425)**.

CADASIL (cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy) is an inherited disorder that presents as small-vessel strokes, progressive dementia, and extensive symmetric white matter changes often including the anterior temporal lobes visualized by MRI. Approximately 40% of patients have migraine with aura, often manifest as transient motor or sensory deficits. Onset is usually in the fourth or fifth decade of life. This autosomal dominant condition is caused by one of several mutations in *Notch-3*, a member of a highly conserved gene family characterized by epidermal growth factor repeats in its extracellular domain. Other monogenic ischemic stroke syndromes include cerebral autosomal recessive arteriopathy with subcortical infarcts and leukoencephalopathy (CARASIL) and hereditary endotheliopathy, retinopathy, nephropathy, and stroke (HERNS). Fabry's disease also produces both a large-vessel arteriopathy and small-vessel infarctions. The COL4A1 mutation is associated with multiple small vessel strokes with hemorrhagic transformation.

TRANSIENT ISCHEMIC ATTACKS

TIAs are episodes of stroke symptoms that last only briefly; the standard definition of duration is <24 h, but most TIAs last <1 h. If a relevant brain infarction is identified on brain imaging, the clinical entity is now classified as stroke regardless of the duration of symptoms. A normal brain imaging study following a TIA does not rule-out TIA; rather, the clinical syndrome is diagnostic. The causes of TIA are similar to the causes of ischemic stroke, but because TIAs may herald stroke, they are an important risk factor that should be considered separately and urgently. TIAs may arise from emboli to the brain or from in situ thrombosis of an intracranial vessel. With a TIA, the occluded blood vessel reopens and neurologic function is restored.

TABLE 420-5 Risk of Stroke Following Transient Ischemic Attack: The ABCD² Score

CLINICAL FACTOR	SCORE
A: Age ≥60 years	1
B: SBP >140 mmHg or DBP >90 mmHg	1
C: Clinical symptoms	
Unilateral weakness	2
Speech disturbance without weakness	1
D: Duration	
>60 min	2
10–59 min	1
D: Diabetes (oral medications or insulin)	1
TOTAL SCORE	**SUM EACH CATEGORY**
ABCD² Score Total	3-Month Rate of Stroke (%)[a]
0	0
1	2
2	3
3	3
4	8
5	12
6	17
7	22

[a]Data ranges are from five cohorts.

Abbreviations: DBP, diastolic blood pressure; SBP, systolic blood pressure.

Source: SC Johnston et al: Validation and refinement of score to predict very early stroke risk after transient ischaemic attack. Lancet 369:283, 2007.

The risk of stroke after a TIA is ~10–15% in the first 3 months, with most events occurring in the first 2 days. This risk can be directly estimated using the well-validated ABCD² score **(Table 420-5)**. Therefore, urgent evaluation and treatment are justified. Because etiologies for stroke and TIA are identical, evaluation for TIA should parallel that of stroke (Fig. 420-2). The improvement characteristic of TIA is a contraindication to thrombolysis. However, because the risk of subsequent stroke in the first few days after a TIA is high, the opportunity to give rtPA rapidly if a subsequent stroke occurs may justify hospital admission for most patients. The combination of aspirin and clopidogrel was found to prevent stroke following TIA better than aspirin alone in a large Chinese randomized trial and is undergoing similar evaluation in an ongoing National Institutes of Health (NIH)-sponsored trial (POINT study). Failure to respond to the combination of aspirin and clopidogrel is linked to carriage of a common CYP2C19 polymorphism that leads to poor metabolism of clopidogrel into its active form. This mutation is common, particularly in Asians.

TREATMENT

Primary and Secondary Prevention of Stroke and TIA

GENERAL PRINCIPLES

Many medical and surgical interventions, as well as lifestyle modifications, are available for preventing stroke. Some of these can be widely applied because of their low cost and minimal risk; others are expensive and carry substantial risk but may be valuable for selected high-risk patients. Identification and control of modifiable risk factors, and especially hypertension, is the best strategy to reduce the burden of stroke, and the total number of strokes could be reduced substantially by these means (Table 420-4).

ATHEROSCLEROSIS RISK FACTORS

The relationship of various factors to the risk of atherosclerosis is described in **Chaps. 232 and 233**. Older age, diabetes mellitus, hypertension, tobacco smoking, abnormal blood cholesterol (particularly, low high-density lipoprotein [HDL] and/or elevated low-density lipoprotein [LDL]), and other factors are either proven or probable

risk factors for ischemic stroke, largely by their link to atherosclerosis. Risk of stroke is much greater in those with prior stroke or TIA. Many cardiac conditions predispose to stroke, including atrial fibrillation and recent MI. Oral contraceptives and hormone replacement therapy increase stroke risk, and although rare, certain inherited and acquired hypercoagulable states predispose to stroke.

Hypertension is the most significant of the risk factors; in general, all hypertension should be treated to a target of <130/80 mmHg. Recent data (the Systolic Blood Pressure Intervention Trial—SPRINT) suggest that lowering systolic blood pressure <120 mmHg reduces stroke and heart attack 43% compared to systolic blood pressure <140 mmHg, without an increased risk of syncope or falls. The presence of known cerebrovascular disease is not a contraindication to treatment aimed at achieving normotension. Data are particularly strong in support of thiazide diuretics and angiotensin-converting enzyme inhibitors.

Several trials have confirmed that statin drugs reduce the risk of stroke even in patients without elevated LDL or low HDL. The Stroke Prevention by Aggressive Reduction in Cholesterol Levels (SPARCL) trial showed benefit in secondary stroke reduction for patients with recent stroke or TIA who were prescribed atorvastatin, 80 mg/d. The primary prevention trial, Justification for the Use of Statins in Prevention: An Intervention Trial Evaluating Rosuvastatin (JUPITER), found that patients with low LDL (<130 mg/dL) caused by elevated C-reactive protein benefitted by daily use of this statin. Primary stroke occurrence was reduced by 51% (hazard ratio 0.49, p = .004), and there was no increase in the rates of intracranial hemorrhage. Meta-analysis has also supported a primary treatment effect for statins given acutely for ischemic stroke. Therefore, a statin should be considered in all patients with prior ischemic stroke. Tobacco smoking should be discouraged in all patients **(Chap. 448)**. The use of pioglitazone (an agonist of peroxisome proliferator-activated receptor gamma) in patients with type 2 diabetes and previous stroke does not lower stroke, MI, or vascular death rates, but is effective in lowering vascular events patients with stroke and insulin resistance alone. Diabetes prevention is likely the most effective strategy for primary and secondary stroke prevention.

ANTIPLATELET AGENTS FOR STROKE PREVENTION

Platelet antiaggregation agents can prevent atherothrombotic events, including TIA and stroke, by inhibiting the formation of intraarterial platelet aggregates. These can form on diseased arteries, induce thrombus formation, and occlude or embolize into the distal circulation. Aspirin, clopidogrel, and the combination of aspirin plus extended-release dipyridamole are the antiplatelet agents most commonly used for this purpose. Ticlopidine has been largely abandoned because of its adverse effects but may be used as an alternative to clopidogrel. Ticagrelor has not been found to be better than aspirin for stroke prevention.

Aspirin is the most widely studied antiplatelet agent. Aspirin acetylates platelet cyclooxygenase, which irreversibly inhibits the formation in platelets of thromboxane A_2, a platelet aggregating and vasoconstricting prostaglandin. This effect is permanent and lasts for the usual 8-day life of the platelet. Paradoxically, aspirin also inhibits the formation in endothelial cells of prostacyclin, an antiaggregating and vasodilating prostaglandin. This effect is transient. As soon as aspirin is cleared from the blood, the nucleated endothelial cells again produce prostacyclin. Aspirin in low doses given once daily inhibits the production of thromboxane A_2 in platelets without substantially inhibiting prostacyclin formation. Higher doses of aspirin have not been proven to be more effective than lower doses.

Ticlopidine and clopidogrel block the adenosine diphosphate (ADP) receptor on platelets and thus prevent the cascade resulting in activation of the glycoprotein IIb/IIIa receptor that leads to fibrinogen binding to the platelet and consequent platelet aggregation. Ticlopidine is more effective than aspirin; however, it has the disadvantage of causing diarrhea, skin rash, and, in rare instances, neutropenia and thrombotic thrombocytopenic purpura (TTP). Clopidogrel rarely causes TTP but does not cause neutropenia. The Clopidogrel versus Aspirin in Patients at Risk of Ischemic Events (CAPRIE) trial, which led to FDA approval, found that it was only marginally more effective than aspirin in reducing risk of stroke. The Management of Atherothrombosis with Clopidogrel in High-Risk Patients (MATCH) trial was a large multicenter, randomized, double-blind study that compared clopidogrel in combination with aspirin to clopidogrel alone in the secondary prevention of TIA or stroke. The MATCH trial found no difference in TIA or stroke prevention with this combination, but did show a small but significant increase in major bleeding complications (3 vs 1%). In the Clopidogrel for High Atherothrombotic Risk and Ischemic Stabilization, Management, and Avoidance (CHARISMA) trial, which included a subgroup of patients with prior stroke or TIA along with other groups at high risk of cardiovascular events, there was no benefit of clopidogrel combined with aspirin compared to aspirin alone. Lastly, the SPS3 trial looked at the long-term combination of clopidogrel and aspirin versus clopidogrel alone in small-vessel stroke and found no improvement in stroke prevention and a significant increase in both hemorrhage and death. Thus, the long-term use of clopidogrel in combination with aspirin is not recommended for stroke prevention.

The short-term combination of clopidogrel with aspirin may be effective in preventing second stroke, however. A trial of 5170 Chinese patients enrolled within 24 h of TIA or minor ischemic stroke found that a clopidogrel-aspirin regimen (clopidogrel 300 mg load then 75 mg/d with aspirin 75 mg for the first 21 days) was superior to aspirin (75 mg/d) alone, with 90-day stroke risk decreased from 11.7 to 8.2% (p <.001) and no increase in major hemorrhage. This benefit was limited to those not carrying the CYP2C19 polymorphism associated with clopidogrel hypometabolism. An international NIH-sponsored trial of similar design is ongoing.

Dipyridamole is an antiplatelet agent that inhibits the uptake of adenosine by a variety of cells, including those of the vascular endothelium. The accumulated adenosine is an inhibitor of aggregation. At least in part through its effects on platelet and vessel wall phosphodiesterases, dipyridamole also potentiates the antiaggregatory effects of prostacyclin and nitric oxide produced by the endothelium and acts by inhibiting platelet phosphodiesterase, which is responsible for the breakdown of cyclic AMP. The resulting elevation in cyclic AMP inhibits aggregation of platelets. Dipyridamole is erratically absorbed depending on stomach pH, but a newer formulation combines timed-release dipyridamole, 200 mg, with aspirin, 25 mg, and has better oral bioavailability. This combination drug was studied in three trials. The European Stroke Prevention Study (ESPS) II showed efficacy of both 50 mg/d of aspirin and extended-release dipyridamole in preventing stroke, and a significantly better risk reduction when the two agents were combined. The open-label ESPRIT (European/Australasian Stroke Prevention in Reversible Ischaemia Trial) trial confirmed the ESPS-II results. After 3.5 years of follow-up, 13% of patients on aspirin and dipyridamole and 16% on aspirin alone (hazard ratio 0.80, 95% confidence index [CI] 0.66–0.98) met the primary outcome of death from all vascular causes. In the Prevention Regimen for Effectively Avoiding Second Strokes (PRoFESS) trial, the combination of extended-release dipyridamole and aspirin was compared directly with clopidogrel with and without the angiotensin receptor blocker telmisartan; there were no differences in the rates of second stroke (9% each) or degree of disability in patients with median follow-up of 2.4 years. Telmisartan also had no effect on these outcomes. This suggests that these antiplatelet regimens are similar and raises questions about default prescription of agents to block the angiotensin pathway in all stroke patients. The principal side effect of dipyridamole is headache. The combination capsule of extended-release dipyridamole and aspirin is approved for prevention of stroke.

Many large clinical trials have demonstrated clearly that most antiplatelet agents reduce the risk of all important vascular atherothrombotic events (i.e., ischemic stroke, MI, and death due to all vascular causes) in patients at risk for these events. The overall *relative* reduction in risk of nonfatal stroke is about 25–30% and of all

vascular events is about 25%. The *absolute* reduction varies considerably, depending on the patient's risk. Individuals at very low risk for stroke seem to experience the same relative reduction, but their risks may be so low that the "benefit" is meaningless. Conversely, individuals with a 10–15% risk of vascular events per year experience a reduction to about 7.5–11%.

Aspirin is inexpensive, can be given in low doses, and could be recommended for all adults to prevent both stroke and MI. However, it causes epigastric discomfort, gastric ulceration, and gastrointestinal hemorrhage, which may be asymptomatic or life threatening. Consequently, not every 40- or 50-year-old should be advised to take aspirin regularly because the risk of atherothrombotic stroke is extremely low and is outweighed by the risk of adverse side effects. Conversely, every patient who has experienced an atherothrombotic stroke or TIA and has no contraindication should be taking an antiplatelet agent regularly because the average annual risk of another stroke is 8–10%; another few percent will experience an MI or vascular death. Clearly, the likelihood of benefit far outweighs the risks of treatment.

The choice of antiplatelet agent and dose must balance the risk of stroke, the expected benefit, and the risk and cost of treatment. However, there are no definitive data, and opinions vary. Many authorities believe low-dose (30–75 mg/d) and high-dose (650–1300 mg/d) aspirin are about equally effective. Some advocate very low doses to avoid adverse effects, and still others advocate very high doses to be sure the benefit is maximal. Most physicians in North America recommend 81–325 mg/d, whereas most Europeans recommend 50–100 mg. Clopidogrel and extended-release dipyridamole plus aspirin are being increasingly recommended as first-line drugs for secondary prevention. Similarly, the choice of aspirin, clopidogrel, or dipyridamole plus aspirin must balance the fact that the latter are more effective than aspirin but the cost is higher, and this is likely to affect long-term patient adherence. The use of platelet aggregation studies in individual patients taking aspirin is controversial because of limited data.

ANTICOAGULATION THERAPY AND EMBOLIC STROKE PREVENTION

Several trials have shown that anticoagulation (INR range, 2–3) in patients with chronic nonvalvular (nonrheumatic) atrial fibrillation (NVAF) prevents cerebral embolism and stroke and is safe. For primary prevention and for patients who have experienced stroke or TIA, anticoagulation with a VKA reduces the risk by about 67%, which clearly outweighs the 1–3% risk per year of a major bleeding complication. VKAs are difficult to dose, their effects vary with dietary intake of vitamin K, and they require frequent blood monitoring of the PTT/INR. Several newer oral anticoagulants (OACs) have recently been shown to be more convenient and efficacious for stroke prevention in NVAF. A randomized trial compared the oral thrombin inhibitor dabigatran to VKAs in a noninferiority trial to prevent stroke or systemic embolization in NVAF. Two doses of dabigatran were used: 110 mg/d and 150 mg/d. Both dose tiers of dabigatran were noninferior to VKAs in preventing second stroke and systemic embolization, and the higher dose tier was superior (relative risk 0.66; 95% CI 0.53–0.82; p <.001) and the rate of major bleeding was lower in the lower dose tier of dabigatran compared to VKAs. Dabigatran requires no blood monitoring to titrate the dose, and its effect is independent of oral intake of vitamin K. Newer oral factor Xa inhibitors have also been found to be equivalent or safer and more effective than VKAs in NVAF stroke prevention. In the Apixaban for Reduction in Stroke and Other Thromboembolic Events in Atrial Fibrillation (ARISTOTLE) trial, patients were randomized between apixaban, 5 mg twice daily, and dose-adjusted warfarin (INR 2–3). The combined endpoint of ischemic or hemorrhagic stroke or system embolism occurred in 1.27% of patients in the apixaban group and in 1.6% in the warfarin group (p <.001 for noninferiority and p <.01 for superiority). Major bleeding was 1% less, favoring apixaban (p <.001). Similar results were obtained in the Rivaroxaban Once Daily Oral Direct Factor Xa Inhibition Compared with Vitamin K Antagonism for Prevention of Stroke and Embolism Trial in Atrial Fibrillation (ROCKET-AF). Here, patients with NVAF were randomized to rivaroxaban versus warfarin: 1.7% of the factor Xa group and 2.2% of the warfarin group reached the endpoint of stroke and systemic embolism (p <.001 for noninferiority); intracranial hemorrhage was also lower with rivaroxaban. Finally, the factor Xa inhibitor edoxaban was also found to be noninferior to warfarin. Thus, oral factor Xa inhibitors are at least a suitable alternative to VKAs, and likely are superior both in efficacy and perhaps compliance.

For patients who cannot take anticoagulant medications, clopidogrel plus aspirin was compared to aspirin alone in the Atrial Fibrillation Clopidogrel Trial with Irbesartan for Prevention of Vascular Events (ACTIVE-A). Clopidogrel combined with aspirin was more effective than aspirin alone in preventing vascular events, principally stroke, but increased the risk of major bleeding (relative risk 1.57, p <.001).

The decision to use anticoagulation for primary prevention is based primarily on risk factors (Table 420-3). The history of a TIA or stroke tips the balance in favor of anticoagulation regardless of other risk factors. Intermittent atrial fibrillation carries the same risk of stroke as chronic atrial fibrillation, and several ambulatory studies of seemingly "cryptogenic" stroke have found evidence of intermittent atrial fibrillation in nearly 20% of patients monitored for a few weeks. Interrogation of implanted pacemakers also confirms an association between subclinical atrial fibrillation and stroke risk. Therefore, for patients with otherwise cryptogenic embolic stroke (no evidence of any other cause for stroke), ambulatory monitoring for 3–4 weeks is a reasonable strategy to determine the best prophylactic therapy.

Because of the high annual stroke risk in untreated rheumatic heart disease with atrial fibrillation, primary prophylaxis against stroke has not been studied in a double-blind fashion. These patients generally should receive long-term anticoagulation. Dabigatran and the oral Xa inhibitors have not been studied in this population.

Anticoagulation also reduces the risk of embolism in acute MI. Most clinicians recommend a 3-month course of anticoagulation when there is anterior Q-wave infarction, substantial left ventricular dysfunction, congestive heart failure, mural thrombosis, or atrial fibrillation. OACs are recommended long-term if atrial fibrillation persists.

Stroke secondary to thromboembolism is one of the most serious complications of prosthetic heart valve implantation. The intensity of anticoagulation and/or antiplatelet therapy is dictated by the type of prosthetic valve and its location. Dabigatran may be less effective than warfarin, and the oral Xa inhibitors have not been studied in this population.

If the embolic source cannot be eliminated, anticoagulation should in most cases be continued indefinitely. Many neurologists recommend combining antiplatelet agents with anticoagulants for patients who "fail" anticoagulation (i.e., have another stroke or TIA), but the evidence basis for this is lacking.

ANTICOAGULATION THERAPY AND NONCARDIOGENIC STROKE

Data do not support the use of long-term VKAs for preventing atherothrombotic stroke for either intracranial or extracranial cerebrovascular disease. The Warfarin-Aspirin Recurrent Stroke Study (WARSS) found no benefit of warfarin sodium (INR 1.4–2.8) over aspirin, 325 mg, for secondary prevention of stroke but did find a slightly higher bleeding rate in the warfarin group; a European study confirmed this finding. The Warfarin and Aspirin for Symptomatic Intracranial Disease (WASID) study (see below) demonstrated no benefit of warfarin (INR 2–3) over aspirin in patients with symptomatic intracranial atherosclerosis and found a higher rate of bleeding complications. Trials are ongoing testing Factor Xa medications for prevention of embolic stroke of unknown source.

TREATMENT

Carotid Atherosclerosis

Carotid atherosclerosis can be removed surgically (endarterectomy) or mitigated with endovascular stenting with or without balloon angioplasty. Anticoagulation has not been directly compared with antiplatelet therapy for carotid disease.

SURGICAL THERAPY

Symptomatic carotid stenosis was studied in the North American Symptomatic Carotid Endarterectomy Trial (NASCET) and the European Carotid Surgery Trial (ECST). Both showed a substantial benefit for surgery in patients with stenosis of ≥70%. In NASCET, the average cumulative ipsilateral stroke risk at 2 years was 26% for patients treated medically and 9% for those receiving the same medical treatment plus a carotid endarterectomy. This 17% *absolute* reduction in the surgical group is a 65% *relative* risk reduction favoring surgery (Table 420-4). NASCET also showed a significant, although less robust, benefit for patients with 50–70% stenosis. ECST found harm for patients with stenosis <30% treated surgically.

A patient's risk of stroke and possible benefit from surgery are related to the presence of retinal versus hemispheric symptoms, degree of arterial stenosis, extent of associated medical conditions (of note, NASCET and ECST excluded "high-risk" patients with significant cardiac, pulmonary, or renal disease), institutional surgical morbidity and mortality, timing of surgery relative to symptoms, and other factors. A recent meta-analysis of the NASCET and ECST trials demonstrated that endarterectomy is most beneficial when performed within 2 weeks of symptom onset. In addition, benefit is more pronounced in patients >75 years, and men appear to benefit more than women.

In summary, a patient with recent symptomatic hemispheric ischemia, high-grade stenosis in the appropriate internal carotid artery, and an institutional perioperative morbidity and mortality rate of ≤6% generally should undergo carotid endarterectomy. If the perioperative stroke rate is >6% for any particular surgeon, however, the benefits of carotid endarterectomy are questionable.

The indications for surgical treatment of *asymptomatic carotid disease* have been clarified by the results of the Asymptomatic Carotid Atherosclerosis Study (ACAS) and the Asymptomatic Carotid Surgery Trial (ACST). ACAS randomized asymptomatic patients with ≥60% stenosis to medical treatment with aspirin or the same medical treatment plus carotid endarterectomy. The surgical group had a risk over 5 years for ipsilateral stroke (and any perioperative stroke or death) of 5.1%, compared to a risk in the medical group of 11%. Although this demonstrates a 53% *relative* risk reduction, the *absolute* risk reduction is only 5.9% over 5 years, or 1.2% annually (Table 420-4). Nearly one-half of the strokes in the surgery group were caused by preoperative angiograms. ACST randomized asymptomatic patients with >60% carotid stenosis to endarterectomy or medical therapy. The 5-year risk of stroke in the surgical group (including perioperative stroke or death) was 6.4%, compared to 11.8% in the medically treated group (46% relative risk reduction and 5.4% absolute risk reduction).

In both ACAS and ACST, the perioperative complication rate was higher in women, perhaps negating any benefit in the reduction of stroke risk within 5 years. It is possible that with longer follow-up, a clear benefit in women will emerge. At present, carotid endarterectomy in asymptomatic women remains particularly controversial.

In summary, the natural history of asymptomatic stenosis is an ~2% per year stroke rate, whereas symptomatic patients experience a 13% per year risk of stroke. Whether to recommend carotid revascularization for an asymptomatic patient is somewhat controversial and depends on many factors, including patient preference, degree of stenosis, age, gender, and comorbidities. Medical therapy for reduction of atherosclerosis risk factors, including cholesterol-lowering agents and antiplatelet medications, is generally recommended for patients with asymptomatic carotid stenosis. As with atrial fibrillation, it is imperative to counsel the patient about TIAs so that therapy can be revised if symptoms develop.

ENDOVASCULAR THERAPY

Balloon angioplasty coupled with stenting is being used with increasing frequency to open stenotic carotid arteries and maintain their patency. These techniques can treat carotid stenosis not only at the bifurcation but also near the skull base and in the intracranial segments. The Stenting and Angioplasty with Protection in Patients at High Risk for Endarterectomy (SAPPHIRE) trial randomized high-risk patients (defined as patients with clinically significant coronary or pulmonary disease, contralateral carotid occlusion, restenosis after endarterectomy, contralateral laryngeal-nerve palsy, prior radical neck surgery or radiation, or age >80) with symptomatic carotid stenosis >50% or asymptomatic stenosis >80% to either stenting combined with a distal emboli-protection device or endarterectomy. The risk of death, stroke, or MI within 30 days and ipsilateral stroke or death within 1 year was 12.2% in the stenting group and 20.1% in the endarterectomy group (p = .055), suggesting that stenting is at the very least comparable to endarterectomy as a treatment option for this patient group at high risk of surgery. However, the outcomes with both interventions may not have been better than leaving the carotid stenoses untreated, particularly for the asymptomatic patients, and much of the benefit seen in the stenting group was due to a reduction in periprocedure MI. Two randomized trials comparing stents to endarterectomy in lower-risk patients have been published. The Carotid Revascularization Endarterectomy versus Stenting Trial (CREST) enrolled patients with either asymptomatic or symptomatic stenosis. The 30-day risk of stroke was 4.1% in the stent group and 2.3% in the surgical group, but the 30-day risk of MI was 1.1% in the stent group and 2.3% in the surgery group, suggesting relative equivalence of risk between the procedures. At median follow-up of 2.5 years, the combined endpoint of stroke, MI, and death was the same (7.2% stent vs 6.8% surgery) and remained so at 10-year follow up. The rate of restenosis at 2 years was also similar in both groups. The International Carotid Stenting Study (ICSS) randomized symptomatic patients to stents versus endarterectomy and found a different result: At 120 days, the incidence of stroke, MI, or death was 8.5% in the stenting group versus 5.2% in the endarterectomy group (p =.006). At median follow-up of 5 years these differences were no longer significant except a small increase in non-disabling stroke in the stenting group but no change in the average disability. In meta-analysis, carotid endarterectomy (CEA) is less morbid in older patients (aged ≥70) then is stenting. Investigation is on-going in asymptomatic patients to compare medical therapy to stenting and CEA. This will likely answer how well medical patients do with more modern medical therapy (statins, close blood pressure control, and life-style modification).

BYPASS SURGERY

Extracranial-to-intracranial (EC-IC) bypass surgery has been proven ineffective for atherosclerotic stenoses that are inaccessible to conventional carotid endarterectomy. In patients with recent stroke, an associated carotid occlusion, and evidence of inadequate perfusion of the brain as measured with positron emission tomography, no benefit from EC-IC bypass was found in a trial stopped for futility.

INTRACRANIAL ATHEROSCLEROSIS

The WASID trial randomized patients with symptomatic stenosis (50–99%) of a major intracranial vessel to either high-dose aspirin (1300 mg/d) or warfarin (target INR, 2.0–3.0), with a combined primary endpoint of ischemic stroke, brain hemorrhage, or death from vascular cause other than stroke. The trial was terminated early because of an increased risk of adverse events related to warfarin anticoagulation. With a mean follow-up of 1.8 years, the primary endpoint was seen in 22.1% of patients in the aspirin group and 21.8% of the warfarin group. Death from any cause was seen in 4.3% of the aspirin group and 9.7% of the warfarin group; 3.2% of patients on aspirin experienced major hemorrhage, compared to 8.3% of patients taking warfarin.

Intracranial stenting of intracranial atherosclerosis was found to be dramatically harmful compared to aspirin in the Stenting and Aggressive Medical Management for Preventing Recurrent Stroke in Intracranial Stenosis (SAMMPRIS) trial. This trial enrolled newly symptomatic TIA or minor stroke patients with associated 70–99% intracranial stenosis to primary stenting with a self-expanding stent or to medical management. Both groups received clopidogrel, aspirin, statin, and aggressive control of blood pressure. The endpoint of stroke or death occurred in 14.7% of the stented group and 5.8% of the medically treated groups (p = .002). This low rate of second stroke was significantly lower than in the WASID trial and suggests that aggressive medical management had a marked influence on secondary stroke risk. A concomitant study of balloon-expandable stenting was halted early at 125 patients because of the negative SAMMPRIS results and due to harm. Therefore, routine use of intracranial stenting is harmful, and medical therapy is superior for intracranial atherosclerosis.

Dural Sinus Thrombosis Limited evidence exists to support short-term use of anticoagulants, regardless of the presence of intracranial hemorrhage, for venous infarction following sinus thrombosis. The long-term outcome for most patients, even those with intracerebral hemorrhage, is excellent.

FURTHER READING

GOYAL M et al: Endovascular thrombectomy after large-vessel ischaemic stroke: A meta-analysis of individual patient data from five randomised trials. Lancet 387:1723, 2016.

GROUP SR et al: A randomized trial of intensive versus standard blood-pressure control. N Engl J Med 373:2103, 2015.

JANUARY CT et al: 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society. J Am Coll Cardiol 64:e1, 2014.

LARSSON SC et al: Prognosis of carotid dissecting aneurysms: Results from CADISS and a systematic review. Neurology 88:646, 2017.

POWERS WJ et al: 2018 Guidelines for the Early Management of Patients With Acute Ischemic Stroke: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. Stroke 49:e46, 2018.

SAVER JL et al: Time to treatment with endovascular thrombectomy and outcomes from ischemic stroke: A meta-analysis. JAMA 316:1279, 2016.

TORBEY MT et al: Evidence-based guidelines for the management of large hemispheric infarction: A statement for health care professionals from the Neurocritical Care Society and the German Society for Neuro-intensive Care and Emergency Medicine. Neurocrit Care 22:146, 2015.

421 Intracranial Hemorrhage

Wade S. Smith, J. Claude Hemphill, III, S. Claiborne Johnston

Intracranial hemorrhage is a form of stroke (see Chap. 419). Compared to ischemic stroke, patients with intracranial hemorrhage are more likely to present with headache; however, brain imaging is required to distinguish these entities. CT imaging of the head is highly sensitive and specific for intracranial hemorrhage and determines the location(s) of bleeding. Hemorrhages are classified by their location and the underlying vascular pathology. Hemorrhage directly into the brain parenchyma, also known as intracerebral hemorrhage (ICH), and arteriovenous malformations (AVMs) of the brain will be considered here. Other categories of hemorrhage include bleeding into subdural and epidural spaces, usually caused by trauma (Chap 435), and subarachnoid hemorrhage due to trauma or the rupture of an intracranial aneurysm (Chap. 302).

DIAGNOSIS

Intracranial hemorrhage is often identified on noncontrast CT imaging of the brain during the acute evaluation of stroke. Because CT is more widely available and may be logistically easier to perform than MRI, CT imaging is generally the preferred method for acute stroke evaluation (Fig. 421-1). The location of the hemorrhage narrows the differential diagnosis to a few entities. Table 421-1 lists the causes and anatomic spaces involved in hemorrhages.

EMERGENCY MANAGEMENT

Close attention should be paid to airway management because a reduction in the level of consciousness is common and often progressive. The initial blood pressure should be maintained until the results of the CT scan are reviewed and demonstrate ICH. In theory, a higher blood pressure should promote hematoma expansion, but it remains unclear if lowering of blood pressure reduces hematoma growth. Recent clinical trials have shown that systolic blood pressure (SBP) can be safely lowered acutely and rapidly to <140 mmHg in patients with spontaneous ICH whose initial SBP was 150–220 mmHg. The INTERACT2 trial was a large phase 3 clinical trial to address the effect of acute blood pressure lowering on ICH functional outcome. INTERACT2 randomized patients with spontaneous ICH within 6 h of onset and a baseline SBP of 150–220 mmHg to two different SBP targets (<140 and <180 mmHg). In those with the target SBP <140 mmHg, 52% had an outcome of death or major disability at 90 days compared with 55.6% of those with a target SBP <180 mmHg (p = .06). There was a significant shift to improved outcomes in the lower blood pressure arm, whereas both groups had a similar mortality. ATACH2 was a similarly designed clinical trial that assessed the same blood pressure targets but demonstrated no difference in outcome between groups. Current U.S. and European guidelines emphasize that blood pressure lowering to a target SBP is likely safe and possibly beneficial. However, these guidelines were completed prior to publication of the ATACH2 results, thus the specific optimal target remains a point of debate. In patients who have higher SBP on presentation or who are deeply comatose with possible elevated intracranial pressure (ICP), it is unclear whether these clinical trial results apply. In patients who have ICP monitors in place, current recommendations are that maintaining the cerebral perfusion

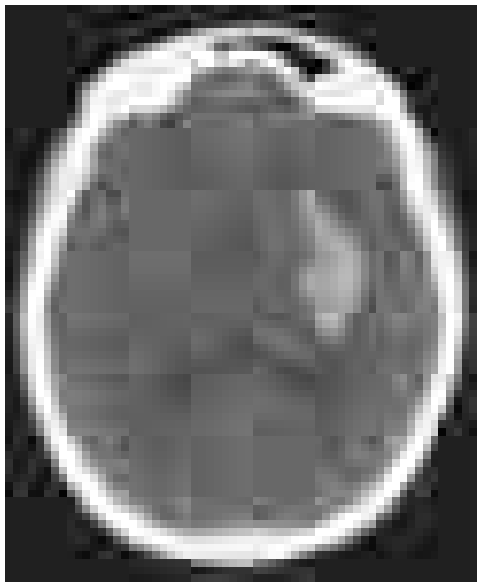


FIGURE 421-1 Hypertensive hemorrhage. Transaxial noncontrast computed tomography scan through the region of the basal ganglia reveals a hematoma involving the left putamen in a patient with rapidly progressive onset of right hemiparesis.

TABLE 421-1 Causes of Intracranial Hemorrhage

CAUSE	LOCATION	COMMENTS
Head trauma	Intraparenchymal: frontal lobes, anterior temporal lobes; subarachnoid; extra-axial (subdural, epidural)	Coup and contrecoup injury during brain deceleration
Hypertensive hemorrhage	Putamen, globus pallidus, thalamus, cerebellar hemisphere, pons	Chronic hypertension produces hemorrhage from small (~30–100 μm) vessels in these regions
Transformation of prior ischemic infarction	Basal ganglion, subcortical regions, lobar	Occurs in 1–6% of ischemic strokes with predilection for large hemispheric infarctions
Metastatic brain tumor	Lobar	Lung, choriocarcinoma, melanoma, renal cell carcinoma, thyroid, atrial myxoma
Coagulopathy	Any	Risk for ongoing hematoma expansion
Drug	Any, lobar, subarachnoid	Cocaine, amphetamine
Arteriovenous malformation	Lobar, intraventricular, subarachnoid	Risk is ~2–3% per year for bleeding if previously unruptured
Aneurysm	Subarachnoid, intraparenchymal, rarely subdural	Mycotic and nonmycotic forms of aneurysms
Amyloid angiopathy	Lobar	Degenerative disease of intracranial vessels; associated with dementia, rare in patients <60 years
Cavernous angioma	Intraparenchymal	Multiple cavernous angiomas linked to mutations in *KRIT1*, *CCM2*, and *PDCD10* genes
Dural arteriovenous fistula	Lobar, subarachnoid	Produces bleeding by venous hypertension
Capillary telangiectasias	Usually brainstem	Rare cause of hemorrhage

pressure (mean arterial pressure [MAP] minus ICP) at 50–70 mmHg is reasonable, depending on the individual patient's cerebral autoregulation status **(Chap. 301)**. Blood pressure should be lowered with nonvasodilating IV drugs such as nicardipine, labetalol, or esmolol. Patients with cerebellar hemorrhages with depressed mental status or radiographic evidence of hydrocephalus should undergo urgent neurosurgical evaluation; these patients require close monitoring because they can deteriorate rapidly. Based on the clinical examination and CT findings, further imaging studies may be necessary, including MRI or conventional x-ray angiography. Stuporous or comatose patients with clinical and imaging signs of herniation are generally treated presumptively for elevated ICP with tracheal intubation and sedation, administration of osmotic diuretics such as mannitol or hypertonic saline, and elevation of the head of the bed while surgical consultation is obtained **(Chap. 301)**. Reversal of coagulopathy and consideration of surgical evacuation of the hematoma (detailed below) are two other principal aspects of initial emergency management.

INTRACEREBRAL HEMORRHAGE

ICH accounts for ~10% of all strokes, and about 35–45% of patients die within the first month. Incidence rates are particularly high in Asians and blacks. Hypertension, coagulopathy, sympathomimetic drugs (cocaine, methamphetamine), and cerebral amyloid angiopathy (CAA) cause most of these hemorrhages. Advanced age and heavy alcohol consumption increase the risk, and cocaine and methamphetamine use is one of the most important causes in the young.

Hypertensive ICH • **PATHOPHYSIOLOGY** Hypertensive ICH usually results from spontaneous rupture of a small penetrating artery deep in the brain. The most common sites are the basal ganglia (especially the putamen), thalamus, cerebellum, and pons. The small arteries in these areas seem most prone to hypertension-induced vascular injury. When hemorrhages occur in other brain areas or in nonhypertensive patients, greater consideration should be given to other causes such as hemorrhagic disorders, neoplasms, vascular malformations, vasculitis, and CAA. The hemorrhage may be small, or a large clot may form and compress adjacent tissue, causing herniation and death. Blood may also dissect into the ventricular space, which substantially increases morbidity and may cause hydrocephalus.

Most hypertensive ICHs initially develop over 30–90 min, whereas those associated with anticoagulant therapy may evolve for as long as 24–48 h. However, it is now recognized that about a third of patients even with no coagulopathy may have significant hematoma expansion with the first day. Within 48 h, macrophages begin to phagocytize the hemorrhage at its outer surface. After 1–6 months, the hemorrhage is generally resolved to a slitlike cavity lined with a glial scar and hemosiderin-laden macrophages.

CLINICAL MANIFESTATIONS ICH generally presents as the abrupt onset of a focal neurologic deficit. Seizures are uncommon. Although clinical symptoms may be maximal at onset, commonly the focal deficit worsens over 30–90 min and is associated with a diminishing level of consciousness and signs of increased ICP such as headache and vomiting.

The putamen is the most common site for hypertensive hemorrhage, and the adjacent internal capsule is usually damaged (Fig. 421-1). Contralateral hemiparesis is therefore the sentinel sign. When mild, the face sags on one side over 5–30 min, speech becomes slurred, the arm and leg gradually weaken, and the eyes deviate away from the side of the hemiparesis. The paralysis may worsen until the affected limbs become flaccid or extend rigidly. When hemorrhages are large, drowsiness gives way to stupor as signs of upper brainstem compression appear. Coma ensues, accompanied by deep, irregular, or intermittent respiration, a dilated and fixed ipsilateral pupil, and decerebrate rigidity. In milder cases, edema in adjacent brain tissue may cause progressive deterioration over 12–72 h.

Thalamic hemorrhages also produce a contralateral hemiplegia or hemiparesis from pressure on, or dissection into, the adjacent internal capsule. A prominent sensory deficit involving all modalities is usually present. Aphasia, often with preserved verbal repetition, may occur after hemorrhage into the dominant thalamus, and constructional apraxia or mutism occurs in some cases of nondominant hemorrhage. There may also be a homonymous visual field defect. Thalamic hemorrhages cause several typical ocular disturbances by extension inferiorly into the upper midbrain. These include deviation of the eyes downward and inward so that they appear to be looking at the nose, unequal pupils with absence of light reaction, skew deviation with the eye opposite the hemorrhage displaced downward and medially, ipsilateral Horner's syndrome, absence of convergence, paralysis of vertical gaze, and retraction nystagmus. Patients may later develop a chronic, contralateral pain syndrome (Déjérine-Roussy syndrome).

In pontine hemorrhages, deep coma with quadriplegia often occurs over a few minutes. Typically, there is prominent decerebrate rigidity and "pinpoint" (1 mm) pupils that react to light. There is impairment of reflex horizontal eye movements evoked by head turning (doll's-head or oculocephalic maneuver) or by irrigation of the ears with ice water **(Chap. 300)**. Hyperpnea, severe hypertension, and hyperhidrosis are common. Most patients with deep coma from pontine hemorrhage ultimately die, or develop a locked-in state, but small hemorrhages are compatible with survival and significant recovery.

Cerebellar hemorrhages usually develop over several hours and are characterized by occipital headache, repeated vomiting, and ataxia of gait. In mild cases, there may be no other neurologic signs except for gait ataxia. Dizziness or vertigo may be prominent. There is often paresis of conjugate lateral gaze toward the side of the hemorrhage, forced deviation of the eyes to the opposite side, or an ipsilateral sixth nerve palsy. Less frequent ocular signs include blepharospasm, involuntary closure of one eye, ocular bobbing, and skew deviation. Dysarthria and dysphagia may occur. As the hours pass, the patient often becomes stuporous and then comatose from brainstem compression or obstructive hydrocephalus; immediate surgical evacuation before severe brainstem

compression occurs may be lifesaving. Hydrocephalus from fourth ventricle compression can be relieved by external ventricular drainage; however, in this situation definitive hematoma evacuation is recommended rather than treatment with ventricular drainage alone. If the deep cerebellar nuclei are spared, full recovery is common.

Lobar Hemorrhage The major neurologic deficit with an occipital hemorrhage is hemianopsia; with a left temporal hemorrhage, aphasia and delirium; with a parietal hemorrhage, hemisensory loss; and with frontal hemorrhage, arm weakness. Large hemorrhages may be associated with stupor or coma if they compress the thalamus or midbrain. Most patients with lobar hemorrhages have focal headaches, and more than one-half vomit or are drowsy. Stiff neck and seizures are uncommon.

Other Causes of ICH CAA is a disease of the elderly in which arteriolar degeneration occurs and amyloid is deposited in the walls of the cerebral arteries. Amyloid angiopathy causes both single and recurrent lobar hemorrhages and is probably the most common cause of lobar hemorrhage in the elderly. It accounts for some intracranial hemorrhages associated with IV thrombolysis given for myocardial infarction. This disorder can be suspected in patients who present with multiple hemorrhages (and infarcts) over several months or years or in patients with "microbleeds" in the cortex, seen on brain MRI sequences sensitive for hemosiderin (iron-sensitive imaging), but it is definitively diagnosed by pathologic demonstration of Congo red staining of amyloid in cerebral vessels. The ε2 and ε4 allelic variations of the apolipoprotein E gene are associated with increased risk of recurrent lobar hemorrhage and may therefore be markers of amyloid angiopathy. Positron emission tomography imaging can image amyloid-beta deposits in CAA using specific antibody labels and may be helpful in diagnosing CAA noninvasively. Although cerebral biopsy is the most definitive method of diagnosis, evidence of inflammation on lumbar puncture should prompt consideration of CAA-associated vasculitis as an underlying cause and oral glucocorticoids may be beneficial. Non-inflammatory CAA has no specific treatment. Oral anticoagulants are typically avoided.

Cocaine and *methamphetamine* are frequent causes of stroke in young (age <45 years) patients. ICH, ischemic stroke, and subarachnoid hemorrhage (SAH) are all associated with stimulant use. Angiographic findings vary from completely normal arteries to large-vessel occlusion or stenosis, vasospasm, or changes consistent with vasculopathy. The mechanism of sympathomimetic-related stroke is not known, but cocaine enhances sympathetic activity causing acute, sometimes severe, hypertension, and this may lead to hemorrhage. Slightly more than one-half of stimulant-related intracranial hemorrhages are intracerebral and the rest are subarachnoid. In cases of SAH, a saccular aneurysm is usually identified. Presumably, acute hypertension causes aneurysmal rupture.

Head injury often causes intracranial bleeding. The common sites are intraparenchymal (especially temporal and inferior frontal lobes) and into the subarachnoid, subdural, and epidural spaces. Trauma must be considered in any patient with an unexplained acute neurologic deficit (hemiparesis, stupor, or confusion), particularly if the deficit occurred in the context of a fall **(Chap. 435)**.

Intracranial hemorrhages associated with *anticoagulant therapy* can occur at any location; they are often lobar or subdural. Anticoagulant-related ICHs may continue to evolve over 24–48 h, especially if coagulopathy is insufficiently reversed. Coagulopathy and thrombocytopenia should be reversed rapidly, as discussed below. ICH associated with *hematologic disorders* (leukemia, aplastic anemia, thrombocytopenic purpura) can occur at any site and may present as multiple ICHs. Skin and mucous membrane bleeding may be evident and offers a diagnostic clue.

Hemorrhage into a *brain tumor* may be the first manifestation of neoplasm. Choriocarcinoma, malignant melanoma, renal cell carcinoma, and bronchogenic carcinoma are among the most common metastatic tumors associated with ICH. Glioblastoma multiforme in adults and medulloblastoma in children may also have areas of ICH.

Hypertensive encephalopathy is a complication of malignant hypertension. In this acute syndrome, severe hypertension is associated with headache, nausea, vomiting, convulsions, confusion, stupor, and coma. Focal or lateralizing neurologic signs, either transitory or permanent, may occur but are infrequent and therefore suggest some other vascular disease (hemorrhage, embolism, or atherosclerotic thrombosis). There are retinal hemorrhages, exudates, papilledema (hypertensive retinopathy), and evidence of renal and cardiac disease. In most cases, ICP and CSF protein levels are elevated. MRI brain imaging shows a pattern of typically posterior (occipital > frontal) brain edema that is reversible and termed *reversible posterior leukoencephalopathy*. The hypertension may be essential or due to chronic renal disease, acute glomerulonephritis, acute toxemia of pregnancy, pheochromocytoma, or other causes. Lowering the blood pressure reverses the process, but stroke can occur, especially if blood pressure is lowered too rapidly. Neuropathologic examination reveals multifocal to diffuse cerebral edema and hemorrhages of various sizes from petechial to massive. Microscopically, there is necrosis of arterioles, minute cerebral infarcts, and hemorrhages. The term *hypertensive encephalopathy* should be reserved for this syndrome and not for chronic recurrent headaches, dizziness, recurrent transient ischemic attacks, or small strokes that often occur in association with high blood pressure. Distinguishing hypertensive encephalopathy with ICH from hypertensive ICH is important since aggressive lowering of SBP to 140–180 mmHg acutely is usually considered in hypertensive ICH but less aggressive measures should be used in hypertensive encephalopathy. Having no alteration in mental status or other prodrome prior to the ICH favors hypertensive ICH as the disease.

Primary intraventricular hemorrhage is rare and should prompt investigation for an underlying vascular anomaly. Sometimes bleeding begins within the periventricular substance of the brain and dissects into the ventricular system without leaving signs of intraparenchymal hemorrhage. Alternatively, bleeding can arise from periependymal veins. Vasculitis, usually polyarteritis nodosa or lupus erythematosus, can produce hemorrhage in any region of the central nervous system; most hemorrhages are associated with hypertension, but the arteritis itself may cause bleeding by disrupting the vessel wall. Nearly one-half of patients with primary intraventricular hemorrhage have identifiable bleeding sources seen using conventional angiography.

Sepsis can cause small petechial hemorrhages throughout the cerebral white matter. *Moyamoya disease* **(Chap. 421)**, mainly an occlusive arterial disease that causes ischemic symptoms, may on occasion produce ICH, particularly in the young. Hemorrhages into the spinal cord are usually the result of an AVM, cavernous malformation, or metastatic tumor. *Epidural spinal hemorrhage* produces a rapidly evolving syndrome of spinal cord or nerve root compression **(Chap. 434)**. Spinal hemorrhages usually present with sudden back pain and some manifestation of myelopathy.

Laboratory and Imaging Evaluation Patients should have routine blood chemistries and hematologic studies. Specific attention to the platelet count and PT/PTT/INR is important to identify coagulopathy. CT imaging reliably detects acute focal hemorrhages in the supratentorial space. Rarely very small pontine or medullary hemorrhages may not be well delineated because of motion and bone-induced artifact that obscure structures in the posterior fossa. After the first 2 weeks, x-ray attenuation values of clotted blood diminish until they become isodense with surrounding brain. Mass effect and edema may remain. In some cases, a surrounding rim of contrast enhancement appears after 2–4 weeks and may persist for months. MRI, although more sensitive for delineating posterior fossa lesions, is generally not necessary for primary diagnosis in most instances. Images of flowing blood on MRI scan may identify AVMs as the cause of the hemorrhage. MRI, CT angiography (CTA), and conventional x-ray angiography are used when the cause of intracranial hemorrhage is uncertain, particularly if the patient is young or not hypertensive and the hematoma is not in one of the usual sites for hypertensive hemorrhage. CTA or postcontrast CT imaging may reveal one or more small

areas of enhancement within a hematoma; this "spot sign" is thought to represent ongoing bleeding. The presence of a spot sign is associated with an increased risk of hematoma expansion, increased mortality, and lower likelihood of favorable functional outcome. Because patients typically have focal neurologic signs and obtundation and often show signs of increased ICP, a lumbar puncture is generally unnecessary and should usually be avoided because it may induce cerebral herniation.

TREATMENT

Intracerebral Hemorrhage

ACUTE MANAGEMENT

After immediate attention to blood pressure and airway protection (see above), focus can switch to medical and surgical management. Approximately 40% of patients with a hypertensive ICH die, but survivors can have a good to complete recovery. The ICH Score **(Table 421-2)** is a validated clinical grading scale that is useful for stratification of mortality risk and clinical outcome. However, a specific ICH clinical grading scale should not be used to precisely prognosticate outcome because of the concern of creating a self-fulfilling prophecy of poor outcome if early aggressive care is withheld. Any identified coagulopathy should be corrected as soon as possible. For patients taking vitamin K antagonists (VKAs), rapid correction of coagulopathy can be achieved by infusing prothrombin complex concentrates (PCCs), which can be administered quickly, with vitamin K administered concurrently. Fresh frozen plasma (FFP) is an alternative but since it requires larger fluid volumes and longer time to achieve adequate reversal than PCC, it is not recommended if PCC is available. Idarucizumab is a monoclonal antibody to dabigatran and the administration of two doses reverses the anticoagulation effect of dabigatran quickly. PCC may partially reverse the effects of oral factor Xa inhibitors and are reasonable to administer if available; targeted drugs to reverse Xa inhibitors are under clinical investigation. When ICH is associated with thrombocytopenia (platelet count <50,000/μL), transfusion of fresh platelets is indicated. A recent clinical trial of platelet transfusions in patients with ICH and without thrombocytopenia who are taking antiplatelet drugs suggested no benefit and possible harm.

Hematomas may expand for several hours following the initial hemorrhage, even in patients without coagulopathy. However, the precise mechanism is unclear. A phase 3 trial of treatment with recombinant factor VIIa reduced hematoma expansion; however, clinical outcomes were not improved, so use of this drug is not recommended. Blood pressure lowering has been considered due to the theoretical risk of acutely elevated blood pressure on hematoma expansion, although recent clinical trials did not find a difference in hematoma expansion between the SBP targets of 140–180 mmHg.

Evacuation of supratentorial hematomas does not appear to improve outcome for most patients. The International Surgical Trial in Intracerebral Haemorrhage (STICH) randomized patients with supratentorial ICH to either early surgical evacuation or initial medical management. No benefit was found in the early surgery arm, although analysis was complicated by the fact that 26% of patients in the initial medical management group ultimately had surgery for neurologic deterioration. The follow-up study STICH-II found that surgery within 24 h of lobar, supratentorial hemorrhage did not improve overall outcome, but might have a role in select severely affected patients. Therefore, existing data do not support routine surgical evacuation of supratentorial hemorrhages in stable patients. However, many centers still consider surgery for patients deemed salvageable and who are experiencing progressive neurologic deterioration due to herniation. Surgical techniques continue to evolve, and minimally invasive endoscopic hematoma evacuation is currently being investigated in clinical trials.

For cerebellar hemorrhages, a neurosurgeon should be consulted immediately to assist with the evaluation; most cerebellar hematomas >3 cm in diameter will require surgical evacuation. If the patient is alert without focal brainstem signs and if the hematoma is <1 cm in diameter, surgical removal is usually unnecessary. Patients with hematomas between 1 and 3 cm require careful observation for signs of impaired consciousness, progressive hydrocephalus, and precipitous respiratory failure. Hydrocephalus due to cerebellar hematoma requires surgical evacuation and should not be treated solely with ventricular drainage.

Tissue surrounding hematomas is displaced and compressed but not necessarily infarcted. Hence, in survivors, major improvement commonly occurs as the hematoma is reabsorbed and the adjacent tissue regains its function. Thus, careful management of the patient during the acute phase of the hemorrhage can lead to considerable recovery.

Surprisingly, ICP is often normal even with large ICHs. However, if the hematoma causes marked midline shift of structures with consequent obtundation, coma, or hydrocephalus, osmotic agents can be instituted in preparation for placement of a ventriculostomy or parenchymal ICP monitor **(Chap. 301)**. Once ICP is recorded, CSF drainage (if available), osmotic therapy, and blood pressure management can be tailored to the individual patient to keep cerebral perfusion pressure (MAP minus ICP) at least 50–70 mmHg. For example, if ICP is found to be high, CSF can be drained from the ventricular space and osmotic therapy continued; persistent or progressive elevation in ICP may prompt surgical evacuation of the clot. Alternately, if ICP is normal or only mildly elevated, interventions such as osmotic therapy may be tapered. Because hyperventilation may actually produce ischemia by cerebral vasoconstriction, induced hyperventilation should be limited to acute resuscitation of the patient with presumptive high ICP and eliminated once other treatments (osmotic therapy or surgical treatments) have been instituted. Glucocorticoids are not helpful for the edema from intracerebral hematoma.

PREVENTION

Hypertension is the leading cause of primary ICH. Prevention is aimed at reducing chronic hypertension, eliminating excessive alcohol use, and discontinuing use of illicit drugs such as cocaine and amphetamines. Current guidelines recommend that patients with CAA should generally avoid oral anticoagulant medications, but antiplatelet agents may be administered if there is an indication based on atherothrombotic vascular disease.

TABLE 421-2 The ICH Score

CLINICAL OR IMAGING FACTOR	POINT SCORE
Age	
<80 years	0
≥80 years	1
Hematoma Volume	
<30 cc	0
≥30 cc	1
Intraventricular Hemorrhage Present	
No	0
Yes	1
Infratentorial Origin of Hemorrhage	
No	0
Yes	1
Glasgow Coma Scale Score	
13–15	0
5–12	1
3–4	2
Total Score	0–6 Sum of each category above

Source: JC Hemphill et al: Stroke 32:891, 2001.

VASCULAR ANOMALIES

Vascular anomalies can be divided into congenital vascular malformations and acquired vascular lesions.

CONGENITAL VASCULAR MALFORMATIONS

True AVMs, venous anomalies, and capillary telangiectasias are lesions that usually remain clinically silent through life. AVMs are probably congenital, but cases of acquired lesions have been reported.

True AVMs are congenital shunts between the arterial and venous systems that may present with headache, seizures, and intracranial hemorrhage. AVMs consist of a tangle of abnormal vessels across the cortical surface or deep within the brain substance. AVMs vary in size from a small blemish a few millimeters in diameter to a large mass of tortuous channels composing an arteriovenous shunt of sufficient magnitude to raise cardiac output and precipitate heart failure. Blood vessels forming the tangle interposed between arteries and veins are usually abnormally thin and histologically resemble both arteries and veins. AVMs occur in all parts of the cerebral hemispheres, brainstem, and spinal cord, but the largest ones are most frequently located in the posterior half of the hemispheres, commonly forming a wedge-shaped lesion extending from the cortex to the ventricle.

Bleeding, headache, and seizures are most common between the ages of 10 and 30, occasionally as late as the fifties. AVMs are more frequent in men, and rare familial cases have been described. Familial AVM may be a part of the autosomal dominant syndrome of hereditary hemorrhagic telangiectasia (Osler-Rendu-Weber) syndrome due to mutations in either endoglin or activin receptor-like kinase 1, both involved in transforming growth factor (TGF) signaling and angiogenesis.

Headache (without bleeding) may be hemicranial and throbbing, like migraine, or diffuse. Focal seizures, with or without generalization, occur in ~30% of cases. One-half of AVMs become evident as ICHs. In most, the hemorrhage is mainly intraparenchymal with extension into the subarachnoid space in some cases. Unlike primary subarachnoid hemorrhages (Chap. 302), blood from a ruptured AVM is usually not deposited in the basal cisterns, and symptomatic cerebral vasospasm is rare. The risk of AVM rupture is strongly influenced by a history of prior rupture. Although unruptured AVMs have a hemorrhage rate of ~2–4% per year, previously ruptured AVMs may have a rate as high as 17% a year, at least for the first year. Hemorrhages may be massive, leading to death, or may be as small as 1 cm in diameter, leading to minor focal symptoms or no deficit. The AVM may be large enough to steal blood away from adjacent normal brain tissue or to increase venous pressure significantly to produce venous ischemia locally and in remote areas of the brain. This is seen most often with large AVMs in the territory of the middle cerebral artery.

Large AVMs of the anterior circulation may be associated with a systolic and diastolic bruit (sometimes self-audible) over the eye, forehead, or neck and a bounding carotid pulse. Headache at the onset of AVM rupture is generally not as explosive as with aneurysmal rupture. MRI is better than CT for diagnosis, although noncontrast CT scanning sometimes detects calcification of the AVM and contrast may demonstrate the abnormal blood vessels. Once identified, conventional x-ray angiography is the gold standard for evaluating the precise anatomy of the AVM.

Surgical treatment of AVMs presenting with hemorrhage often done in conjunction with preoperative embolization to reduce operative bleeding is usually indicated for accessible lesions. Stereotactic radiosurgery, an alternative to conventional surgery, can produce a slow sclerosis of the AVM over 2–3 years.

Several angiographic features can be used to help predict future bleeding risk. Paradoxically, smaller lesions seem to have a higher hemorrhage rate. The presence of deep venous drainage, venous outflow stenosis, and intranidal aneurysms may increase rupture risk. Because of the relatively low annual rate of hemorrhage and the risk of complications due to surgical or endovascular treatment, the indication for surgery in asymptomatic AVMs is debated. The ARUBA (A Randomized Trial of Unruptured Brain Arteriovenous Malformations) trial randomized patients to medical management versus intervention (surgery, endovascular embolization, combination embolization and surgery, or gamma-knife). The trial was stopped prematurely for harm, with the medical arm achieving the combined endpoint of death or symptomatic stroke in 10.1% of patients compared to 30.7% in the intervention group at an average follow-up time of 33 months. This highly significant finding argues against routine intervention for patients presenting without hemorrhage, although debate ensues regarding the generalizability of these results.

Venous anomalies are the result of development of anomalous cerebral, cerebellar, or brainstem venous drainage. These structures, unlike AVMs, are functional venous channels. They are of little clinical significance and should be ignored if found incidentally on brain imaging studies. Surgical resection of these anomalies may result in venous infarction and hemorrhage. Venous anomalies may be associated with cavernous malformations (see below), which do carry some bleeding risk.

Capillary telangiectasias are true capillary malformations that often form extensive vascular networks through an otherwise normal brain structure. The pons and deep cerebral white matter are typical locations, and these capillary malformations can be seen in patients with hereditary hemorrhagic telangiectasia (Osler-Rendu-Weber) syndrome. If bleeding does occur, it rarely produces mass effect or significant symptoms. No treatment options exist.

ACQUIRED VASCULAR LESIONS

Cavernous angiomas are tufts of capillary sinusoids that form within the deep hemispheric white matter and brainstem with no normal intervening neural structures. The pathogenesis is unclear. Familial cavernous angiomas have been mapped to several different genes: *KRIT1*, *CCM2*, and *PDCD10*. Both *KRIT1* and *CCM2* have roles in blood vessel formation, whereas *PDCD10* is an apoptotic gene. Cavernous angiomas are typically <1 cm in diameter and are often associated with a venous anomaly. Bleeding is usually of small volume, causing slight mass effect only. The bleeding risk for single cavernous malformations is 0.7–1.5% per year and may be higher for patients with prior clinical hemorrhage or multiple malformations. Seizures may occur if the malformation is located near the cerebral cortex. Surgical resection eliminates bleeding risk and may reduce seizure risk, but it is usually reserved for those malformations that form near the brain surface. Radiation treatment has not been shown to be of benefit.

Dural arteriovenous fistulas are acquired connections usually from a dural artery to a dural sinus. Patients may complain of a pulse-synchronous cephalic bruit ("pulsatile tinnitus") and headache. Depending on the magnitude of the shunt, venous pressures may rise high enough to cause cortical ischemia or venous hypertension and hemorrhage, particularly SAH. Surgical and endovascular techniques are usually curative. These fistulas may form because of trauma, but most are idiopathic. There is an association between fistulas and dural sinus thrombosis. Fistulas have been observed to appear months to years following venous sinus thrombosis, suggesting that angiogenesis factors elaborated from the thrombotic process may cause these anomalous connections to form. Alternatively, dural arteriovenous fistulas can produce venous sinus occlusion over time, perhaps from the high pressure and high flow through a venous structure.

FURTHER READING

Anderson CS et al: Rapid blood-pressure lowering in patients with acute intracerebral hemorrhage. N Engl J Med 368:2355, 2013.

Hemphill JC et al: Guidelines for the management of spontaneous intracerebral hemorrhage: A guideline for healthcare professionals from the American heart association/American stroke association. Stroke 46:2032, 2015.

Mohr JP et al: Medical management with or without interventional therapy for unruptured brain arteriovenous malformations (ARUBA): A multicentre, non-blinded, randomised trial. Lancet 383:614, 2014.

Qureshi AI et al: Intracerebral haemorrhage. Lancet 373:1632, 2009.

Steiner T et al: European Stroke Organisation (ESO) guidelines for the management of spontaneous intracerebral hemorrhage. Int J Stroke 9:840, 2014.

422 Migraine and Other Primary Headache Disorders

Peter J. Goadsby

The general approach to headache as a cardinal symptom are covered elsewhere **(Chap. 13)**; here, disorders in which headache and associated features occur in the absence of any exogenous cause are discussed. The most common are migraine, tension-type headache (TTH), and the trigeminal autonomic cephalalgias (TACs), notably cluster headache; the complete list is summarized in **Table 422-1**.

MIGRAINE

Migraine, the second most common cause of headache, and the most common headache-related, and indeed neurologic, cause of disability in the world, afflicts ~15% of women and 6% of men over a 1-year period. It is usually an episodic headache associated with certain features such as sensitivity to light, sound, or movement; nausea and vomiting often accompany the headache. A useful description of migraine is a recurring syndrome of headache associated with other symptoms of neurologic dysfunction in varying admixtures **(Table 422-2)**. A migraine attack has three phases: premonitory (prodrome), headache phase, and postdrome; each has distinct and sometimes disabling symptoms. About 20–25% of migraine patients have a fourth, aura, phase. Migraine can often be recognized by its activators, referred to as *triggers*.

Migraineurs are particularly sensitive to environmental and sensory stimuli; migraine-prone patients do not habituate easily to sensory stimuli. This sensitivity is amplified in females during the menstrual cycle. Headache can be initiated or amplified by various triggers, including glare, bright lights, sounds, or other types of afferent stimulation; hunger; let-down from stress; physical exertion; stormy weather or barometric pressure changes; hormonal fluctuations during menses; lack of or excess sleep; and alcohol or other chemical stimulation, such as with nitrates. Knowledge of a patient's susceptibility to specific triggers can be useful in management strategies involving lifestyle adjustments, although it is becoming recognized that some apparent triggers may in fact be part of the initial phase of the attack; i.e., the premonitory phase or prodrome.

Pathogenesis The sensory sensitivity that is characteristic of migraine is probably due to dysfunction of monoaminergic sensory control systems located in the brainstem and hypothalamus **(Fig. 422-1)**.

Activation of cells in the trigeminal nucleus results in the release of vasoactive neuropeptides, particularly calcitonin gene–related peptide (CGRP), at vascular terminals of the trigeminal nerve and within the trigeminal nucleus. Six CGRP receptor antagonists, *gepants*, have now been shown to be effective in the acute treatment of migraine, and four monoclonal antibodies to CGRP or its receptor have been shown to be effective in migraine prevention. Centrally, the second-order trigeminal neurons cross the midline and project to ventrobasal and posterior nuclei of the thalamus for further processing. Additionally, there are projections to the periaqueductal gray and hypothalamus, from which reciprocal descending systems have established antinociceptive effects. Other brainstem regions likely to be involved in descending modulation of trigeminal pain include the nucleus locus coeruleus in the pons and the rostroventromedial medulla.

Pharmacologic and other data point to the involvement of the neurotransmitter 5-hydroxytryptamine (5-HT; also known as serotonin) in migraine. In the late 1950s methysergide was found to antagonize certain peripheral actions of 5-HT and was introduced, based on its anti-inflammation properties, as the first drug capable of preventing migraine attacks. The *triptans* were designed to stimulate selectively subpopulations of 5-HT receptors; at least 14 different 5-HT receptors exist in humans. The triptans are potent agonists of 5-HT_{1B} and 5-HT_{1D} receptors, and some are active at the 5-HT_{1F} receptor; the latter's exclusive agonists are called *ditans*. Triptans arrest nerve signaling in the nociceptive pathways of the trigeminovascular system, at least in the trigeminal nucleus caudalis and trigeminal sensory thalamus, in addition to promoting cranial vasoconstriction, while *ditans*, now shown conclusively to be effective in acute migraine, act only at neural and not vascular targets. A range of neural targets are currently under investigation for the acute and preventive management of migraine.

Data also support a role for dopamine in the pathophysiology of migraine. Most migraine symptoms can be induced by dopaminergic stimulation. Moreover, there is dopamine receptor hypersensitivity in migraineurs, as demonstrated by the induction of yawning, nausea, vomiting, hypotension, and other symptoms of a migraine attack by dopaminergic agonists at doses that do not affect nonmigraineurs. Dopamine receptor antagonists are effective therapeutic agents in migraine, especially when given parenterally or concurrently with other antimigraine agents. Moreover, hypothalamic activation, anterior to that seen in cluster headache, has now been shown in the premonitory (prodromal) phase of migraine using functional imaging, and this may hold a key to understanding some part of the role of dopamine in the disorder.

Migraine genes identified by studying families with familial hemiplegic migraine (FHM) reveal involvement of ion channels, suggesting that alterations in membrane excitability can predispose to migraine. Mutations involving the $Ca_v2.1$ (P/Q)–type voltage-gated calcium channel *CACNA1A* gene are now known to cause FHM 1; this mutation is responsible for about 50% of FHM cases. Mutations in the Na^+-K^+ATPase *ATP1A2* gene, designated FHM 2, are responsible for about 20% of FHMs. Mutations in the neuronal voltage-gated sodium channel *SCN1A* cause FHM 3. Functional neuroimaging has suggested that brainstem regions in migraine **(Fig. 422-2)** and the posterior hypothalamic gray matter region close to the human circadian pacemaker cells of the suprachiasmatic nucleus in cluster headache **(Fig. 422-3)** are good candidates for specific involvement in these primary headaches.

Diagnosis and Clinical Features Diagnostic criteria for migraine headache are listed in **Table 422-3**. A high index of suspicion is required to diagnose migraine: the migraine aura, consisting of visual disturbances with flashing lights or zigzag lines moving across the visual field or of other neurologic symptoms, is reported in only 20–25% of patients. It should be distinguished from the pan-field television static-like disturbance now recognized as the *visual snow syndrome*. The first phase of a migraine attack for most patients is the premonitory (prodromal) phase consisting of some or all of the following: yawning, tiredness, cognitive dysfunction, mood change, neck discomfort, polyuria, and food cravings; this can last from a few hours to days. Typically, the headache phase follows with its associated features, such as nausea, photophobia, and phonophobia as well as allodynia. When questioned, these typical migraine symptoms also emerge in the premonitory phase, and typical premonitory symptoms also continue into the headache phase. As the headache lessens many patients enter a postdrome, most commonly feeling tired/weary, having problems concentrating, and experiencing mild neck discomfort that can last for hours and sometimes up to a day. A headache diary can often be helpful in making the diagnosis; this is also helpful in assessing disability and the frequency of treatment for acute attacks. Patients with episodes of migraine on eight or more days per month and with at least 15 total days of headache per month are considered to have chronic migraine **(see "Chronic Daily Headache" in Chap. 13)**. Migraine must be differentiated from TTH (discussed below), which is reported to be the most common primary headache syndrome. Migraine has several forms that have been defined (Table 422-1): migraine with and without aura and chronic migraine are the most important. *Migraine at its most basic level is headache with associated features, and tension-type headache is headache that is featureless. Most patients with disabling headache probably have migraine.*

Patients with acephalgic migraine (typical aura without headache, 1.2.1.2 in Table 422-1) experience recurrent neurologic symptoms, often with nausea or vomiting, but with little or no headache. Vertigo can be prominent; it has been estimated that one-third of patients referred for vertigo or dizziness have a primary diagnosis of migraine. Migraine aura can have prominent brainstem symptoms, and the terms *basilar*

TABLE 422-1 Primary Headache Disorders, Modified from International Classification of Headache Disorders-III-Beta (Headache Classification Committee of the International Headache Society, 2018)

1. Migraine	1.1 Migraine without aura 1.2 Migraine with aura 1.2.1 Migraine with typical aura 1.2.1.1 Typical aura with headache 1.2.1.2 Typical aura without headache 1.2.2 Migraine with brainstem aura 1.2.3 Hemiplegic migraine 1.2.3.1 Familial hemiplegic migraine (FHM) 1.2.3.1.1 Familial hemiplegic migraine type 1 1.2.3.1.2 Familial hemiplegic migraine type 2 1.2.3.1.3 Familial hemiplegic migraine type 3 1.2.3.1.4 Familial hemiplegic migraine, other loci 1.2.3.2 Sporadic hemiplegic migraine 1.2.4 Retinal migraine 1.3 Chronic migraine 1.4 Complications of migraine 1.4.1 Status migrainosus 1.4.2 Persistent aura without infarction 1.4.3 Migrainous infarction 1.4.4 Migraine aura-triggered seizure 1.5 Probable migraine 1.5.1 Probable migraine without aura 1.5.2 Probable migraine with aura 1.6 Episodic syndromes that may be associated with migraine 1.6.1 Recurrent gastrointestinal disturbance 1.6.1.1 Cyclical vomiting syndrome 1.6.1.2 Abdominal migraine 1.6.2 Benign paroxysmal vertigo 1.6.3 Benign paroxysmal torticollis
2. Tension-type headache	2.1 Infrequent episodic tension-type headache 2.2 Frequent episodic tension-type headache 2.3 Chronic tension-type headache 2.4 Probable tension-type headache
3. Trigeminal autonomic cephalalgias	3.1 Cluster headache 3.1.1 Episodic cluster headache 3.1.2 Chronic cluster headache 3.2 Paroxysmal hemicrania 3.2.1 Episodic paroxysmal hemicrania 3.2.2 Chronic paroxysmal hemicrania 3.3 Short-lasting unilateral neuralgiform headache attacks 3.3.1 Short-lasting unilateral neuralgiform headache attacks with conjunctival injection and tearing (SUNCT) 3.3.1.1 Episodic SUNCT 3.3.1.2 Chronic SUNCT 3.3.2 Short-lasting unilateral neuralgiform headache attacks with cranial autonomic symptoms (SUNA) 3.3.2.1 Episodic SUNA 3.3.2.2 Chronic SUNA 3.4 Hemicrania continua 3.5 Probable trigeminal autonomic cephalalgia
4. Other primary headache disorders	4.1 Primary cough headache 4.2 Primary exercise headache 4.3 Primary headache associated with sexual activity 4.4 Primary thunderclap headache 4.5 Cold-stimulus headache 4.5.1 Headache attributed to external application of a cold stimulus 4.5.2 Headache attributed to ingestion or inhalation of a cold stimulus 4.6 External-pressure headache 4.6.1 External-compression headache 4.6.2 External-traction headache 4.7 Primary stabbing headache 4.8 Nummular headache 4.9 Hypnic headache 4.10 New daily persistent headache (NDPH)

TABLE 422-2 Symptoms Accompanying Severe Migraine Attacks in 500 Patients

SYMPTOM	PATIENTS AFFECTED, %
Nausea	87
Photophobia	82
Lightheadedness	72
Scalp tenderness	65
Vomiting	56
Visual disturbances	36
Paresthesias	33
Vertigo	33
Photopsia	26
Alteration of consciousness	18
Diarrhea	16
Fortification spectra	10
Syncope	10
Seizure	4
Confusional state	4

Source: From NH Raskin: *Headache*, 2nd ed. New York, Churchill Livingston, 1988; with permission.

artery and *basilar-type migraine* have now been replaced by *migraine with brainstem aura* (Table 422-1).

TREATMENT

Migraine Headache

Once a diagnosis of migraine has been established, it is important to assess the extent of a patient's disease and disability. The Migraine Disability Assessment Score (MIDAS) is a well-validated, easy-to-use tool **(Fig. 422-4)**.

Patient education is an important aspect of migraine management. Information for patients is available at websites such as the American Migraine Foundation (www.americanmigrainefoundation.org) and the Migraine Trust (www.migrainetrust.org). It is helpful for patients to understand that migraine is an inherited tendency to headache; that migraine can be modified and controlled by lifestyle adjustments and medications, but it cannot be eradicated; and that, except on some occasions in women on oral estrogens or contraceptives, migraine is not associated with serious or life-threatening illnesses.

NONPHARMACOLOGIC MANAGEMENT

Migraine can often be managed to some degree by a variety of nonpharmacologic approaches. When patients can identify reliable triggers, their avoidance can be useful. A regulated lifestyle is helpful, including a healthy diet, regular exercise, regular sleep patterns, avoidance of excess caffeine and alcohol, and avoidance of acute changes in stress levels, being particularly wary of the let-down effect.

The measures that benefit a given individual should be used routinely because they provide a simple, cost-effective approach to migraine management. Patients with migraine do not encounter more stress than headache-free individuals; over-responsiveness to changes in stress appears to be the issue. Because the stresses of everyday living cannot be eliminated, lessening one's response to stress by various techniques is helpful for many patients. These may include yoga, transcendental meditation, hypnosis, and conditioning techniques such as biofeedback. For most patients seen in clinical practice, this approach is, at best, an adjunct to pharmacotherapy. Nonpharmacologic measures are unlikely to prevent all migraine attacks. If these measures fail to prevent an attack, pharmacologic approaches are then needed.

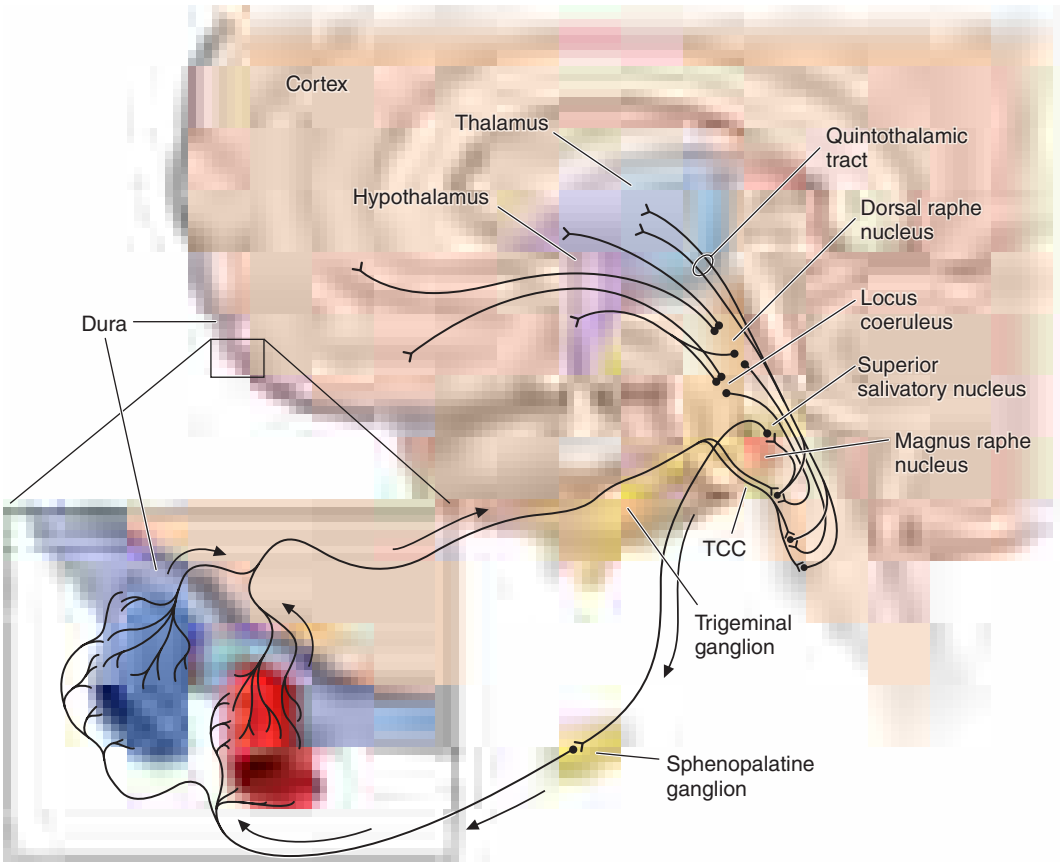


FIGURE 422-1 Brainstem pathways that modulate sensory input. The key pathway for pain in migraine is the trigeminovascular input from the meningeal vessels, which passes through the trigeminal ganglion and synapses on second-order neurons in the trigeminocervical complex (TCC). These neurons in turn project in the quintothalamic tract and, after decussating in the brainstem, synapse on neurons in the thalamus. Important modulation of the trigeminovascular nociceptive input comes from the dorsal raphe nucleus, locus coeruleus, and nucleus raphe magnus.

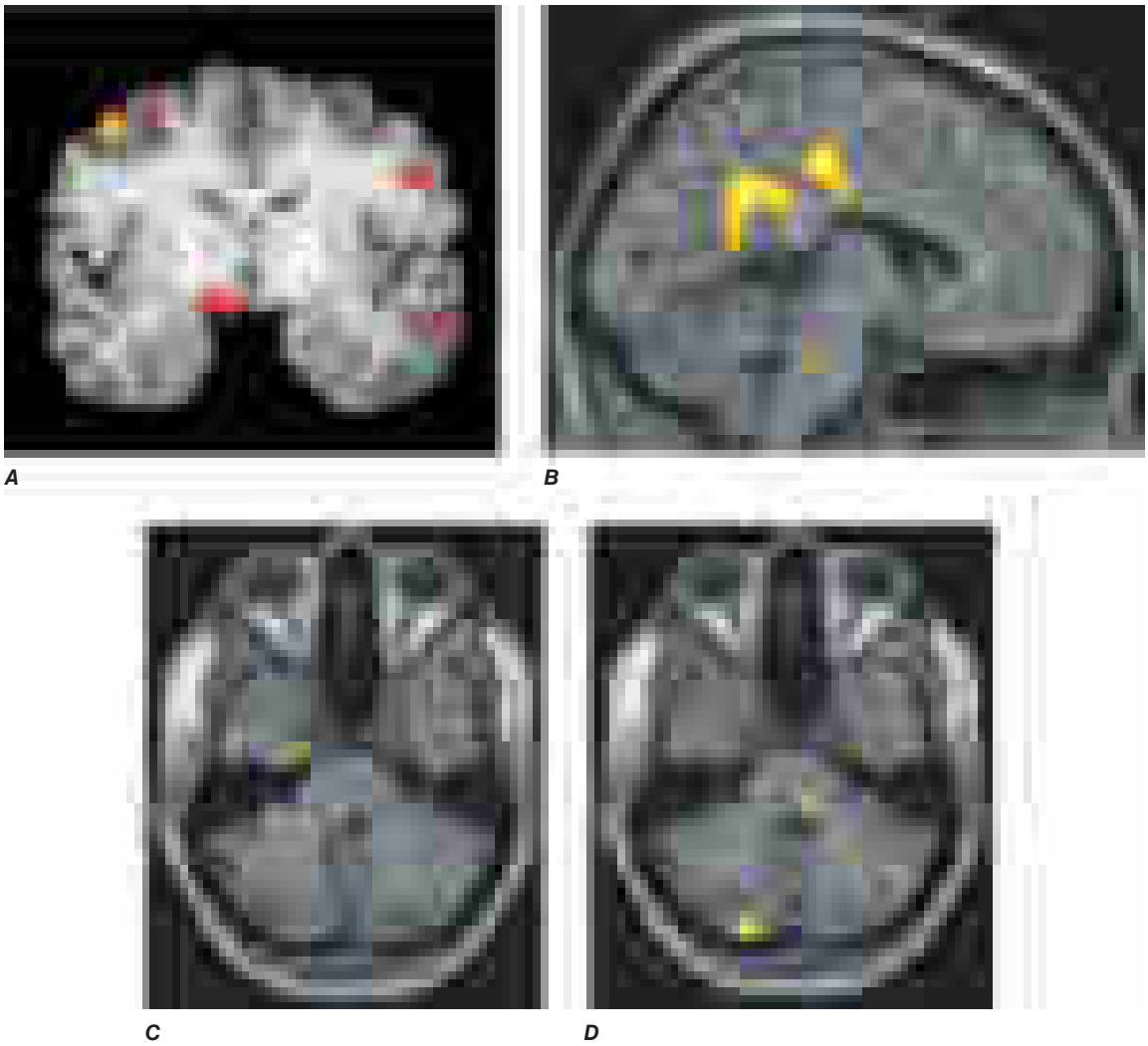


FIGURE 422-2 Positron emission tomography (PET) activation in migraine. Hypothalamic, dorsal midbrain, and dorsolateral pontine activation is seen in triggered attacks in the premonitory phase before pain, whereas in migraine attacks, dorsolateral pontine activation persists, as it does in chronic migraine (not shown). The dorsolateral pontine area, which includes the noradrenergic locus coeruleus, is fundamental to the expression of migraine. Moreover, lateralization of changes in this region of the brainstem correlates with lateralization of the head pain in hemicranial migraine; the scans shown in panels ***C*** and ***D*** are of patients with acute migraine headache on the right and left side, respectively. *(Panel A from FH Maniyar et al: Brain 137:232, 2014; panel B from SK Afridi et al: Arch Neurol 62:1270, 2005; Panels C and D from SK Afridi et al: Brain 128:932, 2005.)*

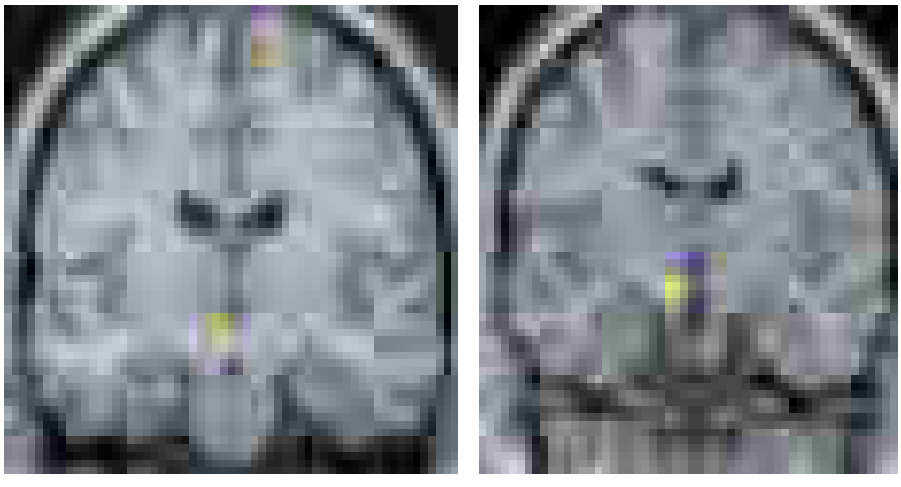


FIGURE 422-3 *A.* Posterior hypothalamic gray matter region activation by positron emission tomography in a patient with acute cluster headache. *(From A May et al: Lancet 352:275, 1998.)* ***B.*** High-resolution T1-weighted magnetic resonance image obtained using voxel-based morphometry demonstrates increased gray matter activity, lateralized to the side of pain in a patient with cluster headache. *(From A May et al: Nat Med 5:836, 1999.)*

TABLE 422-3 Simplified Diagnostic Criteria for Migraine

REPEATED ATTACKS OF HEADACHE LASTING 4–72 h IN PATIENTS WITH A NORMAL PHYSICAL EXAMINATION, NO OTHER REASONABLE CAUSE FOR THE HEADACHE, AND:	
AT LEAST 2 OF THE FOLLOWING FEATURES:	**PLUS AT LEAST 1 OF THE FOLLOWING FEATURES:**
Unilateral pain	Nausea/vomiting
Throbbing pain	Photophobia and phonophobia
Aggravation by movement	
Moderate or severe intensity	

Source: Adapted from the International Headache Society Classification (Headache Classification Committee of the International Headache Society, Cephalalgia 38:1-211, 2018).

ACUTE ATTACK THERAPIES FOR MIGRAINE

The mainstay of pharmacologic therapy is the judicious use of one or more of the many medicines that are effective in migraine **(Table 422-4)**. The selection of the optimal regimen for a given patient depends on a number of factors, the most important of which is the severity of the attack. Mild migraine attacks can usually be managed by oral agents; the average efficacy rate is 50–70%. Severe migraine attacks may require parenteral therapy. Most drugs effective in the treatment of migraine are members of one of three major pharmacologic classes: nonsteroidal anti-inflammatory drugs, 5-$HT_{1B/1D}$ receptor agonists, and dopamine receptor antagonists. Two new classes of therapeutic agents, CGRP receptor antagonists, such as rimegepant and ubrogepant, and 5-HT_{1F} receptor agonists, such as lasmiditan, should soon be available.

In general, an adequate dose of whichever agent is chosen should be used as soon as possible after the onset of an attack. If additional medication is required within 60 min because symptoms return or have not abated, the initial dose should be increased for subsequent attacks or a different class of drug tried as first-line treatment. Migraine therapy must be individualized; a standard approach for all patients is not possible. A therapeutic regimen may need to be constantly refined until one is identified that provides the patient with rapid, complete, and consistent relief with minimal side effects **(Table 422-5)**.

Nonsteroidal Anti-Inflammatory Drugs (NSAIDs) Both the severity and duration of a migraine attack can be reduced significantly by NSAIDs (Table 422-4). Indeed, many undiagnosed migraineurs self-treat with nonprescription NSAIDs. A general consensus is that NSAIDs are most effective when taken early in the migraine attack. However, the effectiveness of these agents in migraine is usually less than optimal in moderate or severe migraine attacks. The combination of acetaminophen (paracetamol), aspirin, and caffeine has been approved for use by the U.S. Food and Drug Administration (FDA) for the treatment of mild to moderate migraine. The combination of aspirin and metoclopramide has been shown to be comparable to a single dose of oral sumatriptan. Important side effects of NSAIDs include dyspepsia and gastrointestinal irritation.

5-$HT_{1B/1D}$ RECEPTOR AGONISTS

Oral Stimulation of 5-$HT_{1B/1D}$ receptors can stop an acute migraine attack. Ergotamine and dihydroergotamine are nonselective receptor agonists, whereas the triptans are selective 5-$HT_{1B/1D}$ receptor agonists. A variety of triptans—sumatriptan, almotriptan, eletriptan, frovatriptan, naratriptan, rizatriptan, and zolmitriptan—are available for the treatment of migraine.

Each drug in the triptan class has similar pharmacologic properties but varies slightly in terms of clinical efficacy. Rizatriptan and eletriptan are, on a population basis, the most efficacious of the triptans currently available in the United States. Sumatriptan and zolmitriptan have similar rates of efficacy as well as time to onset, with an advantage of having multiple formulations, whereas almotriptan has a similar rate of efficacy to sumatriptan and is better tolerated, and frovatriptan and naratriptan are somewhat slower in onset and are also well tolerated. Clinical efficacy appears to be related more to the t_{max} (time to peak plasma level) than to the

***MIDAS Questionnaire**

INSTRUCTIONS: Please answer the following questions about ALL headaches you have had over the last 3 months. Write zero if you did not do the activity in the last 3 months.

1. On how many days in the last 3 months did you miss work or school because of your headaches? ..... ____ days
2. How many days in the last 3 months was your productivity at work or school reduced by half or more because of your headaches *(do not include days you counted in question 1 where you missed work or school)*? ..... ____ days
3. On how many days in the last 3 months did you **not** do household work because of your headaches? ..... ____ days
4. How many days in the last 3 months was your productivity in household work reduced by half or more because of your headaches *(do not include days you counted in question 3 where you did not do household work)*? ..... ____ days
5. On how many days in the last 3 months did you miss family, social, or leisure activities because of your headaches? ..... ____ days

A. On how many days in the last 3 months did you have a headache? *(If a headache lasted more than one day, count each day.)* ..... ____ days

B. On a scale of 0–10, on average how painful were these headaches? *(Where 0 = no pain at all, and 10 = pain as bad as it can be.)* ..... ____

*Migraine Disability Assessment Score
(Questions 1–5 are used to calculate the MIDAS score.)
Grade I—Minimal or Infrequent Disability: 0–5
Grade II—Mild or Infrequent Disability: 6–10
Grade III—Moderate Disability: 11–20
Grade IV—Severe Disability: > 20

FIGURE 422-4 The Migraine Disability Assessment Score (MIDAS) Questionnaire.

TABLE 422-4 Treatment of Acute Migraine

DRUG	TRADE NAME	DOSAGE
Simple Analgesics		
Acetaminophen, aspirin, caffeine	Excedrin Migraine	Two tablets or caplets q6h (max 8 per day)
NSAIDs		
Naproxen	Aleve, Anaprox, generic	220–550 mg PO bid
Ibuprofen	Advil, Motrin, Nuprin, generic	400 mg PO q3–4h
Tolfenamic acid	Clotam Rapid	200 mg PO; may repeat ×1 after 1–2 h
Diclofenac K	Cambia	50 mg PO with water
5-$HT_{1B/1D}$ Receptor Agonists		
Oral		
Ergotamine 1 mg, caffeine 100 mg	Cafergot	One or two tablets at onset, then one tablet q½h (max 6 per day, 10 per week)
Naratriptan	Amerge	2.5-mg tablet at onset; may repeat once after 4 h
Rizatriptan	Maxalt	5–10-mg tablet at onset; may repeat after 2 h (max 30 mg/d)
	Maxalt-MLT	
Sumatriptan	Imitrex	50–100-mg tablet at onset; may repeat after 2 h (max 200 mg/d)
Frovatriptan	Frova	2.5-mg tablet at onset, may repeat after 2 h (max 5 mg/d)
Almotriptan	Axert	12.5-mg tablet at onset, may repeat after 2 h (max 25 mg/d)
Eletriptan	Relpax	40 or 80 mg
Zolmitriptan	Zomig	2.5-mg tablet at onset; may repeat after 2 h (max 10 mg/d)
	Zomig Rapimelt	
Nasal		
Dihydroergotamine	Migranal Nasal Spray	Prior to nasal spray, the pump must be primed 4 times; 1 spray (0.5 mg) is administered, followed in 15 min by a second spray
Sumatriptan	Imitrex Nasal Spray	5–20 mg intranasal spray as 4 sprays of 5 mg or a single 20 mg spray (may repeat once after 2 h, not to exceed a dose of 40 mg/d)
Zolmitriptan	Zomig	5 mg intranasal spray as one spray (may repeat once after 2 h, not to exceed a dose of 10 mg/d)
Parenteral		
Dihydroergotamine	DHE-45	1 mg IV, IM, or SC at onset and q1h (max 3 mg/d, 6 mg per week)
Sumatriptan	Imitrex Injection	6 mg SC at onset (may repeat once after 1 h for max of 2 doses in 24 h)
	Alsuma	
	Sumavel DosePro	
Dopamine Receptor Antagonists		
Oral		
Metoclopramide	Reglan,[a] generic[a]	5–10 mg/d
Prochlorperazine	Compazine,[a] generic[a]	1–25 mg/d
Parenteral		
Chlorpromazine	Generic[a]	0.1 mg/kg IV at 2 mg/min; max 35 mg/d
Metoclopramide	Reglan,[a] generic	10 mg IV
Prochlorperazine	Compazine,[a] generic[a]	10 mg IV
Other		
Oral		
Acetaminophen, 325 mg, *plus* dichloralphenazone, 100 mg, *plus* isometheptene, 65 mg	Midrin, generic	Two capsules at onset followed by 1 capsule q1h (max 5 capsules)
Parenteral		
Opioids	Generic[a]	Multiple preparations and dosages; **see Table 10-1**
Other		
Neuromodulation	SpringTMS	Two pulses at onset followed by two further pulses
Noninvasive Vagus Nerve Stimulation (nVNS)	gammaCore	Two doses each of 120 seconds
Single pulse transcranial magnetic stimulation (sTMS)		

[a]Not all drugs are specifically indicated by the FDA for migraine. Local regulations and guidelines should be consulted.

Note: Antiemetics (e.g., domperidone 10 mg or ondansetron 4 or 8 mg) or prokinetics (e.g., metoclopramide 10 mg) are sometimes useful adjuncts.

Abbreviations: 5-HT, 5-hydroxytryptamine; NSAIDs, nonsteroidal anti-inflammatory drugs.

potency, half-life, or bioavailability. This observation is consistent with a large body of data indicating that faster-acting analgesics are more effective than slower-acting agents.

Unfortunately, monotherapy with a selective oral 5-$HT_{1B/1D}$ receptor agonist does not result in rapid, consistent, and complete relief of migraine in all patients. Triptans are generally not effective in migraine with aura unless given after the aura is completed and the headache initiated. Side effects are common, although often mild and transient. Moreover, 5-$HT_{1B/1D}$ receptor agonists are contraindicated in individuals with a history of cardiovascular and cerebrovascular

TABLE 422-5 Clinical Stratification of Acute Specific Migraine Treatments

CLINICAL SITUATION	TREATMENT OPTIONS
Failed NSAIDs/analgesics	**First tier**
	Sumatriptan 50 mg or 100 mg PO
	Almotriptan 12.5 mg PO
	Rizatriptan 10 mg PO
	Eletriptan 40 mg PO
	Zolmitriptan 2.5 mg PO
	Slower effect/better tolerability
	Naratriptan 2.5 mg PO
	Frovatriptan 2.5 mg PO
	Infrequent headache
	Ergotamine/caffeine 1–2/100 mg PO
	Dihydroergotamine nasal spray 2 mg
Early nausea or difficulties taking tablets	Zolmitriptan 5 mg nasal spray
	Sumatriptan 20 mg nasal spray
	Rizatriptan 10 mg MLT wafer
Headache recurrence	Ergotamine 2 mg (most effective PR/usually with caffeine)
	Naratriptan 2.5 mg PO
	Almotriptan 12.5 mg PO
	Eletriptan 40 mg
Tolerating acute treatments poorly	Naratriptan 2.5 mg
	Almotriptan 12.5 mg
	Single pulse transcranial magnetic stimulation
	Noninvasive vagus nerve stimulation
Early vomiting	Zolmitriptan 5 mg nasal spray
	Sumatriptan 25 mg PR
	Sumatriptan 6 mg SC
Menses-related headache	**Prevention**
	Ergotamine PO at night
	Estrogen patches
	Treatment
	Triptans
	Dihydroergotamine nasal spray
Very rapidly developing symptoms	Zolmitriptan 5 mg nasal spray
	Sumatriptan 6 mg SC
	Dihydroergotamine 1 mg IM

Abbreviation: NSAIDs, nonsteroidal anti-inflammatory drugs.

disease. Recurrence of headache, within usual time course of an attack, is another important limitation of triptan use and occurs at least occasionally in most patients. Evidence from randomized controlled trials shows that coadministration of a longer-acting NSAID, naproxen 500 mg, with sumatriptan will augment the initial effect of sumatriptan and, importantly, reduce rates of headache recurrence.

Ergotamine preparations offer a nonselective means of stimulating 5-HT_1 receptors. A nonnauseating dose of ergotamine should be sought because a dose that provokes nausea is too high and may intensify head pain. Oral (excluding sublingual) formulations of ergotamine also contain 100 mg caffeine (theoretically to enhance ergotamine absorption and possibly to add additional analgesic activity). The average oral ergotamine dose for a migraine attack is 2 mg. Because the clinical studies demonstrating the efficacy of ergotamine in migraine predated the clinical trial methodologies used with the triptans, it is difficult to assess the comparative efficacy of ergotamine versus the triptans. In general, with use of ergotamine there appears to be a much higher incidence of nausea than with triptans but less headache recurrence.

Nasal Nasal formulations of dihydroergotamine, zolmitriptan, or sumatriptan can be useful in patients requiring a nonoral route of administration. The nasal sprays result in substantial blood levels within 30–60 min. Although in theory nasal sprays might provide faster and more effective relief of a migraine attack than oral formulations, their reported efficacy is only ~50–60%. Studies with a new inhalational formulation of dihydroergotamine indicate that its absorption problems can be overcome to produce rapid onset of action with good tolerability.

Parenteral Administration of drugs by injection, such as dihydroergotamine and sumatriptan, is approved by the FDA for the rapid relief of a migraine attack. Peak plasma levels of dihydroergotamine are achieved 3 min after IV dosing, 30 min after IM dosing, and 45 min after SC dosing. If an attack has not already peaked, SC or IM administration of 1 mg of dihydroergotamine is adequate for about 80–90% of patients. Sumatriptan, 4–6 mg SC, is effective in ~50–80% of patients, and can now be administered by a needle-free device.

DOPAMINE RECEPTOR ANTAGONISTS

Oral Oral dopamine receptor antagonists can be considered as adjunctive therapy in migraine. Drug absorption is impaired during migraine because of reduced gastrointestinal motility. Delayed absorption occurs even in the absence of nausea and is related to the severity of the attack and not its duration. Therefore, when oral NSAIDs and/or triptan agents fail, the addition of a dopamine receptor antagonist, such as metoclopramide 10 mg or domperidone 10 mg (not available in the United States), should be considered to enhance gastric absorption. In addition, dopamine receptor antagonists decrease nausea/vomiting and restore normal gastric motility.

Parenteral Dopamine receptor antagonists (e.g., chlorpromazine, prochlorperazine, metoclopramide) by injection can also provide significant acute relief of migraine; they can be used in combination with parenteral 5-$HT_{1B/1D}$ receptor agonists. A common IV protocol used for the treatment of severe migraine is the administration over 2 min of a mixture of 5 mg of prochlorperazine and 0.5 mg of dihydroergotamine.

OTHER OPTIONS FOR ACUTE MIGRAINE

Oral The combination of acetaminophen, dichloralphenazone, and isometheptene, one to two capsules, has been classified by the FDA as "possibly" effective in the treatment of migraine. Because the clinical studies demonstrating the efficacy of this combination analgesic in migraine predated the clinical trial methodologies used with the triptans, it is difficult to compare the efficacy of this sympathomimetic compound to other agents.

Parenteral Opioids are modestly effective in the acute treatment of migraine. For example, IV meperidine (50–100 mg) is given frequently in the emergency room. This regimen "works" in the sense that the pain of migraine is eliminated. Importantly, it is clear from a recent randomized controlled trial that prochlorperazine is superior to hydromorphone in the emergency room setting. However, opioids are clearly suboptimal for patients with recurrent headache. Opioids do not treat the underlying headache mechanism; rather, they act to alter the pain sensation, and there is evidence their use may decrease the likelihood of a response to triptans in the future. Moreover, in patients taking oral opioids, such as oxycodone or hydrocodone, habituation or addiction can greatly confuse the treatment of migraine. Opioid craving and/or withdrawal can aggravate and accentuate migraine. Therefore, it is recommended that opioid use in migraine be limited to patients with severe, but infrequent, headaches that are unresponsive to other pharmacologic approaches or who have contraindications to other therapies.

Neuromodulation Single pulse transcranial magnetic stimulation (sTMS) is FDA-approved for the acute treatment of migraine. Two pulses can be applied at the onset of an attack and this can be repeated. The use of sTMS is safe where there is no cranial metal implant, and offers an option to patients seeking non-pharmaceutical approaches to treatment. Similarly, a noninvasive vagus nerve stimulator (nVNS) is FDA-approved for the treatment of migraine attacks in adults. One to two 120-second doses may be applied for attack treatment.

MEDICATION-OVERUSE HEADACHE

Acute attack medications, particularly opioid or barbiturate-containing compound analgesics, have a propensity to aggravate headache frequency and induce a state of refractory daily or near-daily headache called *medication-overuse headache*. This condition is likely not a separate headache entity but a reaction of the migraine patient's biology to a particular medicine. Migraine patients who have two or more headache days a week should be cautioned about frequent analgesic use **(see "Chronic Daily Headache" in Chap. 13)**.

PREVENTIVE TREATMENTS FOR MIGRAINE

Patients with an increasing frequency of migraine attacks or with attacks that are either unresponsive or poorly responsive to abortive treatments are good candidates for preventive agents. In general, a preventive medication should be considered in patients with four or more attacks a month. Significant side effects are associated with the use of many of these agents; furthermore, determination of dose can be difficult because the recommended doses have been derived for conditions other than migraine. The mechanism of action of these drugs is unclear; it seems likely that the brain sensitivity that underlies migraine is modified. Patients are usually started on a low dose of a chosen treatment; the dose is then gradually increased, up to a reasonable maximum, to achieve clinical benefit.

Treatments that have the capacity to stabilize migraine are listed in **Table 422-6**. Most treatments must be taken daily, and there is usually a lag of between 2 and 12 weeks before an effect is seen. The drugs that have been approved by the FDA for the preventive treatment of migraine include propranolol, timolol, sodium valproate, and topiramate. In addition, a number of other drugs appear to display preventive efficacy. This group includes amitriptyline, nortriptyline, flunarizine, phenelzine, and cyproheptadine. Placebo-controlled trials of onabotulinum toxin type A in episodic migraine were negative, whereas, overall, placebo-controlled trials in chronic migraine were positive. The FDA has approved sTMS for the preventive treatment of migraine. It offers a well-tolerated, effective option for patients. Phenelzine is a monoamine oxidase inhibitor (MAOI); therefore, tyramine-containing foods, decongestants, and meperidine are contraindicated, and it is reserved for only very recalcitrant cases. Methysergide is now of historical interest only, since it is no longer manufactured. Melatonin has been reported to be useful, with controlled trial evidence but is not approved in the U.S. Monoclonal antibodies to the CGRP receptor (erenumab) or to the peptide (eptinezumab, fremanezumab, and galcanezumab) have all proven effective and well tolerated in migraine and should be available soon as preventive agents.

The probability of success with any one of the antimigraine drugs is 50%. Many patients are managed adequately with well-tolerated doses of candesartan, propranolol, amitriptyline, topiramate, or valproate. If these agents fail or produce unacceptable side effects, neuromodulation approaches, such as sTMS, or related agents from the above classes, can be used (Table 422-6). Once effective stabilization is achieved, the drug is continued for ~6 months and then slowly tapered to assess the continued need. Many patients are able to discontinue medication and experience fewer and milder attacks for long periods, suggesting that these drugs may alter the natural history of migraine.

TABLE 422-6 Preventive Treatments in Migraine[a]

DRUG	DOSE	SELECTED SIDE EFFECTS
Beta blocker		
Propranolol	40–120 mg bid	Reduced energy
Metoprolol	25–100 mg bid	Tiredness
		Postural symptoms
		Contraindicated in asthma
Antidepressants		
Amitriptyline	10–75 mg at night	Drowsiness
Dosulepin	25–75 mg at night	
Nortriptyline	25–75 mg at night	***Note:*** Some patients may only need a total dose of 10 mg, although generally 1–1.5 mg/kg body weight is required
Venlafaxine	75–150 mg/d	
Anticonvulsants		
Topiramate	25–200 mg/d	Paresthesias
		Cognitive symptoms
		Weight loss
		Glaucoma
		Caution with nephrolithiasis
Valproate	400–600 mg bid	Drowsiness
		Weight gain
		Tremor
		Hair loss
		Fetal abnormalities
		Hematologic or liver abnormalities
Serotonergic drugs		
Pizotifen[b]	0.5–2 mg qd	Weight gain
		Drowsiness
Other classes		
Flunarizine[b]	5–15 mg qd	Drowsiness
		Weight gain
		Depression
		Parkinsonism
Candesartan	4–24 mg daily	Dizziness
Neuromodulation		
Single pulse transcranial magnetic stimulation (sTMS)	4–24 pulses per day	Lightheadedness
		Tingling
		Tinnitus
Chronic migraine		
Onabotulinum toxin type A	155 U	Loss of brow furrow
No convincing evidence from controlled trials		
Verapamil		
Controlled trials demonstrate *no effect*		
Nimodipine		
Clonidine		
Selective serotonin reuptake inhibitors: fluoxetine		

[a]Commonly used preventives are listed with typical doses and common side effects. Not all listed medicines are approved by the U.S. Food and Drug Administration; local regulations and guidelines should be consulted. [b]Not available in the United States. [c]Not currently available worldwide.

TENSION-TYPE HEADACHE

Clinical Features The term *tension-type headache* is commonly used to describe a chronic head-pain syndrome characterized by bilateral tight, band-like discomfort. The pain typically builds slowly, fluctuates in severity, and may persist more or less continuously for many days. The headache may be episodic or chronic (present >15 days per month).

A useful clinical approach is to diagnose TTH in patients whose headaches are completely without accompanying features such as nausea, vomiting, photophobia, phonophobia, osmophobia, throbbing, and aggravation with movement. Such an approach neatly separates migraine, which has one or more of these features and is the main differential diagnosis, from TTH. The International Headache Society's main definition of TTH allows an admixture of nausea, photophobia, or phonophobia in various combinations, although the appendix definition does not; this illustrates the difficulty in distinguishing these two clinical entities. In clinical practice using the appendix definition to dichotomize patients on the basis of the presence of associated features

(migraine) and the absence of associated features (TTH) is highly recommended. Indeed patients whose headaches fit the TTH phenotype and who have migraine at other times, along with a family history of migraine, migrainous illnesses of childhood, or typical migraine triggers to their migraine attacks, may be biologically different from those who have TTH headache with none of the features. TTH may be infrequent (episodic) or occur on 15 days or more a month (chronic).

Pathophysiology The pathophysiology of TTH is incompletely understood. It seems likely that TTH is due to a primary disorder of central nervous system pain modulation alone, unlike migraine, which involves a more generalized disturbance of sensory modulation. Data suggest a genetic contribution to TTH, but this may not be a valid finding: given the current diagnostic criteria, the studies undoubtedly included many migraine patients. The name *tension-type headache* implies that pain is a product of *nervous tension*, but there is no clear evidence for tension as an etiology. Muscle contraction has been considered to be a feature that distinguishes TTH from migraine, but there appear to be no differences in contraction between the two headache types.

TREATMENT

Tension-Type Headache

The pain of TTH can generally be managed with simple analgesics such as acetaminophen, aspirin, or NSAIDs. Behavioral approaches including relaxation can also be effective. Clinical studies have demonstrated that triptans in pure TTH are not helpful, although triptans are effective in TTH when the patient also has migraine. For chronic TTH, amitriptyline is the only proven treatment (Table 422-6); other tricyclics, selective serotonin reuptake inhibitors, and the benzodiazepines have not been shown to be effective. There is no evidence for the efficacy of acupuncture. Placebo-controlled trials of onabotulinum toxin type A in chronic TTH were negative.

TRIGEMINAL AUTONOMIC CEPHALALGIAS, INCLUDING CLUSTER HEADACHE

The TACs describe a grouping of primary headaches including cluster headache, paroxysmal hemicrania (PH), SUNCT (*s*hort-lasting *u*nilateral *n*euralgiform headache attacks with *c*onjunctival injection and *t*earing)/SUNA (*s*hort-lasting *u*nilateral *n*euralgiform headache attacks with cranial *a*utonomic symptoms), and hemicrania continua (Table 422-1). TACs are characterized by relatively short-lasting attacks of head pain associated with cranial autonomic symptoms, such as lacrimation, conjunctival injection, aural fullness, or nasal congestion **(Table 422-7)**. Pain is usually severe and may occur more than once a day. Because of the associated nasal congestion or rhinorrhea, patients are often misdiagnosed with "sinus headache" and treated with decongestants, which are ineffective.

TACs must be differentiated from short-lasting headaches that do not have prominent cranial autonomic syndromes, notably trigeminal neuralgia (TN), primary stabbing headache, and hypnic headache. The cycling pattern and length, frequency, and timing of attacks are useful in classifying patients. Patients with TACs should undergo pituitary imaging and pituitary function tests because there is an excess of TAC presentations in patients with pituitary tumor–related headache, particularly prolactin and growth hormone secreting tumors.

Cluster Headache Cluster headache is a relatively rare form of primary headache, although nonetheless a common condition, with a population frequency of ~0.1%. The pain is deep, usually retroorbital, often excruciating in intensity, nonfluctuating, and explosive in quality. A core feature of cluster headache is periodicity. At least one of the daily attacks of pain recurs at about the same hour each day for the duration of a cluster bout. The typical cluster headache patient has daily bouts of one to two attacks of relatively short-duration unilateral pain for 8–10 weeks a year; this is usually followed by a pain-free interval that averages a little less than 1 year. Cluster headache is characterized as chronic when there is <3 months of sustained remission without treatment. Patients are generally perfectly well between episodes. Onset of attacks is nocturnal in about 50% of patients, and men are affected three times more often than women. Patients with cluster headache tend to move about during attacks, pacing, rocking, or rubbing their head for relief; some may even become aggressive during attacks. This is in sharp contrast to patients with migraine, who prefer to remain motionless during attacks.

Cluster headache is associated with ipsilateral symptoms of cranial parasympathetic autonomic activation: conjunctival injection or

TABLE 422-7 Clinical Features of the Trigeminal Autonomic Cephalalgias

	CLUSTER HEADACHE	PAROXYSMAL HEMICRANIA	SUNCT/SUNA
Gender	M > F	F = M	F ~ M
Pain			
Type	Stabbing, boring	Throbbing, boring, stabbing	Burning, stabbing, sharp
Severity	Excruciating	Excruciating	Severe to excruciating
Site	Orbit, temple	Orbit, temple	Periorbital
Attack frequency	1/alternate day–8/d	1–20/d (>5/d for more than half the time)	3–200/d
Duration of attack	15–180 min	2–30 min	5–240 s
Autonomic features	Yes	Yes	Yes (prominent conjunctival injection and lacrimation)[a]
Migrainous features[b]	Yes	Yes	Yes
Alcohol trigger	Yes	No	No
Cutaneous triggers	No	No	Yes
Indomethacin effect	—	Yes[c]	—
Abortive treatment	Sumatriptan injection or nasal spray Oxygen nVNS[c]	No effective treatment	Lidocaine (IV)
Prophylactic treatment	Verapamil Topiramate Melatoinin Lithium	Indomethacin[d]	Lamotrigine Topiramate Gabapentin

[a]If conjunctival injection and tearing are not present, consider SUNA. [b]Nausea, photophobia, or phonophobia; photophobia and phonophobia are typically unilateral on the side of the pain. [c]Non-invasive vagus nerve stimulation is FDA approved in episodic cluster headache [d]Indicates complete response to indomethacin.

Abbreviations: SUNA, short-lasting *u*nilateral *n*euralgiform headache attacks with cranial *a*utonomic features; SUNCT, short-lasting *u*nilateral *n*euralgiform headache attacks with conjunctival injection and *t*earing.

lacrimation, aural fullness, rhinorrhea or nasal congestion, or cranial sympathetic dysfunction such as ptosis. The sympathetic deficit is peripheral and likely to be due to parasympathetic activation with injury to ascending sympathetic fibers surrounding a dilated carotid artery as it passes into the cranial cavity. When present, photophobia and phonophobia are far more likely to be unilateral and on the same side of the pain, rather than bilateral, as is seen in migraine. This phenomenon of unilateral photophobia/phonophobia is characteristic of TACs. Cluster headache is likely to be a disorder involving central pacemaker neurons and neurons in the posterior hypothalamic region (Fig. 422-3).

TREATMENT

Cluster Headache

The most satisfactory treatment is the administration of drugs to prevent cluster attacks until the bout is over. However, treatment of acute attacks is required for all cluster headache patients at some time.

ACUTE ATTACK TREATMENT

Cluster headache attacks peak rapidly, and thus a treatment with rapid onset is required. Many patients with acute cluster headache respond very well to oxygen inhalation. This should be given as 100% oxygen at 10–12 L/min for 15–20 min. It appears that high flow and high oxygen content are important. Sumatriptan 6 mg SC is rapid in onset and will usually shorten an attack to 10–15 min; there is no evidence of tachyphylaxis. Sumatriptan (20 mg) and zolmitriptan (5 mg) nasal sprays are both effective in acute cluster headache, offering a useful option for patients who may not wish to self-inject daily. Noninvasive vagus nerve stimulation (nVNS) is FDA approved for the acute treatment of attacks in episodic cluster headache using three 2-min stimulation cycles applied consecutively at the onset of headache on the side of pain; this may be repeated after nine minutes. Oral sumatriptan is not effective for prevention or for acute treatment of cluster headache.

PREVENTIVE TREATMENTS **(TABLE 422-8)**

The choice of a preventive treatment in cluster headache depends in part on the length of the bout. Patients with long bouts or those with chronic cluster headache require medicines that are safe when taken for long periods. For patients with relatively short bouts, limited courses of oral glucocorticoids can be very useful. A 10-day course of prednisone, beginning at 60 mg daily for 7 days and followed by a rapid taper, may interrupt the pain bout for many patients. Greater occipital nerve injection with lidocaine and glucocorticoids has been shown to be effective in randomized controlled trials, with a benefit that lasts up to 6–8 weeks.

Most experts favor verapamil as the first-line preventive treatment for patients with chronic cluster headache or with prolonged bouts. While verapamil compares favorably with lithium in practice, some patients require verapamil doses far in excess of those administered for cardiac disorders. The initial dose range is 40–80 mg twice daily; effective doses may be as high as 960 mg/d. Side effects such as constipation, leg swelling, or gingival hyperplasia can be problematic. Of paramount concern, however, is the cardiovascular safety of verapamil, particularly at high doses. Verapamil can cause heart block by slowing conduction in the atrioventricular node, a condition that can be monitored by following the PR interval on a standard electrocardiogram (ECG). Approximately 20% of patients treated with verapamil develop ECG abnormalities, which can be observed with doses as low as 240 mg/d; these abnormalities can worsen over time in patients on stable doses. A baseline ECG is recommended for all patients. The ECG is repeated 10 days after a dose change in patients whose dose is being increased above 240 mg daily. Dose increases are usually made in 80-mg increments. For patients on long-term verapamil, ECG monitoring every 6 months is advised.

TABLE 422-8 Preventive Management of Cluster Headache

SHORT-TERM PREVENTION	LONG-TERM PREVENTION
EPISODIC CLUSTER HEADACHE	**EPISODIC CLUSTER HEADACHE AND PROLONGED CHRONIC CLUSTER HEADACHE**
Prednisone 1 mg/kg up to 60 mg qd, tapering over 21 days Verapamil 160–960 mg/d Greater occipital nerve injection	Verapamil 160–960 mg/d Topiramate[a] 100–400 mg/d nVNS[b] 6 to 24 stimulations/d Lithium 400–800 mg/d Melatonin[a] 9–12 mg/d Gabapentin[a] 1200–3600 mg/d

[a]Unproven but of potential benefit. [b]Non-invasive vagus nerve stimulation.

NEUROMODULATION THERAPY

When medical therapies fail in chronic cluster headache, neuromodulation strategies can be used. Sphenopalatine ganglion (SPG) stimulation with an implanted battery-free stimulator has been shown in randomized controlled trials to be effective in aborting attacks and reducing their frequency over time. nVNS compares favorably to standard-of-care in open label experience. Similarly, occipital nerve stimulation has been used open label and appears to be beneficial. Deep-brain stimulation of the region of the posterior hypothalamic gray matter is successful in about 50% of patients treated, although its risk-benefit ratio makes it inappropriate before all other less invasive options have been explored.

PAROXYSMAL HEMICRANIA

PH is characterized by frequent unilateral, severe, short-lasting episodes of headache. Like cluster headache, the pain tends to be retroorbital but may be experienced all over the head and is associated with autonomic phenomena such as lacrimation and nasal congestion. Patients with remissions are said to have episodic PH, whereas those with the nonremitting form are said to have chronic PH. The essential features of PH are unilateral; very severe pain; short-lasting attacks (2–45 min); very frequent attacks (usually >5 a day); marked autonomic features ipsilateral to the pain; rapid course (<72 h); and excellent response to indomethacin. In contrast to cluster headache, which predominantly affects males, the male-to-female ratio in PH is close to 1:1.

Indomethacin (25–75 mg tid), which can completely suppress attacks of PH, is the treatment of choice. Although therapy may be complicated by indomethacin-induced gastrointestinal side effects, currently there are no consistently effective alternatives. Topiramate is helpful in some cases. Piroxicam has been used, although it is not as effective as indomethacin. Verapamil, an effective treatment for cluster headache, does not appear to be useful for PH. nVNS can be useful in these patients and can be very effective. In occasional patients, PH can coexist with TN (PH-tic syndrome); similar to cluster-tic syndrome, each component may require separate treatment.

Secondary PH has been reported with lesions in the region of the sella turcica, including arteriovenous malformation, cavernous sinus meningioma, pituitary pathology, and epidermoid tumors. Secondary PH is more likely if the patient requires high doses (>200 mg/d) of indomethacin. In patients with apparent bilateral PH, raised cerebrospinal fluid (CSF) pressure should be suspected. It is important to note that indomethacin reduces CSF pressure. When a diagnosis of PH is considered, magnetic resonance imaging (MRI) is indicated to exclude a pituitary lesion.

SUNCT/SUNA

SUNCT is a rare primary headache syndrome characterized by severe, unilateral orbital or temporal pain that is stabbing or throbbing in quality. Diagnosis requires at least 20 attacks, lasting for 5–240 s; ipsilateral conjunctival injection and lacrimation should be present. In some patients, conjunctival injection or lacrimation is missing, and the diagnosis of SUNA can be made.

DIAGNOSIS The pain of SUNCT/SUNA is unilateral and may be located anywhere in the head. Three basic patterns can be seen: single stabs, which are usually short-lived; groups of stabs; or a longer attack comprising many stabs between which the pain does not completely

resolve, thus giving a "saw-tooth" phenomenon with attacks lasting many minutes. Each pattern may be seen in the context of an underlying continuous head pain. Characteristics that lead to a suspected diagnosis of SUNCT are the cutaneous (or other) triggers of attacks, a lack of refractory period to triggering between attacks, and the lack of a response to indomethacin. Apart from trigeminal sensory disturbance, the neurologic examination is normal in primary SUNCT/SUNA.

The diagnosis of SUNCT/SUNA is often confused with TN particularly in first-division TN (Chap. 433). Minimal or no cranial autonomic symptoms and a clear refractory period to triggering indicate a diagnosis of TN.

SECONDARY (SYMPTOMATIC) SUNCT SUNCT can be seen with posterior fossa or pituitary lesions. All patients with SUNCT/SUNA should be evaluated with pituitary function tests and a brain MRI with pituitary views.

TREATMENT

SUNCT/SUNA

ABORTIVE THERAPY

Therapy of acute attacks is not a useful concept in SUNCT/SUNA because the attacks are of such short duration. However, IV lidocaine, which arrests the symptoms, can be used in hospitalized patients.

PREVENTIVE THERAPY

Long-term prevention to minimize disability and hospitalization is the goal of treatment. The most effective treatment for prevention is lamotrigine, 200–400 mg/d. Topiramate and gabapentin may also be effective. Carbamazepine, 400–500 mg/d, has been reported by patients to offer modest benefit.

Surgical approaches such as microvascular decompression or destructive trigeminal procedures are seldom useful and often produce long-term complications. Greater occipital nerve injection has produced limited benefit in some patients. Occipital nerve stimulation is probably helpful in a subgroup of these patients. For intractable cases, short-term prevention with IV lidocaine can be effective.

Hemicrania Continua The essential features of hemicrania continua are moderate and continuous unilateral pain associated with fluctuations of severe pain; complete resolution of pain with indomethacin; and exacerbations that may be associated with autonomic features, including conjunctival injection, lacrimation, and photophobia on the affected side. The age of onset ranges from 10 to 70 years; women are affected twice as often as men. The cause is unknown.

TREATMENT

Hemicrania Continua

Treatment consists of indomethacin; other NSAIDs appear to be of little or no benefit. The IM injection of 100 mg of indomethacin has been proposed as a diagnostic tool, and administration with a placebo injection in a blinded fashion can be very useful diagnostically. Alternatively, a trial of oral indomethacin, starting with 25 mg tid, then 50 mg tid, and then 75 mg tid, can be given. Up to 2 weeks at the maximal dose may be necessary to assess whether a dose has a useful effect. Topiramate can be helpful in some patients. nVNS can be useful in these patients. Occipital nerve stimulation probably has a role in patients with hemicrania continua who are unable to tolerate indomethacin.

OTHER PRIMARY HEADACHES

Primary Cough Headache Primary cough headache is a generalized headache that begins suddenly, lasts for seconds or several minutes, sometimes up to a few hours, and is precipitated by coughing; it is preventable by avoiding coughing or other precipitating events, which can include sneezing, straining, laughing, or stooping. In all patients with this syndrome, serious etiologies must be excluded before a diagnosis of "benign" primary cough headache can be established. A Chiari malformation or any lesion causing obstruction of CSF pathways or displacing cerebral structures can be the cause of the head pain. Other conditions that can present with cough or exertional headache as the initial symptom include cerebral aneurysm, carotid stenosis, and vertebrobasilar disease. Benign cough headache can resemble benign exertional headache (below), but patients with the former condition are typically older.

TREATMENT

Primary Cough Headache

Indomethacin 25–50 mg two to three times daily is the treatment of choice. Some patients with cough headache obtain complete cessation of their attacks with lumbar puncture; this is a simple option when compared to prolonged use of indomethacin, and it is effective in about one-third of patients. The mechanism of this response is unclear.

Primary Exercise Headache Primary exercise headache has features resembling both cough headache and migraine. It may be precipitated by any form of exercise; it often has the pulsatile quality of migraine. The pain lasts <48 h, is bilateral and often throbbing at onset; migrainous features may develop in patients susceptible to migraine. The duration tends to be shorter in adolescents than in older adults. Primary exercise headache can be prevented by avoiding excessive exertion, particularly in hot weather or at high altitude.

The mechanism of primary exercise headache is unclear. Acute venous distension likely explains one syndrome—the acute onset of headache with straining and breath holding, as in weightlifter's headache. Because exercise can result in headache in a number of serious underlying conditions (Chap. 13), these must be considered in patients with exercise headache. Pain from angina may be referred to the head, probably by central connections of vagal afferents, and may present as exercise headache (cardiac cephalgia). The link to exercise is the main clinical clue that headache is of cardiac origin. Pheochromocytoma may occasionally cause exercise headache. Intracranial lesions and stenosis of the carotid arteries are other possible etiologies.

TREATMENT

Primary Exercise Headache

Exercise regimens should begin modestly and progress gradually to higher levels of intensity. Indomethacin at daily doses from 25 to 150 mg is generally effective in benign exertional headache. Indomethacin (50 mg), ergotamine (1 mg orally), and dihydroergotamine (2 mg by nasal spray) are useful prophylactic measures.

Primary Headache Associated with Sexual Activity Three types of sex headache are reported: a dull bilateral ache in the head and neck that intensifies as sexual excitement increases; a sudden, severe, explosive headache occurring at orgasm; and a postural headache developing after coitus. The last arises from vigorous sexual activity and is a form of low CSF pressure headache and thus not a primary headache disorder (Chap. 13). Headaches developing at the time of orgasm are not always benign; 5–12% of cases of subarachnoid hemorrhage are precipitated by sexual intercourse. Sex headache is reported by men more often than women and may occur at any time during the years of sexual activity. It may appear on several occasions in succession and then not trouble the patient again, even without an obvious change in sexual activity. In patients who stop sexual activity when headache is first noticed, the pain may subside within a period of 5 min to 2 h. In about half of patients, sex headache will subside within 6 months. Most patients with sex headache do not have exercise or

cough headache; this clinical paradox is generally a marker of primary sex headache. Migraine is probably more common in patients with sex headache.

TREATMENT

Primary Sex Headache

Benign sex headaches recur irregularly and infrequently. Management can often be limited to reassurance and advice about ceasing sexual activity if a mild, warning headache develops. Propranolol can be used to prevent headache that recurs regularly or frequently, but the dosage required varies from 40 to 200 mg/d. An alternative is the calcium channel–blocking agent diltiazem, 60 mg tid. Indomethacin (25–50 mg) or frovatriptan (2.5 mg) taken 30–45 min prior to sexual activity can also be helpful.

Primary Thunderclap Headache Sudden onset of severe headache may occur in the absence of any known provocation. The differential diagnosis includes the sentinel bleed of an intracranial aneurysm, cervicocephalic arterial dissection, and cerebral venous thrombosis. Headaches of explosive onset may also be caused by the ingestion of sympathomimetic drugs or of tyramine-containing foods in a patient who is taking MAOIs, or they may be a symptom of pheochromocytoma. Whether thunderclap headache can be the presentation of an unruptured cerebral aneurysm is uncertain. When neuroimaging studies and lumbar puncture exclude subarachnoid hemorrhage, patients with thunderclap headache usually do very well over the long term. In one study of patients whose computed tomography (CT) scans and CSF findings were negative, ~15% had recurrent episodes of thunderclap headache, and nearly half subsequently developed migraine or TTH.

The first presentation of any sudden-onset severe headache should be diligently investigated with neuroimaging (CT or, when possible, MRI with MR angiography) and CSF examination. Reversible segmental cerebral vasoconstriction may be seen in primary thunderclap headache without an intracranial aneurysm, and it is thought that this may be an under-diagnosed condition. In the presence of posterior leukoencephalopathy, the differential diagnosis includes cerebral angiitis, drug toxicity (cyclosporine, intrathecal methotrexate/cytarabine, pseudoephedrine, or cocaine), posttransfusion effects, and postpartum angiopathy. Treatment with nimodipine may be helpful, although the vasoconstriction of primary thunderclap headache resolves spontaneously.

Cold-Stimulus Headache This refers to head pain triggered by application or ingestion/inhalation of something cold. It is bought on quickly and typically resolves within 10–30 min of the stimulus being removed. It is best recognized as "brain-freeze" headache or ice-cream headache when due to ingestion. Although cold may be uncomfortable at some level for many people, it is the reliable, severe, and somewhat prolonged nature of these pains that set them apart. The transient receptor potential cation subfamily M member 8 (TRPM8) channel, a known cold temperature sensor, may be a mediator of this syndrome.

External Pressure Headache External pressure from compression or traction on the head can produce a pain that may have some generalized component, although the pain is largely focused around the site of the pressure. It typically resolves within an hour of the stimulus being removed. Examples of stimuli include helmets, swimming goggles, or very long ponytails. Treatment is to recognize the problem and remove the stimulus.

Primary Stabbing Headache The essential features of primary stabbing headache are stabbing pain confined to the head or, rarely, the face, lasting from 1 to many seconds and occurring as a single stab or a series of stabs; absence of associated cranial autonomic features; absence of cutaneous triggering of attacks; and a pattern of recurrence at irregular intervals (hours to days). The pains have been variously described as "ice-pick pains" or "jabs and jolts." They are more common in patients with other primary headaches, such as migraine, the TACs, and hemicrania continua.

TREATMENT

Primary Stabbing Headache

The response of primary stabbing headache to indomethacin (25–50 mg two to three times daily) is usually excellent. As a general rule, the symptoms wax and wane, and after a period of control on indomethacin, it is appropriate to withdraw treatment and observe the outcome.

Nummular Headache Nummular headache is felt as a round or elliptical discomfort that is fixed in place, ranges in size from 1 to 6 cm, and may be continuous or intermittent. Uncommonly it may be multifocal. It may be episodic but is more often continuous during exacerbations. Accompanying the pain there may be a local sensory disturbance, such as allodynia or hypesthesia. Local dermatologic or bony lesions need to be excluded by examination and investigation. This condition can be difficult to treat when present in isolation; tricyclics, such as amitriptyline, or anticonvulsants, such as topiramate or valproate, are most often tried. This phenotype can be seen in combination with migraine and the TACs, in which cases treatment of the associated condition is often effective for the nummular headache as well.

Hypnic Headache This headache syndrome typically begins a few hours after sleep onset. The headaches last from 15 to 30 min and are typically moderately severe and generalized, although they may be unilateral and can be throbbing. Patients may report falling back to sleep only to be awakened by a further attack a few hours later; up to three repetitions of this pattern occur through the night. Daytime naps can also precipitate head pain. Most patients are female, and the onset is usually after age 60 years. Headaches are bilateral in most, but may be unilateral. Photophobia, phonophobia, and nausea are usually absent. The major secondary consideration in this headache type is poorly controlled hypertension; 24-h blood pressure monitoring is recommended to detect this treatable condition.

TREATMENT

Hypnic Headache

Patients with hypnic headache generally respond to a bedtime dose of lithium carbonate (200–600 mg). For those intolerant of lithium, verapamil (160 mg) is an alternative strategy. One to two cups of coffee or caffeine, 60 mg orally, at bedtime may be effective in approximately one-third of patients. Case reports also suggest that flunarizine, 5 mg nightly, or indomethacin, 25–75 mg nightly, can be effective.

New Daily Persistent Headache Primary new daily persistent headache (NDPH) occurs in both males and females. It can be of the migrainous type, with features of migraine, or it can be featureless, appearing as new-onset TTH. Those with migrainous features are the most common form, and include unilateral headache and throbbing pain; each feature is present in about one-third of patients. Nausea, photophobia, and/or phonophobia occur in about half of patients. Some patients have a previous history of migraine; however, the proportion of NDPH sufferers with preexisting migraine is no greater than the frequency of migraine in the general population. NDPH may be more common in adolesents. Treatment of migrainous-type primary NDPH consists of using the preventive therapies effective in migraine (see above). Featureless NDPH is one of the primary headache forms most refractory to treatment. Standard preventive therapies can be offered but are often ineffective. The secondary NDPHs are discussed elsewhere (Chap. 13).

Acknowledgment

The Editors acknowledge the contributions of Neil H. Raskin to earlier editions of this chapter.

FURTHER READING

Charles A: Migraine. N Engl J Med 377:553, 2017.

Goadsby PJ: Bench to bedside advances in the 21st century for primary headache disorders: Migraine treatments for migraine patients. Brain 139:2571, 2016.

Goadsby PJ et al: Pathophysiology of migraine: A disorder of sensory processing. Physiol Rev 97:553, 2017.

Lipton RB et al: Migraine prevalence, disease burden, and the need for preventive therapy. Neurology 68:343, 2007.

May A: Cluster headache: Pathogenesis, diagnosis, and management. Lancet 366:843, 2005.

Schankin CJ et al: "Visual snow"—a disorder distinct from persistent migraine aura. Brain 137:1419, 2014.

Tolner EA et al: From migraine genes to mechanisms. Pain 156 Suppl 1:S64, 2015.

423 Alzheimer's Disease

William W. Seeley, Bruce L. Miller

ALZHEIMER'S DISEASE

Approximately 10% of all persons aged >70 years have significant memory loss, and in more than half the cause is Alzheimer's disease (AD). It is estimated that the median annual total cost of caring for a single patient with advanced AD is >$50,000, while the emotional toll for family members and caregivers is immeasurable. AD can manifest as early as the third decade of life, but it is the most common cause of dementia in the elderly. Patients most often present with an insidious loss of episodic memory followed by a slowly progressive dementia. In typical amnestic AD, brain atrophy begins in the medial temporal lobes before spreading to lateral and medial parietal and temporal lobes and lateral frontal cortex. Microscopically, there are widespread neuritic plaques containing amyloid beta (Aβ), neurofibrillary tangles (NFTs) composed of hyperphosphorylated tau filaments, and Aβ accumulation in blood vessel walls in cortex and leptomeninges (see "Pathology," below). The identification of causative mutations and susceptibility genes for AD has provided a foundation for rapid progress in understanding the biological basis of the disorder. The major genetic risk for AD is the ε4 allele of the apolipoprotein E (ApoE) gene. Carrying one ε4 allele increases the risk for AD by two- to threefold whereas two alleles increase the risk sixteenfold in both sexes.

CLINICAL MANIFESTATIONS

The cognitive changes of AD tend to follow a characteristic pattern, beginning with memory impairment and progressing to language and visuospatial deficits, followed by executive dysfunction. Yet, ~20% of patients with AD present with non-memory complaints such as word-finding, organizational, or navigational difficulty. In other patients, visual processing dysfunction (referred to as posterior cortical atrophy syndrome) or a progressive "logopenic" aphasia characterized by difficulties with naming and repetition are the primary manifestations of AD for years before progressing to involve memory and other cognitive domains. Still other patients may present with an asymmetric akinetic-rigid-dystonic ("corticobasal") syndrome or a dysexecutive/behavioral, i.e., "frontal" variant of AD.

In the early stages of typical amnestic AD, the memory loss may go unrecognized or be ascribed to benign forgetfulness of aging. Once the memory loss becomes noticeable to the patient and spouse and falls 1.5 standard deviations below normal on standardized memory tests, the term mild cognitive impairment (MCI) is often used. This construct provides useful prognostic information, because ~50% of patients with MCI (roughly 12% per year) will progress to AD over 4 years. Increasingly, the MCI construct is being replaced by the notion of "early symptomatic AD" to signify that AD is considered the underlying disease (based on clinical or biomarker evidence) in a patient who remains functionally compensated. Even earlier in the course, "prodromal AD" refers to a person with biomarker evidence of AD (amyloid imaging positive with positron emission tomography (PET) or low cerebrospinal $A\beta_{42}$ and mildly elevated tau) in the absence of symptoms. These refinements have been developed in anticipation of early-stage treatment and prevention trials that are well underway in humans. New evidence suggests that partial and sometimes generalized seizures herald AD and can occur even prior to dementia onset, especially in younger patients.

Eventually, with AD, the cognitive problems begin to interfere with daily activities, such as keeping track of finances, following instructions on the job, driving, shopping, and housekeeping. Some patients are unaware of these difficulties (*anosognosia*), but most remain acutely attuned to their deficits. Changes in environment (travel, relocation, hospitalization) tend to destabilize the patient. Over time, patients become lost on walks or while driving. Social graces, routine behavior, and superficial conversation may be surprisingly intact, even into the later stages of the illness.

In the middle stages of AD, the patient is unable to work, is easily lost and confused, and requires daily supervision. Language becomes impaired—first naming, then comprehension, and finally fluency. Word-finding difficulties and circumlocution can be evident in the early stages, even when formal testing demonstrates intact naming and fluency. *Apraxia* emerges, manifesting as trouble performing learned sequential motor tasks such as using utensils or appliances. Visuospatial deficits begin to interfere with dressing, eating, or even walking, and patients fail to solve simple puzzles or copy geometric figures. Simple calculations and clock reading become difficult in parallel.

In the late stages, some persons remain ambulatory, wandering aimlessly. Loss of judgment and reasoning is inevitable. Delusions are prevalent and usually simple, with common themes of theft, infidelity, or misidentification. Disinhibition and uncharacteristic belligerence may occur and alternate with passivity and withdrawal. Sleep-wake patterns are disrupted, and nighttime wandering becomes disturbing to the household. Some patients develop a shuffling gait with generalized muscle rigidity associated with slowness and awkwardness of movement. Patients often look parkinsonian **(Chap. 427)** but rarely have a high-amplitude, low-frequency tremor at rest. There is a strong overlap between dementia with Lewy bodies (DLB) **(Chap. 426)** and AD, and some AD patients develop more classical parkinsonian features.

In the end stages, patients with AD become rigid, mute, incontinent, and bedridden, and help is needed with eating, dressing, and toileting. Hyperactive tendon reflexes and myoclonic jerks (sudden brief contractions of various muscles or the whole body) may occur spontaneously or in response to physical or auditory stimulation. Often death results from malnutrition, secondary infections, pulmonary emboli, heart disease, or, most commonly, aspiration. The typical duration of symptomatic AD is 8–10 years, but the course ranges from 1 to 25 years. For unknown reasons, some patients with AD show a steady decline in function while others have prolonged plateaus without major deterioration.

DIFFERENTIAL DIAGNOSIS

A detailed discussion of the diagnosis of dementia is presented in Chap. 25. Early in the disease course, other etiologies of dementia should be excluded **(see Tables 25-1, 25-3, and 25-4)**. Neuroimaging studies (computed tomography [CT] and magnetic resonance imaging [MRI]) do not show a single specific pattern with AD and may be normal early in the disease. As AD progresses, more distributed but usually posterior-predominant cortical atrophy becomes apparent, along with atrophy of the medial temporal memory structures **(see Fig. 25-1)**. The main purpose of imaging is to exclude other disorders, such as primary and secondary neoplasms, vascular dementia, diffuse white matter disease, and normal-pressure hydrocephalus (NPH). Imaging also helps to distinguish AD from other degenerative disorders, such as frontotemporal dementia (FTD) **(Chap. 424)** or the prion disorder Creutzfeldt-Jakob disease (CJD) **(Chap. 430)**, which feature

distinctive imaging patterns. Functional imaging studies, such as PET, reveal hypometabolism in the posterior temporal-parietal cortex in AD **(see Fig. 25-1)**. PET can also be used to detect the presence of fibrillar amyloid in the brain **(see Fig. 25-4)**, and amyloid PET positivity is becoming a criterion for entry into AD treatment trials. Use of amyloid PET in routine clinical evaluation may be limited to specific clinical scenarios, however. Although amyloid PET binding is detected in AD, many asymptomatic healthy older individuals also show amyloid uptake, and the likelihood that these individuals will convert to clinical AD is still under study. Similarly, dementia due to a non-AD disorder can be the underlying etiology in a patient who tests positively on amyloid PET. Electroencephalogram (EEG) is normal or shows nonspecific slowing; prolonged EEG can be used to seek out intermittent nonconvulsive seizures. Routine spinal fluid examination is also normal. Cerebrospinal fluid (CSF) $A\beta_{42}$ level is reduced, whereas phosphorylated tau protein is elevated, but the test characteristics of these assays can still make interpretation challenging in individual patients. *Slowly progressive decline in memory and orientation, normal results on laboratory tests, and an MRI or CT scan showing only distributed or posteriorly predominant cortical and hippocampal atrophy are highly suggestive of AD.* A clinical diagnosis of AD reached after careful evaluation is confirmed at autopsy about 90% of the time, with misdiagnosed cases usually resulting from pathological frontotemporal lobar degeneration (FTLD), DLB, hippocampal sclerosis of the elderly, or a mixture of mild AD changes with vascular or DLB pathology.

Simple clinical clues are useful in the differential diagnosis. Early prominent gait disturbance with only mild memory loss suggests vascular dementia or, rarely, NPH (see below). Resting tremor with stooped posture, bradykinesia, and masked facies suggest PD **(Chap. 427)** or DLB **(Chap. 426)**. When dementia occurs after a well-established diagnosis of PD, PD dementia (PDD) is usually the correct diagnosis, but many patients with this diagnosis will show a mixture of AD and Lewy body disease at autopsy. The early appearance of parkinsonian features in association with fluctuating alertness, visual hallucinations, or delusional misidentification suggests DLB. Chronic alcoholism should prompt the search for vitamin deficiency. Loss of joint position and vibration sensibility accompanied by Babinski signs suggests vitamin B_{12} deficiency, especially in a patient with a history of autoimmune disease, small bowel resection or irradiation, or veganism **(Chap. 95)**. Early onset of a focal seizure suggests a metastatic or primary brain neoplasm **(Chap. 86)**. Previous or ongoing depression raises suspicion for depression-related cognitive impairment, although AD can feature a depressive prodrome. A history of treatment for insomnia, anxiety, psychiatric disturbance, or epilepsy suggests chronic drug intoxication. Rapid progression over a few weeks or months associated with rigidity and myoclonus suggests CJD **(Chap. 430)**. Prominent behavioral changes with intact navigation and focal anterior-predominant atrophy on brain imaging are typical of FTD. A positive family history of dementia suggests either one of the familial forms of AD or one of the other genetic disorders associated with dementia, such as FTD **(Chap. 424)**, Huntington's disease (HD) **(Chap. 428)**, prion disease **(Chap. 430)**, or rare hereditary ataxias **(Chap. 431)**.

EPIDEMIOLOGY

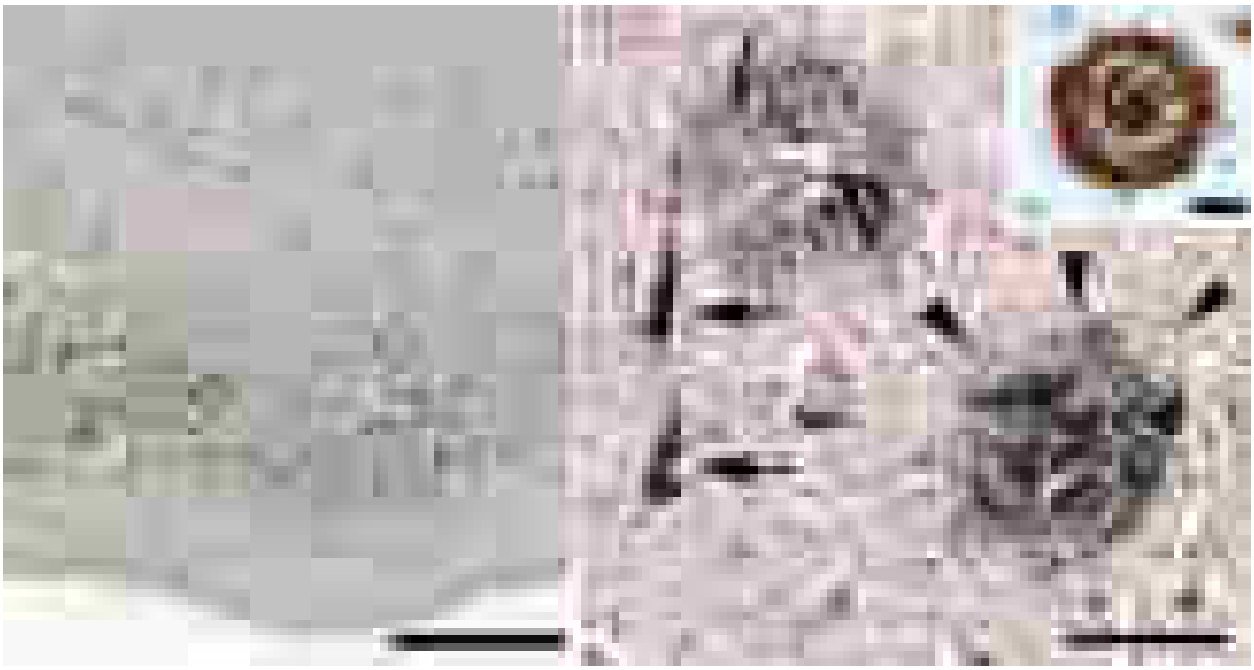
The most important risk factors for AD are age >70 years and a positive family history. The prevalence of AD increases with each decade of adult life, reaching 20–40% of the population aged >85. A positive family history of dementia suggests a genetic contribution to AD, although autosomal dominant inheritance occurs in only 2% of patients. Female sex is a risk factor independent of the greater longevity of women, and women who carry a single Apo ε4 allele are more susceptible than are male ε4 carriers. A history of head trauma with concussion increases the risk for AD. AD is more common in groups with low educational attainment, but education influences test-taking ability, and it is clear that AD can affect persons of all intellectual levels. One study found that the capacity to express complex written language in early adulthood correlated with a decreased risk for AD. Numerous environmental factors, including aluminum, mercury, and viruses, have been proposed as causes of AD, but rigorous studies have failed to demonstrate a significant role for any of these exposures. Similarly, several studies suggest that the use of nonsteroidal anti-inflammatory agents is associated with a decreased risk of AD, but this risk has not been confirmed in large prospective studies. Vascular disease, and stroke in particular, seems to lower the threshold for the clinical expression of AD. Also, in many patients with AD, amyloid angiopathy can lead to microhemorrhages, large lobar hemorrhages, ischemic infarctions most often in the subcortical white matter, or in rare cases an inflammatory leukoencephalopathy. Diabetes increases the risk of AD threefold. Elevated homocysteine and cholesterol levels; hypertension; diminished serum levels of folic acid; low dietary intake of fruits, vegetables, and red wine; and low levels of exercise are all being explored as potential risk factors for AD.

PATHOLOGY

At autopsy, the earliest and most severe degeneration is usually found in the medial temporal lobe (entorhinal/perirhinal cortex and hippocampus), inferolateral temporal cortex, and nucleus basalis of Meynert. The characteristic microscopic findings are neuritic plaques and NFTs **(Fig. 423-1)**. These lesions accumulate in small numbers during normal brain aging but dominate the picture in AD. Increasing evidence suggests that soluble amyloid species called *oligomers* may cause cellular dysfunction and represent the early toxic molecule in AD. Eventually, further amyloid polymerization and fibril formation lead to neuritic plaques, which contain a central core of amyloid, proteoglycans, ApoE, α-antichymotrypsin, and other proteins. Aβ is a protein of 39–42 amino acids that is derived proteolytically from a larger transmembrane protein, *amyloid precursor protein* (APP), when APP is cleaved by β and γ secretases **(Fig. 423-2)**. The normal function of the Aβ peptides remains uncertain. APP has neurotrophic and neuroprotective properties. The plaque core is surrounded by a halo, which contains dystrophic, tau-immunoreactive neurites and activated microglia. The accumulation of Aβ in cerebral arterioles is termed *amyloid angiopathy*. NFTs are composed of silver-staining neuronal cytoplasmic fibrils composed of abnormally phosphorylated tau protein; they appear as paired helical filaments by electron microscopy. Tau binds to and stabilizes microtubules, supporting axonal transport of organelles, glycoproteins, neurotransmitters, and other important cargoes throughout the neuron. Once hyperphosphorylated, tau can no longer bind properly to microtubules and redistributes from the axon to throughout the neuronal cytoplasm and distal dendrites, compromising function. Other theories emphasize that abnormal conformations of tau induce misfolding of native (unfolded) tau into pathological conformations and that this prion-like templating process is responsible for tau spreading **(Chap. 417)**. Finally, patients with AD often show comorbid DLB or vascular pathology. Most prevailing rodent models of AD involve expression of mutant transgenes that leads to $A\beta_{42}$ accumulation in the absence of tauopathy. Even in these models, diminishing neuronal tau ameliorates cognitive deficits and nonconvulsive seizures while $A\beta_{42}$ continues to accumulate, raising hope for tau-lowering therapies in humans. Biochemically, AD is associated with a decrease in the cortical levels of several proteins and neurotransmitters, especially acetylcholine, its synthetic enzyme choline acetyltransferase, and nicotinic cholinergic receptors. Reduction of acetylcholine reflects degeneration of cholinergic neurons in the nucleus basalis of Meynert, located just below the thalamus and adjacent to the third ventricle, that project throughout the cortex. There is also noradrenergic and serotonergic depletion due to degeneration of upper brainstem nuclei such as the locus coeruleus (norepinephrine) and dorsal raphe (serotonin), where tau-immunoreactive neuronal cytoplasmic inclusions can be identified in early adult life, even in individuals lacking entorhinal cortex NFTs.

GENETIC CONSIDERATIONS

Several genes play an important role in the pathogenesis of AD. One is the *APP* gene on chromosome 21. Adults with trisomy 21 (Down's syndrome) consistently develop the typical neuropathologic hallmarks of AD if they survive beyond age 40 years, and many develop a progressive dementia superimposed on their baseline mental retardation. The extra dose of the *APP* gene on chromosome 21 is the initiating cause of AD in adult Down's syndrome and results in excess



A B

FIGURE 423-1 Neuropathology of Alzheimer's disease. *A.* Early neurofibrillary degeneration, consisting of neurofibrillary tangles and neuropil threads, preferentially affects the medial temporal lobes, especially the stellate pyramidal neurons that compose the layer 2 islands of entorhinal cortex, as shown using Gallyas silver staining. ***B.*** Higher magnification view reveals the fibrillar nature of tangles (*arrows*) and the complex structure of neuritic plaques (*arrowheads*), whose major component is Aβ (*inset* shows immunohistochemistry for Aβ). Scale bars are 500 μM in ***A***, 50 μM in ***B***, and 20 μM in ***B*** inset.

cerebral amyloid production. Supporting this hypothesis, some families with early age-of-onset familial AD (FAD) have point mutations in *APP*. Although very rare, these families were the first examples of single-gene autosomal dominant transmission of AD.

Investigation of large families with multigenerational FAD led to the discovery of two additional AD-causing genes, the *presenilins*. Presenilin-1 (*PSEN-1*) is on chromosome 14 and encodes presenilin-1 protein (also known as S182). Mutations in this gene cause an early age-of-onset AD, with onset typically before age 60 and often before age 50, transmitted in an autosomal dominant, highly penetrant fashion. More than 100 different mutations have been found in the *PSEN-1* gene in families from a wide range of ethnic backgrounds. Presenilin-2 (*PSEN-2*) is on chromosome 1 and encodes the presenilin-2 protein (also known as STM2). A mutation in the *PSEN-2* gene was first found in a group of American families with Volga German ethnic background. Mutations in *PSEN-1* are much more common than those in *PSEN-2*. The presenilins are highly homologous and encode similar proteins that at first appeared to have seven transmembrane domains (hence the designation *STM*), but subsequent studies have suggested eight such domains, with a ninth submembrane region. Both presenilins are cytoplasmic neuronal proteins that are widely expressed throughout the nervous system. They are homologous to a cell-trafficking protein, sel 12, found in the nematode *Caenorhabditis elegans*. Patients with mutations in the presenilin genes have elevated plasma levels of $A\beta_{42}$, and *PSEN-1* mutations produce increased $A\beta_{42}$ in the media in cell culture. *PSEN-1* is involved in the cleavage of APP at the γ secretase site and mutations in either gene (*PSEN-1* or *APP*) may disturb γ secretase cleavage. Mutations in *PSEN-1* are the most common cause of early-age-of-onset FAD, representing perhaps 40–70% of all cases. Mutations in *PSEN-1* tend to produce AD with an earlier age of onset (mean onset 45 years) and a shorter, more rapidly progressive course (mean duration 6–7 years) than the disease caused by mutations in *PSEN-2* (mean onset 53 years; duration 11 years). Although some carriers of *PSEN-2* mutations have had onset of dementia after the age of 70, mutations in the presenilins rarely lead to late-age-of-onset AD. Clinical genetic testing for these uncommon mutations is available but likely to be revealing only in early-age-of-onset FAD and should be performed in association with formal genetic counseling.

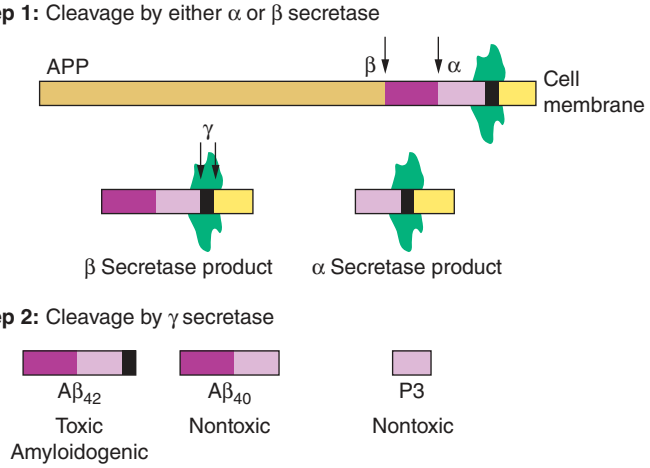


FIGURE 423-2 Amyloid precursor protein (APP) is catabolized by α, β, and γ secretases. A key initial step is the digestion by either β secretase (BASE) or α secretase (ADAM10 or ADAM17 [TACE]), producing smaller nontoxic products. Cleavage of the β secretase product by γ secretase (Step 2) results in either the toxic $A\beta_{42}$ or the nontoxic $A\beta_{40}$ peptide; cleavage of the α secretase product by γ secretase produces the nontoxic P3 peptide. Excess production of $A\beta_{42}$ is a key initiator of cellular damage in Alzheimer's disease (AD). Therapeutics for AD have focused on attempts to reduce accumulation of $A\beta_{42}$ by antagonizing β or γ secretases, promoting α secretase, or clearing $A\beta_{42}$ that has already formed by use of specific antibodies.

The *Apo* ε gene on chromosome 19 is involved in the pathogenesis of AD. The protein product, apolipoprotein E, participates in cholesterol transport **(Chap. 400)**, and the gene has three alleles: ε2, ε3, and ε4. The Apo ε4 allele confers increased risk of AD in the general population, including sporadic and late age-of-onset familial forms. Approximately 24–30% of the non-demented white population has at least one ε4 allele (12–15% allele frequency), and about 2% are ε4/ε4 homozygotes. Among patients with AD, 40–65% have at least one ε4 allele, a highly significant elevation compared with controls. The increased risk associated with a single ε4 allele is especially prominent in women. Conversely, many patients with AD have no ε4 allele, and ε4 carriers may never develop AD. Therefore, ε4 is neither necessary nor sufficient to cause AD. Nevertheless, the Apo ε4 allele represents the

most important genetic risk factor for sporadic AD and acts as a dose-dependent disease modifier, with the earliest age of onset associated with the ε4 homozygosity. Precise mechanisms through which Apo ε4 confers AD risk or hastens onset remain unclear, but ε4 leads to less efficient amyloid clearance and to the production of toxic fragments from cleavage of the molecule. Apo ε can be identified in neuritic plaques and may also be involved in neurofibrillary tangle formation, because it binds to tau protein. Apo ε4 decreases neurite outgrowth in dorsal root ganglion neuronal cultures, perhaps indicating a deleterious role in the brain's response to injury. Some evidence suggests that the ε2 allele may reduce AD risk. Use of Apo ε testing in AD diagnosis remains controversial because its predictive value remains unclear and many individuals with the ε4 allele will never develop dementia. Cognitively normal ε4 heterozygotes and homozygotes may show decreased posterior cerebral cortical metabolic function by PET imaging, suggesting presymptomatic abnormalities due to AD or an inherited vulnerability of the AD-targeted network. In demented persons who meet clinical criteria for AD, finding an ε4 allele increases the reliability of diagnosis; however, the absence of an ε4 allele cannot be considered evidence against AD. Nevertheless, Apo ε4 remains the single most important biologic marker associated with AD risk, and studies of ε4's functional role and diagnostic utility are progressing rapidly. The ε4 allele is not associated with risk for FTD, DLB, or CJD, although some evidence suggests that ε4 may worsen the expression of non-AD degenerative disorders, head trauma, and other brain injuries. Additional genes are also likely to be involved in AD, especially as minor risk alleles for sporadic forms of the disease. Genome-wide association studies have implicated the clusterin (*CLU*), phosphatidylinositol-binding clathrin assembly protein (*PICALM*), and complement component (3b/4b) receptor 1 (*CR1*) genes, among others. *CLU* may play a role in synapse turnover, *PICALM* participates in clathrin-mediated endocytosis, and *CR1* may be involved in amyloid clearance or synapse loss through the complement pathway. *TREM2* is a gene involved with inflammation that increases the likelihood of dementia. Homozygous mutation carriers develop a frontal dementia with bone cysts (Nasu-Hakola disease), whereas heterozygotes are predisposed to the development of AD.

TREATMENT

Alzheimer's Disease

The management of AD is challenging and gratifying despite the absence of a cure or a robust pharmacologic treatment. The primary focus is on long-term amelioration of associated behavioral and neurologic problems, as well as providing caregiver support.

Building rapport with the patient, family members, and other caregivers is essential. In the early stages of AD, memory aids such as notebooks and posted daily reminders can be helpful. Family members should emphasize activities that are pleasant while curtailing those that increase stress on the patient. Kitchens, bathrooms, stairways, and bedrooms need to be made safe, and eventually patients will need to stop driving. Loss of independence and change of environment may worsen confusion, agitation, and anger. Communication and repeated calm reassurance are necessary. Caregiver "burnout" is common, often resulting in nursing home placement of the patient or new health problems for the caregiver. Respite breaks for the caregiver help to maintain a successful long-term therapeutic milieu. Use of adult day care centers can be helpful. Local and national support groups, such as the Alzheimer's Association and the Family Caregiver Alliance, are valuable resources. Internet access to these resources has become available to clinicians and families in recent years.

Donepezil (target dose, 10 mg daily), rivastigmine (target dose, 6 mg twice daily or 9.5-mg patch daily), galantamine (target dose 24 mg daily, extended-release), and memantine (target dose, 10 mg twice daily) are approved by the U.S. Food and Drug Administration (FDA) for the treatment of AD. Due to hepatotoxicity, tacrine is no longer used. Dose escalations for each of these medications must be carried out over 4–6 weeks to minimize side effects. The pharmacologic action of donepezil, rivastigmine, and galantamine is inhibition of the cholinesterases, primarily acetylcholinesterase, with a resulting increase in cerebral acetylcholine levels. Memantine appears to act by blocking overexcited *N*-methyl-D-aspartate (NMDA) glutamate receptors. Double-blind, placebo-controlled, crossover studies with cholinesterase inhibitors and memantine in moderate to severe AD have shown them to be associated with improved caregiver ratings of patients' functioning and with an apparent decreased rate of decline in cognitive test scores over periods of up to 3 years. The average patient on an anticholinesterase inhibitor maintains his or her mini-mental state examination (MMSE) score for close to a year, whereas a placebo-treated patient declines 2–3 points over the same time period. Memantine, used in conjunction with cholinesterase inhibitors or by itself, slows cognitive deterioration and decreases caregiver burden for patients with moderate to severe AD, but is not approved for mild AD. Each of these compounds has only modest efficacy for AD. Cholinesterase inhibitors are relatively easy to administer, and their major side effects are gastrointestinal symptoms (nausea, diarrhea, cramps), altered sleep with unpleasant or vivid dreams, bradycardia (usually benign), and muscle cramps.

In a prospective observational study, the use of estrogen replacement therapy appeared to protect—by about 50%—against development of AD in women. This study seemed to confirm the results of two earlier case-controlled studies. Sadly, a prospective placebo-controlled study of a combined estrogen-progesterone therapy for asymptomatic postmenopausal women increased, rather than decreased, the prevalence of dementia. This study markedly dampened enthusiasm for hormonal treatments to prevent dementia. Additionally, no benefit has been found in the treatment of AD with estrogen alone.

A controlled trial of an extract of *Ginkgo biloba* found modest improvement in cognitive function in subjects with AD and vascular dementia. Unfortunately, a comprehensive 6-year multicenter prevention study using ginkgo found no slowing of progression to dementia in the treated group.

Vaccination against $A\beta_{42}$ has proved highly efficacious in mouse models of AD, helping clear brain amyloid and preventing further amyloid accumulation. In human trials, this approach led to life-threatening complications, including meningoencephalitis, in a minority of patients. Another experimental approach to AD treatment has been the use of β and γ secretase inhibitors that diminish the production of $A\beta_{42}$, but the first two placebo-controlled trials of γ secretase inhibitors, tarenflurbil and semagacestat, were negative, and semagacestat may have accelerated cognitive decline compared to placebo. Passive immunization with monoclonal antibodies against $A\beta_{42}$ has been tried in mild to moderate AD. These studies were negative, leading some to suggest that the patients treated were too advanced to respond to amyloid-lowering therapies. Therefore, new trials have started in asymptomatic individuals with mild AD, in asymptomatic autosomal dominant forms of AD, and in cognitively normal elderly who are amyloid positive with PET. Medications that modify tau phosphorylation and aggregation, including tau antibodies, are beginning to be studied as possible treatments for both AD and non-AD tau-related disorders including FTD and progressive supranuclear palsy (PSP) **(Chap. 424)**.

Several retrospective studies suggest that nonsteroidal anti-inflammatory agents and 3-hydroxy-3-methylglutaryl-coenzyme A (HMG-CoA) reductase inhibitors (statins) may have a protective effect on dementia if used prior to the onset of disease but do not influence clinically symptomatic AD. Finally, there is now a strong interest in the relationship between diabetes and AD, and insulin-regulating studies are being conducted.

Mild to moderate depression is common in the early stages of AD and may respond to antidepressants or cholinesterase inhibitors. Selective serotonin reuptake inhibitors (SSRIs) are commonly used due to their low anticholinergic side effects (for example, escitalopram, target dose 5–10 mg daily). Seizures can be treated with levetiracetam unless the patient had a different regimen that

was effective prior to the onset of AD. Agitation, insomnia, hallucinations, and belligerence are especially troublesome characteristics of some AD patients, and these behaviors can lead to nursing home placement. The newer generation of atypical antipsychotics, such as risperidone, quetiapine, and olanzapine, are being used in low doses to treat these neuropsychiatric symptoms. The few controlled studies comparing drugs against behavioral intervention in the treatment of agitation suggest mild efficacy with significant side effects related to sleep, gait, and cardiovascular complications, including an increased risk of death. All antipsychotics carry a black box FDA warning for use in elderly patients with dementia and thus should be prescribed only with caution; however, careful, daily, nonpharmacologic behavior management is often not available, rendering medications necessary for some patients. Finally, medications with strong anticholinergic effects should be vigilantly avoided, including prescription and over-the-counter sleep aids (e.g., diphenhydramine) or incontinence therapies (e.g., oxybutynin).

A general approach to the symptomatic management of dementia is presented in Chap. 25.

OTHER CAUSES OF DEMENTIA

FTD **(Chap. 424)**, vascular dementia **(Chap. 425)**, and DLB **(Chap. 426)**, are covered in dedicated chapters. Additional important causes of dementia are described here.

Prion diseases such as CJD are rare neurodegenerative conditions (prevalence ~1 per million) that produce dementia. CJD is a rapidly progressive disorder associated with dementia, focal cortical signs, rigidity, and myoclonus, causing death <1 year after first symptoms appear. The rapidity of progression seen with CJD is uncommon in AD so that the distinction between the two disorders is usually straightforward. Cortical basal degeneration (CBD) **(Chap. 424)** and DLB **(Chap. 426)**, more rapid degenerative dementias with prominent movement abnormalities, are more likely to be mistaken for CJD. The differential diagnosis for CJD includes other rapidly progressive dementing conditions such as viral or bacterial encephalitides, Hashimoto's encephalopathy, central nervous system (CNS) vasculitis, lymphoma, or paraneoplastic/autoimmune syndromes. The markedly abnormal periodic complexes on EEG and cortical ribboning and basal ganglia hyperintensities on fluid-attenuate inversion recovery MRI are diagnostic features of CJD, although rarely, prolonged focal or generalized seizures can produce a similar imaging appearance. **Prion diseases are discussed in detail in Chap. 430.**

Huntington's disease (HD) **(Chap. 428)** is an autosomal dominant degenerative brain disorder. HD clinical hallmarks include chorea, behavioral disturbance, and executive impairment. Symptoms typically begin in the fourth or fifth decade, but there is a wide range, from childhood to >70 years. Memory is frequently not impaired until late in the disease, but attention, judgment, self-awareness, and executive functions are often deficient at an early stage. Depression, apathy, social withdrawal, irritability, and intermittent disinhibition are common. Delusions and obsessive-compulsive behavior may occur. Disease duration is variable but typically lasts ~15 years.

Normal-pressure hydrocephalus is a relatively uncommon but treatable syndrome. The clinical, physiologic, and neuroimaging characteristics of NPH must be carefully distinguished from those of other dementias associated with gait impairment. Historically, many patients treated for NPH have suffered from other dementias, particularly AD, vascular dementia, DLB, and PSP. For NPH, the clinical triad includes an abnormal gait (ataxic or apractic), dementia (usually mild to moderate, with an emphasis on executive impairment), and urinary urgency or incontinence. Neuroimaging reveals enlarged lateral ventricles (hydrocephalus) with little or no cortical atrophy, although the sylvian fissures may appear propped open (so-called "boxcarring"), which can be mistaken for perisylvian atrophy. Crowding of dorsal frontal-parietal gyri helps distinguish NPH from other movement disorders, such as PSP and CBD, in which dorsal atrophy with sulcal widening is common. NPH is a communicating hydrocephalus with a patent aqueduct of Sylvius **(see Fig. 25-3)**, in contrast to aqueductal stenosis, in which the aqueduct is small. Lumbar puncture opening pressure falls in the high-normal range, and the CSF protein, glucose, and cell counts are normal. NPH may be caused by obstruction to normal CSF flow over the cerebral convexities and delayed resorption into the venous system. The indolent nature of the process results in enlarged lateral ventricles with relatively little increase in CSF pressure. Presumed edema, stretching, and distortion of subfrontal white matter tracts may lead to clinical symptoms, but the precise underlying pathophysiology remains unclear. Some patients provide a history of conditions that produce meningeal scarring (blocking CSF resorption) such as previous meningitis, subarachnoid hemorrhage, or head trauma. Others with long-standing but asymptomatic congenital hydrocephalus may have adult-onset deterioration in gait or memory that is confused with NPH. In contrast to AD, the patient with NPH complains of an early and prominent gait disturbance without cortical atrophy on CT or MRI.

Numerous attempts to improve NPH diagnosis with various special studies and predict the success of ventricular shunting have been undertaken. These tests include radionuclide cisternography (showing a delay in CSF absorption over the convexity) and various efforts to monitor and alter CSF flow dynamics, including a constant-pressure infusion test. None has proven to be specific or consistently useful. A transient improvement in gait or cognition may follow lumbar puncture (or serial punctures) with removal of 30–50 mL of CSF, but this finding has also not proved to be consistently predictive of postshunt improvement. Perhaps the most reliable strategy is a period of close inpatient evaluation before, during, and after lumbar CSF drainage. Occasionally, when a patient with AD presents with gait impairment (at times due to comorbid subfrontal vascular injury) and absent or only mild cortical atrophy on CT or MRI, distinguishing NPH from AD can be challenging. Hippocampal atrophy on MRI favors AD, whereas a characteristic "magnetic" gait with external hip rotation, low foot clearance, and short strides, along with prominent truncal sway or instability, favors NPH. The diagnosis of NPH should be avoided when hydrocephalus is not detected on imaging studies, even if the symptoms otherwise fit. Thirty to fifty percent of patients identified by careful diagnosis as having NPH will improve with ventricular shunting. Gait may improve more than cognition, but many reported failures to improve cognitively may have resulted from comorbid AD. Short-lasting improvement is common. Patients should be carefully selected for shunting, because subdural hematoma, infection, and shunt failure are known complications and can be a cause for early nursing home placement in an elderly patient with previously mild dementia.

Intracranial hypotension, sometimes called sagging brain syndrome, is a disorder caused by low CSF pressure, leading to downward pressure on the subcortical structures and disruption of cerebral function. It presents in a variable manner with headache, often exacerbated by coughing or a Valsalva maneuver or by moving from lying to standing. Other common symptoms include dizziness, vomiting, disruption of sleep-wake cycles, and sometimes a progressive behavioral variant FTD-like syndrome **(Chap. 424)**. Although sometimes idiopathic, this syndrome can be caused by CSF leaks secondary to lumbar puncture, head trauma, or spinal cord arachnoid cysts. Treatment consists of finding and patching CSF leaks.

Dementia can accompany *chronic alcoholism* **(Chap. 445)** and may result from associated malnutrition, especially of B vitamins, particularly thiamine. Other poorly defined aspects of chronic alcoholism may, however, also produce cerebral damage. A rare idiopathic syndrome of dementia and seizures with degeneration of the corpus callosum has been reported primarily in male Italian red wine drinkers (Marchiafava-Bignami disease).

Thiamine (vitamin B_1) deficiency causes Wernicke's encephalopathy **(Chap. 301)**. The clinical presentation is usually a malnourished patient (frequently but not necessarily alcoholic) with confusion, ataxia, and diplopia resulting from inflammation and necrosis of periventricular midline structures, including dorsomedial thalamus, mammillary bodies, midline cerebellum, periaqueductal gray matter, and trochlear

and abducens nuclei. Damage to the dorsomedial thalamus correlates most closely with the memory loss. Prompt administration of parenteral thiamine (100 mg intravenously for 3 days followed by daily oral dosage) may reverse the disease if given within the first days of symptom onset. Prolonged untreated thiamine deficiency can result in an irreversible and profound amnestic syndrome (Korsakoff's syndrome) or even death.

In *Korsakoff's syndrome*, the patient is unable to recall new information despite normal immediate memory, attention span, and level of consciousness. Memory for new events is seriously impaired, whereas knowledge acquired prior to the illness remains relatively intact. Patients are easily confused, disoriented, and cannot store information for more than a few minutes. Superficially, they may be conversant, engaging, and able to perform simple tasks and follow immediate commands. Confabulation is common, although not always present. There is no specific treatment because the previous thiamine deficiency has produced irreversible damage to the medial thalamic nuclei and mammillary bodies. Mammillary body atrophy may be visible on MRI in the chronic phase **(see Fig. 301-6)**.

Vitamin B_{12} deficiency, as can occur in pernicious anemia, causes a megaloblastic anemia and may also damage the nervous system **(Chaps. 95 and 434)**. Neurologically, it most commonly produces a spinal cord syndrome (myelopathy) affecting the posterior columns (loss of vibration and position sense) and corticospinal tracts (hyperactive tendon reflexes with Babinski signs); it also damages peripheral nerves (neuropathy), resulting in sensory loss with depressed tendon reflexes. Damage to myelinated axons may also cause dementia. The mechanism of neurologic damage is unclear but may be related to a deficiency of *S*-adenosyl methionine (required for methylation of myelin phospholipids) due to reduced methionine synthase activity or accumulation of methylmalonate, homocysteine, and propionate, providing abnormal substrates for fatty acid synthesis in myelin. Use of histamine blockers or metformin, vegan diets, autoimmunity against gastric parietal cells, and various causes of malabsorption are the typical causes for vitamin B_{12} deficiency. The neurologic sequelae of vitamin B_{12} deficiency may occur in the absence of hematologic manifestations, making it critical to avoid using the complete blood count (CBC) and blood smear as a substitute for measuring B_{12} blood levels. Treatment with parenteral vitamin B_{12} (1000 µg intramuscularly daily for a week, weekly for a month, and monthly for life for pernicious anemia) stops progression of the disease if instituted promptly, but complete reversal of advanced nervous system damage will not occur.

Deficiency of nicotinic acid (*pellagra*) is associated with skin rash over sun-exposed areas, glossitis, and angular stomatitis **(Chap. 326)**. Severe dietary deficiency of nicotinic acid along with other B vitamins such as pyridoxine may result in spastic paraparesis, peripheral neuropathy, fatigue, irritability, and dementia. This syndrome has been seen in prisoners of war and in concentration camps but should be considered in any malnourished individual. Low serum folate levels appear to be a rough index of malnutrition, but isolated folate deficiency has not been proved as a specific cause of dementia.

CNS infections usually cause delirium and other acute neurologic syndromes. However, some chronic CNS infections, particularly those associated with chronic meningitis **(Chap. 134)**, may produce a dementing illness. The possibility of chronic infectious meningitis should be suspected in patients presenting with a dementia or behavioral syndrome, who also have headache, meningismus, cranial neuropathy, and/or radiculopathy. Between 20 and 30% of patients in the advanced stages of HIV infection become demented **(Chap. 197)**. Cardinal features include psychomotor retardation, apathy, and impaired memory. This syndrome may result from secondary opportunistic infections but can also be caused by direct infection of CNS neurons with HIV. Neurosyphilis **(Chap. 177)** was a common cause of dementia in the preantibiotic era; it is now uncommon but can still be encountered in patients with multiple sex partners, particularly among patients with HIV. Characteristic CSF changes consist of pleocytosis, increased protein, and a positive Venereal Disease Research Laboratory (VDRL) test.

Primary and metastatic *neoplasms of the CNS* **(Chap. 86)** usually produce focal neurologic findings and seizures rather than dementia, but if tumor growth begins in the frontal or temporal lobes, the initial manifestations may be memory loss or behavioral changes. A paraneoplastic syndrome of dementia associated with occult carcinoma (often small-cell lung cancer) is termed *limbic encephalitis*. In this syndrome, confusion, agitation, seizures, poor memory, emotional changes, and frank dementia may occur. Paraneoplastic *encephalitis associated with NMDA receptor antibodies* presents as a progressive psychiatric disorder with memory loss and seizures; affected patients are often young women with ovarian teratoma **(Chap. 90)**.

A *nonconvulsive seizure disorder* **(Chap. 418)** may underlie a syndrome of confusion, clouding of consciousness, and garbled speech. Often, psychiatric disease is suspected, but an EEG demonstrates the epileptic nature of the illness. If recurrent or persistent, the condition may be termed *complex partial status epilepticus*. The cognitive disturbance often responds to anticonvulsant therapy. The etiology may be previous small strokes or head trauma; some cases are idiopathic. Nonconvulsive temporal lobe seizures can also emerge early in the course of AD.

It is important to recognize *systemic diseases* that indirectly affect the brain and produce chronic confusion or dementia. Such conditions include hypothyroidism; vasculitis; and hepatic, renal, or pulmonary disease. Hepatic encephalopathy may begin with irritability and confusion and slowly progress to agitation, lethargy, and coma.

Isolated vasculitis of the CNS (CNS granulomatous angiitis) **(Chaps. 356 and 419)** occasionally causes a chronic encephalopathy associated with confusion, disorientation, and clouding of consciousness. Headache is common, and strokes and cranial neuropathies may occur. Brain imaging studies may be normal or nonspecifically abnormal. CSF analysis reveals a mild pleocytosis or protein elevation. Cerebral angiography can show multifocal stenoses involving medium-caliber vessels, but some patients have only small-vessel disease that is not revealed on angiography. The angiographic appearance is not specific and may be mimicked by atherosclerosis, infection, or other causes of vascular disease. Brain or meningeal biopsy demonstrates endothelial cell proliferation and mononuclear infiltrates within blood vessel walls. The prognosis is often poor, although the disorder may remit spontaneously. Some patients respond to glucocorticoids or chemotherapy.

Chronic metal exposure represents a rare cause of dementia. The key to diagnosis is to elicit a history of exposure at work or home. Chronic lead poisoning from inadequately fire-glazed pottery has been reported. Fatigue, depression, and confusion may be associated with episodic abdominal pain and peripheral neuropathy. Gray lead lines appear in the gums, usually accompanied by an anemia with basophilic stippling of red blood cells. The clinical presentation can resemble that of acute intermittent porphyria, including elevated levels of urine porphyrins as a result of the inhibition of δ-aminolevulinic acid dehydrase. The treatment is chelation therapy with agents such as ethylenediamine tetraacetic acid (EDTA). Chronic mercury poisoning produces dementia, peripheral neuropathy, ataxia, and tremulousness that may progress to a cerebellar intention tremor or choreoathetosis. The confusion and memory loss of chronic arsenic intoxication is also associated with nausea, weight loss, peripheral neuropathy, pigmentation and scaling of the skin, and transverse white lines of the fingernails (Mees' lines). Treatment is chelation therapy with dimercaprol (BAL). Aluminum poisoning is rare but was documented with the dialysis dementia syndrome, in which water used during renal dialysis was contaminated with excessive amounts of aluminum. This poisoning resulted in a progressive encephalopathy associated with confusion, nonfluent aphasia, memory loss, agitation, and, later, lethargy and stupor. Speech arrest and myoclonic jerks were common and associated with severe and generalized EEG changes. The condition has been eliminated by the use of deionized water for dialysis.

Recurrent head trauma in professional athletes may lead to a dementia previously referred to as "punch-drunk" syndrome or *dementia pugilistica* but now known as chronic traumatic encephalopathy

(CTE) to signify its relevance to contact sport athletes other than boxers **(Chap. 435)**. The symptoms can be progressive, beginning late in an athlete's career or, more often, after retirement. Early in the course, a personality change associated with social instability and sometimes paranoia and delusions occurs. Later, memory loss progresses to full-blown dementia, often associated with parkinsonian signs and ataxia or intention tremor. At autopsy, the cerebral cortex shows tau-immunoreactive NFTs that are more prominent than amyloid plaques (which are usually diffuse or absent rather than neuritic). NFTs and tau-positive reactive astrocytes are often clustered in the depths of cortical sulci and in a perivascular distribution. TDP-43 inclusions **(Chap. 424)** have also been reported, highlighting the overlap with the FTD spectrum. Loss of neurons in the substantia nigra is a variable feature, and some with TDP-43 inclusions also develop motor neuron disease (MND) **(Chap. 429)**.

Chronic subdural hematoma **(Chap. 435)** is also occasionally associated with dementia, often in the context of underlying cortical atrophy from conditions such as AD or HD.

Transient global amnesia (TGA) is characterized by the sudden onset of a severe episodic memory deficit, usually occurring in persons aged >50 years. Often the amnesia occurs in the setting of an emotional stimulus or physical exertion. During the attack, the individual is alert and communicative, general cognition seems intact, and there are no other neurologic signs or symptoms. The patient may seem confused and repeatedly ask about his or her location in place and time. The ability to form new memories returns after a period of hours, and the individual returns to normal with no recall for the period of the attack. Frequently no cause is determined, but cerebrovascular disease, epilepsy (7% in one study), migraine, or cardiac arrhythmias have all been implicated. Approximately one-quarter of patients experience recurrent attacks. Rare instances of permanent memory loss have been reported in patients with TGA-like spells, usually representing ischemic infarction of the hippocampus or dorsomedial thalamic nucleus bilaterally. Seizure activity due to AD should always be suspected in this syndrome.

The *ALS/parkinsonian/dementia complex of Guam* is a rare degenerative disease that has occurred in the Chamorro natives on the island of Guam. Individuals may have any combination of parkinsonian features, dementia, and MND. The most characteristic pathologic features are the presence of NFTs in degenerating neurons of the cortex and substantia nigra and loss of motor neurons in the spinal cord, although recent reanalysis has shown that some patients with this illness also show coexisting TDP-43 pathology. Epidemiologic evidence supports a possible environmental cause, such as exposure to a neurotoxin or an infectious agent with a long latency period. One interesting but unproven candidate neurotoxin is the seed of the false palm tree, which Guamanians traditionally used to make flour. The amyotrophic lateral sclerosis (ALS) syndrome is no longer present in Guam, but a dementing illness with rigidity continues to be seen.

Rarely, adult-onset leukodystrophies, lysosomal storage diseases, and other genetic disorders can present as a dementia in middle to late life. Metachromatic leukodystrophy (MLD) causes a progressive psychiatric or dementia syndrome associated with an extensive, confluent frontal white matter abnormality. MLD is diagnosed by measuring reduced arylsulfatase A enzyme activity in peripheral white blood cells. Adult-onset presentations of adrenoleukodystrophy have been reported in female carriers, and these patients often feature spinal cord and posterior white matter involvement. Adrenoleukodystrophy is diagnosed by demonstrating increased levels of plasma very-long-chain fatty acids. CADASIL is another genetic syndrome associated with white matter disease, often frontally and temporally predominant. Diagnosis is made with skin biopsy, which shows osmophilic granules in arterioles, or, increasingly, through genetic testing for mutations in Notch 3. The neuronal ceroid lipofuscinoses are a genetically heterogeneous group of disorders associated with myoclonus, seizures, vision loss, and progressive dementia. Diagnosis is made by finding eosinophilic curvilinear inclusions within white blood cells or neuronal tissue.

Psychogenic amnesia for personally important memories can be seen. Whether this results from deliberate avoidance of unpleasant memories, outright malingering, or unconscious repression remains unknown and probably depends on the patient. Event-specific amnesia is more likely to occur after violent crimes such as homicide of a close relative or friend or sexual abuse. It may develop in association with severe drug or alcohol intoxication and sometimes with schizophrenia. More prolonged psychogenic amnesia occurs in fugue states that also commonly follow severe emotional stress. The patient with a fugue state suffers from a sudden loss of personal identity and may be found wandering far from home. *In contrast to neurologic amnesia, fugue states are associated with amnesia for personal identity and events closely associated with the personal past.* At the same time, memory for other recent events and the ability to learn and use new information are preserved. The episodes usually last hours or days and occasionally weeks or months while the patient takes on a new identity. On recovery, there is a residual amnesia gap for the period of the fugue. Very rarely does selective loss of autobiographic information reflect a focal injury to the brain areas involved with these functions.

Psychiatric diseases may mimic dementia. Severely depressed or anxious individuals may appear demented, a phenomenon sometimes called *pseudodementia*. Memory and language are usually intact when carefully tested, and a significant memory disturbance usually suggests an underlying dementia, even if the patient is depressed. Patients in this condition may feel confused and unable to accomplish routine tasks. Vegetative symptoms, such as insomnia, lack of energy, poor appetite, and concern with bowel function, are common. Onset is often more abrupt, and the psychosocial milieu may suggest prominent reasons for depression. Such patients respond to treatment of the underlying psychiatric illness. Schizophrenia is usually not difficult to distinguish from dementia, but occasionally the distinction can be problematic. Schizophrenia generally has a much earlier age of onset (second and third decades) than most dementing illnesses and is associated with intact memory. The delusions and hallucinations of schizophrenia are usually more complex, bizarre, and threatening than those of dementia. Some chronic schizophrenics develop an unexplained progressive dementia late in life that is not related to AD. Conversely, FTD, HD, vascular dementia, DLB, AD, or leukoencephalopathy can begin with schizophrenia-like features, leading to the misdiagnosis of a psychiatric condition. Later age of onset, significant deficits on cognitive testing, or the presence of abnormal neuroimaging suggest a degenerative condition. Memory loss may also be part of a *conversion disorder*. In this situation, patients commonly complain bitterly of memory loss, but careful cognitive testing either does not confirm the deficits or demonstrates inconsistent or unusual patterns of cognitive problems. The patient's behavior and "wrong" answers to questions often indicate that he or she understands the question and knows the correct answer.

Clouding of cognition by *chronic drug or medication use*, often prescribed by physicians, is an important cause of dementia. Sedatives, tranquilizers, and analgesics used to treat insomnia, pain, anxiety, or agitation may cause confusion, memory loss, and lethargy, especially in the elderly. Discontinuation of the offending medication often improves mentation.

FURTHER READING

Braak H, Del Tredici K: Where, when, and in what form does sporadic Alzheimer's disease begin? Curr Opin Neurol 25:708, 2012.

Cohen AD, Klunk WE: Early detection of Alzheimer's disease using PiB and FDG PET. Neurobiol Dis 72 Pt A:117, 2014.

Haas C: Strategies, development, and pitfalls of therapeutic options for Alzheimer's disease. J Alzheimers Dis 28:241, 2012.

Musiek ES, Holtzman DM: Origins of Alzheimer's disease: Reconciling cerebrospinal fluid biomarker and neuropathology data regarding the temporal sequence of amyloid-beta and tau involvement. Curr Opin Neurol 25:715, 2012.

Selkoe DJ, Hardy J: The amyloid hypothesis of Alzheimer's disease at 25 years. EMBO Mol Med 8:595, 2016.

424 Frontotemporal Dementia

William W. Seeley, Bruce L. Miller

Frontotemporal dementia (FTD) refers to a group of clinical syndromes united by their links to underlying frontotemporal lobar degeneration (FTLD) pathology. FTD most often begins in the fifth to seventh decades and is nearly as prevalent as AD in this age group. Early studies suggested that FTD may be more common in men than women, however more recent reports cast doubt on this finding. Although a family history of dementia is common, autosomal dominant inheritance is seen in only 10–20% of all FTD cases.

CLINICAL MANIFESTATIONS

The clinical heterogeneity seen in both familial and sporadic forms of FTD is remarkable. Three core clinical syndromes have been described **(Fig. 424-1)**. In the behavioral variant (bvFTD), the most common FTD syndrome, social and emotional systems dysfunction manifests as apathy, disinhibition, compulsivity, loss of empathy, and overeating, often but not always accompanied by deficits in executive control. Two forms of primary progressive aphasia (PPA), the semantic and nonfluent/agrammatic variants, are commonly due to FTLD and included under the FTD umbrella. In the semantic variant, patients slowly lose the ability to decode word, object, person-specific, and emotion meaning, whereas patients with the nonfluent/agrammatic variant develop profound inability to produce words, often with prominent motor speech impairment. Any of these three clinical syndromes, but most often bvFTD, may be accompanied by motor neuron disease (MND) **(Chap. 429)**, in which case the term FTD-MND is applied. In addition, the corticobasal syndrome (CBS) and progressive supranuclear palsy syndrome (PSP-S) can be considered part of the FTLD clinical spectrum. Furthermore, patients may evolve from any of the major syndromes described above to have prominent features of another syndrome.

Findings at the bedside are dictated by the anatomic localization of the disorder. Medial and orbital frontal and anterior insula degeneration predicts bvFTD. Patients with nonfluent/agrammatic PPA show dominant hemisphere lateral frontal and precentral gyrus atrophy. Anterior temporal degeneration presents with semantic variant PPA. Parietal functions such as visuospatial processing and arithmetic may remain normal late into any FTD syndrome. Many patients with nonfluent aphasia or bvFTD later develop aspects of PSP-S, as disease spreads into subcortical or brainstem structures, or CBS-like features, as disease moves into peri-rolandic cortices.

GENETIC CONSIDERATIONS

The most common autosomal dominantly inherited mutations causing FTD involve the *C9ORF72* (chromosome 9), *GRN* (chromosome 17), and *MAPT* (chromosome 17) genes. Hexanucleotide (GGGGCC) expansions in a noncoding exon of *C9ORF72* are the most recently identified and represent the most common genetic cause of familial or sporadic FTD (usually presenting as bvFTD with or without MND) and amyotrophic lateral sclerosis (ALS). The expansion is associated with *C9ORF72* haploinsuffiency, nuclear mRNA foci containing transcribed portions of the expansion and other mRNAs, neuronal cytoplasmic inclusions containing dipeptide repeat proteins translated from the repeat mRNA, and transactive response DNA-binding protein of 43 kDa (TDP-43) neuronal cytoplasmic and glial inclusions. The pathogenic significance of these various features is a topic of vigorous investigation. *MAPT* mutations lead to a change in the alternate splicing of tau or cause loss of function in the tau molecule, thereby altering microtubule binding. With *GRN*, mutations in the coding sequence of the gene encoding progranulin protein result in mRNA degradation due to nonsense-mediated decay, leading to a ~50% reduction in circulating progranulin protein levels. Intriguingly, homozygous *GRN* mutations were recently reported to cause neuronal ceroid lipofuscinosis, focusing investigators on the lysosome as a site of molecular dysfunction in *GRN*-related FTD. Progranulin is a growth factor that binds to tumor necrosis factor (TNF) receptors and participates in tissue repair and tumor growth. How progranulin mutations lead to FTD remains unknown, but the most likely mechanisms include lysosomal dysfunction and enhanced neuroinflammation. Both *MAPT* and *GRN* mutations can be associated with parkinsonian features, whereas ALS is rare. Infrequently, mutations in the valosin-containing protein (*VCP*, chromosome 9), TANK binding kinase 1 (TBK-1), T cell-restricted intracellular antigen-1 (TIA1), and charged multivesicular body protein 2b (*CHMP2b*, chromosome 3) genes also lead to autosomal dominant familial FTD. Mutations in the *TARDBP* (encoding TDP-43) and *FUS* (encoding fused in sarcoma [FUS]) genes (see below) cause familial ALS, sometimes in association with an FTD syndrome, although a few patients presenting with FTD alone have been reported.

NEUROPATHOLOGY

The gross pathologic hallmark of FTLD is a focal atrophy of frontal, insular, and/or temporal cortex, which can be visualized with neuroimaging studies (Fig. 424-1) and is often profound at autopsy. Despite the appearance of advanced disease, however, imaging studies suggest that atrophy often begins focally in one hemisphere before spreading to anatomically interconnected cortical and subcortical regions. Loss of cortical serotonergic innervation is seen in many patients. In contrast to AD, the cholinergic system is relatively spared in FTD, which accounts for the poor efficacy of acetylcholinesterase inhibitors in this group.

Although early studies suggested that 15–30% of patients with FTD showed underlying AD at autopsy, progressive refinement in clinical diagnosis has improved pathologic prediction accuracy, and most patients diagnosed with FTD at a dementia clinic with expertise in FTD will show underlying FTLD pathology. Microscopic findings seen across all patients with FTLD include gliosis, microvacuolation, and neuronal loss, but the disease is subtyped according to the protein composition of neuronal and glial inclusions, which contain either tau or TDP-43 in ~90% of patients, with the remaining ~10% showing inclusions containing FUS **(Fig. 424-2)**.

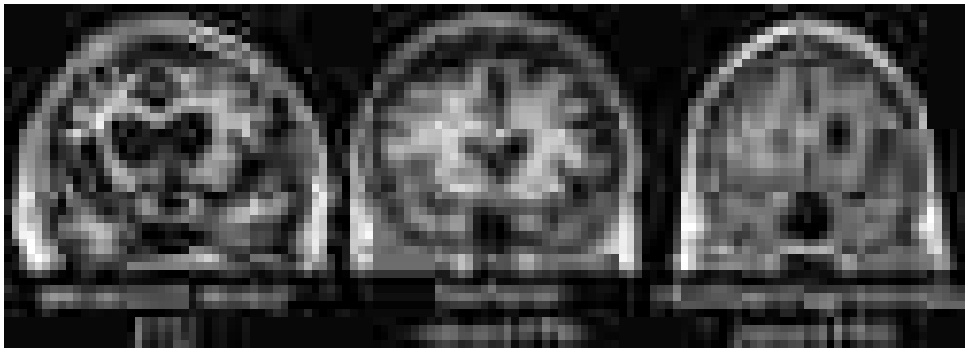


FIGURE 424-1 Three major frontotemporal dementia (FTD) clinical syndromes. Coronal magnetic resonance imaging sections from representative patients with behavioral variant FTD (*left*) and the semantic (*center*) and nonfluent/agrammatic (*right*) variants of primary progressive aphasia (PPA). Areas of early and severe atrophy in each syndrome are highlighted (*white arrowheads*). The behavioral variant features anterior cingulate and frontoinsular atrophy, spreading to orbital and dorsolateral prefrontal cortex. Semantic variant PPA shows prominent temporopolar atrophy, more often on the left. Nonfluent/agrammatic variant PPA is associated with dominant frontal opercular and dorsal insula degeneration.

PATHOGENESIS

The toxicity and spreading capacity of misfolded tau underlies the pathogenesis of many familial cases and is emerging as a key factor in sporadic tauopathies, although loss of tau microtubule stabilizing function may also play a role. TDP-43 and FUS, in contrast, are RNA/DNA binding proteins whose roles in neuronal function are still being actively investigated, but one key role may be the chaperoning of mRNAs to the distal neuron for activity-dependent

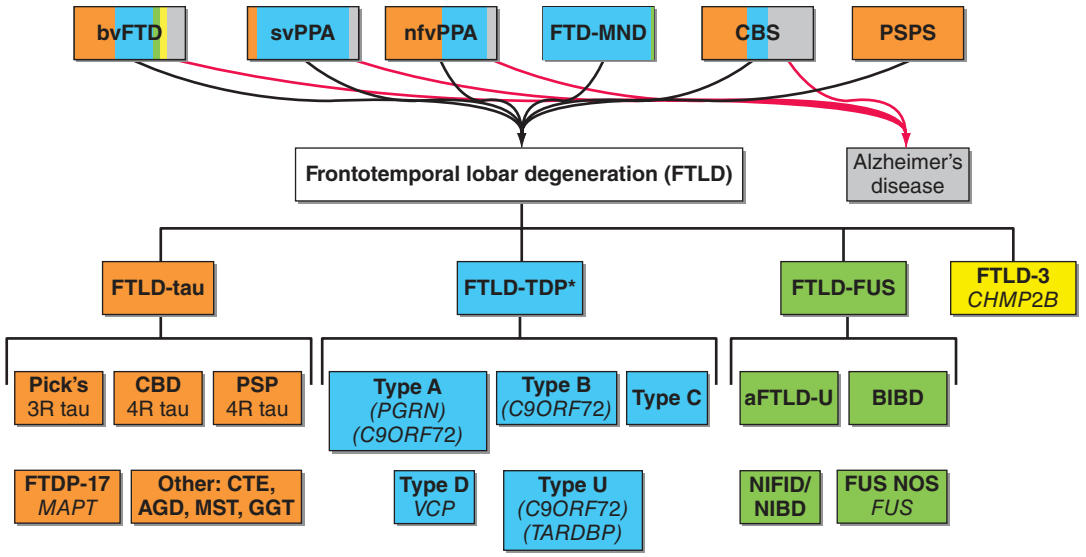


FIGURE 424-2 Frontotemporal dementia syndromes are united by underlying frontotemporal lobar degeneration pathology, which can be divided according to the presence of tau, TDP-43, or FUS-containing inclusions in neurons and glia. Correlations between clinical syndromes and major molecular classes are shown with colored shading. Despite improvements in clinical syndromic diagnosis, a small percentage of patients with some frontotemporal dementia syndromes will show Alzheimer's disease neuropathology at autopsy (*gray shading*). aFTLD-U, atypical frontotemporal lobar degeneration with ubiquitin-positive inclusions; AGD, argyrophilic grain disease; BIBD, basophilic inclusion body disease; bvFTD, behavioral variant frontotemporal dementia; CBD, corticobasal degeneration; CBS, corticobasal syndrome; CTE, chronic traumatic encephalopathy; FTD-MND, frontotemporal dementia with motor neuron disease; FTDP-17, frontotemporal dementia with parkinsonism linked to chromosome 17; FUS, fused in sarcoma; GGT, globular glial tauopathy; MST, multisystem tauopathy; nfvPPA, nonfluent/agrammatic variant primary progressive aphasia; NIBD, neurofilament inclusion body disease; NIFID, neuronal intermediate filament inclusion disease; PSP, progressive supranuclear palsy; PSP-S, progressive supranuclear palsy syndrome; svPPA, semantic variant primary progressive aphasia; Type U, unclassifiable type.

translation within dendritic spines. Because these proteins also form intracellular aggregates and produce similar anatomic progression, protein toxicity and spreading may also factor heavily in the pathogenesis of FTLD-TDP and FTLD-FUS.

Increasingly, misfolded proteins in neurodegenerative disease are being recognized as having "prion-like" properties in that they can template the misfolding of their natively folded (or unfolded) protein counterparts, a process that creates exponential amplification of protein misfolding within a cell and may promote transcellular and even transsynaptic protein propagation between cells. This hypothesis could provide a unifying explanation for the stereotypical patterns of disease spread observed in each syndrome **(Chap. 417)**.

Although the term *Pick's disease* was once used to describe a progressive degenerative disorder characterized by selective involvement of the anterior frontal and temporal neocortex and pathologically by intraneuronal cytoplasmic inclusions (*Pick bodies*), it is now used only in reference to a specific FTLD-tau histopathologic entity. Classical Pick bodies are argyrophilic, staining positively with the Bielschowsky silver method (but not with the Gallyas method) and also with immunostaining for hyperphosphorylated tau. Recognition of the three FTLD major molecular classes has allowed delineation of distinct FTLD subtypes within each class. These subtypes, based on the morphology and distribution of the neuronal and glial inclusions **(Fig. 424-3)**, account for the vast majority of patients, and some subtypes show strong clinical or genetic associations (Fig. 424-2). Despite this progress, clinical features do not allow reliable prediction of the underlying FTLD subtype, or even the major molecular class, for all clinical syndromes. Molecular PET imaging with ligands chosen to bind misfolded tau protein shows great promise and is already being applied to the study of patients with AD and FTD. Because FTLD-tau and FTLD-TDP account for 90% of FTLD patients, the ability to detect pathological tau protein deposition in vivo would greatly improve prediction accuracy, especially when amyloid PET imaging is negative.

TREATMENT

The burden on caregivers of patients with FTD is extremely high, especially when the illness disrupts core emotional and personality functions of the loved one. Treatment is symptomatic, and there are currently no therapies known to slow progression or improve symptoms. Many of the behaviors that may accompany FTD, such as depression, hyperorality, compulsions, and irritability, can be ameliorated with antidepressants, especially SSRIs. Because FTD is often accompanied by parkinsonism, antipsychotics, which can exacerbate this problem, must be used with caution. **A general approach to the symptomatic management of dementia is presented in Chap. 25.**

PROGRESSIVE SUPRANUCLEAR PALSY SYNDROME

PSP-S (also known as Richardson syndrome) is a degenerative disorder that involves the brainstem, basal ganglia, diencephalon, and selected areas of cortex. Clinically, PSP-S begins with falls and executive or subtle personality changes (such as mental rigidity, impulsivity, or apathy). Shortly thereafter, a progressive oculomotor syndrome ensues that begins with square wave jerks, followed by slowed saccades (vertical worse than horizontal) before resulting in progressive supranuclear ophthalmoparesis. Dysarthria, dysphagia, and symmetric axial rigidity can be prominent features that emerge at any point in the illness. A stiff, unstable posture with hyperextension of the neck and a slow, jerky, toppling gait are characteristic. Frequent unexplained and sometimes spectacular falls are common secondary to a combination of axial rigidity, inability to look down, and poor judgment. Even once patients have severely limited voluntary eye movements, they retain oculocephalic reflexes (demonstrated using a vertical doll's head maneuver); thus, the oculomotor disorder is supranuclear. The dementia overlaps with bvFTD, featuring apathy, frontal-executive dysfunction, poor judgment, slowed thought processes, impaired verbal fluency, and difficulty with sequential actions and with shifting from one task to another. These features are common at presentation and often precede the motor syndrome. Some patients with a pathologic diagnosis of PSP begin with a nonfluent aphasia or motor speech disorder and progress to classical PSP-S. Response to L-dopa is limited or absent; no other treatments exist. Death occurs within 5–10 years of onset. Like Pick's disease, increasingly the term PSP is used to refer to a specific histopathologic entity within the FTLD-tau class. In PSP, accumulation of hyperphosphorylated 4-repeat tau is seen within neurons and glia. Tau neuronal inclusions often appear tangle-like and may be large, spherical ("globose") and coarse in subcortical structures. The most prominent involvement is in the subthalamic nucleus, globus pallidus, substantia nigra, locus coeruleus, periaqueductal gray, tectum,

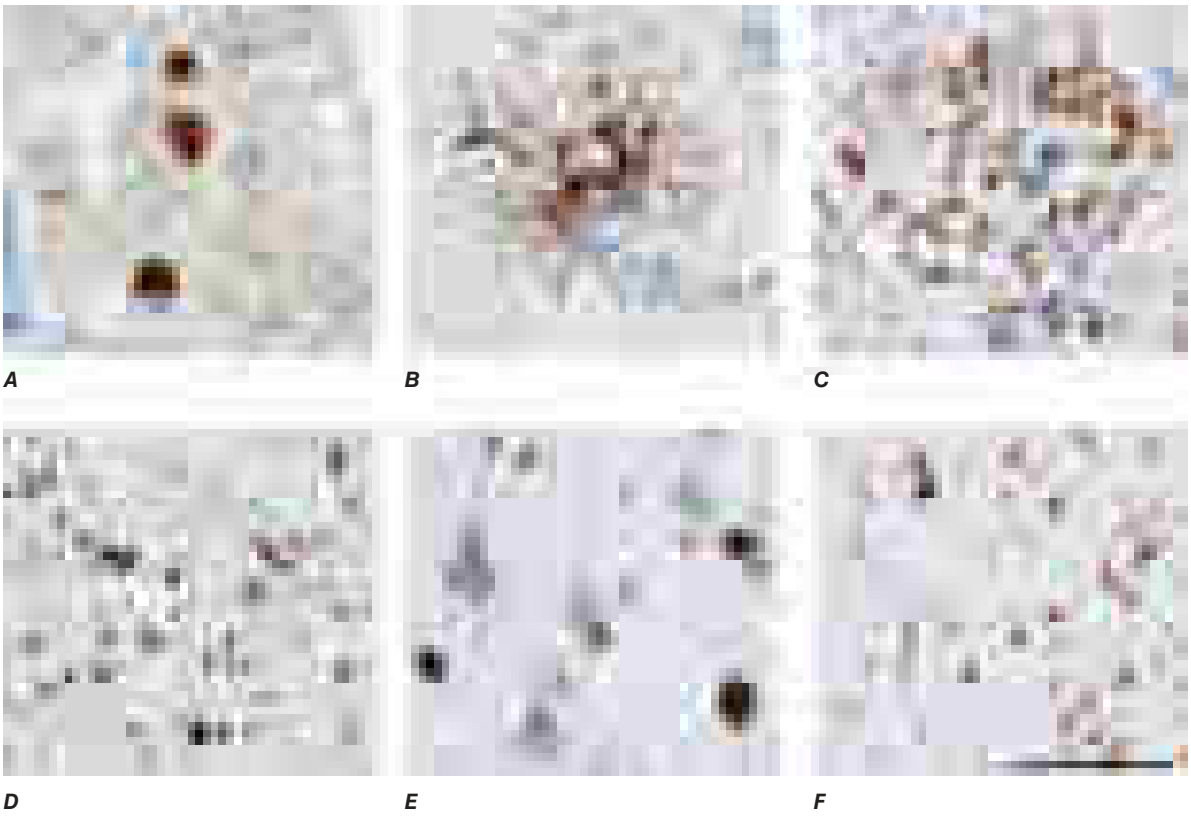


FIGURE 424-3 Neuropathology in frontotemporal lobar degeneration (FTLD). FTLD-tau (***A–C***) and FTLD-TDP (***D–F***) account for >90% of patients with FTLD, and immunohistochemistry reveals characteristic lesions in each of the major histopathologic subtypes within each class: ***A.*** Pick bodies in Pick's disease; ***B.*** a tufted astrocyte in progressive supranuclear palsy; ***C.*** an astrocytic plaque in corticobasal degeneration; ***D.*** small compact or crescentic neuronal cytoplasmic inclusions and short, thin neuropil threads in FTLD-TDP, type A; ***E.*** diffuse/granular neuronal cytoplasmic inclusions (with a relative paucity of neuropil threads) in FTLD-TDP, type B; and ***F.*** long, tortuous dystrophic neurites in FTLD-TDP, type C. TDP can be seen within the nucleus in neurons lacking inclusions but mislocalizes to the cytoplasm and forms inclusions in FTLD-TDP. Immunostains are 3-repeat tau (***A***), phospho-tau (***B*** and ***C***), and TDP-43 (***D–F***). Sections are counterstained with hematoxylin. Scale bar applies to all panels and represents 50 μm in ***A, B, C***, and ***E*** and 100 μm in ***D*** and ***F***.

oculomotor nuclei, and dentate nucleus of cerebellum. Neocortical tangle-like inclusions, like those in AD, often take on a more flame-shaped morphology, but on electron microscopy, PSP tangles can be shown to consist of straight tubules rather than the paired helical filaments found in AD. Furthermore, PSP is associated with prominent tau-positive glial inclusions, such as tufted astrocytes (Fig. 424-3), coiled oligodendroglial inclusions ("coiled bodies"), or, least often, thorny astrocytes. Most patients with PSP-S show PSP at autopsy, although small numbers will show another tauopathy (corticobasal degeneration [CBD] or Pick's disease; Fig. 424-2).

In addition to its overlap with FTD and CBS (see below), PSP is often confused with idiopathic *Parkinson's disease* (PD). Although elderly patients with PD may have restricted upgaze, they do not develop downgaze paresis or other abnormalities of voluntary eye movements typical of PSP. Dementia occurs in ~20% of patients with PD, often due to the emergence of a full-blown DLB-like syndrome. Furthermore, the behavioral syndromes seen with DLB differ from PSP (see below). Dementia in PD becomes more likely with increasing age, increasing severity of extrapyramidal signs, long disease duration, and the presence of depression. Patients with PD who develop dementia also show cortical atrophy on brain imaging. Neuropathologically, there may be AD-related changes in the cortex or LBD-related α-synuclein inclusions in both the limbic system and cerebral cortex. **PD is discussed in detail in Chap. 427.**

CORTICOBASAL SYNDROME

CBS is a slowly progressive dementia-movement disorder associated with severe degeneration in perirolandic cortex and basal ganglia (substantia nigra and striatopallidum). Patients typically present with asymmetric onset of rigidity, dystonia, myoclonus, and apraxia of one limb, at times associated with *alien limb* phenomena in which the limb exhibits unintended motor actions such as grasping, groping, drifting, or undoing. Eventually CBS becomes bilateral and leads to dysarthria, slow gait, action tremor, and a frontal-predominant dementia. Whereas CBS refers to the clinical syndrome, CBD refers to a specific histopathological FTLD-tau entity (Fig. 424-2). Although CBS was once thought to be pathognomonic for CBD, increasingly it has been recognized that CBS can be due to CBD, PSP, FTLD-TDP, or even AD, which accounts for up to 30% of CBS in some series. In CBD, the microscopic features include ballooned, achromatic, tau-positive neurons; astrocytic plaques (Fig. 424-3); and other dystrophic glial tau pathomorphologies that overlap with those seen in PSP. Most specifically, CBD features a severe tauopathy burden in the subcortical white matter, consisting of axonal threads and oligodendroglial coiled bodies. As shown in Fig. 424-2, patients with bvFTD, nonfluent/agrammatic PPA, and PSP-S may also show CBD at autopsy, emphasizing the importance of distinguishing clinical and pathologic constructs and terminology. Treatment of CBS remains symptomatic; no disease-modifying therapies are available.

FURTHER READING

Irwin DJ et al: Frontotemporal lobar degeneration: Defining phenotypic diversity through personalized medicine. Acta Neuropathol 129:469, 2015.

Mackenzie I et al: Nomenclature and nosology for neuropathologic subtypes of frontotemporal lobar degeneration: An update. Acta Neuropathol 119:1, 2010.

Olney NT et al: Frontotemporal dementia. Neurol Clin 35:339, 2017.

Onyike CU, Diehl-Schmid J: The epidemiology of frontotemporal dementia. Int Rev Psychiatry 25:130, 2013.

Roberson ED: Mouse models of frontotemporal dementia. Ann Neurol 72:837, 2012.

425 Vascular Dementia

Joel Kramer, William W. Seeley, Bruce L. Miller

Vascular cognitive impairment and vascular dementia (VCI-VaD) denote deficits in cognition and behavior, along a spectrum of severity, that are associated with cerebrovascular disease (CVD). A dementia syndrome results when CVD is severe enough to cause significant deficits in occupational, social, or functional abilities. VaD is among the most common causes of dementia in the elderly, although its prevalence is disputed. Vascular disease can disrupt structural cognitive networks with lesions such as microinfarcts, microbleeds, macroinfarcts, large hemorrhages, and chronic progressive white matter degeneration, as well as altered cerebral hemodynamics, such as hypoperfusion, disrupted cerebrovascular autoregulation **(Chap. 301)** neurovascular decoupling (loss of normal hemodynamic responses to neural activity), and blood brain barrier dysfunction. The pathophysiological underpinnings of VCI-VaD remain an active area of research.

Age remains the strongest risk factor for CVD and stroke. By the age of 70, 70% of the population has white matter disease and lesions on neuroimaging, with small infarcts (lacunar infarcts) found in 11–24% of the population. In addition to genetic predisposition, risk factors that directly contribute to CVD include chronic hypertension, hyperlipidemia, diabetes, and smoking. Cardiac disease, such as atrial fibrillation or heart failure, can also cause cognitive impairment via embolic infarcts and hypoxemia due to inadequate cerebral blood flow.

A review of data from across the globe indicates good evidence for variability in CVD and stroke risk. Intracranial atherosclerosis, for example, is higher in Asians, Hispanics, and American blacks than it is in European and American whites, while whites may have more extracranial disease. The causes of these disparities remain under investigation, but likely include genetics, lifestyle, and access to health care.

VaD is strongly associated with hemorrhagic and ischemic strokes, with an estimated one-third of stroke survivors affected by post-stroke dementia or cognitive impairment. Hemorrhages, including subdural, intracerebral and subarachnoid bleeds, account for roughly 20% of all strokes. The disruption of cerebral networks caused by hemorrhage depends to a certain extent on size and location. Subarachnoid hemorrhage (SAH) has a more intricate relation with cognitive deficits. For instance, a history of SAH can triple the lifetime risk of developing a dementia syndrome; the molecular underpinnings of this observation are being actively studied. Of note, hemorrhagic strokes may occur as a result of vessel wall damage and inflammation associated with cerebral amyloid angiopathy (CAA). The build-up of β-amyloid protein in cerebral blood vessels that increases their susceptibility to rupture. Although CAA is frequently present in patients with Alzheimer's disease (AD) **(Chap. 423)**, it can also be found in the absence of neocortical amyloid plaques or in individuals with specific genetic predispositions and lead to cognitive impairment in the absence of AD.

Ischemic strokes compose 80% of all strokes. Large vessel and small vessel disease (SVD) can lead to dementia, although the mechanisms and clinical presentation vary. In a cross-sectional study of 706 VaD cases, large vessel disease, often referred to as multi-infarct dementia, made up roughly 18% of all cases. Neurobehavioral symptoms vary as a function of cerebral lesion size and location, and can include aphasia, apraxia, agnosia, and inattention syndromes. Some ambiguity between lesion location and cognitive symptoms continues to persist. This may be a result of the interconnected nature of cognitive, behavioral functional, and structural networks of the brain, as well as the remote consequences of cerebral inflammation and blood-brain barrier disruption caused by strokes.

Of all CVD subtypes, chronic cerebral SVD shows the strongest association with cognitive impairment. SVD accounts for 36–67% of all VaDs. Nonetheless, the relationship between SVD and dementia remains complex and the neuropsychology of dementia due to SVD remains controversial. SVD typically causes occlusion of the deep-penetrating arterioles and disease of draining venules, causing damage to subcortical structures such as the basal ganglia, thalami, and white matter tracts. In addition, cerebral microinfarcts, first observed microscopically in autopsy series, then visualized on 7T MRI, have also showed strong associations with cognitive impairment. Larger lesions, referred to as lacunar infarcts, can result in either a stepwise or gradual decline in cognition sometimes referred to as etat lacunaire. Lacunes average 2 mm in volume, but can range from 0.2 to 15 mm.

Although most VCI-VaD cases caused by SVD are sporadic, there are also several genetic SVD-related VaD syndromes. The most prevalent is cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL), a genetic disorder linked to a mutation in the *NOTCH3* gene on chromosome 19. CADASIL presents as small vessel strokes, progressive dementia and extensive white matter disease often beginning in mid-adult life **(Chap. 420)**. Through altered extracellular molecular signaling pathways and protein elimination failure and accumulation, pericytes and endothelia of small vessels are involved. CADASIL may offer a unique opportunity to study "pure" VaD, as individuals diagnosed with CADASIL tend to display cognitive decline at an early age, when the likelihood of comorbid neurodegenerative diseases is lower.

In addition to infarcts and hemorrhages, SVD is also associated with blood-brain-barrier compromise, acculmulating white matter disease, and a state of chronic cerebral hypoperfusion, hippocampal atrophy and sclerosis. Neuroimaging markers of SVD, such as white matter hyperintensities (WMH) of presumed vascular origin, microbleeds, cortical microinfarcts, and enlarged Virchow-Robin spaces (eVRS) increase the risk of dementia, with conventional vascular risk factors explaining little of the variance in the absence of these changes. Otto Binswanger described the drastic effect of white matter injury when he observed eight patients with gradual cognitive deterioration and notable white matter changes. Binswanger's disease, commonly referred to as subcortical arteriosclerotic encephalopathy, is considered a prototypical clinical syndrome of VCI and a pathologically homogenous subgroup. On neuroimaging, a progressive confluent subcortical and periventricular white matter disease is seen **(see Fig. 25-2)**, with hypoperfusion and hypometabolism. Novel neuroimaging techniques for assessment of blood brain barrier integrity, such as dynamic contrast-enhanced MRI (DCE-MRI), in combination with biofluid markers, have shown a characteristic unrelenting progressive course of hypoxic injury with inflammatory disruption of blood brain barrier. Individuals with Binswanger's disease typically have hypertension or systemic vascular disease, and the clinical course may include gradual accumulation of focal neurological deficits. Neuropathological features include extensive demyelination and destruction of white matter with relative sparing of the subcortical U fibers. The resulting dementia syndrome is largely dysexecutive. From a cognitive and behavioral standpoint, affected individuals typically display apathy, mood alterations with mild depression, executive dysfunction, and slowed information processing. They can also have marked motor slowing and symmetric parkinsonism.

A strategically placed infarct, usually in the thalamus, basal ganglia or angular gyrus can also result in marked cognitive dysfunction and dementia. For example, a single paramedian artery (artery of Percheron) supplies both anteromedial thalamic regions, and occlusion of this artery can lead to bilateral infarction of the dorsomedial nucleus and the mammillothalamic tracts, resulting in altered mental status, vertical gaze palsy, and memory impairment. Similarly, an infarct in the inferior genu of the internal capsule may strategically disrupt the inferior and medial thalamic peduncles carrying thalamo-cortical fibers important for cognition and memory.

A consensus on frequency, causal factors, underlying neuropathology, clinical symptoms, characteristic neuropsychological presentations, and developmental course of VaD has yet to emerge. Issues related to comorbidity are particularly challenging. A large number of dementia patients with significant CVD show multiple pathologies, most frequently a mix of AD and/or Lewy body disease. Despite

variability in reported prevalence, the consistent decrease of pure VaD with age and the related increase in mixed pathologies is well established and reflect the complex interactions between CVD, brain aging, and accumulation of abnormal proteins in neurodegeneration. Unsurprisingly, post-stroke dementia risk factors extensively overlap with those identified for AD, including older age, low education, hypertension, diabetes, smoking, and cerebral atrophy and hippocampal sclerosis. Additionally, concurrent CVD lowers the threshold for dementia due to AD and Lewy body disease **(Chap. 426)**.

Treatment of VaD must be focused on accurate diagnosis of VaD so that new ischemic injury can be prevented by stabilizing or removing the underlying causes. Effective control of modifiable risk factors **(Chap. 2)**, including hypertension **(Chap. 271)**, smoking **(Chap. 448)**, alcohol intake **(Chap. 445)**, sodium consumption, diabetes mellitus **(Chap. 397)**, obesity **(Chap. 395)**, and the metabolic syndrome **(Chap. 401)**, are key to slowing the rising global prevalence of this condition. There is a great need for sensitive and specific in vivo biomarkers of VCI-VaD for early diagnosis and ultimately a surrogate marker of therapeutic efficacy in clinical trials.

FURTHER READING

GARDENER H et al: Brain health and shared risk factors for dementia and stroke. Nat Rev Neurol 11:651, 2015.

GORELICK PB et al: Vascular contributions to cognitive impairment and dementia. Stroke 42:2672, 2011.

MCALEESE KE et al: Post-mortem assessment in vascular dementia: Advances and aspirations. BMC Medicine 14:129, 2016.

RIZZI L et al: Global epidemiology of dementia: Alzheimer's and vascular types. Biomed Res Int 2014:908915, 2014.

ROSENBERG GA et al: Consensus statement for diagnosis of subcortical small vessel disease. J Cereb Blood Flow Metab 36:6, 2016.

SACHDEV P et al: Diagnostic criteria for vascular cognitive disorders: A VASCOG statement. Alzheimer Dis Assoc Disord 28:206, 2014.

426 Dementia with Lewy Bodies

William W. Seeley, Caroline M. Tanner, Bruce L. Miller

Although Lewy body disease (LBD) was once conceptualized as a disease of the substantia nigra, modern human postmortem studies with alpha-synuclein immunohistochemistry revealed that Lewy body and Lewy neurite pathology most often begins in the enteric and autonomic nervous systems before ascending through the brainstem to the substantia nigra, limbic system, and ultimately the cerebral cortex. In other individuals, disease may begin in the olfactory bulb and spread inward through olfactory system connections. These sites of onset, positioned as they are at neural interfaces with the environment, have suggested to some that a toxic environmental exposure may trigger the disease. Individual patients vary in their adherence to these general patterns, and the tempo and topology of progression dictate the clinical syndrome. Some patients with long-standing Parkinson's disease (PD) **(Chap. 427)** without cognitive impairment slowly develop a dementia that is associated with visual hallucinations and fluctuating alertness. In this scenario, the term *Parkinson's disease dementia* (PDD) is often used. In other patients, dementia and a neuropsychiatric syndrome precede or co-emerge with the parkinsonism, and the patient is diagnosed with *dementia with Lewy bodies* (DLB).

CLINICAL MANIFESTATIONS

The DLB syndrome is characterized by visual hallucinations, parkinsonism, fluctuating alertness, neuroleptic sensitivity, rapid eye movement (REM) sleep behavior disorder (RBD), and often hyposmia and excessive daytime sleepiness. Delusions related to persecution, invasion, and person or place identity (reduplicative paramnesia) are common. Both PDD and DLB may be accompanied or preceded by symptoms referable to brainstem pathology below the substantia nigra including constipation, orthostatic lightheadedness, depression/anxiety, and RBD, and most researchers now conceptualize DLB and PDD as points on a spectrum of LBD pathology. When orthostatic hypotension is present, DLB must be distinguished from multiple system atrophy with parkinsonism (MSA-P) **(Chap. 432)**. Recurrent, disabling syncope early in the course, accompanied by laryngeal spasms and anterocollis, suggest MSA-P. In DLB, orthostasis can appear early but rarely becomes disabling until well into the course, when neuropsychiatric symptoms and cognitive dysfunction are well-established.

Patients with DLB and PDD are highly sensitive to metabolic perturbations, and in some patients the first manifestation of illness is a delirium, often precipitated by an infection, new medicine, or other systemic disturbance. A hallucinatory delirium induced by L-dopa, prescribed for parkinsonian symptoms attributed to PD, may likewise provide the initial clue to a DLB or PDD diagnosis. Conversely, patients with mild cognitive deficits and hallucinations may receive typical or atypical antipsychotic medications, which induce profound parkinsonism at low doses due to a subclinical LBD-related nigral dopaminergic neuron loss. Minor day-to-day variation in cognitive functioning is common across dementias, but in DLB fluctuations can be marked, with bouts of confusion, lethargy, or even stupor that may rapidly resolve. Cognitively, DLB features relative preservation of episodic memory but often more severe visuospatial and executive deficits than seen in patients with early stages of Alzheimer's disease. Iodine-123-meta-iodobenzylguanidine (MIBG) cardiac scintigraphy to illustrate cardiac postganglionic sympathetic denervation and dopamine transporter imaging using SPECT or PET can both be useful in distinguishing dementia with parkinsonism due to LBD from that due to AD alone.

NEUROPATHOLOGY

The key neuropathological feature in LBD is the presence of Lewy bodies and Lewy neurites throughout specific brainstem nuclei, substantia nigra, amygdala, cingulate gyrus, and, ultimately, the neocortex. Formal staging criteria identify three stages of ascension: (1) brainstem predominant, (2) transitional limbic, and (3) diffuse neocortical, although healthy older individuals may also show isolated scattered Lewy body pathology in the substantia nigra, amygdala, or olfactory bulb. Lewy bodies are intraneuronal cytoplasmic inclusions that stain with periodic acid–Schiff (PAS) but are now identified with antibodies to the presynaptic protein, α-synuclein. Lewy bodies are composed of straight neurofilaments 7–20 nm long with surrounding amorphous material and contain epitopes recognized by antibodies against phosphorylated and nonphosphorylated neurofilament proteins, ubiquitin, and α-synuclein. Generally neuronal and synaptic loss, rather than Lewy pathology per se, best predict the clinical deficits. A profound cholinergic deficit, owing to basal forebrain and pedunculopontine nucleus involvement, is present in many patients with DLB and may be a factor responsible for the fluctuations, inattention, and visual hallucinations. Adrenergic deficits from locus ceruleus involvement further undermine arousal and alerting.

TREATMENT

Dementia with Lewy Bodies

Due to the frequent comorbidity with AD and the cholinergic deficit in DLB, cholinesterase inhibitors often provide significant benefit, reducing hallucinosis, stabilizing delusional symptoms, and even helping with RBD in some patients. Exercise programs maximize motor function and protect against fall-related injury. Antidepressants are often necessary. Atypical antipsychotics may be required for psychosis but can worsen extrapyramidal syndromes, even at low doses, and should be used cautiously, given the side effects including an increased risk of death. Patients with DLB are extremely sensitive to dopaminergic medications **(Chap. 427)**, which

must be carefully titrated; tolerability may be improved by concomitant use of a cholinesterase inhibitor. A general approach to the symptomatic management of dementia is presented in Chap. 25.

FURTHER READING

GOEDERT M et al: 100 years of Lewy pathology. Nat Rev Neurol 9:13, 2013.

MCKEITH IG et al: Diagnosis and management of dementia with Lewy bodies: Fourth consensus report of the DLB Consortium. Neurology 89:88, 2017.

MORRA LF, DONOVICK PJ: Clinical presentation and differential diagnosis of dementia with Lewy bodies: A review. Int J Geriatr Psychiatry 29:569, 2014.

SONNI I et al: Clinical validity of presynaptic dopaminergic imaging with 123I-ioflupane and noradrenergic imaging with 123I-MIBG in the differential diagnosis between Alzheimer's disease and dementia with Lewy bodies in the context of a structured 5-phase development framework. Neurobiol Aging 52:228, 2017.

VELAYUDHAN L et al: New therapeutic strategies for Lewy body dementias. Curr Neurol Neurosci Rep 17:68, 2017.

427 Parkinson's Disease

C. Warren Olanow, Christine Klein, Anthony H.V. Schapira

PARKINSON'S DISEASE AND RELATED DISORDERS

Parkinson's disease (PD) is the second most common age-related neurodegenerative disease, exceeded only by Alzheimer's disease (AD). Its cardinal clinical features were first described by the English physician James Parkinson in 1817. It is noteworthy that James Parkinson was a general physician who captured the essence of this condition based on a visual inspection of a mere handful of patients, several of whom he only observed and did not formally examine. It is estimated that the number of people with PD in the most populous nations worldwide was ~4 million persons in 2005, and this number is expected to more than double to ~9 million by the year 2030 based on the aging of the population. The mean age of onset of PD is about 60 years, and the lifetime risk is ~2% for men and 1.3% for women. The frequency of PD increases with aging, but cases can be seen in individuals in their twenties and even younger, particularly in association with a gene mutation.

Clinically, PD is characterized by rest tremor, rigidity (stiffness), bradykinesia (slowing), and gait dysfunction with postural instability. These are known as the "cardinal features" of the disease. Additional clinical features can include freezing of gait, speech difficulty, swallowing impairment, autonomic disturbances, and a series of nonmotor features that include sensory alterations, mood disorders, sleep dysfunction, cognitive impairment, and dementia (see Table 427-1 and discussion below).

Pathologically, the hallmark features of PD are degeneration of dopaminergic neurons in the substantia nigra pars compacta (SNc), reduced striatal dopamine, and intraneuronal proteinaceous inclusions known as Lewy bodies and Lewy neurites that primarily contain the protein α-synuclein (Fig. 427-1). While interest has primarily focused on the dopamine system, neuronal degeneration with inclusion body formation can also affect cholinergic neurons of the nucleus basalis of Meynert (NBM), norepinephrine neurons of the locus coeruleus (LC), serotonin neurons in the raphe nuclei of the brainstem, and neurons of the olfactory system, cerebral hemispheres, spinal cord, and peripheral autonomic nervous system. This "nondopaminergic" pathology is likely responsible for the development of the nondopaminergic clinical features listed in Table 427-1. There is some evidence that Lewy body pathology can begin in the peripheral autonomic nervous system, olfactory system, and dorsal motor nucleus of the vagus nerve in the lower brainstem, and then spread in a predictable and sequential manner to affect the upper brainstem (SNc) and cerebral hemispheres (Braak staging). These studies suggest that the classic degeneration of SNc dopamine neurons and the cardinal motor features of PD develop at a mid-stage of the disease. Indeed, epidemiologic studies suggest that clinical symptoms reflecting early involvement of nondopaminergic neurons such as constipation, anosmia, rapid eye movement (REM) behavior sleep disorder, and cardiac denervation can precede the onset of the classic motor features of PD by several years if not decades. Based on these findings, efforts are underway to accurately define a premotor stage of PD.

TABLE 427-1 Clinical Features of Parkinson's Disease

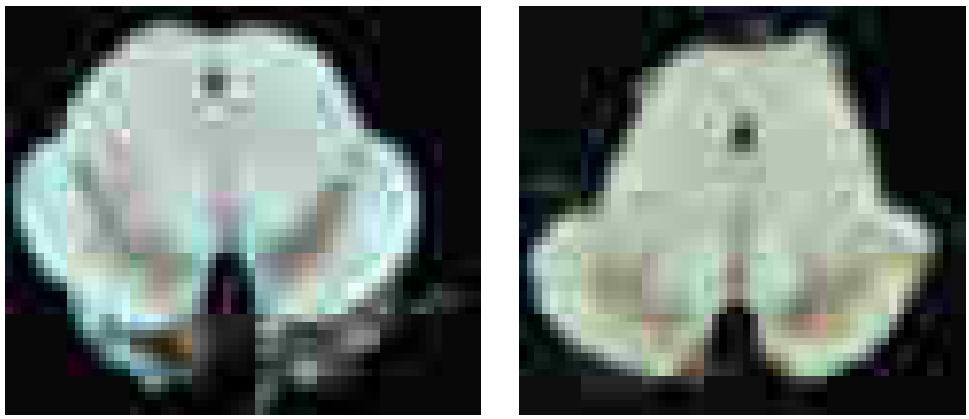
CARDINAL MOTOR FEATURES	OTHER MOTOR FEATURES	NONMOTOR FEATURES
Bradykinesia	Micrographia	Anosmia
Rest tremor	Masked facies (hypomimia)	Sensory disturbances (e.g., pain)
Rigidity	Reduced eye blinking	Mood disorders (e.g., depression)
Postural instability	Drooling	Sleep disturbances (e.g., RBD)
	Soft voice (hypophonia)	Autonomic disturbances
	Dysphagia	Orthostatic hypotension
	Freezing	Gastrointestinal disturbances
		Genitourinal disturbances
		Sexual dysfunction
		Cognitive impairment/ Dementia

Abbreviations: RBD, rapid eye movement behavior disorder.

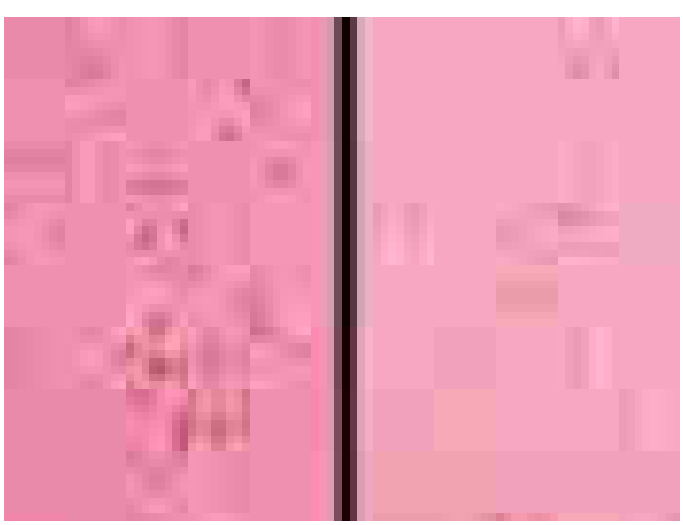
DIAGNOSIS AND DIFFERENTIAL DIAGNOSIS

Parkinsonism is a generic term that is used to define a syndrome manifest by bradykinesia with rigidity and/or tremor. It has a differential diagnosis (Table 427-2) that reflects differences in the site of damage and pathology in the various components of the basal ganglia. The basal ganglia comprise a group of subcortical nuclei that include the striatum (putamen and caudate nucleus), subthalamic nucleus (STN), globus pallidus pars externa (GPe), globus pallidus pars interna (GPi), and the SNc (Fig. 427-2). Among the different forms of parkinsonism, PD is the most common (~75% of cases). Historically, PD was diagnosed based on the presence of two of three parkinsonian features (tremor, rigidity, bradykinesia). However, postmortem studies found a 24% error rate when diagnosis was based solely on these criteria. Clinicopathologic correlation studies subsequently determined that parkinsonism associated with rest tremor, asymmetry of motor impairment, and a good response to levodopa was more likely to predict the correct pathologic diagnosis. With these revised criteria (known as the U.K. Brain Bank Criteria), a clinical diagnosis of PD could be confirmed pathologically in as many as 99% of cases. The International Parkinson's Disease and Movement Disorder Society (MDS) has recently suggested revised clinical criteria for PD (known as the MDS Clinical Diagnostic Criteria for Parkinson's disease) that are currently undergoing international validation. While motor parkinsonism has been retained as the core feature of the disease, the diagnosis of PD as the cause of parkinsonism relies on three additional categories of diagnostic features: supportive criteria (features that increase confidence in the diagnosis of PD), absolute exclusion criteria, and red flags (which must be counterbalanced by supportive criteria to permit a diagnosis of PD). Utilizing these criteria, two levels of certainty have been delineated; clinically established PD, and probable PD. The key diagnostic criteria for PD based on MDS criteria are illustrated in Table 427-2.

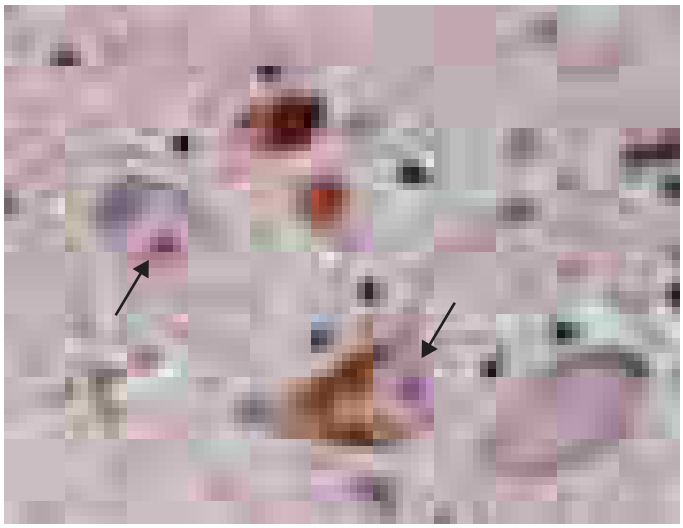
Imaging of the brain dopamine system in patients with PD can be performed using positron emission tomography (PET) or single-photon emission computed tomography (SPECT). These studies



A



B



C

FIGURE 427-1 Pathologic specimens from a patient with Parkinson's disease (PD) compared to a normal control demonstrating (***A***) reduction of pigment in SNc in PD (*right*) versus control (*left*), (***B***) reduced numbers of cells in SNc in PD (*right*) compared to control (*left*), and (***C***) Lewy bodies (*arrows*) within melanized dopamine neurons in PD. SNc, substantia nigra pars compacta.

typically show reduced and asymmetric uptake of striatal dopaminergic biomarkers, particularly in the posterior putamen with relative sparing of the caudate nucleus **(Fig. 427-3)**. These findings reflect the degeneration of nigrostriatal dopaminergic neurons and the loss of striatal terminals. Imaging can be useful in patients where there is diagnostic uncertainty (e.g., essential tremor, dystonic tremor, psychogenic tremor) or in research studies, but is rarely necessary in routine practice because the diagnosis can usually be established on clinical criteria alone. This may change in the future when there is a disease-modifying therapy and it is critically important to make a correct diagnosis as early as possible. Genetic testing can be helpful for establishing a diagnosis, but is not routinely employed as monogenic forms are rare and likely account for no more than 5% of cases (see discussion below). A genetic form of PD should be considered in patients with a positive family history, early age of onset (<40 years), a specific clinical picture or a particular ethnic background, and in research studies. Mutations of the *LRRK2* gene have attracted particular interest because they are the most common known cause of familial PD and are responsible for ~1% of typical sporadic cases of the disease. Mutations in *LRRK2* are a particularly frequent cause (~25%) of PD in Ashkenazi Jews and North African Berber Arabs; however, there is considerable variability in penetrance and many carriers never develop clinical features of PD.

TABLE 427-2 Differential Diagnosis of Parkinsonism

Parkinson's Disease	Atypical Parkinsonism	Secondary Parkinsonism	Neurodegenerative Disorders and other forms of parkinsonism
Sporadic	Multiple-system atrophy (MSA)	Drug-induced	Wilson's disease
Genetic	Cerebellar type (MSA-c)	Tumor	Huntington's disease
Dementia with Lewy bodies	Parkinson type (MSA-p)	Infection	Neurodegeneration with brain iron accumulation
	Progressive supranuclear palsy	Vascular	SCA 3 (spinocerebellar ataxia)
	Parkinsonism	Normal-pressure hydrocephalus	Fragile X–associated ataxia-tremor-parkinsonism
	Richardson variant	Trauma	Prion disease
	Corticobasal Syndrome	Liver failure	X-linked Dystonia-parkinsonism
	Frontotemporal dementia	Toxins (e.g., carbon monoxide, manganese, MPTP, cyanide, hexane, methanol, carbon disulfide)	Alzheimer's disease with parkinsonism
			Dopa-Responsive Dystonia

Abbreviations: MPTP, 1-methyl-4-phenyl-1,2,5,6-tetrahydropyridine.

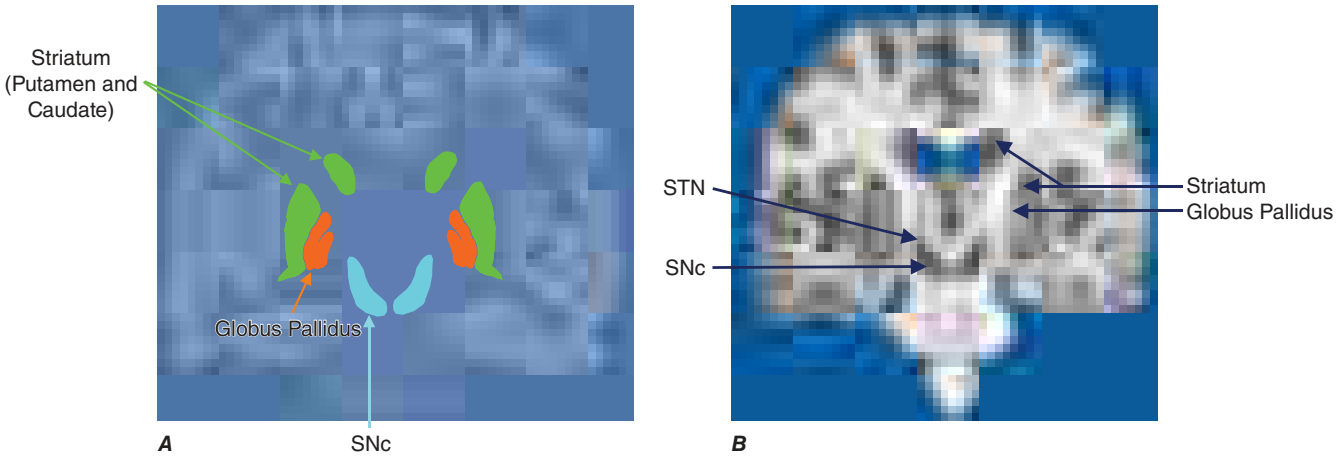


FIGURE 427-2 Basal ganglia nuclei. Schematic (***A***) and postmortem (***B***) coronal sections illustrating the various components of the basal ganglia. SNc, substantia nigra pars compacta; STN, subthalamic nucleus.

Genetic testing is of particular interest to identify at-risk individuals in a research setting. There is also some evidence that diagnosis of PD, and even pre-PD, may possible based on the presence of increased iron accumulation in the SNc using transcranial sonography or special MRI protocols.

Atypical, Secondary and Other Forms of Parkinsonism

Atypical parkinsonism refers to a group of neurodegenerative conditions that usually are associated with more widespread pathology than found in PD (potentially with degeneration of striatum, globus pallidus, cerebellum and brainstem as well as the SNc). These include Multiple System Atrophy (MSA), Progressive Supranuclear Palsy (PSP), and Corticobasal syndrome (CBS). As a group, they present with parkinsonism (rigidity and bradykinesia) but manifest clinical differences from PD reflecting the differences in their underlying pathology. In comparison to PD, the atypical parkinsonisms are characterized clinically by early involvement of speech and gait, absence of rest tremor, lack of motor asymmetry, poor or no response to levodopa, and a more aggressive clinical course. In the early stages, they may show a modest benefit from levodopa and can be difficult to distinguish from PD, but the diagnosis becomes clearer with disease evolution. Pathologically, neurodegeneration involves the SNc (typically without Lewy bodies) and has more extensive neurodegeneration than occurs in PD (see below for individual conditions). Neuroimaging of the dopamine system is usually not helpful, as striatal dopamine depletion can be seen in both PD and atypical parkinsonism. By contrast, metabolic imaging of the basal ganglia/thalamus network (using 2-F-deoxyglucose) may be helpful, showing a pattern of decreased activity in the GPi with increased activity in the thalamus, the reverse of what is seen in PD.

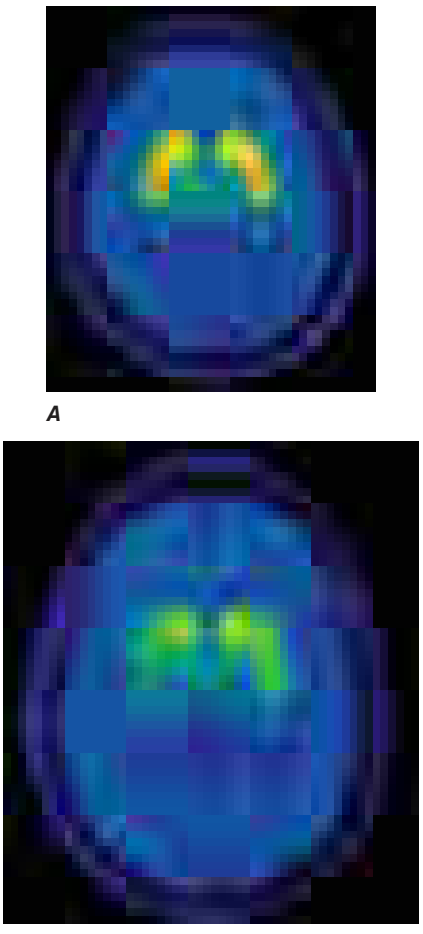


FIGURE 427-3 [^{11}C]Dihydrotetrabenazine positron emission tomography (a marker of VMAT2) in healthy control (*A*) and Parkinson's disease (*B*) patient. Note the reduced striatal uptake of tracer, which is most pronounced in the posterior putamen and tends to be asymmetric. *(Courtesy of Dr. Jon Stoessl.)*

MSA manifests as a combination of parkinsonian, cerebellar, and autonomic features and can be divided into a predominant parkinsonian (MSA-p) or cerebellar (MSA-c) form. Clinically, MSA is suspected when a patient presents with features of atypical parkinsonism as described above in conjunction with cerebellar signs and/or prominent autonomic dysfunction, usually orthostatic hypotension **(Chap. 432)**. Pathologically, MSA is characterized by degeneration of the SNc, striatum, cerebellum, and inferior olivary nuclei coupled with characteristic glial cytoplasmic inclusions (GCIs) that stain positively for α-synuclein. Magnetic resonance imaging (MRI) can show pathologic iron accumulation in the striatum on T2-weighted scans, high signal change in the region of the external surface of the putamen (putaminal rim) in MSA-p, or cerebellar and brainstem atrophy (the pontine "hot cross bun" sign **[Fig. 432-2]**) in MSA-c. There is currently no established evidence for any gene mutation/genetic risk factor for MSA. Recent studies suggest the possibility that MSA may be a prion disorder (see discussion below).

PSP is a form of atypical parkinsonism that is characterized by slow ocular saccades, eyelid apraxia, and restricted vertical eye movements with particular impairment of downward gaze. Patients frequently experience hyperextension of the neck with early gait disturbance and falls. In later stages, speech and swallowing difficulty and cognitive impairment may become evident. There is usually little or no response to levodopa. Two clinical forms of PSP have been identified; a "Parkinson" form that can closely resemble PD in the early stages including a positive response to levodopa, and the more classic "Richardson" form that is characterized by the features described above. MRI may reveal a characteristic atrophy of the midbrain with relative preservation of the pons on midsagittal images (the so-called "hummingbird sign"). Pathologically, PSP is characterized by degeneration of the SNc, striatum,

STN, midline thalamic nuclei, and pallidum, coupled with neurofibrillary tangles and inclusions that stain for the tau protein. Mutations in the MAPT gene which encodes for the tau protein have been detected in some familial cases.

CBS is the least common of the three atypical parkinsonisms and usually presents with asymmetric dystonic contractions and clumsiness of one hand coupled with cortical sensory disturbances manifest as apraxia, agnosia, focal limb myoclonus, or alien limb phenomenon (where the limb assumes a position in space without the patient being aware of the position or recognizing that the limb belongs to him/her). Dementia may occur at any stage of the disease. Both cortical and basal ganglia features are required to make this diagnosis. MRI frequently shows asymmetric cortical atrophy but this must be carefully sought. Pathologic findings include achromatic neuronal degeneration with tau deposits. Considerable overlap may occur both clinically and pathologically between CBS and PSP, and they may be difficult to distinguish without pathologic confirmation.

Secondary parkinsonisms occur as a result of a variety of primary conditions including drugs, stroke, tumor, infection, or exposure to toxins such as carbon monoxide or manganese that can cause damage to specific regions of the basal ganglia. Clinical features reflect the region of the basal ganglia that has been damaged. For example, strokes or tumors that affect the SNc may have a clinical picture identical to PD, whereas toxins such as carbon monoxide or manganese that damage the globus pallidus more closely resemble atypical parkinsonism. Dopamine-blocking agents such as neuroleptics are the most common cause of secondary parkinsonism. These drugs are most widely used in psychiatry, *but medical physicians should be aware that drugs such as metoclopramide which are primarily used to treat gastrointestinal problems are also neuroleptic agents* and may induce secondary parkinsonism. These drugs can also cause acute and tardive dyskinesias **(see Chap. 428)**. Other drugs that can cause secondary parkinsonism include tetrabenazine, calcium channel blockers (flunarizine, cinnarizine), amiodarone, and lithium.

Parkinsonism can also be seen in Dopa-Responsive Dystonia, a condition that results from a mutation in the *GTP-Cyclohydrolase 1 gene* which can lead to a defect in a cofactor for tyrosine hydroxylase and the impaired manufacture of dopa and dopamine. While it typically presents as dystonia **(Chap. 428)**, it can present as a biochemically based form of parkinsonism (due to reduced synthesis of dopamine) which closely resembles PD and responds to levodopa, but is not associated with abnormalities on fluoro-dopa positron emission tomography (FD-PET) nor neurodegeneration. This diagnosis should be considered in individuals aged <20 years who present with a clinical picture resembling PD.

Finally, parkinsonism can be seen as a feature of a variety of other degenerative disorders such as Wilson's disease, Huntington's disease (especially the juvenile form known as the Westphal variant), certain forms of spinocerebellar ataxias, and neurodegenerative disorders with brain iron accumulation such as pantothenate kinase (PANK)–associated neurodegeneration (formerly known as Hallervorden-Spatz disease).

Some features that suggest that parkinsonism might be due to a condition other than PD are shown in **Table 427-3**.

TABLE 427-3 Features Suggesting an Atypical or Secondary Cause of Parkinsonism

SYMPTOMS/SIGNS	ALTERNATIVE DIAGNOSIS TO CONSIDER
History	
Early speech and gait impairment (Lack of tremor, lack of motor asymmetry)	Atypical parkinsonism
Exposure to neuroleptics	Drug-induced parkinsonism
Onset prior to age 40	Genetic form of PD
Liver disease	Wilson's disease, non-Wilsonian hepatolenticular degeneration
Early hallucinations and dementia with later development of PD features	Dementia with Lewy bodies
Diplopia, impaired down gaze	PSP
Poor or no response to an adequate trial of levodopa	Atypical or secondary parkinsonism
Physical Examination	
Dementia as first or early feature	Dementia with Lewy bodies
Prominent orthostatic hypotension	MSA-p
Prominent cerebellar signs	MSA-c
Slow saccades with impaired down gaze	PSP
High-frequency (6–10 Hz) symmetric postural tremor with a prominent kinetic component	Essential tremor

Abbreviations: MSA-c, multiple-system atrophy–cerebellar type; MSA-p, multiple-system atrophy–Parkinson's type; PD, Parkinson's disease; PSP, progressive supranuclear palsy.

ETIOLOGY AND PATHOGENESIS

Most PD cases occur sporadically (~85–90%) and are of unknown cause. Gene mutations (see below) are the only known causes of PD. Twin studies performed several decades ago suggested that environmental factors might play an important role in patients with an age of onset ≥50 years, with genetic factors being more important in younger-onset patients. However, the demonstration of later onset genetic variants (e.g., *LRRK2* and *GBA*) argues against the emphasis on environmental factors, even in individuals >50 years of age. The environmental hypothesis received some support in the 1980s with the demonstration that MPTP (1-methyl-4-phenyl-1,2,5,6-tetrahydropyridine), a by-product of the illicit manufacture of a heroin-like drug, caused a PD-like syndrome in addicts in northern California. MPTP is transported into the central nervous system, where it is oxidized to form MPP^+, a mitochondrial toxin that is selectively taken up by, and damages, dopamine neurons. However, MPTP or MPTP-like compounds have not been linked to sporadic PD. Epidemiologic studies have reported an increased risk of developing PD in association with exposure to pesticides, rural living, farming, and drinking well water. Dozens of other associations have also been reported in individual studies but results have been inconsistent, and no environmental factor has yet been proven to be a cause or to contribute to the cause of PD. Some possible protective factors have also been identified in epidemiologic studies including caffeine, smoking, intake of nonsteroidal anti-inflammatory drugs, and calcium channel blockers. The validity of these findings and the responsible mechanism also remain to be established.

About 5–15% of cases are familial in origin, and mutations in several PD-linked genes have been identified **(Table 427-4)**. While monogenic mutations have been shown to be causative of PD, genetic risk factors that increase the risk of developing PD have also been identified. Large-size genome-wide association studies (GWASs) have identified 26 independent gene variants (single nucleotide polymorphisms) as PD risk factors including variants in the *SNCA, LRRK2, MAPT*, and *GBA* genes as well as in the HLA region on chromosome 6. It has been proposed that many cases of PD may be due to a "double hit" involving an interaction between (a) one or more genetic risk factors that induce susceptibility coupled with (b) exposure to a toxic environmental factor that may induce epigenetic or somatic DNA alterations or has the potential to directly damage the dopaminergic system. In this scenario, both factors are required for PD to ensue, while the presence of either one alone is not sufficient to cause the disease. Notably, however, even if a genetic or environmental risk factor doubles the risk to develop PD, this only results in a lifetime risk of 4% or lower, and thus cannot presently be used for individual patient counseling.

Several factors have been implicated in the pathogenesis of cell death in PD, including oxidative stress, inflammation, excitotoxicity, mitochondrial dysfunction, and the accumulation of misfolded proteins with consequent proteolytic stress. Recent studies have demonstrated that with aging, dopamine neurons switch from sodium to calcium pacing through calcium channels, potentially making these high-energy neurons vulnerable to calcium-mediated neurotoxicity. Whatever the pathogenic mechanism, cell death appears to occur, at

TABLE 427-4 Confirmed Genetic Causes of Parkinson's Disease*

DESIGNATION* AND REFERENCE	GENEREVIEWS AND OMIM REFERENCE	CLINICAL CLUES	INHERITANCE	PREVIOUS LOCUS SYMBOL
1. Classical PD				
PARK-*SNCA*	GeneReviews *http://www.ncbi.nlm.nih.gov/books/NBK1223/* OMIM 168601	Missense mutations cause classical parkinsonism. Duplication or triplication mutations in this gene cause early onset parkinsonism with prominent dementia.	AD	PARK1
PARK-*LRRK2*	GeneReviews *http://www.ncbi.nlm.nih.gov/books/NBK1208/* OMIM 607060	Clinically typical PD	AD	PARK8
PARK-*VPS35*	GeneReviews *http://www.ncbi.nlm.nih.gov/books/NBK1223/* OMIM 614203	Clinically typical PD	AD	PARK17
2. Early-onset PD				
PARK-*Parkin*	GeneReviews *http://www.ncbi.nlm.nih.gov/books/NBK1155/* OMIM 600116	Often presents with dystonia, typically in a leg	AR	PARK2
PARK-*PINK1*	GeneReviews *http://www.ncbi.nlm.nih.gov/books/NBK1223/* OMIM 605909	Often presents with psychiatric features	AR	PARK6
PARK-*DJ1*	GeneReviews *http://www.ncbi.nlm.nih.gov/books/NBK1223/* OMIM 606324		AR	PARK7
3. Parkinsonism				
PARK-*ATP13A2*	GeneReviews *http://www.ncbi.nlm.nih.gov/books/NBK1223/* OMIM 606693	Kufor-Rakeb syndrome with parkinsonism and dystonia; additional features: Supranuclear gaze palsy, spasticity/pyramidal signs, dementia, facial-faucial-finger mini-myoclonus, dysphagia, dysarthria, olfactory dysfunction	AR	PARK9
PARK-*FBXO7*	GeneReviews *http://www.ncbi.nlm.nih.gov/books/NBK1223/* OMIM: 260300	Early onset parkinsonism with pyramidal signs	AR	PARK15
PARK-*DNAJC6*	GeneReviews: n/a OMIM 615528	May present with mental retardation and seizures	AR	PARK19
PARK-*SYNJ1*	GeneReviews: n/a OMIM 615530	May have seizures, cognitive decline, abnormal eye movements, and dystonia	AR	PARK20

*According to the recommendations of the International Parkinson's and Movement Disorder Society (C Marras: Mov Disord 31:436, 2016).

least in part, by way of a signal-mediated apoptotic or "suicidal" process. Each of these mechanisms offers a potential target for putative neuroprotective drugs. However, it is not clear which of these factors is primary, if they are the same in all cases or specific to individual (genetic) patient subgroups, if they act by way of a network such that multiple insults are required for neurodegeneration to ensue, or if the findings to date merely represent an epiphenomenon unrelated to the true cause of cell death that still remains undiscovered (Fig. 427-4).

Although gene mutations cause only a minority of cases of PD, they may be helpful in pointing to specific pathogenic pathways and molecular mechanisms that are central to a neurodegenerative process that might be relevant to all forms of the disease. To date, most interest has focused on pathways implicated by mutations in α-*synuclein (SNCA)*, *GBA*, *LRRK2*, and *PINK1/Parkin*.

Although mutations in *SNCA* are an extremely rare cause of PD, *SNCA* was the first PD-linked and most intensely investigated PD gene, with respect to causative mutations but also risk variants, function of the gene and of the encoded protein. Shared clinical features of patients with *SNCA* mutations include earlier age of disease onset than in nongenetic PD, a faster progression of motor signs that are mostly levodopa-responsive, early occurrence of motor fluctuations, and presence of prominent nonmotor features. Intriguingly, SNCA constitutes the major component of Lewy bodies in patients with both monogenic

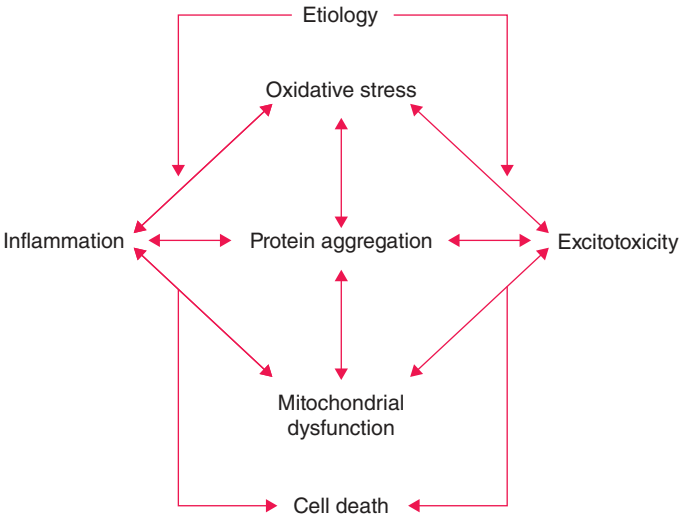


FIGURE 427-4 Schematic representation of how pathogenetic factors implicated in Parkinson's disease interact in a network manner, ultimately leading to cell death. This figure illustrates how interference with any one of these factors may not necessarily stop the cell death cascade. *(Adapted from CW Olanow: Movement Disorders 22:S-335, 2007.)*

and sporadic forms of PD (Fig. 427-1). Duplication or triplication of the wild-type *SNCA* gene also causes PD with triplication carriers being more severely affected than carriers of duplications. These findings indicate that increased production of the normal protein alone can cause the disease in a dose-dependent fashion. More recently, Lewy pathology was discovered to have developed in healthy embryonic dopamine neurons that had been implanted into the striatum of PD patients, suggesting that the abnormal protein had transferred from affected cells to healthy unaffected dopamine neurons. Based on these findings, it has been proposed that the SNCA protein may be a prion, and PD a prion disorder **(Chaps. 417 and 430)**. Like the prion protein PrP^C, SNCA can misfold to form β-rich sheets, join to form toxic oligomers and aggregates, polymerize to form amyloid plaques (i.e., Lewy bodies), and cause neurodegeneration with spread to involve unaffected neurons. Indeed, injection of SNCA fibrils into the striatum of both transgenic and wild-type rodents leads to the development of Lewy pathology in host neurons, neurodegeneration, behavioral abnormalities, with spread of SNCA pathology to anatomically connected sites. Further support for this hypothesis comes from the demonstration that inoculation of SNCA derived from human Lewy bodies induces dopamine cell degeneration and widespread Lewy pathology in mice and primates. Collectively, this evidence supports the possibility that neuroprotective therapies for PD might be developed based on inhibiting the accumulation or accelerating the removal of SNCA aggregates, knocking down levels of host SNCA, or blocking the templating phenomenon whereby misfolded SNCA promotes misfolding of the native protein in a prion-like chain reaction.

Mutations in the glucocerebrosidase (*GBA*) gene represent the most important risk factor in terms of effect size for the development of PD, and experimentally there is a direct pathophysiological link between increased levels of SNCA and reduced levels of GBA. GBA encodes the enzyme glucocerebrosidase (GCase) which promotes lysosomal function and enhances the clearance of misfolded proteins. The identification of *GBA* as a risk gene for PD resulted from the clinical observation that patients with Gaucher's disease (GD) and their relatives commonly show signs of parkinsonism. This clinical observation of a link between GD and PD led to the discovery that several mutations in *GBA*, which cause Gaucher's disease in an autosomal recessive manner, confer risk for the development of PD, also in a heterozygous state. Further, reduced GCase activity due to GBA mutations impairs lysosomal function which results in the accumulation of SNCA. Accumulation of SNCA can also lead to inhibition of lysosomal function and a further reduction in levels of wild-type GBA by interfering with endoplasmic reticulum-to-Golgi trafficking. This in turn, leads to decreased GBA activity and a further increase in the accumulation of SNCA. In this regard, it is noteworthy that lysosomal function is impaired and levels of GCase are reduced in patients with sporadic PD. These findings suggest that this molecular pathway may apply not only to patients with GD or with a *GBA* mutation, but also to patients with sporadic PD or other synucleinopathies who have two wild-type *GBA* alleles. These bidirectional effects of SNCA and GBA form a positive feedback loop that, after surpassing a theoretical threshold, could lead to self-propagating disease. Studies of drugs that enhance GCase activity are currently underway.

Seven different *LRRK2* mutations have now been clearly linked to PD, with p.G2019S being the most common due to a founder effect in the Ashkenazi Jewish and North African Arab populations. Mutations in *LRRK2* account for 3–41% of familial PD cases (depending on specific population) and are also found in apparently sporadic cases, albeit at a lower rate. The phenotype of LRRK2 p.G2019S mutations is indistinguishable from that of sporadic PD, although tremor appears to be more common, and leg tremor may be a useful diagnostic clue. The mechanism responsible for cell death with this mutation is not conclusively known but is thought to involve changes in kinase activity with altered phosphorylation of target proteins (including autophosphorylation) with possible impairment of lysosomal function. Kinase inhibitors can block toxicity associated with *LRRK2* mutations in laboratory models, and there has been much interest in developing drugs directed at this target. However, kinase inhibitors are potentially toxic, and the majority of PD patients do not carry a *LRRK2* mutation.

Mutations in *Parkin* and *PINK1* have also been identified as a cause of PD. *Parkin* mutations are the more common, and the major cause of autosomal recessive and early-onset PD, accounting for up to 77% of cases of juvenile PD with an age of onset <20 years, and for 10–20% of early-onset PD patients in general. The disease is slowly progressive, responds well to antiparkinsonian treatment, and is commonly complicated by dystonia, but very rarely by dementia. At pathology, neurodegeneration tends to be restricted to the SNc and LC in patients with *Parkin* mutations, and Lewy bodies are typically absent. The reason for these differences from classic PD are not known, but may related to impaired ubiquitination of damaged proteins (parkin is a ubiquitin ligase). The clinical phenotypes of *Parkin*- and *PINK1*-linked PD are similar. Recent studies suggest a role for Parkin and PINK1 proteins in the turnover and clearance of damaged mitochondria (mitophagy), and mutations in *Parkin* and *PINK1* cause mitochondrial dysfunction in transgenic animals that can be corrected with overexpression of Parkin or with drugs acting on the mitochondrial electron transfer chain, such as Vitamin K2. Improving mitochondrial function is a particularly attractive potential therapeutic target because postmortem studies in PD patients show a defect in complex I of the respiratory chain in SNc neurons.

Thus, evidence is accumulating that genetics plays an important role in both familial and "sporadic" forms of PD. It is anticipated that better understanding of the pathways responsible for cell death caused by these mutations will permit the development of more relevant animal models of PD and targets for the development of gene-specific neuroprotective drugs.

PATHOPHYSIOLOGY OF PD

The classic model of the organization of the basal ganglia in the normal and PD states is provided in **Fig. 427-5**. With respect to motor function, a series of neuronal circuits or loops link the basal ganglia nuclei with corresponding cortical motor regions in a somatotopic manner. The striatum is the major input region of the basal ganglia, while the GPi and SNr are the major output regions. The input and output regions are connected via direct and indirect pathways that have reciprocal effects on the activity of the basal ganglia output pathway. The output of the basal ganglia provides inhibitory (GABAergic) tone to thalamic and brainstem neurons that in turn connect to motor systems in the cerebral cortex and spinal cord that control motor function. An increase in neuronal activity in the output regions of the basal ganglia (GPi/SNr) is associated with poverty of movement or parkinsonism, while decreased output results in movement facilitation and involuntary movements. Dopaminergic projections from SNc neurons serve to modulate neuronal firing and to stabilize the basal ganglia network. Normal dopamine innervation thus serves to facilitate the selection of the desired movement and reject unwanted movements. Cortical loops integrating the cortex and the basal ganglia are now thought to also play an important role in regulating behavioral, emotional, and cognitive functions.

In PD, dopamine denervation with loss of dopaminergic tone leads to increased firing of neurons in the STN and GPi, excessive inhibition of the thalamus, reduced activation of cortical motor systems, and the development of parkinsonian features (Fig. 427-5). The current role of surgery in the treatment of PD is based on this model, which predicted that lesions or high-frequency stimulation of the STN or GPi might reduce this neuronal overactivity and improve PD features.

TREATMENT

Parkinson's Disease

LEVODOPA

Since its introduction in the late 1960s, levodopa has been the mainstay of therapy for PD. Experiments in the late 1950s by Carlsson and colleagues demonstrated that blocking dopamine uptake with reserpine caused rabbits to become parkinsonian; this could be reversed with the dopamine precursor, levodopa. Subsequently, Hornykiewicz demonstrated a dopamine deficiency in the striatum

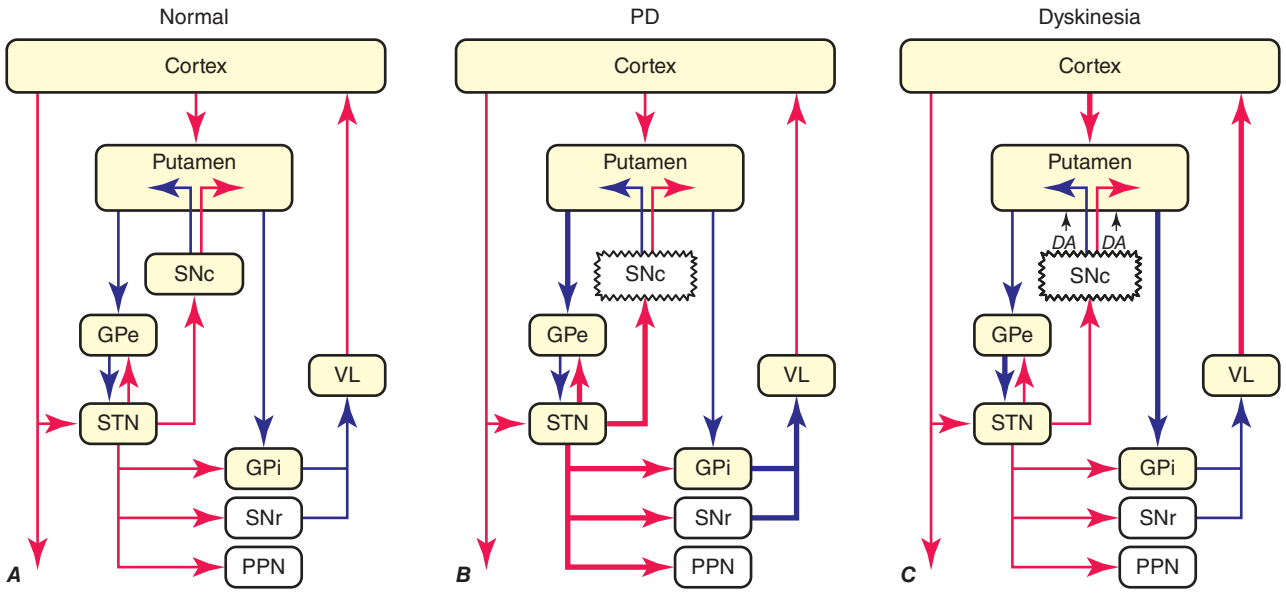


FIGURE 427-5 Basal ganglia organization. Classic model of the organization of the basal ganglia in the normal (***A***), Parkinson's disease (PD) (***B***), and levodopa-induced dyskinesia (***C***) state. Inhibitory connections are shown as *blue arrows* and excitatory connections as *red arrows*. The striatum is the major input region and receives its major input from the cortex. The GPi and SNr are the major output regions, and they project to the thalamocortical and brainstem motor regions. The striatum and GPi/SNr are connected by direct and indirect pathways. This model predicts that parkinsonism results from increased neuronal firing in the STN and GPi and that lesions or DBS of these targets might provide benefit. This concept led to the rationale for surgical therapies for PD. The model also predicts that dyskinesia results from decreased firing of the output regions, resulting in excessive cortical activation by the thalamus. This component of the model is not completely correct because lesions of the GPi ameliorate rather than increase dyskinesia in PD, suggesting that firing frequency is just one of the components that lead to the development of dyskinesia. DBS, deep brain stimulation; GPe, external segment of the globus pallidus; GPi, internal segment of the globus pallidus; PPN, pedunculopontine nucleus; SNc, substantia nigra, pars compacta; SNr, substantia nigra, pars reticulata; STN, subthalamic nucleus; VL, ventrolateral thalamus. *(Derived from JA Obeso et al: Trends Neurosci 23:S8, 2000.)*

of PD patients, and suggested the potential benefit of dopamine replacement therapy. Dopamine does not cross the blood-brain barrier (BBB), so clinical trials were initiated with levodopa, the precursor of dopamine. Studies over the course of the next decade confirmed the value of levodopa and revolutionized the treatment of PD.

Levodopa is routinely administered in combination with a peripheral decarboxylase inhibitor to prevent its peripheral metabolism to dopamine and the development of nausea, vomiting, and orthostatic hypotension due to activation of dopamine receptors in the area postrema that are not protected by the BBB. In the United States, levodopa is combined with the decarboxylase inhibitor carbidopa (Sinemet®), whereas in many other countries it is combined with benserazide (Madopar®). Levodopa plus a decarboxylase inhibitor is also available in a methylated formulation, a controlled-release formulation (Sinemet CR® or Madopar HP®) and in combination with a catechol-*O*-methyltransferase (COMT) inhibitor (Stalevo®). A long-acting formulation of levodopa (Rytary®) has also recently been approved. An inhaled form of levodopa that is rapidly and reliably absorbed is currently in late stage investigation as a rescue therapy for the treatment of individual "off" episodes (see below).

Levodopa remains the most effective symptomatic treatment for PD and the gold standard against which new therapies are compared. No current medical or surgical treatment provides antiparkinsonian benefits superior to what can be achieved with levodopa. Levodopa benefits the classic motor features of PD, prolongs independence and employability, improves quality of life, and increases life span. Almost all PD patients experience improvement, and failure to respond to an adequate trial of levodopa should cause the diagnosis to be questioned.

There are, however, important limitations of levodopa therapy. Acute dopaminergic side effects include nausea, vomiting, and orthostatic hypotension as indicated above. These are usually transient and can generally be avoided by starting with low doses and gradual titration. If they persist, they can be treated with additional doses of a peripheral decarboxylase inhibitor (e.g., carbidopa), administering with food, or adding a peripheral dopamine-blocking agent such as domperidone (not available in the United States). More important are motor complications (see below) that develop in the majority of patients treated long-term with levodopa. In addition, the disease continues to progress, and features such as neuropsychiatric problems, falling, freezing, autonomic dysfunction, sleep disorders, and dementia may emerge that are not adequately controlled by levodopa. Indeed, these nondopaminergic features (especially falling and dementia) are the primary source of disability and the main reason in the present era for nursing home placement for patients with advanced PD.

Levodopa-induced motor complications consist of fluctuations in motor response ("on" episodes when the drug is working and "off" episodes when parkinsonian features return) and involuntary movements known as dyskinesias which typically complicate "on" periods **(Fig. 427-6)**. When patients initially take levodopa, benefits are long-lasting (many hours) even though the drug has a relatively short half-life (60–90 min). With continued treatment, however, the duration of benefit following an individual dose becomes progressively shorter until it approaches the half-life of the drug. This loss of benefit is known as the *wearing-off effect*. In more severe cases, the response to a given dose may be variable with patients potentially experiencing a delay in turning on (delayed-on) or no response at all (no-on). Peak-dose dyskinesias occur at the time of levodopa peak plasma concentration and maximal clinical benefit. They are usually choreiform, but can manifest as dystonic movements, myoclonus, or other movement disorders. They are not troublesome when mild, but can be disabling when severe, and can limit the ability to use higher doses of levodopa to better control PD motor features. In more advanced states, patients may cycle between "on" periods complicated by disabling dyskinesias and "off" periods in which they suffer from severe parkinsonism and painful dystonic postures. Patients may also experience "diphasic dyskinesias," which occur as the levodopa dose begins to take effect and again as it wears off. These dyskinesias typically consist of transient, stereotypic, rhythmic movements that predominantly involve the lower extremities and are frequently associated with parkinsonism in other body regions. They can be relieved by increasing the dose of levodopa,

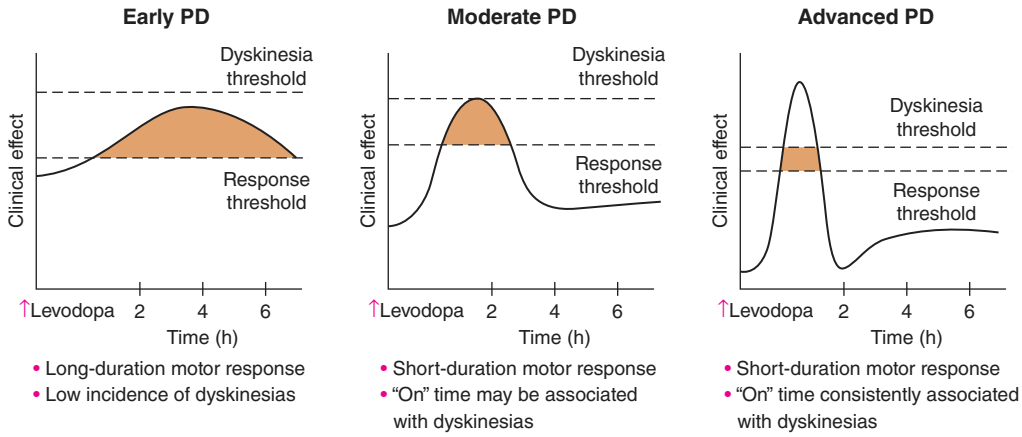


FIGURE 427-6 Changes in motor response associated with chronic levodopa treatment. Levodopa-induced motor complications. Schematic illustration of the gradual shortening of the duration of a beneficial motor response to levodopa (wearing off) and the appearance of dyskinesias complicating "on" time. PD, Parkinson's disease.

although higher doses may induce more severe peak-dose dyskinesia. Long-term double blind studies show that motor complications are dose related, and can be minimized by using the lowest dose of levodopa that provides satisfactory benefit and through the use of polypharmacy to avoid raising the dose of levodopa.

The cause of levodopa-induced motor complications is not precisely known. They are more likely to occur in females, younger individuals with more severe disease, and with the use of higher doses of levodopa. The classic model of the basal ganglia has been useful for understanding the origin of motor features in PD, but has proved less valuable for understanding levodopa-induced dyskinesias (Fig. 427-5). The model predicts that dopamine replacement might excessively inhibit the pallidal output system, thereby leading to increased thalamocortical activity, enhanced stimulation of cortical motor regions, and the development of dyskinesia. However, lesions of the pallidum are associated with amelioration rather than induction of dyskinesia as would be suggested by the classic model. It is now thought that dyskinesia results from alterations in the GPi/SNr neuronal firing pattern (pauses, bursts, synchrony, etc.) and not simply the firing frequency alone. This in turn leads to the transmission of "misinformation" from pallidum to thalamus/cortex, resulting in dyskinesia. Surgical lesions or high-frequency stimulation targeted at the GPi or STN can ameliorate dyskinesia by interfering with (blocking or masking) this abnormal neuronal activity and preventing the transfer of misinformation to motor systems. There has also been recent interest in the use of ultrasound to lesion these target regions in a relatively noninvasive manner.

Current information suggests that altered neuronal firing patterns and motor complications develop in response to nonphysiologic levodopa replacement. Striatal dopamine levels are normally maintained at a relatively constant level. In PD, dopamine neurons degenerate and striatal dopamine is dependent on the peripheral availability of levodopa. Intermittent oral doses of levodopa result in fluctuating plasma levels because of variability in the transit of the drug from the stomach to the duodenum where it is absorbed and the short half-life of the drug. This variability is translated to the brain and results in exposure of striatal dopamine receptors to alternating high and low concentrations of dopamine. It has been hypothesized that more continuous delivery of levodopa might prevent the development of motor complications. Indeed, a recent double-blind, double-dummy, double titration study demonstrated that continuous intraintestinal infusion of levodopa/carbidopa is associated with significant improvement in "off" time and in "on" time without dyskinesia in advanced PD patients compared with optimized standard oral levodopa. These benefits are superior to what has been observed in double blind controlled studies with other dopaminergic agents, and this therapy is now approved in the United States and Europe (Duodopa®, Duopa®). The treatment is, however, complicated by potentially serious adverse events related to the surgical procedure and the tubing, and the inconvenience of the infusion system. New approaches are currently being tested in which levodopa is continuously administered by subcutaneous infusion or by long-acting oral levodopa formulations in an effort to avoid the need for a surgical procedure. An inhaled formulation of levodopa is in late stage development as an acute rescue therapy for individual off episodes.

Behavioral complications can also be encountered in levodopa-treated patients. A dopamine dysregulation syndrome has been described where patients have a craving for levodopa and take frequent and unnecessary doses of the drug in an addictive manner. PD patients taking high doses of levodopa can also develop purposeless, stereotyped behaviors such as the assembly and disassembly or collection and sorting of objects. This is known as punding, a term taken from the Swedish description of the meaningless behaviors seen in chronic amphetamine users. Hypersexuality and other impulse-control disorders are occasionally encountered with levodopa, although these are more commonly seen with dopamine agonists.

DOPAMINE AGONISTS

Dopamine agonists are a diverse group of drugs that act directly on dopamine receptors. Unlike levodopa, they do not require metabolic conversion to an active product and do not undergo oxidative metabolism. Initial dopamine agonists were ergot derivatives (e.g., bromocriptine, pergolide, cabergoline) and were associated with potentially serious ergot-related side effects such as cardiac valvular damage and pulmonary fibrosis. They have largely been replaced by a second generation of nonergot dopamine agonists (e.g., pramipexole, ropinirole, rotigotine). In general, dopamine agonists do not have comparable efficacy to levodopa. They were initially introduced as adjuncts to levodopa to enhance motor function and reduce "off" time in fluctuating patients. Subsequently, it was shown that dopamine agonists are less prone than levodopa to induce dyskinesia, possibly because they are relatively long-acting. For this reason, many physicians initiate therapy with a dopamine agonist particularly in younger patients, although supplemental levodopa is eventually required in virtually all patients. This view has been tempered by the recognition that dopamine agonists are associated with potentially serious adverse effects such as unwanted sleep episodes and impulse control disorders (see below). Both ropinirole and pramipexole are available as orally administered immediate (tid) and extended-release (qd) formulations. Rotigotine is administered as a once-daily transdermal patch, and may be useful in managing surgical patients who are NPO. Apomorphine is a dopamine agonist with efficacy comparable to levodopa, but it must be administered parenterally as it is rapidly and extensively metabolized if taken orally. It has a short half-life and duration of activity (45 min). It can be administered by subcutaneous injection as a rescue agent for the treatment of severe "off" episodes, but can also be administered by continuous subcutaneous infusion where it has been demonstrated

to reduce both "off" time and dyskinesia in advanced patients. This latter approach has not been approved in the United States. A sublingual bilayer formulation of apomorphine is in late stage development as a rapid and reliable therapy for individual "off" periods that avoids the need for a subcutaneous (SC) injection.

Dopamine agonist use is associated with a variety of side effects. Acute side effects are primarily dopaminergic and include nausea, vomiting, and orthostatic hypotension. As with levodopa, these can usually be avoided by starting with low doses and using slow titration. Side effects associated with chronic use include hallucinations and cognitive impairment. Sedation with sudden unintended episodes of falling asleep that can occur in dangerous situations such as while driving a motor vehicle have been reported. Patients should be informed about this potential problem and should not drive when tired. Dopamine agonists can also be associated with impulse-control disorders, including pathologic gambling, hypersexuality, and compulsive eating and shopping. Patients should also be advised of these risks and specifically questioned for their occurrence at follow-up examinations. The precise cause of these problems, and why they appear to occur more frequently with dopamine agonists than levodopa, remains to be resolved, but reward systems associated with dopamine and alterations in the ventral striatum and orbitofrontal regions have been implicated. In general, chronic side effects are dose-related and can be avoided or minimized with lower doses. Injections of apomorphine can be complicated by skin lesions at sites of administration, but this has not been a problem with the sublingual bilayer formulation.

MAO-B INHIBITORS

Inhibitors of monoamine oxidase type B (MAO-B) block central dopamine metabolism and increase synaptic concentrations of the neurotransmitter. Selegiline and rasagiline are relatively selective suicide inhibitors of the MAO-B isoform of the enzyme. Clinically, these agents provide antiparkinsonian benefits when used as monotherapy in early disease stages and reduced "off" time when used as an adjunct to levodopa in patients with motor fluctuations. MAO-B inhibitors are generally safe and well tolerated. They may increase dyskinesia in levodopa-treated patients, but this can usually be controlled by down-titrating the dose of levodopa. Inhibition of the MAO-A isoform prevents metabolism of tyramine in the gut, leading to a potentially fatal hypertensive reaction known as a "cheese effect" because it can be precipitated by foods rich in tyramine such as some cheeses, aged meats, and red wine. Selegiline and rasagiline do not functionally inhibit MAO-A and are not associated with a cheese effect with doses used in clinical practice. There are theoretical risks of a serotonin reaction in patients receiving concomitant selective serotonin reuptake inhibitor (SSRI) antidepressants, but these are rarely encountered. Safinamide (Xadago®) is a reversible MAO-B inhibitor that has recently been approved as an adjunct to levodopa in advanced PD patients with motor fluctuations. The drug also acts to block activated sodium channels and inhibit glutamate release, and is currently being studied as a possible antidyskinetic agent.

Interest in MAO-B inhibitors has also focused on their potential to have disease-modifying effects. MPTP toxicity can be prevented experimentally by coadministration of an MAO-B inhibitor that blocks its conversion to the toxic pyridinium ion MPP^+ that selectively damages dopamine neurons. MAO-B inhibitors also have the potential to block the oxidative metabolism of dopamine and prevent oxidative stress. In addition, both selegiline and rasagiline incorporate a propargyl ring within their molecular structure that provides antiapoptotic effects in laboratory models. The DATATOP study showed that selegiline significantly delayed the time until the emergence of disability necessitating the introduction of levodopa in untreated PD patients. However, it could not be definitively determined whether this was due to a neuroprotective effect that slowed disease progression or a symptomatic effect that masked ongoing neurodegeneration. More recently, the ADAGIO study used a two-period delayed-start design and demonstrated that early treatment with rasagiline 1 mg/d, but not 2 mg/d, provided benefits that could not be achieved when treatment with the same drug was initiated at a later time point. This benefit is consistent with a disease-modifying effect; however, the long-term significance of these findings is uncertain.

COMT INHIBITORS

When levodopa is administered with a decarboxylase inhibitor, it is primarily metabolized in the periphery by the catechol-O-methyl transferase (COMT) enzyme. Inhibitors of COMT increase the elimination half-life of levodopa and enhance its brain availability. Combining levodopa with a COMT inhibitor reduces "off" time and prolongs "on" time in fluctuating patients while enhancing motor scores. Two COMT inhibitors, tolcapone and entacapone, have been approved for use. More recently, opicapone (a long-acting, once daily COMT inhibitor) has been approved in Europe. There is also a combination tablet of levodopa, carbidopa, and entacapone (Stalevo®).

Side effects of COMT inhibitors are primarily dopaminergic (nausea, vomiting, increased dyskinesia) and can usually be controlled by down-titrating the dose of levodopa by 20–30%. Severe diarrhea has been described with tolcapone, and to a lesser degree with entacapone, and necessitates stopping the medication in 5–10% of individuals. Cases of fatal hepatic toxicity have been reported with tolcapone. It is still used because it is the most effective of the COMT inhibitors, but periodic monitoring of liver function is required. This problem has not been encountered with entacapone. Discoloration of urine can be seen with COMT inhibitors due to accumulation of a metabolite, but it is of no clinical concern.

It has been proposed that initiating levodopa in combination with a COMT inhibitor to enhance its elimination half-life could provide more continuous levodopa delivery and reduce the risk of motor complications if administered at frequent intervals. While this result has been demonstrated in a preclinical MPTP model, and continuous infusion reduces both "off" time and dyskinesia in advanced PD patients, no benefit of initiating levodopa with a COMT inhibitor compared to levodopa alone was detected in early PD patients in the STRIDE-PD study. This may have been because the combination was not administered at frequent enough intervals to provide continuous levodopa availability. For now, the main value of COMT inhibitors continues to be in patients who experience motor fluctuations.

OTHER MEDICAL THERAPIES

Centrally acting anticholinergic drugs such as trihexyphenidyl and benztropine were used historically for the treatment of PD, but they lost favor with the introduction of dopaminergic agents. Their major clinical effect is on tremor, although it is not certain that this benefit is superior to what can be obtained with agents such as levodopa and dopamine agonists. Still, they can be helpful in individual patients with severe tremor. Their use is limited particularly in the elderly, due to their propensity to induce a variety of side effects including urinary dysfunction, glaucoma, and particularly cognitive impairment.

Amantadine was originally introduced as an antiviral agent, but was appreciated to also have antiparkinsonian effects that are thought to be due to *N*-methyl-D-aspartate (NMDA) receptor antagonism. While some physicians use amantadine in patients with early disease for its mild symptomatic effects, it is most widely used as an antidyskinesia agent in patients with advanced PD. Indeed, it is the only oral agent that has been demonstrated in controlled studies to reduce dyskinesia without worsening parkinsonian features, although benefits may be relatively transient. Cognitive impairment is a major concern. Other side effects include livedo reticularis and weight gain. Amantadine should always be discontinued gradually because patients can experience withdrawal-like symptoms. An extended release formulation of amantadine has recently been approved in the United States.

The anticonvulsant zonisamide has also been shown to have antiparkinsonian effects and is approved for use in Japan. Its mechanism of action is unknown.

Several new classes of drugs are currently being investigated in an attempt to enhance antiparkinsonian effects, reduce off time, and treat or prevent dyskinesia. These include adenosine A_{2A} antagonists, nicotinic agonists, glutamate antagonists, and 5-HT_{1A} agonists. The A_{2A} antagonist Istradefylline is approved in Japan.

A list of the major drugs and available dosage strengths currently available to treat PD is provided in **Table 427-5**.

NEUROPROTECTION

Despite the many therapeutic agents available for the treatment of PD, patients continue to progress and to develop intolerable disability. A neuroprotective therapy that slows or stops disease progression remains the major unmet therapeutic need. Numerous trials have shown positive results (e.g., selegiline, rasagiline, pramipexole, ropinirole) consistent with a disease-modifying effect. However, it has not been possible to determine with certainty if the positive results were due to neuroprotection with slowing of disease progression or confounding symptomatic or pharmacologic effects that mask disease progression. There is a flurry of clinical activity testing interventions targeting etiopathogenic factors; these include calcium channel blockers, urate, and agents that enhance glucocerebrocidase (GCase) or interfere with SNCA or LRRK2 in the hope that they might provide disease-modifying effects. A major limitation is the uncertainty as to a specific clinical development plan and trial design that will prove acceptable to both clinicians and regulatory authorities.

TABLE 427-5 Drugs Commonly Used for Treatment of Parkinson's Disease[a]

AGENT	AVAILABLE DOSAGES	TYPICAL DOSING
Levodopa[a]		
Carbidopa/levodopa	10/100, 25/100, 25/250 mg	200–1000 mg levodopa/day
Benserazide/levodopa	25/100, 50/200 mg	
Carbidopa/levodopa CR	25/100, 50/200 mg	
Benserazide/levodopa MDS	25/200, 25/250 mg	
Parcopa	10/100, 25/100, 25/250	
Rytary (carbidopa/levodopa)	23.75/95, 36.25/145, 48.75/195, 61.25/245	See conversion tables
Carbidopa/levodopa/entacapone	12.5/50/200, 18.75/75/200, 25/100/200, 31.25/125/200, 37.5/150/200, 50/200/200 mg	
Dopamine agonists		
Pramipexole	0.125, 0.25, 0.5, 1.0, 1.5 mg	0.25–1.0 mg tid
Pramipexole ER	0.375, 0.75, 1.5. 3.0, 4.5 mg	1–3 mg/d
Ropinirole	0.25, 0.5, 1.0, 3.0 mg	6–24 mg/d
Ropinirole XL	2, 4, 6, 8 mg	6–24 mg/d
Rotigotine patch	2-, 4-, 6-, 8-mg patches	4–24 mg/d
Apomorphine SC	2-8 mg	2–8 mg
COMT inhibitors		
Entacapone	200 mg	200 mg with each levodopa dose
Tolcapone	100, 200 mg	100–200 mg tid
Opicapone	50 mg	50 mg HS
MAO-B inhibitors		
Selegiline	5 mg	5 mg bid
Rasagiline	0.5, 1.0 mg	mg QAM
Safinamide	100 mg	100 mg QAM

[a]Treatment should be individualized. Generally, drugs should be started in low doses and titrated to optimal dose.

Note: Drugs should not be withdrawn abruptly but should be gradually lowered or removed as appropriate.

Abbreviations: COMT, catechol-*O*-methyltransferase; MAO-B, monoamine oxidase type B; QAM, every morning.

SURGICAL TREATMENT

Surgical treatments for PD have been used for more than a century. Lesions were initially placed in the motor cortex and improved tremor but were associated with motor deficits, and this approach was abandoned. Subsequently, it was appreciated that lesions placed into the ventral intermediate (VIM) nucleus of the thalamus reduced contralateral tremor without inducing hemiparesis, but these lesions did not meaningfully help other more disabling features of PD. In the 1990s, it was shown that lesions placed in the posteroventral portion of the GPi (motor territory) improved rigidity and bradykinesia as well as tremor. Importantly, pallidotomy was also associated with marked improvement in contralateral dyskinesia. This procedure gained favor with greater understanding of the pathophysiology of PD (see above). However, this procedure is not optimal, because bilateral lesions are associated with side effects such as dysphagia, dysarthria, and impaired cognition. Lesions of the STN are associated with antiparkinsonian benefit and reduced levodopa requirement, but there is a concern about the risk of hemiballismus, and this procedure is not commonly performed.

Most surgical procedures for PD performed today use deep brain stimulation (DBS). Here, an electrode is placed into the target area and connected to a stimulator inserted subcutaneously over the chest wall. DBS simulates the effects of a lesion without necessitating making a brain lesion. The precise mechanism whereby DBS works is not fully resolved but may act by disrupting the abnormal neurophysiological signals associated with PD and motor complications. The stimulation variables can be adjusted with respect to electrode configuration, voltage, frequency, and pulse duration in order to maximize benefit and minimize adverse side effects. The procedure does not require making a lesion in the brain and is thus suitable for performing bilateral procedures with relative safety. In cases with intolerable side effects, stimulation can be stopped and the system removed.

DBS for PD primarily targets the STN or the GPi. It provides dramatic results, particularly with respect to tremor and reducing both "off" time and dyskinesias, but does not provide superior clinical benefits or improve features that do not respond to levodopa such as freezing, falling, and dementia. The procedure is thus primarily indicated for patients who suffer disability resulting from severe tremor, or levodopa-induced motor complications that cannot be satisfactorily controlled with drug manipulation. In such patients, DBS has been shown to provide benefits in comparison to best medical therapy. Side effects can be seen with respect to the surgical procedure (hemorrhage, infarction, infection), the DBS system (infection, lead break, lead displacement, skin ulceration), or the stimulation (ocular and speech abnormalities, muscle twitches, paresthesias, depression, and rarely suicide). Recent studies indicate that benefits following DBS of the STN and GPi are comparable, but that GPi stimulation may be associated with a reduced frequency of depression. Although not all PD patients are candidates, the procedure can be profoundly beneficial for many. Long-term studies demonstrate continued benefits with respect to the classic motor features of PD, but DBS does not prevent the development of nondopaminergic features, which continue to evolve and to be a source of disability. Studies continue to evaluate the optimal way to use DBS (low- vs high-frequency stimulation, closed loop systems, etc.). Studies of DBS in early PD patients show benefits in comparison to medical therapy, but this must be weighed against the cost of the procedure and the risk of side effects in patients who might otherwise be well controlled with medical therapies. Controlled studies comparing DBS to other therapies aimed at improving motor function without causing dyskinesia, such as Duodopa® and apomorphine infusions, remain to be performed. The utility of DBS may also be reduced in future years if new medical therapies are developed that provide the benefits of levodopa without motor complications. New targets for

DBS that might benefit gait dysfunction, depression, and cognitive impairment are being actively explored **(Chap. 477)**.

MRI-guided ultrasound is also now being used as a means of damaging critical target regions such as the GPi in PD patients with motor complications in a noninvasive manner that avoids the needs for a surgical procedure. Preliminary results suggest good target localization and safety.

EXPERIMENTAL THERAPIES FOR PD

There has been considerable scientific and public interest in a number of novel interventions that are being investigated as possible treatments for PD. These include cell-based therapies (such as transplantation of fetal nigral dopamine cells or dopamine neurons derived from stem cells), gene therapies, trophic factors, and therapies directed against gene-specific targets. Transplant strategies are based on the concept of implanting dopaminergic cells into the striatum to replace degenerating SNc dopamine neurons. Fetal nigral mesencephalic cells have been demonstrated to survive implantation, re-innervate the striatum in an organotypic manner, and restore motor function in PD models. However, two double-blind studies failed to show significant benefit of fetal nigral transplantation in comparison to a sham operation with respect to their primary endpoints. Additionally, grafting of fetal nigral cells is associated with a previously unrecognized form of dyskinesia (graft-induced dyskinesia) that persists after lowering or even stopping levodopa. This has been postulated to be related to suboptimal release of dopamine from grafted cells leading to a sustained form of diphasic dyskinesia. In addition, there is evidence that after many years, transplanted healthy embryonic dopamine neurons from unrelated donors develop PD pathology and become dysfunctional, suggesting transfer of α-synuclein from affected to unaffected neurons in a prion-like manner (see discussion above). Perhaps most importantly, it is not clear how replacing dopamine cells alone will improve nondopaminergic features such as falling and dementia, which are the major sources of disability for patients with advanced disease. While stem cells, and specifically induced pluripotent stem cells derived from the recipient, may overcome problems related to immunity, type and number of cells, and physiologic integration, many of these same concerns still apply. To date, stem cells have not yet been properly tested in PD patients and bear the additional concern of tumors and other unanticipated side effects. While there remains a need for scientifically based studies attempting to evaluate the potential role of cell-based therapies in PD, there is no scientific basis to warrant routine treatment of PD patients with stem cells as is being marketed in some countries.

Trophic factors are a series of proteins that enhance neuronal growth and restore function to damaged neurons. There are several different trophic factors that have been demonstrated to have beneficial effects on dopamine neurons in laboratory studies. Glial-derived neurotrophic factor (GDNF) and neurturin have attracted particular attention as possible therapies for PD. However, double-blind trials of intraventricular and intraputaminal infusions of GDNF failed to show benefits compared to placebo in PD patients, possibly because of inadequate delivery of the trophic molecule to the target region.

Gene therapy offers the potential of providing long-term expression of a therapeutic protein with a single procedure. Gene therapy involves placing the DNA of a therapeutic protein into a viral vector that can then be taken up and incorporated into the genome of host cells and then synthesized and released on a continual basis. The AAV2 virus has been most often used as the vector because it does not promote an inflammatory response, is not incorporated into the host genome, and is associated with long-lasting transgene expression. Clinical trials of AAV2 delivery of the trophic factor neurturin showed promising results in open label trials but failed in double-blind trials, even when injected into both the putamen and the SNc. This likely reflects α-synuclein-mediated downregulation of Nurr1 and RET receptors, thereby limiting the potential of the trophic factor to interact with its receptor and induce upregulation of repair genes. Gene delivery is also being explored as a means of delivering aromatic amino acid decarboxylase with or without tyrosine hydroxylase to the striatum to facilitate the conversion of orally administered levodopa to dopamine. Animal studies suggest that this approach can provide antiparkinsonian benefits with reduced motor complications, and clinical trials in PD patients are underway. Although gene delivery technology has great potential and will likely be used to deliver novel therapies in the future (e.g. Parkin), current approaches carry the risk of unanticipated side effects and do not address the nondopaminergic features of the illness.

MANAGEMENT OF THE NONMOTOR AND NONDOPAMINERGIC FEATURES OF PD

Although PD management has primarily focused on dopaminergic features, management of the nondopaminergic features should not be ignored. Some nonmotor features, although not thought to reflect dopaminergic pathology, nonetheless benefit from dopaminergic drugs. For example, problems such as anxiety, panic attacks, depression, pain, sweating, sensory problems, freezing, and constipation all tend to be worse during "off" periods and have been reported to improve with better dopaminergic control. Approximately 50% of PD patients suffer depression during the course of the disease, and depression is frequently underdiagnosed and undertreated. Antidepressants should not be withheld, particularly for patients with major depression, although dopaminergic agents such as pramipexole may prove helpful for both depression and PD motor features. Serotonin syndromes have been a theoretical concern with the combined use of SSRIs and MAO-B inhibitors but these problems are rarely encountered. Anxiety is also a common problem, and if not adequately controlled with better antiparkinsonian drugs can be treated with short-acting benzodiazepines.

Psychosis can be a problem for some PD patients, and is often a harbinger of developing dementia. In contrast to AD, hallucinations are typically visual, formed, and nonthreatening. Importantly, they can limit the use of dopaminergic agents necessary to obtain satisfactory motor control. They can be associated with dopaminergic drugs, and the first approach is typically to withdraw agents that are less effective than levodopa such as anticholinergics, amantadine, and dopamine agonists followed by lowering the dose of levodopa if possible. Psychosis in PD often responds to low doses of atypical neuroleptics and may permit higher doses of levodopa to be tolerated. Clozapine is an effective drug, but it can be associated with agranulocytosis, and regular monitoring is required. Quetiapine avoids these problems but it has not been established to be effective in placebo-controlled trials. Pimavanserin (Nuplazid®) differs from other atypical neuroleptics in that it is also an inverse agonist of the serotonin 5-HT_{2A} receptor. It has been shown to be effective in double-blind trials with a relatively good safety profile. It was recently approved for use in the United States.

Dementia in PD (PDD) is common, ultimately affecting as many as 80% of patients. Its frequency increases with aging and, in contrast to AD, primarily affects executive functions and attention, with relative sparing of language, memory, and calculation domains. When dementia precedes or develops within 1 year after the onset of motor dysfunction, it is by convention referred to as dementia with Lewy bodies (DLB; **Chap. 426**). These patients are particularly prone to have hallucinations and diurnal fluctuations. Pathologically, DLB is characterized by Lewy bodies distributed throughout the cerebral cortex (especially the hippocampus and amygdala) and is often also associated with AD pathology. It is likely that DLB and PD with dementia represent a spectrum of PD rather than separate disease entities. Mild cognitive impairment (MCI) frequently precedes the onset of dementia and is a more reliable index of impending PDD than in the general population. Dopaminergic drugs can worsen cognitive function in demented patients and should be stopped or reduced to try and provide a compromise between antiparkinsonian benefit and preserved cognitive function. Drugs are usually discontinued in the following sequence: anticholinergics, amantadine, dopamine agonists, COMT inhibitors, and MAO-B inhibitors.

Eventually, patients with cognitive impairment should be managed with the lowest dose of standard levodopa that provides meaningful antiparkinsonian effects and does not worsen mental function. Anticholinesterase agents such as memantine, rivastigmine, and donepezil reduce the rate of deterioration of measures of cognitive function and can improve attention in PD, but do not typically improve cognitive function in any meaningful way. More effective therapies that treat or prevent dementia are a critical unmet need in the therapy of PD.

Autonomic disturbances are common and frequently require attention. Orthostatic hypotension can be problematic and contribute to falling. Initial treatment should include adding salt to the diet and elevating the head of the bed to prevent overnight sodium natriuresis. Low doses of fludrocortisone (Florinef) or midodrine provide control for most cases. The norepinephrine precursor 3-0-methylDOPS (Droxidopa®) has been shown to provide mild and transient benefits for patients with orthostatic hypotension, and was recently approved by the U.S. Food and Drug Administration. Vasopressin and erythropoietin can be used in more severe or refractory cases. If orthostatic hypotension is prominent in early parkinsonian cases, a diagnosis of MSA should be considered **(Chap. 432)**. Sexual dysfunction can be helped with sildenafil or tadalafil. Urinary problems, especially in males, should be treated in consultation with a urologist to exclude prostate problems. Anticholinergic agents, such as oxybutynin (Ditropan), may be helpful. Constipation can be a very important problem for PD patients. Mild laxatives or enemas can be useful, but physicians should first ensure that patients are drinking adequate amounts of fluid and consuming a diet rich in bulk with green leafy vegetables and bran. Agents that promote gastrointestinal (GI) motility can also be helpful.

Sleep disturbances are common in PD patients, with many experiencing fragmented sleep with excess daytime sleepiness. Restless leg syndrome, sleep apnea, and other sleep disorders should be treated as appropriate. REM behavior disorder (RBD) is a syndrome composed of violent movements and vocalizations during REM sleep, possibly representing acting out of dreams due to a failure of motor inhibition that typically accompanies REM sleep **(Chap. 27)**. Low doses of clonazepam (0.5–1 mg at bedtime) are usually effective in controlling this problem. Consultation with a sleep specialist and polysomnography may be necessary to identify and optimally treat sleep problems. Many PD patients have a history of RBD preceding the onset of the classic motor features of PD, and most cases of RBD go on to develop an α-synucleinopathy (PD or MSA).

NONPHARMACOLOGIC THERAPY

Gait dysfunction with falling is an important cause of disability in PD. Dopaminergic therapies can help patients whose gait is worse in "off" time, but there are currently no specific therapies for gait dysfunction. Canes and walkers may become necessary to increase stability and reduce the risk of falling. An effective therapy for gait impairment is an important unmet need in PD.

Freezing, where patients suddenly become stuck in place for seconds to minutes as if their feet were glued to the ground, is a major cause of falling. Freezing may occur during "on" or "off" periods. Freezing during "off" periods may respond to dopaminergic therapies, but there are no specific treatments for "on" period freezing. Some patients will respond to sensory cues such as marching in place, singing a song, or stepping over an imaginary line.

Speech impairment is another source of disability for many advanced PD patients. Speech therapy programs may be helpful, but benefits are generally transient.

Exercise has been shown to maintain and even improve function for PD patients, and active and passive exercises with full range of motion reduce the risk of arthritis and frozen joints. Some laboratory studies suggest the possibility that exercise might also have neuroprotective effects, but this has not been confirmed in PD. Exercise is generally recommended for all PD patients. It is less clear that physical therapy or specific exercise programs such as tai chi or dance offer any specific advantage. It is important for patients to maintain social and intellectual activities to the extent possible. Education, assistance with financial planning, social services, and attention to home safety are important elements of the overall care plan. Information is available through numerous PD foundations and on the web, but should be reviewed with physicians to ensure accuracy. The needs of the caregiver should not be neglected. Caring for a person with PD involves a substantial work effort and there is an increased incidence of depression among caregivers. Support groups for patients and caregivers may be useful.

CURRENT MANAGEMENT OF PD

The management of PD should be tailored to the needs of the individual patient, and there is no single treatment approach that is universally accepted and applicable to all individuals. Clearly, if an agent could be demonstrated to have disease-modifying effects, it should be initiated at the time of diagnosis. Indeed, recent studies suggest that dopamine terminal degeneration may be complete within 4 years of diagnosis. Epidemiologic and pathologic studies suggest that constipation, RBD, and anosmia may represent premotor features of PD and could permit diagnosis and the initiation of a disease-modifying therapy even prior to the onset of the classical motor features of the disease. However, no therapy has yet been conclusively proven to be a disease-modifying agent. For now, physicians must use their judgment in deciding whether or not to introduce a drug such as rasagiline (see above) for its possible disease-modifying effects based on available preclinical and clinical information.

The next important issue to address is when to initiate symptomatic therapy. Several studies suggest that it may be best to start therapy at the time of diagnosis in order to preserve beneficial compensatory mechanisms and possibly provide functional benefits even in the early stage of the disease. Levodopa remains the most effective symptomatic therapy for PD, and some recommend starting it immediately using low doses (≤400 mg/d), as motor complications have now clearly been shown to be dose-related. Others, however, prefer to delay levodopa treatment, particularly in younger patients, in order to reduce the risk of inducing motor complications entirely. An alternate approach is to begin with a MAO-B inhibitor and/or a dopamine agonist, and reserve levodopa for later stages when these drugs no longer provide satisfactory control. In making this decision, the age, degree of disability, and side effect profile of the drug must all be considered. In patients with more severe disability, the elderly, those with cognitive impairment, those with significant comorbidities, or those in whom the diagnosis is uncertain, most physicians would initiate therapy with levodopa. Regardless of initial choice, most patients ultimately require polypharmacy (combination of levodopa, an MAO-B inhibitor, and a dopamine agonist). While it is important to use low doses of each agent in order to reduce the risk of side effects, it is important not to deny patients levodopa when they cannot be adequately controlled with alternative medications.

If motor complications develop, patients can initially be treated by manipulating the frequency and dose of levodopa or by combining lower doses of levodopa with a dopamine agonist, a COMT inhibitor, or a MAO-B inhibitor. Amantadine is the only drug that has been demonstrated to treat dyskinesia without worsening parkinsonism, but benefits may be short-lasting, and there are important side effects related to cognitive function. In advanced cases, it may be necessary to consider a surgical therapy such as DBS or Duodopa® if the patient is a suitable candidate, but as described above, these procedures have their own set of complications. The use of DBS in early PD patients has been advocated by some, but there is considerable skepticism about this approach considering the costs and potential side effects, when inexpensive, well tolerated, and effective medical alternatives are available. Continuous intraintestinal infusion of levodopa/carbidopa intestinal gel (Duodopa) appears to offer similar benefits to DBS, but also requires a surgical intervention with potentially serious complications. Continuous infusion of apomorphine is a treatment option that does not require

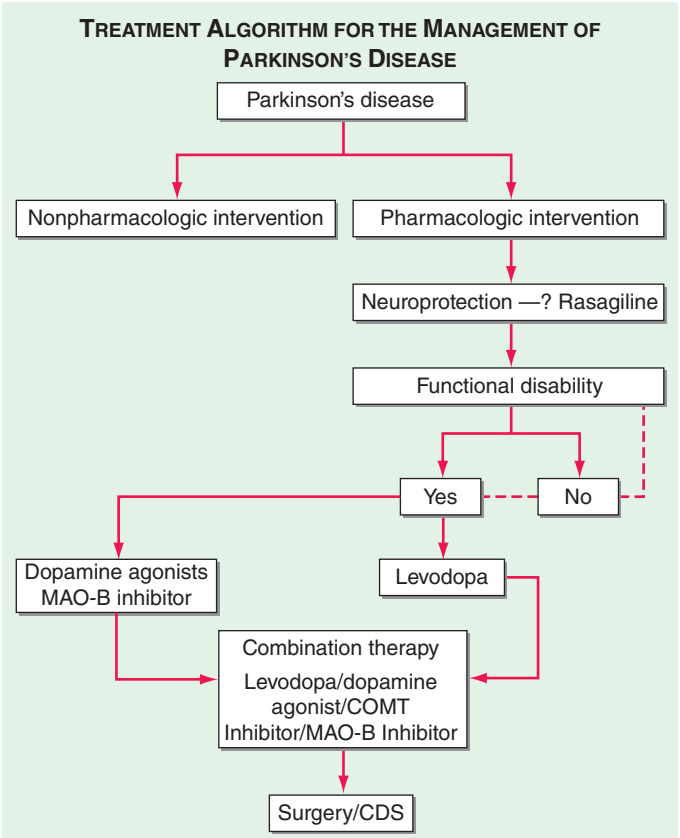


FIGURE 427-7 Treatment options for the management of Parkinson's disease (PD). Decision points include: (1) Introduction of a neuroprotective therapy: no drug has been established to have or is currently approved for neuroprotection or disease modification, but there are several agents that have this potential based on laboratory and preliminary clinical studies (e.g., rasagiline 1 mg/d, coenzyme Q10 1200 mg/d, the dopamine agonists ropinirole, and pramipexole). (2) When to initiate symptomatic therapy: There is a trend toward initiating therapy at the time of diagnosis or early in the course of the disease because patients may have some disability even at an early stage, and there is the possibility that early treatment may preserve beneficial compensatory mechanisms; however, some experts recommend waiting until there is functional disability before initiating therapy. (3) What therapy to initiate: many experts favor starting with a monoamine oxidase type B (MAO-B) inhibitor in mildly affected patients because of the good safety profile of the drug and the potential for a disease-modifying effect; dopamine agonists for younger patients with functionally significant disability to reduce the risk of motor complications; and levodopa for patients with more advanced disease, the elderly, or those with cognitive impairment. Recent studies suggest the early employment of polypharmacy using low doses of multiple drugs to avoid side effects associated with high doses of any one agent. (4) Management of motor complications: motor complications are typically approached with combination therapy to try and reduce dyskinesia and enhance the "on" time. When medical therapies cannot provide satisfactory control, surgical therapies such as DBS or continuous infusion of levodopa/carbidopa intestinal gel can be considered. (5) Nonpharmacologic approaches: interventions such as exercise, education, and support should be considered throughout the course of the disease. CDS, continuous dopaminergic stimulation; COMT, catechol-O-methyltransferase. *(Adapted from CW Olanow et al: Neurology 72:S1, 2009.)*

surgery but is associated with potentially troublesome skin nodules. Comparative studies of these approaches in more advanced patients are awaited. There are ongoing efforts aimed at developing a long-acting oral or subcutaneous formulation of levodopa that mirrors the pharmacokinetic properties of a levodopa infusion. Such a formulation might provide all of the benefits of levodopa without motor complications and avoid the need for polypharmacy and surgical intervention.

A decision tree that considers the various treatment options and decision points for the management of PD is provided in Fig. 427-7.

FURTHER READING

ASCHERIO A, SCHWARZSCHILD MA: The epidemiology of Parkinson's disease: Risk factors and prevention. Lancet Neurol 15:1257, 2016.

BERG D et al: MDS research criteria for prodromal Parkinson's disease. Mov Disord 12:1600, 2015.

DORSEY ER et al: Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. Neurology 68:384, 2007.

HERNANDEZ DG et al: Genetics in Parkinson disease: Mendelian versus non-Mendelian inheritance. J Neurochem 139 Suppl 1:59, 2016.

HÖGLINGER GU et al: Clinical diagnosis of progressive supranuclear palsy: The movement disorder society criteria. Movement Disorder Society-endorsed PSP Study Group. Mov Disord 32:853, 2017.

MARRAS C et al: Nomenclature of genetic movement disorders: Recommendations of the International Parkinson and Movement Disorder Society task force. Mov Disord 32:724, 2017.

OBESO JA et al: Past, present and future of Parkinson's disease: A special essay on the 200th Anniversary of the Shaking Palsy. Mov Disord 32:1264, 2017.

OLANOW CW et al: Scientific and clinical basis for the treatment of PD—2009. Neurology 72:S1, 2009.

POSTUMA RB et al: MDS clinical diagnostic criteria for Parkinson's disease. Mov Disord 12:1591, 2015.

SCHAPIRA AH et al: Slowing of neurodegeneration in Parkinson's disease and Huntington's disease: Future therapeutic perspectives. Lancet 384:545, 2014.

SCHAPIRA AHV et al: Non-motor features of Parkinson disease. Nat Rev Neurosci 18:435, 2017.

428 Tremor, Chorea, and Other Movement Disorders

C. Warren Olanow, Christine Klein, Jose A. Obeso

HYPERKINETIC MOVEMENT DISORDERS

Hyperkinetic movement disorders are characterized by involuntary movements unaccompanied by weakness (Table 428-1). This term is somewhat arbitrary and potentially misleading as hypokinetic disorders such as Parkinson's disease (PD) are often accompanied by tremor which is a hyperkinetic feature, and hyperkinetic disorders such as dystonia may manifest slow movements because of the severe muscle contractions. Nonetheless, the terms continue to be used because of convention. The major hyperkinetic movement disorders and the diseases with which they are associated are considered in this section.

TREMOR

CLINICAL FEATURES

Tremor consists of alternating contractions of agonist and antagonist muscles in an oscillating, rhythmic manner. It can be most prominent at rest (rest tremor), on assuming a posture (postural tremor), on actively reaching for a target (kinetic tremor), or on carrying out a movement (action tremor). Tremor may also be characterized based on distribution, frequency, amplitude, and related neurologic dysfunction.

PD (Chap. 427) is characterized by a resting tremor, essential tremor (ET) by a tremor that typically occurs while trying to sustain a posture coupled with an action tremor, and cerebellar dysfunction by a kinetic tremor and is usually associated with hypotonia and past pointing. Normal individuals can have a physiologic tremor that typically manifests as a mild, high-frequency (10–12 Hz), postural or action tremor typically affecting the upper extremities. This tremor is usually of no clinical consequence and often is only appreciated with an

TABLE 428-1 Hyperkinetic Movement Disorders	
Tremor	Rhythmic oscillation of a body part due to intermittent muscle contractions
Dystonia	Involuntary, patterned, sustained, or repeated muscle contractions often associated with twisting movements and abnormal posture
Athetosis	Slow, distal, writhing, involuntary movements with a propensity to affect the arms and hands (this represents a form of dystonia with increased mobility)
Chorea	Rapid, semi-purposeful, graceful, dance-like nonpatterned involuntary movements involving distal or proximal muscle groups. When the movements are of large amplitude and predominant proximal distribution, the term *ballism* is used.
Myoclonus	Sudden, brief (<100 ms), jerk-like, arrhythmic muscle twitches
Tic	Brief, repeated, stereotyped muscle contractions that can often be suppressed for a short time. These can be simple and involve a single muscle group or complex and affect a range of motor activities.

accelerometer or under stress. An enhanced physiologic tremor (EPT) can be seen in up to 10% of the population, and tends to occur in association with anxiety, fatigue, a metabolic disturbance (e.g., hyperthyroidism, electrolyte abnormalities), drugs (e.g., valproate, lithium), or toxins (e.g., caffeine, smoking, alcohol). Treatment is initially directed at control of any underlying disorder and, if necessary, can often be improved with a beta blocker.

ESSENTIAL TREMOR

ET is the most common movement disorder, affecting ~5% of the population (an estimated 5–10 million persons in the United States or Western Europe). It can present in childhood but dramatically increases in prevalence in those aged >70 years. ET is characterized by a high-frequency tremor (6–10 Hz) that predominantly affects the upper extremities. The tremor is most often manifest as a postural or action tremor and, in severe cases, can interfere with functions such as eating and drinking. It is typically bilateral and symmetric but may begin on one side and remain asymmetric. Patients with severe ET can have an intention tremor with overshoot and slowness of movement suggesting the possibility of a cerebellar origin. Tremor involves the head in ~30% of cases, voice in ~20%, tongue in ~20%, face/jaw in ~10%, and lower limbs in ~10%. Multiple body parts are involved in at least 50% of cases. The tremor is characteristically improved by alcohol and worsened by stress. Subtle impairment of coordination or tandem walking may be present, and disturbances of hearing, cognition, personality, mood, and olfaction have been described, but usually the neurologic examination is normal aside from tremor. The differential diagnosis includes dystonic tremor (see below) or PD. PD can usually be differentiated from ET because the former stops at the onset of a voluntary action and is typically associated with bradykinesia with progressive slowing of sequential movements (sequence effect), rigidity, gait and postural instability, and other parkinsonian features. However, the examiner should be aware that PD patients may have a postural tremor and ET patients may develop a rest tremor, and that these typically begin after a latency of a few seconds (emergent tremor). In contrast to the micrographia of PD, ET patients have relatively large handwriting with evidence of the effect of tremor. The examiner must also differentiate the effect of tremor when assessing tone in ET to distinguish this from the cogwheel rigidity found in PD.

ETIOLOGY AND PATHOPHYSIOLOGY

The etiology and pathophysiology of ET are not known. Approximately 50% of cases have a positive family history with an autosomal dominant pattern of inheritance. Linkage studies have detected possibly linked loci in large ET families, but no independently confirmed causative genes have been identified to date. It is likely, however, that there are undiscovered genes for ET that have escaped detection to date because of the heterogeneity of the syndrome and the high population frequency of ET likely resulting in a large number of phenocopies, (i.e., family members with a similar clinical syndrome, but not carrying the causative mutation). The cerebellum and inferior olives have been implicated as possible sites of a "tremor pacemaker" based on the presence of cerebellar signs in about 10% of ET patients, and increased metabolic activity and blood flow in these regions in some patients. Some pathologic studies have described cerebellar pathology with a loss of Purkinje cells and axonal torpedoes, but these findings are controversial, and the precise pathologic correlate of ET remains to be defined. It is likely that multiple causes of ET will ultimately be identified.

TREATMENT

Many cases are mild and require no treatment other than reassurance. Occasionally, tremor can be severe and interfere with eating, writing, and activities of daily living. This is more likely to occur as the patient ages and is often associated with a reduction in tremor frequency. Beta blockers and primidone are the standard drug therapies for ET and help in about 50% of cases. Propranolol (20–120 mg daily, given in divided doses) is usually effective at relatively low doses, but higher doses may be needed in some patients. The drug is contraindicated in patients with bradycardia or asthma. Hand tremor tends to be most improved, while head tremor is often refractory. Primidone can be helpful but should be started at low doses (12.5 mg) and gradually increased (125–250 mg tid) to avoid sedation, nausea, and dizziness. Benefits have also been reported with gabapentin and topiramate, but these drugs have not been widely employed. Botulinum toxin injections may be helpful for limb or voice tremor, but treatment can be associated with muscle weakness. Surgical therapies targeting the ventro-intermediate (VIM) nucleus of the thalamus can be very effective for severe and drug-resistant cases. Recently, focal ultrasound (which is a procedure that does not require surgery) has also been shown to be an effective therapy against tremor in ET.

DYSTONIA

CLINICAL FEATURES

Dystonia is a movement disorder characterized by sustained or intermittent muscle contractions of antagonist muscles causing abnormal often repetitive movements and postures. Dystonic movements are typically patterned and twisting and may be associated with a "dystonic tremor". Dystonia is often initiated or worsened by voluntary action and associated with overflow muscle activation. Dystonia can range from focal minor contractions affecting only an individual muscle group to severe and disabling involvement of multiple muscle groups. The frequency is estimated to be 16 per 100,000 (~50,000 cases in the United States) but is likely to be much higher because many cases are not recognized. Dystonia is often brought out by voluntary movements (action dystonia) and can extend to involve other muscle groups and body regions not required for a given action (overflow). It can be aggravated by stress and fatigue and attenuated by relaxation and sensory tricks such as touching the affected body part (geste antagoniste).

Historically, dystonia has been described as primary or secondary. However, because of a confusing and not always congruent combination of phenotypic and etiologic features, the older terms are no longer recommended. A Movement Disorder Society Task Force charged with redefining dystonia recommends classifying dystonia along two main axes: clinical and etiologic. On clinical grounds, dystonia can be categorized by age of onset (infancy, childhood, adolescence, early and late adulthood), body distribution (focal, segmental, multifocal, and generalized), temporal pattern (static or progressive, action-specific [diurnal and paroxysmal]), and association with additional features. Clinical description along these lines enables formulating specific dystonia syndromes (e.g., early-onset generalized isolated dystonia).

Etiology of dystonia primarily reflects genetic abnormalities, although occasionally there may be other causes such as trauma and stroke. Genetic features used for classification include mode of inheritance or identification of a specific genetic defect. In the past three decades, more than 200 genes have been linked to different, mainly childhood-onset and generalized forms of dystonia. These include

forms in which dystonia is the only disease manifestation with the exception of tremor ("isolated dystonia"), forms in which dystonia co-occurs with another movement disorder such as parkinsonism or myoclonus ("combined dystonia") and disorders in which dystonia is only one of several clinical manifestations and may be a less prominent or even inconsistent feature ("complex dystonia"). Most of the genetic forms belong to the latter phenotypic group, which also represents the most heterogeneous class in terms of clinical expression.

ISOLATED DYSTONIAS

Focal (Multifocal, Segmental) Dystonia Adult-onset, focal dystonia is by far the most frequent form of isolated dystonia, with women being affected about twice as often as men. Focal dystonia typically presents in the fourth to sixth decade and can be focal, multifocal, or segmental. The major clinical phenotypes are as follows: (1) *Cervical dystonia*—dystonic contractions of neck muscles causing the head to deviate to one side (*laterocollis*), twist (*torticollis*), move in a forward direction (*anterocollis*), or move in a backward direction (*retrocollis*). Muscle contractions can be painful and occasionally can be complicated with a secondary cervical radiculopathy. (2) *Blepharospasm*—dystonic contractions of the eyelids with increased blinking that can interfere with reading, watching television, working on a computer, and driving. This can sometimes be so severe as to cause functional blindness. (3) *Oromandibular dystonia* (OMD)—contractions of muscles of the lower face, lips, tongue, and jaw (opening or closing). Meige's syndrome is a combination of OMD and blepharospasm that predominantly affects women aged >60 years. (4) *Spasmodic dysphonia*—dystonic contractions of the vocal cords during phonation, causing impaired speech. Most cases affect the adductor muscles and cause speech to have a choking or strained quality. Less commonly, the abductors are affected, leading to speech with a breathy or whispering quality. (5) *Limb dystonias*—these can be present in either arms or legs and are often brought out by task-specific activities such as handwriting (writer's cramp), playing a musical instrument (musician's cramp), or putting in golf (the yips). The vast majority of patients have dystonia of the neck (cervical dystonia; ~50%) or the eye lid (blepharospasm; ~20%). Focal hand or leg dystonia (~5%), spasmodic dysphonia (~2%), musician's dystonia (~3%), or OMD (~1%) are much less common. Focal dystonias can extend to involve other body regions (about 30% of cases) and are frequently misdiagnosed as psychiatric or orthopedic in origin. Their cause is usually not known, but genetic factors, autoimmunity, and trauma have been suggested. Focal dystonias are often associated with a high-frequency tremor that can resemble ET. Dystonic tremor can usually be distinguished from ET because it tends to occur in conjunction with the dystonic contraction and disappears when the dystonia is relieved (e.g., turning the head in the opposite direction of the dystonia).

GENERALIZED DYSTONIA

Generalized dystonia is often hereditary in nature and, unlike focal dystonia, generally has an age of onset in childhood or adolescence. There are currently at least four well established genes for *isolated* dystonia; *TOR1A, THAP1, ANO3*, and *GNAL*. According to the recommendations of the International Parkinson's Disease and Movement Disorder Society, confirmed monogenic forms are classified according to absence or presence of accompanying clinical features and preceded by a "DYT" prefix, e.g., DYT-TOR1A. These genetic forms are all inherited in an autosomal dominant fashion and found in <5% of dystonia patients. Not all mutation carriers develop generalized dystonia; about 35% remain unaffected despite harboring a pathogenic mutation (reduced penetrance), and rarely they present with dystonia that remains focal or segmental in nature.

Mutations in the *TOR1A* gene (torsin family 1 member A—*formerly known as the DYT1 gene*) are the most common cause of early-onset generalized dystonia. The first, and currently the only clearly established mutation, is a 3-base pair deletion in the *TOR1A* gene. The mutation is frequently found among Ashkenazi Jewish patients due to a founder effect. Mutation carriers usually present with dystonia in an extremity in childhood that later progress to other body parts, but typically spare the face and neck.

THAP1 gene (*THAP domain containing, apoptosis associated protein 1*) mutations have been linked to adolescent-onset dystonia with mixed phenotype. About 100 different mutations have been reported in *THAP1*. Mutations typically manifest with dysphonia or writer's cramp beginning in late childhood or adolescence. Over the course of the disease, dystonia spreads to other body parts with prominent craniocervical involvement.

Mutations in the *ANO3* gene (*anoctamin 3*) were first reported in patients with predominantly craniocervical dystonia with a broad range of ages of onset. While a large number of missense variants can be found in healthy individuals, a pathogenic role of *ANO3* mutations has recently been supported by the description of additional families with dystonia and myoclonic jerks.

Mutations in the *GNAL* gene (*guanine nucleotide-binding protein subunit alpha L*) are a rare cause of cervical or cranial dystonia with a mean age of onset in the thirties. About 30 different *GNAL* mutations have been reported in dystonia patients.

In addition to the above, missense mutations in *KMT2B* (*lysine methyltransferase 2B*) have recently been identified, and confirmed to be a cause of an early-onset generalized dystonia which may be accompanied by other syndromic features including intellectual disability, microcephaly, psychiatric features, dysmorphia, or skin lesions. The majority of the mutations occurred de novo. *KMT2B* mutations may account for up to 10% of early-onset generalized dystonia but further validation is warranted and placement into the group of isolated vs complex dystonias is currently under debate.

Combined Dystonia A number of other well-established genes have been described for combined forms of dystonia in which dystonia occurs in conjunction with a different movement disorder, such as parkinsonism or myoclonus.

Dopa-responsive dystonia (DRD; also known as Segawa syndrome) is caused by mutations in the *GCH1* gene (*GTP cyclohydrolase-1*) that encodes for the rate-limiting enzyme in the biosynthesis of dopamine via the biopterin pathway. It is manifest as a childhood-onset form of dystonia with diurnal fluctuations and is important to recognize as the condition dramatically responds to low doses of levodopa. Parkinsonism can be a major, or even the only finding, and there may be a presynaptic dopaminergic deficit as evidenced by SPECT. To date, more than 100 different mutations have been reported with a penetrance of around 50% which is considerably higher in women compared to men. Recessively inherited (biallelic) mutations in *GCH1* result in a much more severe clinical phenotype with developmental delay and infantile onset. Due to the enzymatic defect in the levodopa biosynthesis, there is a lifelong and dramatic response to levodopa therapy. Indeed, all young onset forms of dystonia should be tested with levodopa to exclude the possibility of DRD.

X-linked dystonia-parkinsonism (Lubag) is a combined form of dystonia and parkinsonism that is found exclusively in patients of Filipino origin due to a founder effect and seems to be fully penetrant. Patients usually develop focal (cranial) dystonia first that rapidly generalizes and, after 5–10 years, is gradually replaced by a form of L-dopa-unresponsive parkinsonism. The exact mutation causing X-linked dystonia-parkinsonism (Lubag) is not yet known but several variants in a disease haplotype segregate with the disease and a retrotransposon insertion in the *TAF1* (*TATA-Box Binding Protein Associated Factor 1*) gene has been suggested as the most likely disease cause.

Biallelic mutations in the *PRKRA* (*protein activator of interferon-induced protein kinase EIF2AK2*) gene are linked to a dystonia-parkinsonism syndrome and mostly due to the same missense mutation that seems to result from a shared founder. The phenotype includes early-onset generalized dystonia, often with laryngeal dystonia, tongue protrusion, prominent oromandibular involvement, dysphagia, and retrocollis. Parkinsonian features are mild (or even absent) and do not respond to levodopa therapy.

Mutations in the *ATP1A3* (*ATPase Na^+/K^+ transporting subunit alpha 3*) gene present with a characteristic, sudden onset usually in adolescence or young adulthood, often triggered by high fever, physical exertion, or emotional stress. Dystonic symptoms frequently show a rostrocaudal

gradient with a strong involvement of the bulbar region and are often accompanied by bradykinesia as a parkinsonian feature. In addition, mutations in *ATP1A3* have also been linked to a variety of clinical syndromes (pleiotropy) including epileptic or hemiplegic attacks, ataxia, cognitive decline, and other neurological disorders, often with a more severe course and an earlier age at onset.

Myoclonic-dystonia is characterized by action-induced, alcohol-responsive myoclonic jerks predominantly involving the upper body half. Onset is usually in childhood or adolescence. Additionally, many individuals develop psychiatric features such as depression, anxiety-related disorders, and alcohol dependence. The disorder is primarily related to mutations in the *SGCE* gene (*sarcoglycan epsilon*) which codes for the ε member of the sarcoglycan family. About 80 different mutations have been reported in *SGCE* including deletions of the entire gene. The latter type of mutation often also involves loss of adjacent genes leading to additional clinical features such as joint problems. *SGCE* mutations are incompletely penetrant and only manifest when inherited from the father due to the epigenetic effect of maternal imprinting of *SGCE*.

A number of additional monogenic causes have been suggested for isolated and combined forms of dystonia but still await independent confirmation. **Table 428-2** provides a list of the confirmed Monogenic Forms of Isolated and Combined Dystonias.

COMPLEX DYSTONIAS

In the complex dystonias, dystonia is one part of a syndrome with multiple different disease manifestations. Most frequently, they are hereditary such as Wilson's disease (WD), Huntington's disease (HD), Lesh Nyhan syndrome, corticobasal ganglionic disorders, and a variety of other neurologic, neurometabolic, and mitochondrial disorders. Complex dystonias may also develop as a consequence of drugs or toxins (previously referred to as secondary dystonias). Drug-induced dystonia may be acute or chronic, and is most commonly seen with neuroleptic drugs or after chronic levodopa treatment in PD patients. Dystonia can also be observed following discrete lesions in the striatum, and occasionally in the pallidum, thalamus, cortex, and brainstem due to infarction, hemorrhage anoxia, trauma, tumor, infection, or toxins such as manganese or carbon monoxide. In these cases, dystonia often assumes a segmental distribution, but may be generalized when lesions are bilateral or widespread. More rarely, dystonia can develop following peripheral nerve injury and be associated with features of complex regional pain syndrome **(Chap. 432)**. A psychogenic origin is responsible for some cases of dystonia; these typically present with fixed, immobile dystonic postures (see below).

PATHOPHYSIOLOGY OF DYSTONIA

The pathophysiologic basis of dystonia is not completely known. The phenomenon is characterized by co-contracting synchronous bursts of agonist and antagonist muscle groups with recruitment of muscle groups that are not required for a given movement (overflow). Dystonia is characterized by derangement of the basic physiological principle of action-selection, leading to abnormal recruitment of inappropriate muscles for a given action with inadequate inhibition of this undesired motor activity. Physiologically, loss of surround inhibition is observed at multiple levels of the motor system (e.g., cortex, brainstem, spinal cord) accompanied by increased cortical excitability and reorganization. Attention has focused on the basal ganglia as the site of origin of at least some types of dystonia because there are alterations in blood flow and metabolism in these structures. Further, lesions of the basal ganglia (particularly the putamen) can induce dystonia, and surgical ablation or deep brain stimulation (DBS) of specific regions of the globus pallidus may ameliorate dystonia. The dopamine system has also been implicated, because dopaminergic therapies can both induce and treat some forms of dystonia in different circumstances. Interestingly, no specific pathology has been consistently identified in dystonia.

TREATMENT

Dystonia

Treatment of dystonia is for the most part symptomatic except in rare cases where correction of a primary underlying condition is possible. Wilson's disease should be ruled out in young patients with dystonia. Levodopa should be tried in all cases of childhood-onset dystonia to test for DRD. High-dose anticholinergics (e.g., trihexyphenidyl 20–120 mg/d) may be beneficial in children, but adults can rarely tolerate high doses because of side effects related to cognitive impairment and hallucinations. Oral baclofen (20–120 mg) may also be helpful, but benefits, if present, are usually modest, and side effects of sedation, weakness, and memory loss can be problematic. Intrathecal infusion of baclofen is more likely to be useful, particularly for leg and trunk dystonia, but benefits are frequently not sustained, and complications can be serious and include infection, seizures, and coma. Tetrabenazine is another consideration (the usual starting dose is 12.5 mg/d and the average treating dose is 25–75 mg/d), but its use may be limited by sedation and the development of parkinsonism. Neuroleptics can improve as well as induce dystonia, but they are typically not recommended because of their potential to induce parkinsonism and other movement disorders, including tardive dystonia. Clonazepam and diazepam are rarely effective.

Botulinum toxin has become the preferred treatment for patients with focal dystonia, particularly where involvement is limited to small muscle groups such as in blepharospasm, torticollis, and spasmodic dysphonia. Botulinum toxin acts by blocking the release of acetylcholine at the neuromuscular junction, leading to reduced dystonic muscle contractions. However, treatment with botulinum toxin can be complicated by excessive weakness that can be troublesome,

TABLE 428-2 Confirmed Monogenic Forms of Isolated and Combined Dystonia[a]

FORM OF DYSTONIA		GENE	LOCUS NAME	DESIGNATION AND PHENOTYPIC SUBGROUP[a]	ADDITIONAL DISTINGUISHING FEATURES	MOI
Isolated		*TOR1A*	DYT1	DYT-TOR1A	Childhood or adolescent-onset, generalized	AD
		THAP1	DYT6	DYT-THAP1	Adolescent-onset, cranial or generalized	AD
		ANO3	DYT24	DYT-ANO3	Adult-onset, focal or segmental	AD
		GNAL	DYT25	DYT-GNAL	Mostly adult-onset, focal or segmental	AD
		KMT2B[b]	DYT28	DYT-KMT2B	Early-onset, generalized, mild syndromic features	AD
Combined	**Dystonia plus Parkinsonism**	*GCH1*	DYT5a	DYT-GCH1	Dopa-responsive	AD
		TAF1	DYT3	DYT-TAF1	Neurodegeneration	XL
		PRKRA	DYT16	DYT-PRKRA	Dystonia with mild parkinsonism	AR
		ATP1A3	DYT12	DYT-ATP1A3	Rapid-onset	AD
	Dystonia plus Myoclonus	*SGCE*	DYT11	DYT-SGCE	Psychiatric disease	AD

[a]According to C Marras et al: Mov Disord 31:436, 2016. [b]Several, but not all, patients show syndromic features; DYT-KMT2B may thus be better placed with the complex dystonias.

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; MOI, mode of inheritance; XL, X-linked.

particularly if it involves neck and swallowing muscles. Two serotypes of botulinum toxin are currently available (A and B). Both are effective, and it is not clear that there are advantages of one over the other. No systemic side effects are encountered with the doses typically used, but benefits are transient, and repeat injections are required at 2–5 month intervals. Some patients fail to respond after having experienced an initial benefit. This has been attributed to antibody formation, but improper muscle selection, injection technique, and inadequate dose should be excluded.

Surgical therapy is an alternative for patients with severe dystonia who are not responsive to other treatments. Peripheral procedures such as rhizotomy and myotomy were used in the past to treat cervical dystonia, but are now rarely employed. DBS of the pallidum can provide dramatic benefits for some patients with various forms of hereditary and nonhereditary generalized dystonia. This represents a major therapeutic advance because previously there was no consistently effective therapy, especially for patients with severe disability. Benefits tend to be obtained with a lower frequency of stimulation and often occur after a relatively longer latency (weeks to months) than in PD. Better results are typically obtained in younger patients with shorter disease duration. Recent studies suggest that DBS may also be valuable for patients with focal and secondary dystonias, although results are less consistent. Supportive treatments such as physical therapy and education should be a part of the treatment regimen.

Physicians should be aware of dystonic storm, a rare but potentially fatal condition that can occur in response to a stress situation such as surgery or a systemic infection in patients with preexisting dystonia. It consists of the acute onset of generalized and persistent dystonic contractions that can involve the vocal cords or laryngeal muscles, leading to airway obstruction. Patients may experience rhabdomyolysis with renal failure and should be managed in an intensive care unit with airway protection if required. Treatment can be instituted with one or a combination of anticholinergics, diphenhydramine, baclofen, benzodiazepines, and dopaminergic agents. Spasms may be difficult to control, and anesthesia with muscle paralysis may be required.

CHOREAS

HUNTINGTON'S DISEASE

HD is a progressive, fatal, highly penetrant autosomal dominant disorder characterized by motor, behavioral, oculomotor, and cognitive dysfunction. The disease is named for George Huntington, a family physician who described cases on Long Island, New York, in the nineteenth century. Onset is typically between the ages of 25 and 45 years (range, 3–70 years) with a prevalence of 2–8 cases per 100,000 and an average age at death of 60 years. It is prevalent in Europe, North America, South America, and Australia but is rare in African blacks and Asians. HD is characterized by rapid, nonpatterned, semi-purposeful, involuntary choreiform movements, and for this reason was formerly referred to as Huntington's chorea. However, dysarthria, gait disturbance, oculomotor abnormalities, behavioral disturbance, and cognitive impairment with dementia are also common features, thus the condition is currently referred to as HD. In the early stages, chorea tends to be focal or segmental, but progresses over time to involve multiple body regions. With advancing disease, there tends to be a reduction in chorea and the emergence of dystonia, rigidity, bradykinesia, and myoclonus. Functional decline is often predicted by progressive weight loss despite adequate calorie intake. In younger patients (~10% of cases), HD can present as an akinetic-rigid or parkinsonian syndrome (Westphal variant). HD patients eventually develop behavioral and cognitive disturbances, and the majority progress to dementia. Depression with suicidal tendencies, aggressive behavior, and psychosis can be prominent features. HD patients may also develop noninsulin-dependent diabetes mellitus and neuroendocrine abnormalities (e.g., hypothalamic dysfunction). A clinical diagnosis of HD can be strongly suspected in cases of chorea with a positive family history, but genetic testing provides the ultimate confirmation of the diagnosis. The disease predominantly affects the striatum but progresses to involve the cerebral cortex and other brain regions. Progressive atrophy of the head of the caudate nucleus, which form the lateral margin of the lateral ventricle, can be visualized by MRI **(Fig. 428-1)**, but the putamen can be equally or even more severely affected. More diffuse cortical atrophy can be seen in the middle and late stages of the disease. Supportive studies include reduced metabolic activity in the caudate nucleus and putamen, and reduced brain metabolites on MR spectroscopy. Genetic testing can be used to confirm the diagnosis and to detect at-risk individuals in the family, but must be performed with caution and in conjunction with trained counselors, because positive results can worsen depression and generate suicidal reactions. The neuropathology of HD consists of prominent neuronal loss and gliosis in the caudate nucleus and putamen; similar changes are also widespread in the cerebral cortex. Intraneuronal inclusions containing aggregates of ubiquitin and the mutant protein huntingtin are found in the nuclei of affected neurons.

In anticipation of developing neuroprotective therapies, there has been an intensive effort to define the premanifest stage of HD. Subtle motor impairment, cognitive alterations, and imaging changes can be detected in at-risk individuals who later go on to develop the manifest form of the disease. Defining the rate of progression of these features is

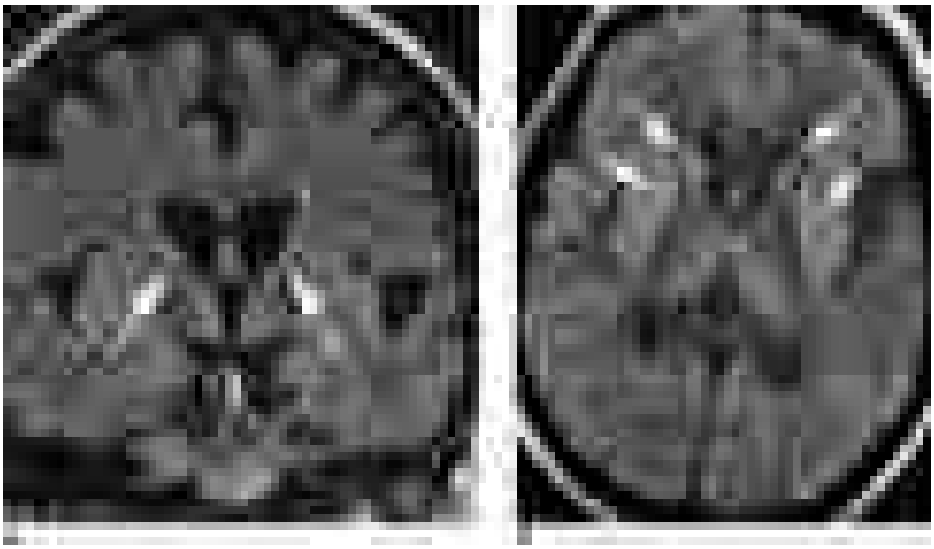


FIGURE 428-1 Huntington's disease. ***A.*** Coronal fluid attenuated inversion recovery (FLAIR) magnetic resonance imaging shows enlargement of the lateral ventricles reflecting typical atrophy (*arrows*). ***B.*** Axial FLAIR image demonstrates abnormal high signal in the caudate and putamen (*arrows*).

paramount for future studies of putative disease-modifying therapies designed to slow the rate of disease progression and the development of cumulative disability.

ETIOLOGY

HD is caused by an increase in the number of polyglutamine (CAG) repeats (>40) in the coding sequence of the *huntingtin* gene located on the short arm of chromosome 4. The larger the number of repeats, the earlier the disease is manifest. Intermediate forms of the disease with 36–39 repeats are described in some patients, typically with less severe clinical involvement. Acceleration of the process tends to occur, particularly in males, with subsequent generations having larger numbers of repeats and earlier age of disease onset, a phenomenon referred to as anticipation. The gene encodes the highly conserved cytoplasmic protein huntingtin, which is widely distributed in neurons throughout the central nervous system (CNS), but whose function is largely unknown. Mitochondrial dysfunction has been demonstrated in the striatum and skeletal muscle of symptomatic and presymptomatic individuals. Fragments of the mutant huntingtin protein can be toxic, possibly by translocating into the nucleus and interfering with transcriptional regulation of proteins. Neuronal inclusions found in affected regions in HD may represent a protective mechanism aimed at segregating and facilitating the clearance of these toxic proteins. There is also interest in the possibility that protein accumulation and aggregation in HD, like Alzheimer's disease **(Chap. 423)** and PD **(Chap. 427)**, may be critical to the disease process and reflect a prion-like disorder **(Chap. 430)**. Models of HD with striatal pathology can be induced in transgenic animals that express the mutant gene and by excitotoxic agents such as kainic acid and 3-nitropropionic acid which promote calcium entry into the cell and cytotoxicity.

TREATMENT

Huntington's Disease

Although the gene for HD was identified 25 years ago, there is still no disease-modifying therapy for this disorder and symptomatic treatment is limited. Current treatment involves a multidisciplinary approach, with medical, neuropsychiatric, social, and genetic counseling for patients and their families. Dopamine-blocking agents may control the choreatic movements. Tetrabenazine (a presynaptic dopamine depleting agent) has been approved for the treatment of chorea, but can cause secondary parkinsonism. More recently, deuterated tetrabenazine (Austedo™) has been approved as a treatment for chorea in HD. Deuteration interferes with the metabolism of tetrabenazine and avoids a high Cmax. In clinical trials, it has been shown to have fewer dose-related side effects than tetrabenazine, and therefore can be administered in higher doses with potentially superior clinical benefits. Neuroleptics are generally not recommended because of their potential to induce other troubling movement disorders and because HD chorea tends to be self-limited and is usually not disabling. These drugs may be used however in patients with severe and disabling chorea. Unfortunately, no medications have been developed as yet that interfere with the nonchoreic aspects of motor dysfunction in HD, although many promising agents are currently in clinical trials. Depression and anxiety can be major problems, and patients should be treated with appropriate antidepressant and antianxiety drugs and monitored for mania and suicidal ideations. Psychosis can be treated with atypical antipsychotics such as clozapine (50–600 mg/d), quetiapine (50–600 mg/d), and risperidone (2–8 mg/d). There is no adequate treatment for the cognitive or motor decline. A neuroprotective therapy that slows or stops disease progression is the major unmet medical need in HD. Drugs that enhance mitochondrial function and increase the clearance of defective mitochondria are being tested as possible disease-modifying therapies. Other investigative approaches include antiglutamate agents, dopamine stabilizers, caspase inhibitors, neurotrophic factors, anti-inflammatory agents, transplantation of fetal striatal cells or stem cells, and DBS of the globus pallidus pars interna (GPi), but none has as yet been demonstrated to have a beneficial effect in HD. The potential to block/edit the mutant huntingtin gene with small interfering RNAs (siRNAs) or CRISPR/cas9 technology is an exciting area of research that is currently being investigated as a possible future therapy.

HUNTINGTON'S DISEASE-LIKE DISORDERS

A group of rare inherited conditions that can mimic HD, designated HD-like (HDL) disorders, have also been identified. HDL-1, 2, and 4 are autosomal dominant conditions that typically present in adulthood. HDL-1 is due to expansion of an octapeptide repeat in *PRNP*, the gene encoding the prion protein **(Chap. 430)**. Thus HDL-1 is properly considered a prion disease. Patients exhibit onset of personality change in the third or fourth decade, followed by chorea, rigidity, myoclonus, ataxia, and epilepsy. HDL-2 manifests in the third or fourth decade with a variety of movement disorders, including chorea, dystonia, or parkinsonism and dementia. Most patients are of African descent. Acanthocytosis can sometimes be seen in these patients, and this condition must be distinguished from neuroacanthocytosis (below). HDL-2 is caused by an abnormally expanded CTG/CAG trinucleotide repeat expansion in the *junctophilin-3* (*JPH3*) gene. The pathology of HDL-2 consists of intranuclear inclusions immunoreactive for ubiquitin and expanded polyglutamine repeats. HDL-4, the most common condition in this group, is caused by expansion of trinucleotide repeats in *TBP*, the gene that encodes the TATA box-binding protein involved in regulating transcription; this condition is identical to spinocerebellar ataxia (SCA) 17 **(Chap. S10)**, and most patients present primarily with ataxia rather than chorea. Mutations of the *C9Orf72* gene associated with amyotrophic lateral sclerosis **(Chap. 429)** have also been reported in some individuals with an HDL phenotype.

OTHER CHOREAS

Chorea can be seen in a number of additional disorders related to genetic mutations or other disease states.

Among the hereditary forms of childhood-onset chorea, mutations in the *ADCY5* (adenylate cyclase 5) gene are an increasingly recognized and probably relatively common cause of childhood-onset chorea, often in combination with dystonia and developmental delay in some cases. Characteristic perioral movements are a hallmark of the disorder.

Chorea-acanthocytosis (neuroacanthocytosis) is a progressive and typically fatal autosomal recessive disorder that is characterized by chorea coupled with red cell abnormalities on peripheral blood smear (acanthocytes). The chorea can be severe and associated with self-mutilating behavior, dystonia, tics, seizures, and a polyneuropathy. Mutations in the *VPS13A* gene encoding chorein have been described. A phenotypically similar X-linked form of the disorder has been described in older individuals who have reactivity with Kell blood group antigens (McLeod syndrome). A benign hereditary chorea of childhood (BHC1) due to mutations in the gene for thyroid transcription factor 1 and a late-onset benign senile chorea (BHC2) have also been described. It is important to ensure that patients with these types of choreas do not have HD.

Chorea may also occur in association with a variety of infections and degenerative disorders as well as vascular diseases and hypo- and hyperglycemia. Sydenham's chorea (originally called St. Vitus's dance) is more common in females and is typically seen in childhood (5–15 years). It often develops in association with prior exposure to group A streptococcal infection **(Chap. 143)** and is thought to be autoimmune in nature. It is characterized by the acute onset of choreiform movements and behavioral disturbances. With the reduction in the incidence of rheumatic fever, the incidence of Sydenham's chorea has fallen, but it can still be seen in developing countries. The chorea generally responds to dopamine-blocking agents, valproic acid, and carbamazepine, but is self-limited, and treatment is generally restricted to those with severe chorea. Chorea may recur in later life, particularly in association with pregnancy (chorea gravidarum) or treatment with sex hormones. Several reports have documented cases of chorea associated with N-methyl-D-aspartate (NMDA) receptor antibody–positive encephalitis **(Chap. 90)** following herpes simplex virus encephalitis.

Systemic lupus erythematosus (Chap. 349) is the most common systemic disorder that is associated with chorea. The chorea can last for days to years. Chorea can also be seen with hyperthyroidism, autoimmune disorders including Sjögren's syndrome, infectious disorders including HIV disease, metabolic alterations, and polycythemia rubra vera. Chorea has also been described following open-heart surgery in the pediatric population and in association with many medications (especially anticonvulsants, cocaine, CNS stimulants, estrogens, and lithium). Chorea is commonly seen in association with chronic levodopa treatment (Parkinson's Disease, Chap. 427). Chorea may also be encountered in paraneoplastic syndromes associated with anti-CRMP-5 or anti-Hu antibodies (Chap. 90).

HEMIBALLISMUS

Ballism is a violent form of choreiform movement composed of wild, flinging, large-amplitude movements most frequently affecting proximal limb muscles on one side of the body (hemiballism). The movements may only affect one limb (monoballism) or, more exceptionally, both upper or lower limbs (paraballism). The movements may be so severe as to cause exhaustion, dehydration, local injury, and, in extreme cases, death. Fortunately, dopamine-blocking drugs can be very helpful, and importantly, hemiballismus is usually self-limiting and tends to resolve spontaneously after weeks or months. The most common cause is a partial lesion (infarct or hemorrhage) in the subthalamic nucleus (STN), but in 30–40% of cases the lesion is found in the putamen, thalamus, or parietal cortex. Hemiballismus is also a common feature of the paroxysmal dyskinesias (see below). In extreme cases, pallidotomy or DBS of the GPi can be effective and abolish the involuntary movements. Interestingly, surgically induced lesions and DBS of the STN in PD patients are usually not associated with hemiballismus.

TICS

A *tic* is a brief, rapid, recurrent, and seemingly purposeless stereotyped motor contraction. Motor tics can be simple, with movement only affecting an individual muscle group (e.g., blinking, twitching of the nose, jerking of the neck), or complex, with coordinated involvement of multiple muscle groups (e.g., jumping, sniffing, head banging, and echopraxia [mimicking movements]). Phonic (or vocal) tics can also be simple (e.g., grunting) or complex (e.g., echolalia [repeating other people's words], palilalia [repeating one's own words], and coprolalia [expression of obscene words]). Patients may also experience sensory tics, composed of unpleasant focal sensations in the face, head, or neck. These can be mild and of little clinical consequence or severe and disabling to the patient. Tics may present in adulthood and can be seen in association with a variety of disorders, including PD, HD, trauma, dystonia, drugs (e.g., levodopa, neuroleptics), and toxins.

TOURETTE'S SYNDROME (TS)

TS is a neurobehavioral disorder named after the French neurologist Georges Gilles de la Tourette. It predominantly affects males, and the prevalence is estimated to be 0.03–1.6%, but it is likely that many mild cases do not come to medical attention. TS is characterized by multiple motor tics often accompanied by vocalizations (phonic tics). Patients characteristically can voluntarily suppress tics for short periods of time, but then experience an irresistible urge to express them. Tics vary in intensity and may be absent for days or weeks only to recur, occasionally in a different pattern. Tics tend to present between ages 2 and 15 years (mean 7 years) and often lessen or even disappear in adulthood, particularly in males. Associated behavioral disturbances include anxiety, depression, attention deficit hyperactivity disorder, and obsessive-compulsive disorder. Patients may experience personality disorders, self-destructive behaviors, difficulties in school, and impaired interpersonal relationships.

Etiology and Pathophysiology TS is thought to be a genetic disorder, but no specific monogenic cause has yet been identified. Current evidence supports a complex inheritance pattern with an important contribution of de-novo, likely gene-disrupting variants. Four likely risk genes with multiple de novo damaging variants in unrelated probands include *WWC1, CELSR3, NIPBL*, and *FN1*. The risk of a family with one affected child having a second is about 25%. The pathophysiology of TS is not known, but alterations in dopamine neurotransmission, opioids, and second-messenger systems have been proposed. Some cases of TS may be the consequence of an autoimmune response to β-hemolytic streptococcal infection (pediatric autoimmune neuropsychiatric disorder associated with streptococcal infection [PANDAS]); however, this entity remains controversial.

TREATMENT

Tics

Patients with mild disease often only require education and counseling (for themselves and family members). In a high proportion of patients the severity of tics wane in adult life becoming less of a medical problem, thus arguing for a conservative management when possible during the first decades of life. Drug treatment is indicated when the tics are disabling and interfere with quality of life. Therapy is individualized, and there is no singular treatment regimen that has been properly evaluated in double-blind trials. Some physicians use the α-agonist clonidine, starting at low doses and gradually increasing the dose and frequency until satisfactory control is achieved. Guanfacine (0.5–2 mg/d) is an α-agonist that is preferred by some because it only requires once-a-day dosing. Other physicians prefer to use neuroleptics. Atypical neuroleptics are usually used initially (risperidone, olanzapine, ziprasidone) because they are thought to be associated with a reduced risk of tardive dyskinesia. If they are not effective, low doses of classical neuroleptics such as haloperidol, fluphenazine, pimozide, or tiapride can be tried because the risk of tardive dyskinesia in young people is relatively low. Tetrabenazine and deuterated tetrabenazine are also currently being evaluated. Botulinum toxin injections can be effective in controlling focal tics that involve small muscle groups. Behavioral features, and particularly anxiety and compulsions, can be a disabling feature of TS and should be treated. The potential value of DBS targeting the anterior portion of the internal capsule, the GPi, or the thalamus is currently being explored.

MYOCLONUS

Myoclonus is a brief, rapid (<100 ms), shock-like, jerky movement consisting of single or repetitive muscle discharges. Myoclonic jerks can be focal, multifocal, segmental, or generalized and can occur spontaneously, in association with voluntary movement (action myoclonus) or in response to an external stimulus (reflex myoclonus). Negative myoclonus consists of a brief loss of muscle activity (e.g., asterixis in hepatic failure). Myoclonic jerks can be severe and interfere with normal movement or benign and of no clinical consequence as is commonly observed in normal people when waking up or falling asleep (hypnagogic jerks).

Myoclonic jerks differ from tics in that they are not typically repetitive, can severely interfere with normal voluntary movement, and are not suppressible. They can arise in association with abnormal neuronal discharges in cortical, subcortical, brainstem, or spinal cord regions, particularly in association with hypoxemia (especially following cardiac arrest), encephalopathy, and neurodegeneration. Reversible myoclonus can be seen with metabolic disturbances (renal failure, electrolyte imbalance, hypocalcemia), toxins, and many medications. The combination of action myoclonus (cortical origin) with ataxia and generalized epilepsy is associated with several recognized causes. The most common is myoclonic epilepsy or Unverricht-Lundborg disease (EPM-1) which can have a variable but often progressive course. This is an autosomal recessive disease caused by mutations in the *CSBT* gene. Other causes are Lafora body epilepsy or progressive myoclonic epilepsy (PME-2) caused by mutations in in the *EPM2A* gene or the *NHLRC1* gene and ceroid lipofuscinosis. In patients with less severe or absent epilepsy, mitochondrial disorders and neurodegenerative disorders affecting the cerebellum (i.e., SCAs) should be considered. Essential myoclonus is a relatively benign familial condition characterized by multifocal, very brief, lightning-like movements that are

frequently alcohol-sensitive. Mutations in the *epsilon-sarcoglycan* gene have been associated with myoclonus seen in association with dystonia (myoclonic-dystonia).

TREATMENT

Myoclonus

Treatment primarily consists of managing the underlying condition or removing an offending agent. Pharmacologic therapy involves one or a combination of GABAergic agents such as valproic acid (800–3000 mg/d), piracetam (8–20 g/d), clonazepam (2–15 mg/d), levetiracetam (1000–3000 mg/d), or primidone (500–1000 mg/d) and may be associated with striking clinical improvement in chronic cases (e.g., postanoxic myoclonus, progressive myoclonic epilepsy) in which a cortical origin for the myoclonic discharges has been identified. The serotonin precursor 5-hydroxytryptophan (plus carbidopa) may be useful in some cases of postanoxic myoclonus.

DRUG-INDUCED MOVEMENT DISORDERS

This important group of movement disorders is primarily associated with drugs that block dopamine receptors (neuroleptics) or central dopaminergic transmission. These drugs are widely used in psychiatry, but it is important to appreciate that drugs used in the treatment of nausea or vomiting (e.g., prochlorperazine [Compazine]) or gastroesophageal disorders (e.g., metoclopramide) are neuroleptic agents and can also cause these disorders. Hyperkinetic movement disorders secondary to neuroleptic drugs can be divided into those that present acutely, subacutely, or after prolonged exposure (tardive syndromes). Dopamine-blocking drugs can also be associated with a reversible parkinsonian syndrome for which anticholinergics are often concomitantly prescribed, but there is concern that this may increase the risk of developing a tardive syndrome and an underlying subclinical PD syndrome should be considered.

ACUTE

Dystonia is the most common acute hyperkinetic drug reaction. It is typically generalized in children and focal in adults (e.g., blepharospasm, torticollis, or OMD). The reaction can develop within minutes of exposure and can be successfully treated in most cases with parenteral administration of anticholinergics (benztropine or diphenhydramine), benzodiazepines (lorazepam, clonazepam, or diazepam), or dopamine agonists. The abrupt onset of severe spasms may occasionally be confused with a seizure; however, there is no loss of consciousness, automatisms, or postictal features typical of epilepsy. The acute onset of chorea, stereotypic behavior, and tics may also be seen, particularly following exposure to CNS stimulants such as methylphenidate, cocaine, or amphetamines.

SUBACUTE

Akathisia is the most common reaction in this category. It consists of motor restlessness with a need to move that is alleviated by movement. Therapy consists of removing the offending agent. When this is not possible, symptoms may be ameliorated with benzodiazepines, anticholinergics, beta blockers, or dopamine agonists.

TARDIVE SYNDROMES

These disorders develop months to years after initiation of the neuroleptic agent. Tardive dyskinesias (TD) are most common, and typically present with choreiform and/or dystonic movements involving the mouth, lips, and tongue. In severe cases, the trunk, limbs, and respiratory muscles may also be affected. In approximately one-third of patients, TD remit within 3 months of stopping the drug, and most patients gradually improve over the course of several years. However, abnormal movements may also develop or worsen after stopping the offending agent. The movements are often mild and more upsetting to the family than to the patient, but they can be severe and disabling, particularly in the context of an underlying psychiatric disorder. Atypical antipsychotics (e.g., clozapine, risperidone, olanzapine, quetiapine, ziprasidone, and aripiprazole) are associated with a lower risk of causing TD in comparison to traditional antipsychotics. Younger patients have a lower risk of developing neuroleptic-induced TD, whereas elderly, females, and those with underlying organic cerebral dysfunction have been reported to be at greater risk. Chronic use is associated with increased risk, and specifically, the U.S. Food and Drug Administration has warned that use of metoclopramide for more than 12 weeks increases the risk of TD. Because TD can be permanent and resistant to treatment, antipsychotics should be used judiciously, atypical neuroleptics should be the preferred agent when possible, and the need for continued use should be regularly monitored.

Treatment primarily consists of stopping the offending agent. If the patient is receiving a traditional antipsychotic, and withdrawal is not possible, replacement with an atypical antipsychotic should be tried. Abrupt cessation of a neuroleptic should be avoided because acute withdrawal can induce worsening. TD can persist after withdrawal of antipsychotics and can be difficult to treat. Valbenazine (Ingrezza™) is an ester of tetrabenazine that has recently been approved for the treatment of tardive dyskinesia based on results of efficacy in double blind trials, but it is associated with sleepiness and QT prolongation. It acts as a vesicular monoamine transporter type 2 (VMAT-2) inhibitor and blocks storage of dopamine. Deuterated tetrabenazine is also being studied for this indication. Benefits in open label studies have been reported with valproic acid (750–3000 mg/d), anticholinergics, or botulinum toxin injections. Other approaches that have been tried include baclofen (40–80 mg/d) or clonazepam (1–8 mg/d). In some cases, the abnormal movement is refractory to therapy.

Chronic neuroleptic exposure can also be associated with tardive dystonia, with preferential involvement of axial muscles and characteristic rocking movements of the trunk and pelvis. Tardive dystonia can be more troublesome than tardive dyskinesia and frequently persists despite stopping medication. Valproic acid, anticholinergics, and botulinum toxin may occasionally be beneficial, but patients are frequently refractory to medical therapy. Tardive akathisia, tardive TS, and tardive tremor syndromes are rare but may also occur after chronic neuroleptic exposure.

Neuroleptic medications can also be associated with a neuroleptic malignant syndrome (NMS). NMS is characterized by the acute or subacute onset of muscle rigidity, elevated temperature, altered mental status, hyperthermia, tachycardia, labile blood pressure, renal failure, and markedly elevated creatine kinase levels. Symptoms typically evolve within days or weeks after initiating the drug. NMS can also be precipitated by the abrupt withdrawal of dopaminergic medications in PD patients. Treatment involves immediate cessation of the offending antipsychotic drug and the introduction of a dopaminergic agent (e.g., a dopamine agonist or levodopa), dantrolene, or a benzodiazepine. In very severe cases, when oral intake is not possible, a patch (delivering rotigotine subcutaneously) or an infusion pump (delivering apomorphine subcutaneously) may be the best approach to provide dopaminergic treatment. Treatment may need to be undertaken in an intensive care setting and include supportive measures such as control of body temperature (antipyretics and cooling blankets), hydration, electrolyte replacement, and control of renal function and blood pressure.

Drugs that have serotonin-like activity (tryptophan, MDMA or "ecstasy," meperidine) or that block serotonin reuptake can induce a rare, but potentially fatal, serotonin syndrome that is characterized by confusion, hyperthermia, tachycardia, and coma as well as rigidity, ataxia, and tremor. Myoclonus is often a prominent feature, in contrast to NMS, which it resembles in other respects. Patients can be managed with propranolol, diazepam, diphenhydramine, chlorpromazine, or cyproheptadine as well as supportive measures.

A variety of drugs can also be associated with parkinsonism and other hyperkinetic movement disorders. Some examples include phenytoin (chorea, dystonia, tremor, myoclonus), carbamazepine (tics and dystonia), tricyclic antidepressants (dyskinesias, tremor, myoclonus), fluoxetine (myoclonus, chorea, dystonia), oral contraceptives (dyskinesia), β-adrenergics (tremor), buspirone (akathisia, dyskinesias, myoclonus), and digoxin, cimetidine, diazoxide, lithium, methadone, and fentanyl (dyskinesias).

PAROXYSMAL DYSKINESIAS

Paroxysmal dyskinesias are a group of rare disorders characterized by episodic, brief involuntary movements that can manifest as various types of hyperkinetic movements, including chorea, dystonia, tremor, myoclonus, and ballism. There are three main types: (1) *paroxysmal kinesigenic dyskinesia (PKD)*, where the involuntary movements are triggered by sudden movement, (2) *paroxysmal nonkinesigenic dyskinesias (PNKD)*, where the attacks are not induced by movement, and (3) rare cases of *exertion-induced dyskinesia (PED)*, where attacks are induced by prolonged exercise.

PKD are characterized by brief, self-limited attacks induced by movement onset such as running but also occasionally by unexpected sound or photic stimulation. Attacks may affect one side of the body, last seconds to minutes at a time, and recur several times a day. They usually manifest as a mixed hyperkinetic movement disorder with dystonic posturing of a limb, ballismus, and chorea, which may also become generalized. PKD is most commonly familial with an autosomal dominant pattern of inheritance and mutations in the *proline-rich transmembrane protein 2* (*PRRT2*) gene, but may also occur secondary to various brain disorders such as multiple sclerosis or hyperglycemia. PKD is more frequent in males (4:1), and the onset is typically in the first or second decade of life. About 70% report sensory symptoms such as tingling or numbness of the affected limb preceding the attack by a few milliseconds. The evolution is relatively benign, and there is a trend toward resolution of the attacks over time. Treatment with low-dose anticonvulsant therapy such as carbamazepine or phenytoin is advised when the attacks are frequent and interfere with daily life activities, and is effective in about 80% of patients. Some clinical features of PKD (abrupt and short-lasting attacks preceded by an "aura"), the association with true seizure episodes, and its favorable response to anticonvulsant drugs have led to speculation that it is epileptic in origin, but this has not been established.

PNKD involve attacks of generalized dyskinesias precipitated by alcohol, caffeine, stress, or fatigue. In comparison to PKD, the episodes have a relatively longer duration (minutes to hours) and are less frequent (one to three per day). PNKD is inherited as an autosomal dominant condition with high (~80%) but incomplete penetrance. A missense mutation in the *myofibrillogenesis regulator* (*PNKD*) gene has been identified in several families. Recognition of the condition and elimination of the underlying precipitating factors, where possible, are the first priorities. Tetrabenazine, neuroleptics, dopamine-blocking agents, propranolol, clonazepam, and baclofen may be helpful. Treatment may not be required if the condition is mild and self-limited. Most patients with PNKD do not benefit from anticonvulsant drugs, but some may respond to clonazepam or other benzodiazepines.

The *SLC2A1* (solute carrier family 2 member 1) gene, previously linked to GLUT1 (glucose transporter of the blood brain barrier) deficiency syndrome, has been identified to also cause paroxysmal PED. The attacks in this disorder are characterized by a combination of chorea, athetosis, and dystonia in excessively exercised body regions with the legs being most frequently affected. A single attack lasts from a few minutes to an hour and occurs after prolonged physical exercise. In addition to the movement disorder, several patients have other disease manifestations such as epilepsy, hemolytic anemia, and migraine. A ketogenic diet is an effective therapeutic option.

RESTLESS LEGS SYNDROME (RLS)

RLS is a neurologic disorder that affects ~10% of the adult population (it is rare in Asians) and can cause significant morbidity in some individuals. It was first described in the seventeenth century by the English physician Thomas Willis, but has only recently been recognized as being a bona fide movement disorder. The four core symptoms required for diagnosis are as follows: an urge to move the legs usually caused or accompanied by an unpleasant sensation in the legs; symptoms that begin or worsen with rest; partial or complete relief by movement; and worsening during the evening or night.

Symptoms most commonly begin in the legs, but can spread to or even begin in the upper limbs. The unpleasant sensation is often described as a creepy-crawly feeling, paresthesia, or burning. In about 80% of patients, RLS is associated with periodic leg movements (PLMs) during sleep and occasionally while awake. These involuntary movements are usually brief, lasting no more than a few seconds, and recur every 5–90 s. The restlessness and PLMs are a major cause of sleep disturbance in patients, leading to poor-quality sleep and daytime sleepiness.

Primary RLS has a strong genetic component, and several loci have been associated with an autosomal dominant pattern of inheritance, although penetrance may be variable and no specifically causative gene has been identified to date. The mean age of onset in familial forms is 27 years, although pediatric cases are recognized. The severity of symptoms is variable. Secondary RLS may be associated with pregnancy or a range of underlying disorders, including anemia, ferritin deficiency, renal failure, and peripheral neuropathy. The pathogenesis probably involves disordered dopamine function, which may be peripheral or central, possibly in association with an abnormality of iron metabolism. Diagnosis is made on clinical grounds but can be supported by polysomnography and the demonstration of PLMs. The neurologic examination is normal. Secondary causes of RLS should be excluded, and ferritin levels, glucose, and renal function should be measured.

Most RLS sufferers have mild symptoms that do not require specific treatment. General measures to improve sleep hygiene and quality should be attempted first. If symptoms remain intrusive, low doses of dopamine agonists, e.g., pramipexole (0.25–0.5 mg), ropinirole (1–2 mg), or patch rotigotine (2–3 mg), taken 1–2 h before bedtime are generally effective. Levodopa may also be effective but is more likely to be associated with augmentation (spread and worsening of restlessness and its appearance earlier in the day) or rebound (reappearance sometimes with worsening of symptoms at a time related to the drug's short half-life). Augmentation can also be seen with dopamine agonists, particularly if higher doses are employed. Other drugs that can be effective include anticonvulsants, analgesics, and opiates. Management of secondary RLS should be directed to correcting the underlying disorder; for example, iron replacement for anemia.

OTHER DISORDERS THAT MAY PRESENT WITH A COMBINATION OF PARKINSONISM AND HYPERKINETIC MOVEMENTS

WILSON'S DISEASE

WD is an autosomal recessive inherited disorder of copper metabolism that manifests with neurologic, psychiatric, and liver disorders, alone or in combination. It is caused by mutations in the *ATP7B* gene encoding a P-type ATPase. The disease was first described by the English neurologist Kinnier Wilson at the beginning of the twentieth century, although at around the same time the German physicians Kayser and Fleischer separately noted the characteristic association of corneal pigmentation with hepatic and neurologic features. WD has a worldwide prevalence of ~1 in 30,000, with a mutation carrier frequency of 1 in 90. About half of WD patients (especially younger patients) manifest with liver abnormalities. The remainder present with neurologic disease (with or without underlying liver abnormalities), and a small proportion have hematologic or psychiatric problems at disease onset.

Neurologic onset usually manifests in the second decade with tremor, rigidity, and dystonia. The tremor is usually in the upper limbs, bilateral, and asymmetric. Tremor can be on intention or occasionally at rest and, in advanced disease, can take on a wing-beating characteristic (a flapping movement when the arms are held outstretched with the fingers opposed). Other features can include parkinsonism with bradykinesia, dystonia (particularly facial grimacing), dysarthria, and dysphagia. More than half of those with neurologic features have a history of psychiatric disturbances, including depression, mood swings, and overt psychosis. Kayser-Fleischer (KF) rings are seen virtually in all patients with neurologic features and 80% of those with hepatic presentations. KF rings represent the deposition of copper in Descemet's membrane around the cornea. They consist of a characteristic grayish rim or circle at the limbus of the cornea and are best detected by slit-lamp examination. Neuropathologic examination is characterized by

neurodegeneration and astrogliosis in the basal ganglia, particularly in the striatum.

WD should always be considered in the differential diagnosis of a movement disorder in the first decades of life. Low levels of blood copper and ceruloplasmin and high levels of urinary copper may be present, but normal levels do not exclude the diagnosis. Brain imaging usually reveals generalized brain atrophy in established cases, and ~50% have signal hypointensity in the caudate head, putamen, globus pallidus, substantia nigra, and red nucleus on T2-weighted MRI scans. However, correlation of imaging changes with clinical features is not good. Liver biopsy with demonstration of high copper levels and genetic testing remain the gold standard for the diagnosis.

In the absence of treatment, the course is progressive and leads to severe neurologic dysfunction and early death in the majority of patients, although a small proportion experience a relatively benign course. Treatment is directed at reducing tissue copper levels and maintenance therapy to prevent reaccumulation. There is no clear consensus on optimal treatment, and patients should be managed in a unit with expertise in WD. Penicillamine is frequently used to increase copper excretion, but may lead to a worsening of symptoms in the initial stages of therapy. Side effects are common and can to some degree be attenuated by coadministration of pyridoxine. Tetrathiomolybdate blocks the absorption of copper and can be used instead of penicillamine. Trientine and zinc are useful drugs for maintenance therapy. Effective treatment can reverse the neurologic features in most patients, particularly when started early. However, some patients may still progress, especially those with hepatocerebral disease. KF rings tend to decrease after 3–6 months and disappear by 2 years. Adherence to maintenance therapy is a major challenge in long-term care. Patients with advanced hepatic disease may require a liver transplant, and research is looking into the potential role of organ-specific chelators.

NEURODEGENERATION WITH BRAIN IRON ACCUMULATION (NBIA)

NBIA represents a group of inherited disorders characterized by iron accumulation in the basal ganglia. Clinically, they can manifest as a progressive neurologic disorder with a variety of clinical features including parkinsonism, dystonia, neuropsychiatric abnormalities, and retinal degeneration. Cognitive disorders and cerebellar dysfunction may also be seen. Presentation is usually in childhood, but adult cases have been described. Multiple genes have been identified to date. Pantothenate kinase–associated neurodegeneration (PKAN) formerly known as Hallervorden-Spatz disease is caused by a mutation in the *PANK2* gene, and is the most common form of NBIA accounting for about 50% of cases. Onset is usually in early childhood and is manifest as a combination of dystonia, parkinsonism, and spasticity. MRI shows a characteristic low signal abnormality in the center of the globus pallidus on T2-weighted scans caused by iron accumulation known as the "eye of the tiger" sign. Numerous other gene mutations have been described associated with iron accumulation including mutations in *PLA2G6, C19orf12, FA2H, ATP13A2, WDR45, FTL, CP*, and *DCAF17*. One must be cautious, however, not to assume that all cases with iron accumulation in the basal ganglia represent an NBIA, because iron accumulation in specific basal ganglia regions is normal, and excess iron accumulation may occur in the basal ganglia region as a consequence of neurodegeneration associated with multiple causes unrelated to a defect in iron metabolism.

PSYCHOGENIC (FUNCTIONAL) DISORDERS

Virtually all movement disorders including tremor, tics, dystonia, myoclonus, chorea, ballism, and parkinsonism can be psychogenic in origin. Tremor affecting the upper limbs is the most common psychogenic movement disorder. Psychogenic movements can result from a somatoform or conversion disorder, malingering (e.g., seeking financial gain), or a factitious disorder (e.g., seeking psychological gain). Psychogenic movement disorders are relatively common (estimated to be 2–3% of patients seen in a movement disorder clinic), more frequent in women, disabling for the patient and family, and expensive for society. Clinical features suggesting a psychogenic movement disorder include an acute onset with a pattern of abnormal movement that is inconsistent with a known movement disorder. Diagnosis is based on the nonorganic quality of the movement, the absence of findings of an organic disease process, and positive features that specifically point to a psychogenic illness such as variability and distractibility. For example, the magnitude of a psychogenic tremor is increased with attention and diminishes or even disappears when the patient is distracted by being asked to perform a different task or is unaware that he or she is being observed. This is the opposite of an organic tremor where the magnitude is increased with distraction and tends to be reduced when observed. Other positive features suggesting a psychogenic problem include a tremor frequency that is variable or that entrains with the frequency of a designated movement in the contralateral limb, or a response to placebo interventions. Associated features can include nonanatomic sensory findings, give-way weakness, astasia-abasia (an odd, gyrating gait or posture); **(Chap. 23)**, and multiple somatic complaints with no underlying pathology (somatoform disorder). Comorbid psychiatric problems such as anxiety, depression, and emotional trauma may be present but are not necessary for the diagnosis of a psychogenic movement disorder to be made. Psychogenic movement disorders can occur as an isolated entity or in association with an underlying organic problem. The diagnosis can often be made based on clinical features alone, and unnecessary tests or medications can be avoided. Underlying psychiatric problems may be present and should be identified and treated, but many patients with psychogenic movement disorders have no obvious psychiatric pathology. Psychotherapy and hypnosis may be of value for patients with conversion reaction, and cognitive behavioral therapy may be helpful for patients with somatoform disorders. Patients with hypochondriasis, factitious disorders, and malingering have a poor prognosis.

FURTHER READING

Albanese A et al: Therapeutic advances in dystonia. Mov Disord 30:1547, 2015.

Albanese A et al: Phenomenology and classification of dystonia: A consensus update. Mov Disord 28:863, 2013.

Balint B, Bhatia KP: Dystonia: An update on phenomenology, classification, pathogenesis and treatment. Curr Opin Neurol 27:468, 2014.

Elble RJ: The essential tremor syndromes. Curr Opin Neurol 29:507, 2016.

Espay AJ et al: Essential pitfalls in "essential" tremor. Mov Disord 32:325, 2017.

Kieburtz K et al: Huntington's disease: Current and future therapeutic prospects. Mov Disord, 2018.

Krack P et al: Current applications and limitations of surgical treatments for movement disorders. Mov Disord 32:36, 2017.

Marras C et al: Nomenclature of genetic movement disorders: Recommendations of the International Parkinson and Movement Disorder Society task force. Mov Disord 32:724, 2017.

Schapira AH et al: Slowing of neurodegeneration in Parkinson's disease and Huntington's disease: Future therapeutic perspectives. Lancet 384:545, 2014.

429 Amyotrophic Lateral Sclerosis and Other Motor Neuron Diseases

Robert H. Brown, Jr.

AMYOTROPHIC LATERAL SCLEROSIS (ALS)

ALS is the most common progressive motor neuron disease. It is a prime example of a neurodegenerative disease and is arguably the most devastating of the neurodegenerative disorders.

PATHOLOGY

The pathologic hallmark of motor neuron degenerative disorders is death of lower motor neurons (consisting of anterior horn cells in the spinal cord and their brainstem homologues innervating bulbar muscles) and upper, or corticospinal, motor neurons (originating in layer five of the motor cortex and descending via the pyramidal tract to synapse with lower motor neurons, either directly or indirectly via interneurons) **(Chap. 21)**. Although at its onset ALS may involve selective loss of function of only upper or lower motor neurons, it ultimately causes progressive loss of both categories of motor neurons. Indeed, in the absence of clear involvement of both motor neuron types, the diagnosis of ALS is questionable. In a subset of cases, ALS arises concurrently with frontotemporal dementia **(Chap. 424)**; in these instances, there is degeneration of frontotemporal cortical neurons and corresponding cortical atrophy.

Other motor neuron diseases involve only particular subsets of motor neurons **(Tables 429-1 and 429-2)**. Thus, in bulbar palsy and spinal muscular atrophy (SMA; also called *progressive muscular atrophy*), the lower motor neurons of brainstem and spinal cord, respectively, are most severely involved. By contrast, pseudobulbar palsy, primary lateral sclerosis (PLS), and hereditary spastic paraplegia (HSP) affect only upper motor neurons innervating the brainstem and spinal cord.

In each of these diseases, the affected motor neurons undergo shrinkage, often with accumulation of the pigmented lipid (lipofuscin) that normally develops in these cells with advancing age. In ALS, the motor neuron cytoskeleton is typically affected early in the illness. Focal enlargements are frequent in proximal motor axons; ultrastructurally, these "spheroids" are composed of accumulations of neurofilaments and other proteins. Commonly in both sporadic and familial ALS, the affected neurons demonstrate ubiquitin-positive aggregates, typically associated with the protein TDP43 (see below). Also seen is proliferation of astroglia and microglia, the inevitable accompaniment of all degenerative processes in the central nervous system (CNS).

The death of the peripheral motor neurons in the brainstem and spinal cord leads to denervation and atrophy of the corresponding muscle fibers. Histochemical and electrophysiologic evidence indicates that in the early phases of the illness denervated muscle can be reinnervated by sprouting of nearby distal motor nerve terminals, although reinnervation in this disease is considerably less extensive than in most other disorders affecting motor neurons (e.g., poliomyelitis, peripheral neuropathy). As denervation progresses, muscle atrophy is readily recognized in muscle biopsies and on clinical examination. This is the basis for the term *amyotrophy*. The loss of cortical motor neurons results in thinning of the corticospinal tracts that travel via the internal capsule **(Fig. 429-1)** and pyramidal tracts in the brainstem to the lateral and anterior white matter columns of the spinal cord. The loss of fibers in the lateral columns and resulting fibrillary gliosis impart a particular firmness (*lateral sclerosis*). A remarkable feature of the disease is the selectivity of neuronal cell death. By light microscopy, the entire sensory apparatus, the regulatory mechanisms for the control and coordination of movement, remains intact. Except in cases of frontotemporal dementia, the components of the brain required for cognitive processing are also preserved. However, immunostaining indicates that neurons bearing ubiquitin, a marker for degeneration, are also detected in nonmotor systems. Moreover, studies of glucose metabolism in the illness also indicate that there is neuronal dysfunction outside of the motor system. Pathological studies reveal proliferation of microglial cells and astrocytes in affected regions; in some cases, this phenomenon, designated neuroinflamation, can be visualized using positron emission tomography (PET) scanning for ligands that are recognized by activated microglia. Within the motor system, there is some selectivity of involvement. Thus, motor neurons required for ocular motility remain unaffected, as do the parasympathetic neurons in the sacral spinal cord (the nucleus of Onufrowicz, or Onuf) that innervate the sphincters of the bowel and bladder.

CLINICAL MANIFESTATIONS

The manifestations of ALS are somewhat variable depending on whether corticospinal neurons or lower motor neurons in the brainstem and spinal cord are more prominently involved. With lower motor neuron dysfunction and early denervation, typically the first evidence of the disease is insidiously developing asymmetric weakness, usually first evident distally in one of the limbs. A detailed history often discloses recent development of cramping with volitional movements, typically in the early hours of the morning (e.g., while stretching in bed). Weakness caused by denervation is associated with progressive wasting and atrophy of muscles and, particularly early in the illness, spontaneous twitching of motor units, or fasciculations. In the hands, a preponderance of extensor over flexor weakness is common. When the initial denervation involves bulbar rather than limb muscles, the problem at onset is difficulty with chewing, swallowing, and movements of the face and tongue. Rarely, early involvement of the muscles of respiration may lead to death before the disease is far advanced elsewhere. With prominent corticospinal involvement, there is hyperactivity of the muscle-stretch reflexes (tendon jerks) and, often, spastic

TABLE 429-1 Etiology of Motor Neuron Disorders

DIAGNOSTIC CATEGORY	INVESTIGATION
Structural lesions	MRI scan of head (including foramen magnum and cervical spine)
Parasagittal or foramen magnum tumors	
Cervical spondylosis	
Chiari malformation of syrinx	
Spinal cord arteriovenous malformation	
Infections	CSF exam, culture
Bacterial—tetanus, Lyme	Lyme titer
Viral—poliomyelitis, herpes zoster	Anti-viral antibody
Retroviral—myelopathy	HTLV-1 titers
Intoxications, physical agents	24-h urine for heavy metals
Toxins—lead, aluminum, others	Serum lead level
Drugs—strychnine, phenytoin	
Electric short, x-irradiation	
Immunologic mechanisms	Complete blood count[a]
Plasma cell dyscrasias	Sedimentation rate[a]
Autoimmune polyradiculopathy	Total protein[a]
Motor neuropathy with conduction block	Anti-GM1 antibodies[a]
	Anti-Hu antibody
Paraneoplastic	MRI scan, bone marrow biopsy
Paracarcinomatous	
Metabolic	Fasting blood sugar[a]
Hypoglycemia	Routine chemistries including calcium[a]
Hyperparathyroidism	PTH
Hyperthyroidism	Thyroid function[a]
Deficiency of folate, vitamin B_{12}, vitamin E	Vitamin B_{12}, vitamin E, folate[a]
Malabsorption	Serum zinc, copper[a]
Deficiency of copper, zinc	24-h stool fat, carotene, prothrombin time
Mitochondrial dysfunction	Fasting lactate, pyruvate, ammonia
	Consider mtDNA
Hyperlipidemia	Lipid electrophoresis
Hyperglycinuria	Urine and serum amino acids
	CSF amino acids
Hereditary disorders	WBC DNA for mutational analysis
C9orf72	
Superoxide dismutase	
TDP43	
FUS/TLS	
Androgen receptor defect (Kennedy's disease)	

[a]Should be obtained in all cases.

Abbreviations: CSF, cerebrospinal fluid; FUS/TLS, fused in sarcoma/translocated in liposarcoma; HTLV-1, human T-cell lymphotropic virus; MRI, magnetic resonance imaging; PTH, parathyroid; WBC, white blood cell.

TABLE 429-2 Sporadic Motor Neuron Diseases	
CHRONIC	**ENTITY**
Upper and lower motor neuron	Amyotrophic lateral sclerosis
Predominantly upper motor neuron	Primary lateral sclerosis
Predominantly lower motor neuron	Multifocal motor neuropathy with conduction block
	Motor neuropathy with paraproteinemia or cancer
	Motor predominant peripheral neuropathies
Other	
Associated with other neurodegenerative disorders	
Secondary motor neuron disorders (see Table 429-1)	
Acute	
Poliomyelitis	
Herpes zoster	
Coxsackie virus	
West Nile virus	

resistance to passive movements of the affected limbs. Patients with significant reflex hyperactivity complain of muscle stiffness often out of proportion to weakness. Degeneration of the corticobulbar projections innervating the brainstem results in dysarthria and exaggeration of the motor expressions of emotion. The latter leads to involuntary excess in weeping or laughing (pseudobulbar affect).

Virtually any muscle group may be the first to show signs of disease, but, as time passes, more and more muscles become involved until ultimately the disorder takes on a symmetric distribution in all regions. It is characteristic of ALS that, regardless of whether the initial disease involves upper or lower motor neurons, both will eventually be implicated. Even in the late stages of the illness, sensory, bowel and bladder, and cognitive functions are preserved. Even when there is severe brainstem disease, ocular motility is spared until the very late stages of the illness. As noted, in some cases (particularly those that are familial), ALS develops concurrently with frontotemporal dementia, characterized by early behavioral abnormalities with prominent behavioral features indicative of frontal lobe dysfunction.

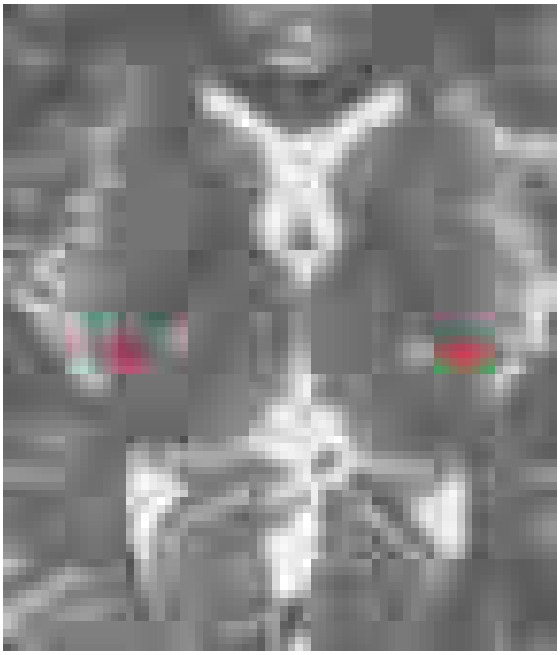


FIGURE 429-1 Amyotrophic lateral sclerosis. Axial T2-weighted magnetic resonance imaging (MRI) scan through the lateral ventricles of the brain reveals abnormal high signal intensity within the corticospinal tracts (*arrows*). This MRI feature represents an increase in water content in myelin tracts undergoing Wallerian degeneration secondary to cortical motor neuronal loss. This finding is commonly present in ALS, but can also be seen in AIDS-related encephalopathy, infarction, or other disease processes that produce corticospinal neuronal loss in a symmetric fashion.

A committee of the World Federation of Neurology has established diagnostic guidelines for ALS. Essential for the diagnosis is simultaneous upper and lower motor neuron involvement with progressive weakness and the exclusion of all alternative diagnoses. The disorder is ranked as "definite" ALS when three or four of the following are involved: bulbar, cervical, thoracic, and lumbosacral motor neurons. When two sites are involved, the diagnosis is "probable," and when only one site is implicated, the diagnosis is "possible." An exception is made for those who have progressive upper and lower motor neuron signs at only one site and a mutation in the gene encoding superoxide dismutase (SOD1; see below).

EPIDEMIOLOGY

The illness is relentlessly progressive, leading to death from respiratory paralysis; the median survival is from 3 to 5 years. There are very rare reports of stabilization or even regression of ALS. In most societies, there is an incidence of 1–3 per 100,000 and a prevalence of 3–5 per 100,000. It is striking that at least 1 in 1000 deaths in North America and Western Europe (and probably elsewhere) are due to ALS; this finding predicts that more than 300,000 individuals now alive in the United States will die of ALS. Several endemic foci of higher prevalence exist in the western Pacific (e.g., in specific regions of Guam or Papua New Guinea). In the United States and Europe, males are somewhat more frequently affected than females. Epidemiologic studies have incriminated risk factors for this disease including exposure to pesticides and insecticides, smoking, and possibly service in the military. Although ALS is overwhelmingly a sporadic disorder, some 10% of cases are inherited as an autosomal dominant trait.

FAMILIAL ALS

Several forms of selective motor neuron disease are inheritable (Table 429-3). Familial ALS (FALS) involves both corticospinal and lower motor neurons. Apart from its inheritance as an autosomal dominant trait, it is clinically indistinguishable from sporadic ALS. Genetic studies have identified mutations in multiple genes, including those encoding the protein C9orf72 (open reading frame 72 on chromosome 9), cytosolic enzyme SOD1 (superoxide dismutase), the RNA binding proteins TDP43 (encoded by the TAR DNA binding protein gene), and fused in sarcoma/translocated in liposarcoma (FUS/TLS), as the most common causes of FALS. Mutations in C9orf72 account for ~45–50% of FALS and perhaps 5% of sporadic ALS cases. Mutations in SOD1 explain another 20% of cases of FALS, whereas TDP43 and FUS/TLS each represent about 5% of familial cases. Mutations in several other genes (such as optineurin, TBK1 and profilin-1) each cause about ~1% of cases.

Rare mutations in other genes are also clearly implicated in ALS-like diseases. Thus, a familial, dominantly inherited motor disorder that in some individuals closely mimics the ALS phenotype arises from mutations in a gene that encodes a vesicle-binding protein. Mutations in senataxin, a helicase, cause an early adult-onset, slowly evolving ALS variant. Kennedy's syndrome is an X-linked, adult-onset disorder that may mimic ALS, as described below. Tau gene mutations usually underlie frontotemporal dementia, but in some instances may be associated with prominent motor neuron findings.

Genetic analyses are also beginning to illuminate the pathogenesis of some childhood-onset motor neuron diseases. For example, a slowly disabling degenerative, predominantly upper motor neuron disease that starts in the first decade is caused by mutations in a gene that expresses a novel signaling molecule with properties of a guanine-exchange factor, termed *alsin*.

DIFFERENTIAL DIAGNOSIS

Because ALS is currently untreatable, it is imperative that potentially remediable causes of motor neuron dysfunction be excluded (Table 429-1). This is particularly true in cases that are atypical by virtue of (1) restriction to either upper or lower motor neurons,

TABLE 429-3 Genetic Motor Neuron Diseases

DISEASE	GENE SYMBOL	GENE NAME	INHERITANCE	FREQUENCY (IN THE UNITED STATES)	USUAL ONSET	PROTEIN FUNCTION	UNUSUAL FEATURES
I. Upper and Lower Motor Neurons (Familial ALS)							
ALS1	SOD1	Cu/Zn superoxide dismutase 1	AD	20% FALS	Adult	Protein antioxidant	
ALS2	ALS2	Alsin	AR	<1% FALS	Juvenile	GEF signaling	Severe corticobulbar, corticospinal featuresk may mimic PLS
ALS4	SETX	Senataxin	AD	~1% FALS	Late juvenile	DNA helicase	Late childhood onset
ALS6	FUS/TLS	Fused in Sarcoma/ Translocated in liposarcoma	AD	5% FALS	Adult	DNA, RNA binding	
ALS8 /SMA	VAPB	Vesicle associated protein B	AD	<1%	Adult	Vesicular trafficking	
ALS9	ANG	Angiogenin	AD	<1%	Adult	RNAse, angiogenesis	
ALS10	TARDBP	TAR DNA binding protein	AD	5% FALS	Adult	DNA, RNA binding	
ALS12	OPTN	Optineurin	AD/AR	~1% FALS	Adult	Attenuates NF-κB	
ALS13	ATXN2	Ataxin 2	AD	<1%	Adult	Cytotoxic expanded CAG repeat	
ALS14	VCP	Valosin-containing protein	AD	~ 1% FALS	Adult	ATPase	
ALS18	PFN1	Profilin 1	AD	~1% FALS	Adult	Involved in actin polymerization	
ALS19	ERB4	v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 4	AD		Adult	Signaling molecule	
ALS20	HNRNPA1	Heterogeneous nuclear ribonucleoprotein A1	AD	<1%	Adult	Heteronuclear RNA binding protein	
ALS	DCTN1	Dynactin	AD	<1%	Adult	Axonal transport	May cause vocal cord paralysis or PLSe
ALS-FTD	TBK1	TankBinding Protein 1	AD		Adult	NF-κB signalling	also mimics PLS
ALS-FTD	UBQLN2	Ubiquilin 2	X-LD	<1%	Adult or Juvenile	Protein degradation	
ALS-FTD	CHMP2B	Chromatin modifying protein 2B	AD	<1% FALS	Adult	Chromatin binding protein	
ALS-FTD	C9ORF72	Chromosome 9 Open Reading Frame 72	AD	40-50% FALS	Adult	Regulates vessicle trafficking	May also be associated with Parkinosonism, PLS
ALS-FTD	MAPT	Microtubule Associated Protein Tau	AD		Adult	cytoskeletal protein	Usually causes only FTD
ALS	COX	Cytochrome c oxidase	Maternally inherited		Adult	Mitochondrial: ATP generation	
ALS	tRNA-isoleucine		Maternally inherited		Adult	Mitochondrial: ATP generation	
II. Lower Motor Neurons							
Spinal muscular atrophies	SMN	Survival motor neuron	AR	1/10,000 live births	Infancy	RNA metabolism	
GM2-gangliosidosis							
1. Sandhoff's disease	HEXB	Hexosaminidase B	AR		Childhood	Ganglioside recycling	
2. AB variant	GM2A	GM2-activator protein	AR		Childhood	Ganglioside recycling	
3. Adult Tay-Sachs disease	HEXA	Hexosaminidase A	AR		Childhood	Ganglioside recycling	
X-linked spinobulbar muscular atrophy	AR	Androgen receptor	XR		Adult	Nuclear signaling	
III. Upper Motor Neuron (Selected HSPs)							
SPG3A	ATL1	Atlastin	AD	10% AD FSP	Childhood	GTPase—vesicle recycling	
SPG4	SPAST	Spastin	AD	50-60% AD FSP	Early adulthood	ATPase family—microtubule associate	Some sensory loss

(*Continued*)

TABLE 429-3 Genetic Motor Neuron Diseases (Continued)

DISEASE	GENE SYMBOL	GENE NAME	INHERITANCE	FREQUENCY (IN THE UNITED STATES)	USUAL ONSET	PROTEIN FUNCTION	UNUSUAL FEATURES
SPG6	NIPA1	Non imprinted in Prader-Willi/Angelman syndrome 1	AD		Early adulthood	Membrane transporter or receptor	Deleted in Prader-Willi, Angelman's
SPG8	WASHC5	Strumpellin	AD		Early adulthood	Ubiquitous, spectrin-like	
SPG10	KIF5A	Kinesin heavy chain isoform 5A	AD	10% AD FSP	Second–third decade	Motor-associated protein	± Peripheral neuropathy, retardation
SPG12	RTN2	Reticulon 2	AD		Childhood	ER protein, interacts with spastin	
SPG13	HSP60	Heat shock protein 60	AD		Early adulthood	Chaperone protein	
SPG17/ variants of charcot Marie Tooth type 2/ Silver syndrome	BSCL2	Seipin lipid droplet biogenesis associated	AD		Variable	Membrane protein in ER	Amyotrophy hands, feet
SPG31	REEP1	Receptor Expression Enhancing Protein 1	AD	10% AD FSP	Early	Mitochondrial protein	Rarely, amyotrophy
SPG33	ZFYVE27	Zinc Finger FYVE-Type Containing 27	AD		Adult	Interacts with spastin	Pes equinus
SPG42	SLC33A1	Acetyl-CoA-transporter	AD		Variable	Solute carrier	
SPG72	REEP2	Receptor Expression Enhancing Protein 2	AD		Childhood	ER protein	
SPG5	CYP7B1	Cytochrome P450	AR	5-10% AR FSP	Variable	Degrades endogenous substances	Sensory loss
SPG7	SPG7	Paraplegin	AR	5-10% AR FSP	Variable	Mitochondrial protein	Rarely, optic atrophy, ataxia, rarely PLS
SPG11	SPG11	Spatacsin	AR	20-70% AR FSP depends on ethnicity	Predominantly childhood	Cytosolic, ? membrane-associated	Some sensory loss, thin corpus callosum; may mimic ALS (ALS5)
SPG15	ZFYVE26	Spastizin	AR		Childhood	Zinc finger protein	Some amyotrophy, some CNS features
SPG20	SPG20	Spartin	AR		Childhood	Endosomal trafficking protein	
SPG21	SPG21	Maspardin	AR		Childhood	Endosomal trafficking protein	
SPG35	Fatty acid 2 hydrolase		AR		Childhood	Membrane protein	Multiple CNS features
SPG39	PNPLA6	patatin-like phospholipase domain-containing protein 6 / Neuropathy target esterase	AR		Early childhood	Esterase	May have PLS=like phenotype
SPG44	GJC2	Gap junction protein gamma 2/ Connexin 47	AR		Childhood	Gap junction protein	Possible mild CNS features
SPG46	GBA2	β-Glucosidase 2	AR		Childhood	Glycoside hydrolase	Thin corpus callosum, mental retardation
SPG2	PLP	Proteolipid protein	XR		Early childhood	Myelin protein	Sometimes multiple CNS features
SPG1	L1-CAM	Neural cell adhesion molecule L1 precursor	XR		Infancy	Cell adhesion molecule	
SPG22	SLC16A2	Solute Carrier Family 16 Member 2	XR		Infancy	Monocarboxylic acid transporter	
Adrenoleukodystrophy	ALDP	Adrenoleukodystrophy protein	XR		Early adulthood	ATP binding transporter protein	Possible adrenal insufficiency, CNS inflammation

Abbreviations: AD, autosomal dominant; ALS, amyotrophic lateral sclerosis; AR, autosomal recessive; CNS, central nervous system; BSCL2, Bernadelli-Seip congenital lipodystrophy 2B; FUS/TLS, fused in sarcoma/translocated in liposarcoma; GEF, Guanidine nucleotide exchange factor; HSP, hereditary spastic paraplegia; TDP43, Tar DNA binding protein 43 kd; XR, X-linked recessive.

(2) involvement of neurons other than motor neurons, and (3) evidence of motor neuronal conduction block on electrophysiologic testing. Compression of the cervical spinal cord or cervicomedullary junction from tumors in the cervical regions or at the foramen magnum or from cervical spondylosis with osteophytes projecting into the vertebral canal can produce weakness, wasting, and fasciculations in the upper limbs and spasticity in the legs, closely resembling ALS. The absence of cranial nerve involvement may be helpful in differentiation, although some foramen magnum lesions may compress the twelfth cranial (hypoglossal) nerve, with resulting paralysis of the tongue. Absence of pain or of sensory changes, normal bowel and bladder function, normal radiologic studies of the spine, and normal cerebrospinal fluid (CSF) all favor ALS. Where doubt exists, magnetic resonance imaging (MRI) scans and possibly contrast myelography should be performed to visualize the cervical spinal cord.

Another important entity in the differential diagnosis of ALS is *multifocal motor neuropathy with conduction block* (MMCB), discussed below. A diffuse, lower motor axonal neuropathy mimicking ALS sometimes evolves in association with hematopoietic disorders such as lymphoma or multiple myeloma. In this clinical setting, the presence of an M-component in serum should prompt consideration of a bone marrow biopsy. Lyme disease **(Chap. 181)** may also cause an axonal, lower motor neuropathy, although typically with intense proximal limb pain and a CSF pleocytosis.

Other treatable disorders that occasionally mimic ALS are chronic lead poisoning and thyrotoxicosis. These disorders may be suggested by the patient's social or occupational history or by unusual clinical features. When the family history is positive, disorders involving the genes encoding C9orf72, cytosolic SOD1, TDP43, FUS/TLS, and adult hexosaminidase A or α-glucosidase deficiency must be excluded **(Chap. 411)**. These are readily identified by appropriate laboratory tests. Benign fasciculations are occasionally a source of concern because on inspection they resemble the fascicular twitchings that accompany motor neuron degeneration. The absence of weakness, atrophy, or denervation phenomena on electrophysiologic examination usually excludes ALS or other serious neurologic disease. Patients who have recovered from poliomyelitis may experience a delayed deterioration of motor neurons that presents clinically with progressive weakness, atrophy, and fasciculations. Its cause is unknown, but it is thought to reflect sublethal prior injury to motor neurons by poliovirus **(Chap. 199)**.

Rarely, ALS develops concurrently with features indicative of more widespread neurodegeneration. Thus, one infrequently encounters otherwise typical ALS patients with a parkinsonian movement disorder or frontotemporal dementia, particularly in instances of C9orf72 mutations, which strongly suggests that the simultaneous occurrence of two disorders is a direct consequence of the gene mutation. As another example, prominent amyotrophy has been described as a dominantly inherited disorder in individuals with bizarre behavior and a movement disorder suggestive of parkinsonism; many such cases have now been ascribed to mutations that alter the expression of tau protein in brain **(Chap. 424)**. In other cases, ALS develops simultaneously with a striking frontotemporal dementia. An ALS-like disorder has also been described in some individuals with chronic traumatic encephalopathy, associated with deposition of TDP43 and neurofibrillary tangles in motor neurons.

PATHOGENESIS

The cause of sporadic ALS is not well defined. Several mechanisms that impair motor neuron viability have been elucidated in rodents induced to develop motor neuron disease by SOD1 or profilin-1 transgenes with ALS-associated mutations. One may loosely group the genetic causes of ALS into three categories. In one group, the primary problem is inherent instability of the mutant proteins, with subsequent perturbations in protein degradation (SOD1, ubiquilin-1 and 2, p62). In the second category, the causative mutant genes perturb RNA processing, transport, and metabolism (C9orf73, TDP43, FUS). In the case of C9orf72, the molecular pathology is an expansion of an intronic hexanucleotide repeat (-GGGGCC-) beyond an upper normal of 30 repeats to hundreds or even thousands of repeats. As observed in other intronic repeat disorders such as myotonic dystrophy **(Chap. 441)** and spinocerebellar atrophy type 8 **(Chap. 431)**, data suggest that the expanded intronic repeats generate expanded RNA repeats that form intranuclear foci and may confer toxicity by sequestering transcription factors or by undergoing noncanonical protein translation across all possible reading frames of the expanded RNA tracts. Importantly, the latter process generates lengthy dipeptides that are detected in the spinal fluid and are a unique biomarker for C9orf72 ALS. TDP43 and FUS are multifunctional proteins that bind RNA and DNA and shuttle between the nucleus and the cytoplasm, playing multiple roles in the control of cell proliferation, DNA repair and transcription, and gene translation, both in the cytoplasm and locally in dendritic spines in response to electrical activity. How mutations in FUS/TLS provoke motor neuron cell death is not clear, although this may represent loss of function of FUS/TLS in the nucleus or an acquired, toxic function of the mutant proteins in the cytosol. In the third group of ALS genes, the primary problem is defective axonal cytoskeleton and transport (dynactin, profilin-1). It is striking that variants in other genes (e.g., EphA4) influence survival in ALS but not ALS susceptibility. Data indicate that intermediate-length polyglutamine-coding expansions (-CAG-) in the gene *ataxin-2* confer increased ALS susceptibility; suppression of ataxin-2 expression extends survival in transgenic ALS mice. Beyond the upstream, primary defects, it is also evident that the ultimate neuronal cell death process is complex, involving multiple cellular processes acting in diverse components of the motor neuron (dendrites, cell body, axons, neuromuscular junction) to accelerate cell death. These include but are not limited to excitotoxicity, defective autophagy, impairment of axonal transport, oxidative stress, activation of endoplasmic reticulum stress and the unfolded protein response, and mitochondrial dysfunction. As well, the hexanucleotide expansions that cause C9orf72 ALS disrupt nucleocytoplasmic transport; the importance of this observation is underscored by the finding that mutations in the gene encoding GLE1, a protein that mediates mRNA export, cause an aggressive, infantile motor neuron disease.

Multiple recent studies have convincingly demonstrated that proliferating, activated nonneuronal cells such as microglia and astrocytes importantly influence the disease course, at least in ALS transgenic mice. A striking additional finding in ALS and most neurodegenerative disorders is that miscreant proteins arising from gene defects in familial forms of these diseases are often implicated in sporadic forms of the same disorder. For example, some reports propose that nonheritable, posttranslational modifications in SOD1 are pathogenic in sporadic ALS; indeed, SOD1 aggregates are sometimes observed in spinal cord in sporadic ALS without SOD1 mutations. Germline mutations in the genes encoding β-amyloid and α-synuclein cause familial forms of Alzheimer's and Parkinson's diseases, and posttranslational, noninherited abnormalities in these proteins are also central to sporadic Alzheimer's and Parkinson's diseases.

TREATMENT

Amyotrophic Lateral Sclerosis

No treatment arrests the underlying pathologic process in ALS. The drug riluzole (100 mg/d) was approved for ALS because it produces a modest lengthening of survival. In one trial, the survival rate at 18 months with riluzole was similar to placebo at 15 months. The mechanism of this effect is not known with certainty; riluzole may reduce excitotoxicity by diminishing glutamate release. Riluzole is generally well tolerated; nausea, dizziness, weight loss, and elevated liver enzymes occur occasionally. A second drug, edaravone, has also been approved by the U.S. Food and Drug Administration based on a single 6-month study in a highly selected ALS population that demonstrated a modest reduction in the trajectory of worsening on an ALS disability scale; survival was not included as an endpoint. This drug, which is believed to act as an antioxidant, is administered via recurring monthly 10-day series of daily intravenous infusions. Pathophysiologic studies of mutant SOD1–related ALS in mice have disclosed diverse targets for therapy; consequently,

multiple therapies are presently in clinical trials for ALS including experimental trials of small molecules, mesenchymal stem cells, and immunosuppression. Interventions such as antisense oligonucleotides (ASO) that diminish expression of mutant SOD1 protein prolong survival in transgenic ALS mice and rats are also now in clinical trials for SOD1-mediated ALS.

In the absence of a primary therapy for ALS, a variety of rehabilitative aids may substantially assist ALS patients. Foot-drop splints facilitate ambulation by obviating the need for excessive hip flexion and by preventing tripping on a floppy foot. Finger extension splints can potentiate grip. Respiratory support may be life-sustaining. For patients electing against long-term ventilation by tracheostomy, positive-pressure ventilation by mouth or nose provides transient (weeks to months) relief from hypercarbia and hypoxia. Also extremely beneficial for some patients is a respiratory device (Cough Assist Device) that produces an artificial cough. This is highly effective in clearing airways and preventing aspiration pneumonia. When bulbar disease prevents normal chewing and swallowing, gastrostomy is uniformly helpful, restoring normal nutrition and hydration. Fortunately, an increasing variety of speech synthesizers are now available to augment speech when there is advanced bulbar palsy. These facilitate oral communication and may be effective for telephone use.

In contrast to ALS, several of the disorders (Tables 429-1 and 429-3) that bear some clinical resemblance to ALS are treatable. For this reason, a careful search for causes of secondary motor neuron disease is warranted.

OTHER MOTOR NEURON DISEASES

SELECTED LOWER MOTOR NEURON DISORDERS

In these motor neuron diseases, the peripheral motor neurons are affected without evidence of involvement of the corticospinal motor system (Tables 429-1–429-3).

X-Linked Spinobulbar Muscular Atrophy (Kennedy's Disease) This is an X-linked lower motor neuron disorder in which progressive weakness and wasting of limb and bulbar muscles begins in males in mid-adult life and is conjoined with androgen insensitivity manifested by gynecomastia and reduced fertility **(Chap. 384)**. In addition to gynecomastia, which may be subtle, two findings distinguishing this disorder from ALS are the absence of signs of pyramidal tract disease (spasticity) and the presence of a subtle sensory neuropathy in some patients. The underlying molecular defect is an expanded trinucleotide repeat (-CAG-) in the first exon of the androgen receptor gene on the X chromosome. An inverse correlation appears to exist between the number of CAG- repeats and the age of onset of the disease.

Adult Tay-Sachs Disease Several reports have described adult-onset, predominantly lower motor neuropathies arising from deficiency of the enzymeβ-hexosaminidase (hex A). These tend to be distinguishable from ALS because they are very slowly progressive and in some cases may have been symptomatic for years; dysarthria and radiographically evident cerebellar atrophy may be prominent. In rare cases, spasticity may also be present, although it is generally absent **(Chap. 411)**.

Spinal Muscular Atrophy The SMAs are a family of selective lower motor neuron diseases of early onset. Despite some phenotypic variability (largely in age of onset), the defect in the majority of families with SMA is loss of a protein (SMN, for survival motor neuron) that is important in the formation and trafficking of RNA complexes across the nuclear membrane. Neuropathologically these disorders are characterized by extensive loss of large motor neurons; muscle biopsy reveals evidence of denervation atrophy. Several clinical forms exist.

Infantile SMA (SMA I, Werdnig-Hoffmann disease) has the earliest onset and most rapidly fatal course. In some instances, it is apparent even before birth, as indicated by decreased fetal movements late in the third trimester. Though alert, afflicted infants are weak and floppy (hypotonic) and lack muscle stretch reflexes. Death generally ensues within the first year of life. *Chronic childhood SMA* (SMA II) begins later in childhood and evolves with a more slowly progressive course. *Juvenile SMA* (SMA III, Kugelberg-Welander disease) manifests during late childhood and runs a slow, indolent course. Unlike most denervating diseases, in this chronic disorder weakness is greatest in the proximal muscles; indeed, the pattern of clinical weakness can suggest a primary myopathy such as limb-girdle dystrophy. Electrophysiologic and muscle biopsy evidence of denervation distinguish SMA III from the myopathic syndromes. Remarkably, two treatments have shown dramatic benefit in infantile SMA. One, nusinersen, now an approved therapy, entails administering small oligonucleotides that alter mRNA splicing of one of the SMN genes, generating sufficient normal SMN protein to provide clinical benefit (including prolonged survival). The other uses systemically administered adeno-associated virus (AAV) to deliver the missing SMN gene to motor neurons and other cells.

Multifocal Motor Neuropathy with Conduction Block In this disorder, lower motor neuron function is regionally and chronically disrupted by focal blocks in conduction. Many cases have elevated serum titers of mono- and polyclonal antibodies to ganglioside GM1; it is hypothesized that the antibodies produce selective, focal, paranodal demyelination of motor neurons. MMCB is not typically associated with corticospinal signs. In contrast with ALS, MMCB may respond dramatically to therapy such as IV immunoglobulin or chemotherapy; thus, it is imperative that MMCB be excluded when considering a diagnosis of ALS.

Other Forms of Lower Motor Neuron Disease In individual families, other syndromes characterized by selective lower motor neuron dysfunction in an SMA-like pattern have been described. There are rare X-linked and autosomal dominant forms of apparent SMA. There is an ALS variant of juvenile onset, the Fazio-Londe syndrome, that involves mainly the musculature innervated by the brainstem. A component of lower motor neuron dysfunction is also found in degenerative disorders such as Machado-Joseph disease and the related olivopontocerebellar degenerations **(Chap. 431)**.

SELECTED DISORDERS OF THE UPPER MOTOR NEURON

Primary Lateral Sclerosis This rare disorder arises sporadically in adults in mid to late life. Clinically, PLS is characterized by progressive spastic weakness of the limbs, preceded or followed by spastic dysarthria and dysphagia, indicating combined involvement of the corticospinal and corticobulbar tracts. Fasciculations, amyotrophy, and sensory changes are absent; neither electromyography nor muscle biopsy shows denervation. On neuropathologic examination, there is selective loss of the large pyramidal cells in the precentral gyrus and degeneration of the corticospinal and corticobulbar projections. The peripheral motor neurons and other neuronal systems are spared. The course of PLS is variable; although long-term survival is documented, the course may be as aggressive as in ALS, with ~3-year survival from onset to death. Early in its course, PLS raises the question of multiple sclerosis or other demyelinating diseases as diagnostic considerations **(Chap. 436)**. A myelopathy suggestive of PLS is infrequently seen with infection with the retrovirus human T cell lymphotropic virus 1 (HTLV-1) **(Chap. 434)**. The clinical course and laboratory testing will distinguish these possibilities.

Hereditary Spastic Paraplegia In its pure form, HSP is usually transmitted as an autosomal trait; most adult-onset cases are dominantly inherited. There are more than 80 genetic types of HSP for which causative mutations in more than 60 genes have been identified. Table 429-3 lists more commonly identified genetic types of HSP. Symptoms usually begin in the third or fourth decade, presenting as progressive spastic weakness beginning in the lower extremities; however, there are variants with onset so early that the differential diagnosis includes cerebral palsy. HSP typically has a long survival, presumably because respiratory function is spared. Late in the illness, there may be urinary urgency and incontinence and sometimes fecal incontinence; sexual function tends to be preserved.

In pure forms of HSP, the spastic leg weakness is often accompanied by posterior column (vibration and position) abnormalities and disturbance of bowel and bladder function. Some family members may have spasticity without clinical symptoms.

By contrast, particularly when recessively inherited, HSP may have complex or complicated forms in which altered corticospinal and dorsal column function is accompanied by significant involvement of other regions of the nervous system, including amyotrophy, mental retardation, optic atrophy, and sensory neuropathy.

Neuropathologically, in HSP there is degeneration of the corticospinal tracts, which appear nearly normal in the brainstem but show increasing atrophy at more caudal levels in the spinal cord; in effect, this pathologic picture is of a dying-back or distal axonopathy of long neuronal fibers within the CNS.

Defects at numerous loci underlie both dominantly and recessively inherited forms of HSP (Table 429-3). The gene most commonly implicated in dominantly inherited HSP is *spastin*, which encodes a microtubule interacting protein. The most common childhood-onset dominant form arises from mutations in the *atlastin* gene.

An infantile-onset form of X-linked, recessive HSP arises from mutations in the gene for myelin proteolipid protein. This is an example of rather striking allelic variation, as most other mutations in the same gene cause not HSP but Pelizaeus-Merzbacher disease, a widespread disorder of CNS myelin. Another recessive variant is caused by defects in the *paraplegin* gene. Paraplegin has homology to metalloproteases that are important in mitochondrial function in yeast.

FURTHER READING

Brown RH, Al-Chalabi A: Review article: Amyotrophic lateral sclerosis. N Engl J Med 377:162, 2017.

Finkel RS et al: Treatment of infantile-onset spinal muscular atrophy with nusinersin: A phase 2, open-label, dose-escalation study. Lancet 388:3017, 2016.

Gendron TF et al: Poly(GP) proteins are a useful pharmacodynamic marker for C9ORF72-associated amyotrophic lateral sclerosis. Sci Transl Med 9:pii:eaai7866, 2017.

Miller TM et al: An antisense oligonucleotide against SOD1 delivered intrathecally for patients with SOD1 familial amyotrophic lateral sclerosis: A phase 1, randomised, first-in-man study. Lancet Neurol 12:435, 2013.

Robberecht W, Philips T: The changing scene of amyotrophic lateral sclerosis. Nat Rev Neurosci 14:248, 2013.

Schüle R et al: Hereditary spastic paraplegia: Clinicogenetic lessons from 608 patients. Ann Neurol 79:646, 2016.

Taylor JP et al: Decoding ALS: From genes to mechanism. Nature 539:197, 2016.

The Writing Group on Behalf of the Edaravone (MCI-186) ALS 19 Study Group: Safety and efficacy of edaravone in well defined patients with amyotrophic lateral sclerosis: A randomised, double-blind, placebo controlled trial. Lancet Neurol 16:505, 2017.

WEBSITES

Several websites provide valuable information on ALS including those offered by the Muscular Dystrophy Association (*www.mdausa.org*), the Amyotrophic Lateral Sclerosis Association (*www.alsa.org*), and the World Federation of Neurology and the Neuromuscular Unit at Washington University in St. Louis (*www.neuro.wustl.edu*).

430 Prion Diseases

Stanley B. Prusiner, Bruce L. Miller

Prions are proteins that adopt an alternative conformation, which becomes self-propagating. Some prions cause degeneration of the central nervous system (CNS). Once relegated to causing a group of rare disorders of the CNS such as Creutzfeldt-Jakob disease (CJD), prions also appear to play a role in more common illnesses such as Alzheimer's disease (AD) and Parkinson's disease (PD). While CJD is caused by the accumulation of PrP^{Sc} prions, recent investigations demonstrate unequivocally that α-synuclein prions cause multiple system atrophy (MSA). Infectious MSA prions have been recovered from human brain samples stored in formalin for up to 20 years. Similar resistance to formalin was demonstrated for brain samples from sheep with scrapie. Increasing data argue that Aβ prions contribute to AD, α-synuclein prions to PD, and tau prions to some types of frontotemporal dementia (FTD). In this chapter, we confine our discussion to CJD, which typically presents with a rapidly progressive dementia as well as motor abnormalities. The illness is relentlessly progressive and generally causes death within 9 months of onset. Most CJD patients are between 50 and 75 years of age; however, patients as young as 17 and as old as 83 have been recorded. The role of prions in the pathogenesis of neurodegenerative diseases is reviewed in **Chap. 417**.

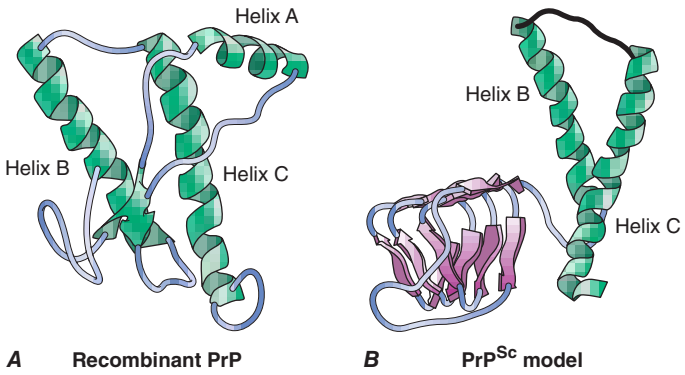


FIGURE 430-1 Structures of PrP prion proteins. *A.* Nuclear magnetic resonance structure of Syrian hamster recombinant (rec) PrP(90–231). Presumably, the structure of the α-helical form of recPrP(90–231) resembles that of PrP^{C}. recPrP(90–231) is viewed from the interface where PrP^{Sc} is thought to bind to PrP^{C}. Shown are α-helices A (residues 144–157), B (172–193), and C (200–227). Flat ribbons depict β-strands S1 (129–131) and S2 (161–163). ***B.*** Structural model of PrP^{Sc}. The 90–160 region has been modeled onto a β-helical architecture while the COOH terminal helices B and C are preserved as in PrP^{C}.

CJD is one malady in a group of disorders caused by prions composed of the prion protein (PrP). PrP prions reproduce by binding to the normal, *c*ellular isoform of the *pr*ion *p*rotein (PrP^{C}) and stimulating conversion of PrP^{C} into the disease-causing isoform PrP^{Sc}. PrP^{C} is rich in α-helix and has little β-structure, whereas PrP^{Sc} has less α-helix and a high amount of β-structure **(Fig. 430-1)**. This α-to-β structural transition in PrP is the fundamental event underlying this group of prion diseases **(Table 430-1)**.

Four new concepts have emerged from studies of PrP prions: (1) Prions are the only known transmissible pathogens that are devoid of nucleic acid; all other infectious agents possess genomes composed of either RNA or DNA that direct the synthesis of their progeny. (2) Prion diseases may manifest as infectious, genetic, or sporadic disorders; no other group of illnesses with a single etiology presents with such a wide spectrum of clinical manifestations. (3) Prion diseases result from the accumulation of PrP^{Sc}, the conformation of which differs substantially from that of its precursor, PrP^{C}. (4) Distinct strains of prions exhibit different biologic properties, which are epigenetically inherited. In other words, PrP^{Sc} can exist in a variety of different conformations, many of which seem to specify particular disease phenotypes.

How a specific conformation of a PrP^{Sc} molecule is imparted to PrP^{C} during prion replication to produce nascent PrP^{Sc} with the same conformation is unknown. Additionally, it is unclear what factors determine where in the CNS a particular PrP^{Sc} molecule will be deposited.

SPECTRUM OF PrP PRION DISEASES

The sporadic form of CJD is the most common prion disorder in humans. Sporadic CJD (sCJD) accounts for ~85% of all cases of human PrP prion disease, whereas inherited prion diseases account for 10–15% of all cases **(Table 430-2)**. Familial CJD (fCJD), Gerstmann-Sträussler-Scheinker (GSS) disease, and fatal familial insomnia (FFI) are all dominantly inherited prion diseases that are caused by mutations in the PrP gene.

TABLE 430-1 Glossary of PrP Prion Terminology	
Prion	*Pro*teinaceous *in*fectious particle that lacks nucleic acid. Prions are composed entirely of alternatively folded proteins that undergo self-propagation. Distinct strains of prions exhibit different biologic properties, which are epigenetically heritable. PrP prions cause scrapie in sheep and goats, mad cow disease, and related neurodegenerative diseases of humans such as Creutzfeldt-Jakob disease (CJD).
PrP^{Sc}	Disease-causing isoform of the prion protein. This protein is the only identifiable macromolecule in purified preparations of scrapie prions.
PrP^{C}	Cellular isoform of the prion protein. PrP^{C} is the precursor of PrP^{Sc}.
PrP 27-30	A fragment of PrP^{Sc}, generated by truncation of the NH_2-terminus by limited digestion with proteinase K. PrP 27-30 retains prion infectivity and polymerizes into amyloid.
PRNP	PrP gene located on human chromosome 20.
Prion rod	An aggregate of prions composed largely of PrP 27-30 molecules. Created by detergent extraction and limited proteolysis of PrP^{Sc}. Morphologically and histochemically indistinguishable from many amyloids.
PrP amyloid	Amyloid containing PrP in the brains of animals or humans with prion disease; often accumulates as plaques.

Although infectious PrP prion diseases account for <1% of all cases and infection does not seem to play an important role in the natural history of these illnesses, the transmissibility of PrP prions is an important biologic feature. Kuru of the Fore people of New Guinea is thought to have resulted from the consumption of brains from dead relatives during ritualistic cannibalism. After the cessation of ritualistic cannibalism in the late 1950s, kuru nearly disappeared, with the exception of a few recent patients exhibiting incubation periods of >40 years. Iatrogenic CJD (iCJD) seems to be the result of the accidental inoculation of patients with prions. Variant CJD (vCJD) in teenagers and young adults in Europe is the result of exposure to tainted beef from cattle with bovine spongiform encephalopathy (BSE). Although occasional cases of iatrogenic CJD still occur, this form of CJD is currently on the decline due to public health measures aimed at preventing the spread of PrP prions.

Six diseases of animals are caused by prions (Table 430-2). Scrapie of sheep and goats is the prototypic PrP prion disease. Mink encephalopathy, BSE, feline spongiform encephalopathy, and exotic ungulate encephalopathy are all thought to occur after the consumption of prion-infected foodstuffs. The BSE epidemic emerged in Britain in the late 1980s and was shown to be due to industrial cannibalism. Whether BSE began as a sporadic case of BSE in a cow or started with scrapie in sheep is unknown. The origin of chronic wasting disease (CWD), a prion disease endemic in deer and elk in regions of North America, is uncertain. In contrast to other prion diseases, CWD is highly communicable. Feces from asymptomatic, infected cervids contain prions that are likely to be responsible for the spread of CWD.

TABLE 430-2 The PrP Prion Diseases		
DISEASE	**HOST**	**MECHANISM OF PATHOGENESIS**
Human		
Kuru	Fore people	Infection through ritualistic cannibalism
iCJD	Humans	Infection from prion-contaminated hGH, dura mater grafts, etc.
vCJD	Humans	Infection from bovine prions
fCJD	Humans	Germline mutations in *PRNP*
GSS	Humans	Germline mutations in *PRNP*
FFI	Humans	Germline mutation in *PRNP* (D178N, M129)
sCJD	Humans	Somatic mutation or spontaneous conversion of PrP^{C} into PrP^{Sc}?
sFI	Humans	Somatic mutation or spontaneous conversion of PrP^{C} into PrP^{Sc}?
Animal		
Scrapie	Sheep, goats	Infection in genetically susceptible sheep
BSE	Cattle	Infection with prion-contaminated MBM
TME	Mink	Infection with prions from sheep or cattle
CWD	Mule deer, elk	Unknown
FSE	Cats	Infection with prion-contaminated beef
Exotic ungulate encephalopathy	Greater kudu, nyala, or oryx	Infection with prion-contaminated MBM

Abbreviations: BSE, bovine spongiform encephalopathy; CJD, Creutzfeldt-Jakob disease; CWD, chronic wasting disease; fCJD, familial Creutzfeldt-Jakob disease; FFI, fatal familial insomnia; FSE, feline spongiform encephalopathy; GSS, Gerstmann-Sträussler-Scheinker disease; hGH, human growth hormone; iCJD, iatrogenic Creutzfeldt-Jakob disease; MBM, meat and bone meal; sCJD, sporadic Creutzfeldt-Jakob disease; sFI, sporadic fatal insomnia; TME, transmissible mink encephalopathy; vCJD, variant Creutzfeldt-Jakob disease.

EPIDEMIOLOGY

CJD is found throughout the world. The incidence of sCJD is ~1 case per million population, and accounts for ~1 in every 10,000 deaths. Because sCJD is an age-dependent neurodegenerative disease, its incidence is expected to increase steadily as older segments of populations in developed and developing countries continue to expand. Although many geographic clusters of CJD have been reported, each has been shown to segregate with a PrP gene mutation. Attempts to identify common exposure to some etiologic agent have been unsuccessful for both the sporadic and familial cases. Ingestion of scrapie-infected sheep or goat as a cause of CJD in humans has not been demonstrated by epidemiologic studies, although speculation about this potential route of infection continues. Of particular interest are deer hunters who develop CJD, because up to 90% of culled deer in some game herds have been shown to harbor CWD prions. Whether prion disease in deer or elk has passed to cows, sheep, or directly to humans remains unknown. Studies with rodents demonstrate that oral infection with prions can occur, but the process is inefficient compared to intracerebral inoculation.

PATHOGENESIS

The human PrP prion diseases were initially classified as neurodegenerative disorders of unknown etiology on the basis of pathologic changes being confined to the CNS. With the transmission of kuru and CJD to apes, investigators began to view these diseases as infectious CNS illnesses caused by slow viruses. Even though the familial nature of a subset of CJD cases was well described, the significance of this observation became more obscure with the transmission of CJD to animals. Eventually, the meaning of heritable CJD became clear with the discovery of mutations in the *PRNP* gene of these patients. The prion concept explains how a disease can manifest as a heritable as well as an infectious illness. Moreover, the hallmark of all PrP prion diseases, whether sporadic, dominantly inherited, or acquired by infection, is that they involve the aberrant metabolism of PrP.

A major feature that distinguishes PrP prions from viruses is the finding that both PrP isoforms are encoded by a chromosomal gene. In humans, the PrP gene is designated *PRNP* and is located on the short arm of chromosome 20. Limited proteolysis of PrP^{Sc} produces a smaller, protease-resistant molecule of ~142 amino acids designated PrP 27-30; PrP^{C} is completely hydrolyzed under the same conditions **(Fig. 430-2)**. In the presence of detergent, PrP 27-30 polymerizes into amyloid. Prion rods formed by limited proteolysis and detergent extraction are indistinguishable from the filaments that aggregate to form PrP amyloid plaques in the CNS. Both the rods and the PrP amyloid filaments found in brain tissue exhibit similar ultrastructural morphology and green-gold birefringence after staining with Congo red dye.

Prion Strains Distinct strains of PrP prions exhibit different biologic properties, which are epigenetically heritable. The existence of prion strains raised the question of how heritable biologic information can be

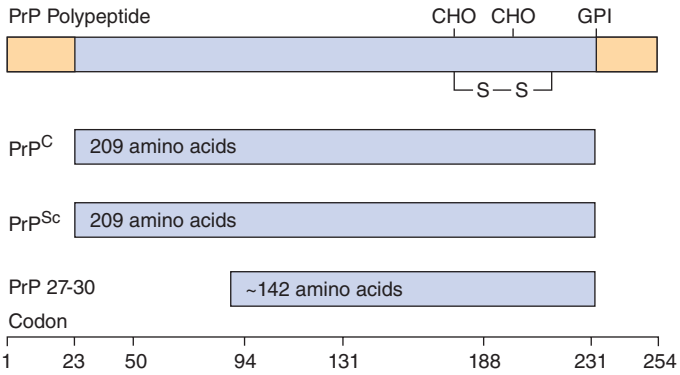


FIGURE 430-2 PrP prion protein isoforms. Bar diagram of Syrian hamster PrP, which consists of 254 amino acids. After processing of the NH_2 and COOH termini, both PrP^C and PrP^{Sc} consist of 209 residues. After limited proteolysis, the NH_2 terminus of PrP^{Sc} is truncated to form PrP 27–30 composed of ~142 amino acids. GPI, glycosylphosphatidylinositol anchor attachment site; S—S, disulfide bond; CHO, N-linked sugars.

enciphered in a molecule other than nucleic acid. Various strains of PrP prions have been defined by incubation times, distribution of neuronal vacuolation, and stabilities of PrP^{Sc} to denaturation. Subsequently, the patterns of PrP^{Sc} deposition were found to correlate with vacuolation profiles, and these patterns were also used to characterize prion strains.

Persuasive evidence that strain-specific information is enciphered in the tertiary structure of PrP^{Sc} comes from transmission of two different inherited human prion diseases to mice expressing a chimeric human-mouse PrP transgene. In FFI, the protease-resistant fragment of PrP^{Sc} after deglycosylation has a molecular mass of 19 kDa, whereas in fCJD and most sporadic prion diseases, it is 21 kDa **(Table 430-3)**. This difference in molecular mass was shown to be due to different sites of proteolytic cleavage at the NH_2 termini of the two human PrP^{Sc} molecules, reflecting different tertiary structures. These distinct conformations were not unexpected because the amino acid sequences of the PrP fragments differ. Extracts from the brains of patients with FFI transmitted disease to the mice expressing the chimeric human-mouse PrP transgene and induced formation of 19-kDa PrP^{Sc}, whereas brain extracts from fCJD and sCJD patients produced 21-kDa PrP^{Sc} in mice expressing the same transgene. On second passage, these differences were maintained, demonstrating that chimeric PrP^{Sc} can exist in two different conformations based on the sizes of the protease-resistant fragments, even though the amino acid sequence of PrP^{Sc} is invariant.

This analysis was extended when patients with sporadic fatal insomnia (sFI) were identified. Although they did not carry a *PRNP* gene mutation, the patients demonstrated a clinical and pathologic phenotype that was indistinguishable from that of patients with FFI. Furthermore, 19-kDa PrP^{Sc} was found in their brains, and on passage of prion disease to mice expressing the chimeric human-mouse PrP transgene, 19-kDa PrP^{Sc} was also found. These findings indicate that the disease phenotype is dictated by the conformation of PrP^{Sc} and not the amino acid sequence. PrP^{Sc} acts as a template for the conversion of PrP^C into nascent PrP^{Sc}. On the passage of prions into mice expressing a chimeric hamster-mouse PrP transgene, a change in the conformation of PrP^{Sc} was accompanied by the emergence of a new strain of prions.

Many new strains of prions were generated using recombinant (rec) PrP produced in bacteria; recPrP was polymerized into amyloid fibrils and inoculated into transgenic mice expressing high levels of wild-type mouse PrP^C. Approximately 500 days later, the mice died of prion disease. The incubation times of the "synthetic prions" in mice were dependent on the conditions used for polymerization of the amyloid fibrils. Highly stable amyloids gave rise to stable prions with long incubation times; low-stability amyloids led to prions with short incubation times. Amyloids of intermediate stability gave rise to prions with intermediate stabilities and intermediate incubation times. Such findings are consistent with earlier studies showing that the incubation times of synthetic and naturally occurring prions are directly proportional to the stability of the prion.

Species Barrier Studies on the role of the primary and tertiary structures of PrP in the transmission of prion disease have provided new insights into the pathogenesis of these maladies. The amino acid sequence of PrP encodes the species of the prion, and the prion derives its PrP^{Sc} sequence from the last mammal in which it was passaged. While the primary structure of PrP is likely to be the most important or even the sole determinant of the tertiary structure of PrP^C, PrP^{Sc} seems to function as a template in determining the tertiary structure of nascent PrP^{Sc} molecules as they are formed from PrP^C. In turn, prion diversity appears to be enciphered in the conformation of PrP^{Sc}, and thus prion strains seem to represent different conformers of PrP^{Sc}.

In general, transmission of PrP prion disease from one species to another is inefficient, in that not all intracerebrally inoculated animals develop disease, and those that fall ill do so only after long incubation times that can approach the natural life span of the animal. This "species barrier" to transmission is correlated with the degree of similarity between the amino acid sequences of PrP^C in the inoculated host and of PrP^{Sc} in the inoculum. The importance of sequence similarity between the host and donor PrP argues that PrP^C directly interacts with PrP^{Sc} in the prion conversion process.

SPORADIC AND INHERITED PrP PRION DISEASES

Several different scenarios might explain the initiation of sporadic prion disease: (1) A somatic mutation may be the cause and thus follow a path similar to that for germline mutations in inherited disease. In this situation, the mutant PrP^{Sc} must be capable of targeting wild-type PrP^C, a process known to be possible for some mutations but less likely for others. (2) The activation energy barrier separating wild-type PrP^C from PrP^{Sc} could be crossed on rare occasions when viewed in the context of a population. Most individuals would be spared, while presentations in the elderly with an incidence of ~1 per million would be seen. (3) PrP^{Sc} may be present at low levels in some normal cells, where it performs an important, but as yet unknown, function. The level of PrP^{Sc} in such cells is hypothesized to be sufficiently low as not to be detected by routine bioassay. In some altered metabolic states, the cellular mechanisms for clearing PrP^{Sc} might become compromised, and the rate of PrP^{Sc} formation would then begin to exceed the capacity of the cell to clear it. The third possible mechanism is attractive because it suggests that PrP^{Sc} is not simply a misfolded protein, as proposed for the first and second mechanisms, but that it is an alternatively folded

TABLE 430-3 Distinct Prion Strains Generated in Humans with Inherited Prion Diseases and Transmitted to Transgenic Mice[a]

INOCULUM	HOST SPECIES	HOST PrP GENOTYPE	INCUBATION TIME [days ± SEM] (n/n_0)	PrP^{Sc} (kDa)
None	Human	FFI(D178N, M129)		19
FFI	Mouse	Tg(MHu2M)	206 ± 7 (7/7)	19
FFI → Tg(MHu2M)	Mouse	Tg(MHu2M)	136 ± 1 (6/6)	19
None	Human	fCJD(E200K)		21
fCJD	Mouse	Tg(MHu2M)	170 ± 2 (10/10)	21
fCJD → Tg(MHu2M)	Mouse	Tg(MHu2M)	167 ± 3 (15/15)	21

[a]Tg(MHu2M) mice express a chimeric mouse-human PrP gene.

Notes: Clinicopathologic phenotype is determined by the conformation of PrP^{Sc} in accord with the results of the transmission of human prions from patients with FFI to transgenic mice.

Abbreviations: fCJD, familial Creutzfeldt-Jakob disease; FFI, fatal familial insomnia; SEM, standard error of the mean.

molecule with a function. Moreover, the multitude of conformational states that PrP^{Sc} can adopt, as described above, raises the possibility that PrP^{Sc} or another prion-like protein might function in a process like short-term memory where information storage occurs in the absence of new protein synthesis.

More than 40 different mutations resulting in nonconservative substitutions in the human *PRNP* gene have been found to segregate with inherited human prion diseases. Missense mutations and expansions in the octapeptide repeat region of the gene are responsible for familial forms of prion disease. Five different mutations of the *PRNP* gene have been linked genetically to heritable prion disease.

Although phenotypes may vary dramatically within families, specific phenotypes tend to be observed with certain mutations. A clinical phenotype indistinguishable from typical sCJD is usually seen with substitutions at codons 180, 183, 200, 208, 210, and 232. Substitutions at codons 102, 105, 117, 198, and 217 are associated with the GSS variant of prion disease with prominent Parkinsonian and cerebellar features. The normal human PrP sequence contains five repeats of an eight-amino-acid sequence. Insertions from two to nine extra octarepeats frequently cause variable phenotypes ranging from a condition indistinguishable from sCJD to a slowly progressive dementing illness of many years in duration to an early-age-of-onset disorder that is similar to AD. A mutation at codon 178 that results in substitution of asparagine for aspartic acid produces FFI if a methionine is encoded at the polymorphic residue 129 on the same allele. Typical CJD is seen if the D178N mutation occurs with a valine encoded at position 129 of the same allele.

HUMAN *PRNP* GENE POLYMORPHISMS

Polymorphisms influence the susceptibility to sporadic, inherited, and infectious forms of PrP prion disease. The methionine/valine polymorphism at position 129 not only modulates the age of onset of some inherited prion diseases but also can determine the clinical phenotype. The finding that homozygosity at codon 129 predisposes an individual to sCJD supports a model of prion production that favors PrP interactions between homologous proteins.

Substitution of the basic residue lysine at position 218 in mouse PrP produced dominant-negative inhibition of prion replication in transgenic mice. This same lysine at position 219 in human PrP has been found in 12% of the Japanese population, a group that appears to be resistant to prion disease. Dominant-negative inhibition of prion replication was also found with substitution of the basic residue arginine at position 171; sheep with arginine were resistant to scrapie prions but were susceptible to BSE prions that were inoculated intracerebrally.

INFECTIOUS PrP PRION DISEASES

IATROGENIC CJD

Accidental transmission of CJD to humans appears to have occurred with corneal transplantation, contaminated electroencephalogram (EEG) electrode implantation, and surgical procedures. Corneas from donors with unsuspected CJD have been transplanted to apparently healthy recipients who developed CJD after variable incubation periods. The same improperly decontaminated EEG electrodes that caused CJD in two young patients with intractable epilepsy caused CJD in a chimpanzee 18 months after their experimental implantation.

Surgical procedures may have resulted in accidental inoculation of patients with prions, presumably because some instrument or apparatus in the operating theater became contaminated when a CJD patient underwent surgery. Although the epidemiology of these studies is highly suggestive, no proof for such episodes exists.

Dura Mater Grafts More than 160 cases of CJD after implantation of dura mater grafts have been recorded. All of the grafts appear to have been acquired from a single manufacturer whose preparative procedures were inadequate to inactivate human prions. One case of CJD occurred after repair of an eardrum perforation with a pericardium graft.

Human Growth Hormone and Pituitary Gonadotropin Therapy The transmission of CJD prions from contaminated human growth hormone (hGH) preparations derived from human pituitaries has been responsible for fatal cerebellar disorders with dementia in >180 patients ranging in age from 10 to 41 years. These patients received injections of hGH every 2–4 days for 4–12 years. If it is thought that these patients developed CJD from injections of prion-contaminated hGH preparations, the possible incubation periods range from 4 to 30 years. Only recombinant hGH is now used therapeutically so that possible contamination with prions is no longer an issue. Four cases of CJD have occurred in women receiving human pituitary gonadotropin.

VARIANT CJD

The restricted geographic occurrence and chronology of vCJD raised the possibility that BSE prions had been transmitted to humans through the consumption of tainted beef. More than 190 cases of vCJD have occurred, with >90% of these in Britain. Variant CJD has also been reported in people either living in or originating from France, Ireland, Italy, the Netherlands, Portugal, Spain, Saudi Arabia, the United States, Canada, and Japan.

The steady decline in the number of vCJD cases over the past decade argues that there will not be a prion disease epidemic in Europe, similar to those seen for BSE and kuru. What is certain is that prion-tainted meat should be prevented from entering the human food supply.

The most compelling evidence that vCJD is caused by BSE prions was obtained from experiments in mice expressing the bovine PrP transgene. Both BSE and vCJD prions were efficiently transmitted to these transgenic mice and with similar incubation periods. In contrast to sCJD prions, vCJD prions did not transmit disease efficiently to mice expressing a chimeric human-mouse PrP transgene. Earlier studies with nontransgenic mice suggested that vCJD and BSE might be derived from the same source because both inocula transmitted disease with similar but very long incubation periods.

Attempts to determine the origin of BSE and vCJD prions have relied on passaging studies in mice, some of which are described above, as well as studies of the conformation and glycosylation of PrP^{Sc}. One scenario suggests that a particular conformation of bovine PrP^{Sc} was selected for heat resistance during the rendering process and was then reselected multiple times as cattle infected by ingesting prion-contaminated meat and bone meal (MBM) were slaughtered and their offal rendered into more MBM. Variant CJD cases have virtually disappeared with protection of the beef supply in Europe.

NEUROPATHOLOGY

Frequently, the brains of patients with CJD have no recognizable abnormalities on gross examination. Patients who survive for several years have variable degrees of cerebral atrophy.

On light microscopy, the pathologic hallmarks of CJD are spongiform degeneration and astrocytic gliosis. The lack of an inflammatory response in CJD and other prion diseases is an important pathologic feature of these degenerative disorders. Spongiform degeneration is characterized by many 1- to 5-μm vacuoles in the neuropil between nerve cell bodies. Generally, the spongiform changes occur in the cerebral cortex, putamen, caudate nucleus, thalamus, and molecular layer of the cerebellum. Astrocytic gliosis is a constant but nonspecific feature of PrP prion diseases. Widespread proliferation of fibrous astrocytes is found throughout the gray matter of brains infected with CJD prions. Astrocytic processes filled with glial filaments form extensive networks.

Amyloid plaques have been found in ~10% of CJD cases. Purified CJD prions from humans and animals exhibit the ultrastructural and histochemical characteristics of amyloid when treated with detergents during limited proteolysis. On first passage of samples from some human Japanese CJD cases into mice, amyloid plaques were found. These plaques stain with antibodies raised against PrP.

The amyloid plaques of GSS disease are morphologically distinct from those seen in kuru or scrapie. GSS plaques consist of a central dense core of amyloid surrounded by smaller globules of amyloid. Ultrastructurally, they consist of a radiating fibrillar network of amyloid fibrils, with scant or no neuritic degeneration. The plaques can be distributed throughout the brain but are most frequently found in the cerebellum. They are often located adjacent to blood vessels. Congophilic angiopathy has been noted in some cases of GSS disease.

In vCJD, a characteristic feature is the presence of "florid plaques." These are composed of a central core of PrP amyloid, surrounded by vacuoles in a pattern suggesting petals on a flower.

CLINICAL FEATURES

Nonspecific prodromal symptoms occur in approximately a third of patients with CJD and may include fatigue, sleep disturbance, weight loss, headache, anxiety, vertigo, malaise, and ill-defined pain. Most patients with CJD present with deficits in higher cortical function. Similarly, psychiatric symptoms, such as depression, psychosis, and visual hallucinations, are often the defining features of the illness. These deficits almost always progress over weeks or months to a state of profound dementia characterized by memory loss, impaired judgment, and a decline in virtually all aspects of intellectual function. A few patients present with either visual impairment or cerebellar gait and coordination deficits. Frequently, the cerebellar deficits are rapidly followed by progressive dementia. Visual problems often begin with blurred vision and diminished acuity, rapidly followed by dementia.

Other symptoms and signs include extrapyramidal dysfunction manifested as rigidity, masklike facies, or (less commonly) choreoathetoid movements; pyramidal signs (usually mild); seizures (usually major motor) and, less commonly, hypoesthesia; supranuclear gaze palsy; optic atrophy; and vegetative signs such as changes in weight, temperature, sweating, or menstruation.

Myoclonus Most patients (~90%) with CJD exhibit myoclonus that appears at various times throughout the illness. Unlike other involuntary movements, myoclonus persists during sleep. Startle myoclonus elicited by loud sounds or bright lights is frequent. It is important to stress that myoclonus is neither specific nor confined to CJD and tends to occur later in the course of CJD. Dementia with myoclonus can also be due to AD **(Chap. 423)**, dementia with Lewy bodies **(Chap. 426)**, corticobasal degeneration **(Chap. 424)**, cryptococcal encephalitis **(Chap. 210)**, or the myoclonic epilepsy disorder Unverricht-Lundborg disease **(Chap. 418)**.

Clinical Course In documented cases of accidental transmission of CJD to humans, an incubation period of 1.5–2 years preceded the development of clinical disease. In other cases, incubation periods of up to 40 years have been suggested. Most patients with CJD live 6–12 months after the onset of clinical signs and symptoms, whereas some live for up to 5 years.

DIAGNOSIS

The constellation of dementia, myoclonus, and periodic electrical bursts in an afebrile 60-year-old patient generally indicates CJD. Clinical abnormalities in CJD are confined to the CNS. Fever, elevated sedimentation rate, leukocytosis in blood, or a pleocytosis in cerebrospinal fluid (CSF) should alert the physician to another etiology to explain the patient's CNS dysfunction, although there are rare cases of CJD in which mild CSF pleocytosis is observed.

Variations in the typical course appear in inherited and transmitted forms of the disease. Familial CJD has an earlier mean age of onset than sCJD. In GSS disease, ataxia is usually a prominent and presenting feature, with dementia occurring late in the disease course. GSS disease presents earlier than CJD (mean age 43 years) and is typically more slowly progressive than CJD; death usually occurs within 5 years of onset. FFI is characterized by insomnia and dysautonomia; dementia occurs only in the terminal phase of the illness. Rare sporadic cases have been identified. Variant CJD has an unusual clinical course, with a prominent psychiatric prodrome that may include visual hallucinations and early ataxia, whereas frank dementia is usually a late sign of vCJD.

DIFFERENTIAL DIAGNOSIS

Many conditions mimic CJD. Dementia with Lewy bodies **(Chap. 426)** is the most common disorder to be mistaken for CJD. It can present in a subacute fashion with delirium, myoclonus, and extrapyramidal features. Other neurodegenerative disorders to consider include AD, FTD, corticobasal degeneration, progressive supranuclear palsy, ceroid lipofuscinosis, and myoclonic epilepsy with Lafora bodies. The absence of abnormalities on diffusion-weighted and fluid-attenuated inversion recovery (FLAIR) magnetic resonance imaging (MRI) will almost always distinguish these dementing conditions from CJD.

Hashimoto's encephalopathy, which presents as a subacute progressive encephalopathy with myoclonus and periodic triphasic complexes on the EEG, should be excluded in every case of suspected CJD. It is diagnosed by the finding of high titers of antithyroglobulin or antithyroid peroxidase (antimicrosomal) antibodies in the blood and improves with glucocorticoid therapy. Unlike CJD, fluctuations in severity typically occur in Hashimoto's encephalopathy.

Intracranial vasculitides **(Chap. 356)** may produce nearly all of the symptoms and signs associated with CJD, sometimes without systemic abnormalities. Myoclonus is exceptional with cerebral vasculitis, but focal seizures may confuse the picture. Prominent headache, absence of myoclonus, stepwise change in deficits, abnormal CSF, and focal white matter change on MRI or angiographic abnormalities all favor vasculitis.

Paraneoplastic conditions **(Chap. 90)**, particularly limbic encephalitis and cortical encephalitis, can also mimic CJD. In many of these patients, dementia appears prior to the diagnosis of a tumor, and in some, no tumor is ever found. Detection of the paraneoplastic antibodies is often the only way to distinguish these cases from CJD.

Other diseases that can simulate CJD include neurosyphilis **(Chap. 177)**, AIDS dementia complex **(Chap. 197)**, progressive multifocal leukoencephalopathy **(Chap. 132)**, subacute sclerosing panencephalitis, progressive rubella panencephalitis, herpes simplex encephalitis **(Chap. 132)**, diffuse intracranial tumor (gliomatosis cerebri; **Chap. 86**), anoxic encephalopathy, dialysis dementia, uremia, hepatic encephalopathy, voltage-gated potassium channel (VGkC) autoimmune encephalopathy, and lithium or bismuth intoxication.

LABORATORY TESTS

The only specific diagnostic tests for CJD and other human PrP prion diseases measure PrP^{Sc}. The most widely used method involves limited proteolysis that generates PrP 27-30, which is detected by immunoassay after denaturation. The conformation-dependent immunoassay (CDI) is based on immunoreactive epitopes that are exposed in PrP^{C} but buried in PrP^{Sc}. In humans, the diagnosis of CJD can be established by brain biopsy if PrP^{Sc} is detected although biopsy is rarely indicated. If no attempt is made to measure PrP^{Sc}, but the constellation of pathologic changes frequently found in CJD is seen in a brain biopsy, then the diagnosis is reasonably secure (see "Neuropathology," above). The high sensitivity and specificity of cortical ribboning and basal ganglia hyperintensity on FLAIR and diffusion-weighted MRI for the diagnosis of CJD have greatly diminished the need for brain biopsy in patients with suspected CJD. Because PrP^{Sc} is not uniformly distributed throughout the CNS, the apparent absence of PrP^{Sc} in a limited sample such as a biopsy does not rule out prion disease. At autopsy, sufficient brain samples should be taken for both PrP^{Sc} immunoassay, preferably by CDI, and immunohistochemistry of tissue sections.

To establish the diagnosis of either sCJD or familial prion disease, sequencing the *PRNP* gene must be performed. Finding the wild-type *PRNP* gene sequence permits the diagnosis of sCJD if there is no history to suggest infection from an exogenous source of prions. The identification of a mutation in the *PRNP* gene sequence that encodes a nonconservative amino acid substitution argues for familial prion disease.

CT may be normal or show cortical atrophy. MRI is valuable for distinguishing sCJD from most other conditions. On FLAIR sequences and diffusion-weighted imaging, ~90% of patients show increased intensity in the basal ganglia and cortical ribboning **(Fig. 430-3)**. This pattern is not seen with other neurodegenerative disorders but has been seen infrequently with viral encephalitis, paraneoplastic syndromes, or seizures. When the typical MRI pattern is present, in the proper clinical setting, diagnosis is facilitated. However, some cases of sCJD do not show this typical pattern, and other early diagnostic approaches are still needed.

CSF is nearly always normal but may show protein elevation and, rarely, mild pleocytosis. Although the stress protein 14-3-3 is elevated in the CSF of some patients with CJD, similar elevations of 14-3-3 are found in patients with other disorders; thus this elevation is not

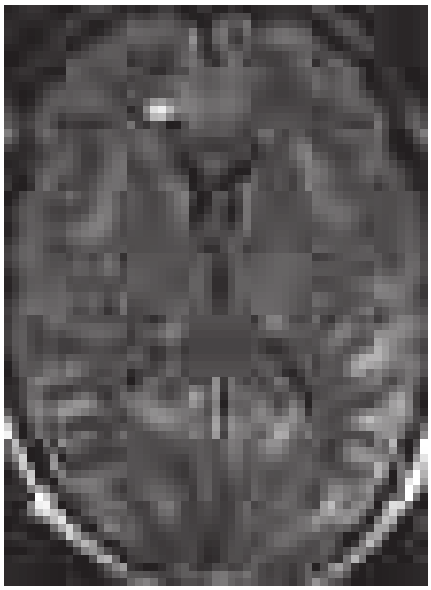


FIGURE 430-3 T2-weighted FLAIR MRI showing hyperintensity in the cortex in a patient with sCJD. This so-called "cortical ribboning" along with increased intensity in the basal ganglia on T2- or diffusion-weighted imaging can aid in the diagnosis of CJD.

specific. Similarly, elevations of CSF neuron-specific enolase and tau occur in CJD but lack specificity for diagnosis.

The EEG is often useful in the diagnosis of CJD, although only ~60% of individuals show the typical pattern, which appears quite late in the clinical course. During the early phase of CJD, the EEG is usually normal or shows only scattered theta activity. In most advanced cases, repetitive, high-voltage, triphasic, and polyphasic sharp discharges are seen, but in many cases their presence is transient. The presence of these stereotyped periodic bursts of <200 ms in duration, occurring every 1–2 s, makes the diagnosis of CJD very likely. These discharges are frequently but not always symmetric; there may be a one-sided predominance in amplitude. As CJD progresses, normal background rhythms become fragmentary and slower.

CARE OF CJD PATIENTS

Although CJD should not be considered either contagious or communicable, it is transmissible. The risk of accidental inoculation by aerosols is very small; nonetheless, procedures producing aerosols should be performed in certified biosafety cabinets. Biosafety level 2 practices, containment equipment, and facilities are recommended by the Centers for Disease Control and Prevention and the National Institutes of Health. The primary worry in caring for patients with CJD is the inadvertent infection of health care workers by needle and stab wounds, although with the possible exception of vCJD even blood transfusions appear to carry little risk for transmission. Electroencephalographic and electromyographic needles should not be reused after studies on patients with CJD have been performed.

There is no reason for pathologists or other morgue employees to resist performing autopsies on patients whose clinical diagnosis was CJD. Standard microbiologic practices outlined here, along with specific recommendations for decontamination, seem to be adequate precautions for the care of patients with CJD and the handling of infected specimens.

DECONTAMINATION OF CJD PRIONS

Prions are extremely resistant to common inactivation procedures, and there is some disagreement about the optimal conditions for sterilization. Some investigators recommend treating CJD-contaminated materials once with 1 *N* NaOH at room temperature, but we believe this procedure may be inadequate for sterilization. Autoclaving at 134°C for 5 h or treatment with 2 *N* NaOH for several hours is recommended for sterilization of prions. The term *sterilization* implies complete destruction of prions; any residual infectivity can be hazardous. Transgenic mouse studies show that sCJD prions bound to stainless steel surfaces are resistant to inactivation by autoclaving at 134°C for 2 h; exposure of bound prions to an acidic detergent solution prior to autoclaving rendered prions susceptible to inactivation. Recent studies show that α-synuclein prions in brain homogenates prepared from MSA patients bind to stainless steel wires and that the bound prions can be transmitted to transgenic mice expressing mutant human α-synuclein.

PREVENTION AND THERAPEUTICS

There is no known effective therapy for preventing or treating CJD. The finding that phenothiazines and acridines inhibit PrP^{Sc} formation in cultured cells led to clinical studies of quinacrine in CJD patients. Unfortunately, quinacrine failed to slow the rate of cognitive decline in CJD, possibly because therapeutic concentrations in the brain were not achieved. Although inhibition of the P-glycoprotein (Pgp) transport system resulted in substantially increased quinacrine levels in the brains of mice, the prion incubation times were not extended by treatment with the drug. Whether such an approach can be used to treat CJD remains to be established.

Like the acridines, anti-PrP antibodies have been shown to eliminate PrP^{Sc} from cultured cells. Additionally, such antibodies in mice, either administered by injection or produced from a transgene, have been shown to prevent prion disease when prions are introduced by a peripheral route, such as intraperitoneal inoculation. Unfortunately, the antibodies were ineffective in mice inoculated intracerebrally with prions. Several drugs, including pentosan polysulfate as well as porphyrin and phenylhydrazine derivatives, delay the onset of disease in animals inoculated intracerebrally with prions if the drugs are given intracerebrally beginning soon after inoculation.

DIFFERENT PRIONS CAUSING OTHER NEURODEGENERATIVE DISEASES

There is a rapidly expanding body of literature demonstrating that besides PrP, other proteins including amyloid beta (Aβ), tau, α-synuclein, and huntingtin can all become prions **(Chap. 417)**. Experimental studies have shown that transgenic mice expressing mutant amyloid precursor protein (APP) develop amyloid plaques containing fibrils composed of the Aβ peptide ~6 months after inoculation with synthetic Aβ peptides polymerized into amyloid fibrils or extracts prepared from the brains of patients with AD. Mutant tau aggregates in transgenic mice and cultured cells can trigger the aggregation of tau into fibrils that resemble those found in neurofibrillary tangles and Pick bodies. Such tangles have been found in AD, FTDs, Pick's disease, and some cases of posttraumatic brain injury called chronic traumatic encephalopathy, all of which are thought to be caused by the prion isoforms of Aβ and/or tau.

In patients with advanced PD who received grafts of fetal substantia nigral neurons, Lewy bodies containing β-sheet-rich α-synuclein were identified in grafted cells ~10 years after transplantation, arguing for the axonal transport of misfolded α-synuclein crossing into grafted neurons, where it initiated aggregation of nascent α-synuclein into fibrils that coalesced into Lewy bodies. These findings combined with MSA studies argue that the synucleinopathies are caused by prions. Brain homogenates from MSA patients injected into transgenic mice transmitted lethal neurodegeneration in ~3 months; moreover, recombinant synuclein injected into wild-type mice initiated the deposition of synuclein fibrils.

In summary, a wealth of evidence continues to accumulate arguing that proteins causing AD, PD, FTDs, amyotrophic lateral sclerosis (ALS), and even Huntington's disease (HD) acquire alternative conformations that become self-propagating. Each of these neurodegenerative diseases is thought to be caused by the abnormal aggregation of a different protein that undergoes a self-replicating conformational change to become a prion. Prions explain many of the features that the neurodegenerative diseases have in common: (1) incidence increases with age, (2) steady progression over years, (3) spread from one region of the CNS to another, (4) protein deposits often but not always consisting of amyloid fibrils, and (5) late onset of inherited forms. Notably, amyloid plaques containing PrP^{Sc} are a nonobligatory feature of PrP prion disease in humans and animals. Furthermore, amyloid plaques

 in AD do not correlate with the level of dementia; however, the level of soluble (oligomeric) Aβ peptide does correlate with memory loss and other intellectual deficits.

FURTHER READING

COLLINGE J: Mammalian prions and their wider relevance in neurodegenerative diseases. Nature 539:217, 2016.

PRUSINER SB et al: Evidence for α-synuclein prions causing multiple system atrophy in humans with parkinsonism. Proc Natl Acad Sci U S A 112:E5308, 2015.

PRUSINER SB (ed): *Prion Biology*. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press, 2017.

PRUSINER SB (ed): *Prion Diseases*. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press, 2017.

ZWECKSTETTER M, REQUENA JR, WILLE H: Elucidating the structure of an infectious protein. PLoS Pathog 13:e1006229, 2017.

431 Ataxic Disorders

Roger N. Rosenberg

APPROACH TO THE PATIENT

Ataxic Disorders

Symptoms and signs of ataxia consist of gait impairment, unclear ("scanning") speech, visual blurring due to nystagmus, hand incoordination, and tremor with movement. These result from the involvement of the cerebellum and its afferent and efferent pathways, including the spinocerebellar pathways, and the frontopontocerebellar pathway originating in the rostral frontal lobe. True cerebellar ataxia must be distinguished from ataxia associated with vestibular nerve or labyrinthine disease, as the latter results in a disorder of gait associated with a significant degree of dizziness, light-headedness, or the perception of movement **(Chap. 19)**. True cerebellar ataxia is devoid of these vertiginous complaints and is clearly an unsteady gait due to imbalance. Sensory disturbances can also on occasion simulate the imbalance of cerebellar disease; with sensory ataxia, imbalance dramatically worsens when visual input is removed (Romberg sign). Rarely, weakness of proximal leg muscles mimics cerebellar disease. In the patient who presents with ataxia, the rate and pattern of the development of cerebellar symptoms help to narrow the diagnostic possibilities **(Table 431-1)**. A gradual and progressive increase in symptoms with bilateral and symmetric involvement suggests a genetic, metabolic, immune, or toxic etiology. Conversely, focal, unilateral symptoms with headache and impaired level of consciousness accompanied by ipsilateral cranial nerve palsies and contralateral weakness imply a space-occupying cerebellar lesion.

SYMMETRIC ATAXIA

Progressive and symmetric ataxia can be classified with respect to onset as acute (over hours or days), subacute (weeks or months), or chronic (months to years). Acute and reversible ataxias include those caused by intoxication with alcohol, phenytoin, lithium, barbiturates, and other drugs. Intoxication caused by toluene exposure, gasoline sniffing, glue sniffing, spray painting, or exposure to methyl mercury or bismuth are additional causes of acute or subacute ataxia, as is treatment with cytotoxic chemotherapeutic drugs such as fluorouracil and paclitaxel. Patients with a postinfectious syndrome (especially after varicella) may develop gait ataxia and mild dysarthria, both of which are reversible **(Chap. 436)**. Rare infectious causes of acquired ataxia include poliovirus, coxsackievirus, echovirus, Epstein-Barr virus, toxoplasmosis, *Legionella*, and Lyme disease.

The subacute development of ataxia of gait over weeks to months (degeneration of the cerebellar vermis) may be due to the combined effects of alcoholism and malnutrition, particularly with deficiencies of vitamins B_1 and B_{12}. Hyponatremia has also been associated with ataxia. Paraneoplastic cerebellar ataxia is associated with a number of different tumors (and autoantibodies) such as breast and ovarian cancers (anti-Yo), small-cell lung cancer (anti-PQ-type voltage-gated calcium channel), and Hodgkin's disease (anti-Tr) **(Chap. 90)**. Another paraneoplastic syndrome associated with myoclonus and opsoclonus occurs with breast (anti-Ri) and lung cancers and neuroblastoma. Elevated serum anti-glutamic acid decarboxylase (GAD) antibodies have been associated with a progressive ataxic syndrome affecting speech and gait. For all of these paraneoplastic ataxias, the neurologic syndrome may be the presenting symptom of the cancer. Another immune-mediated progressive ataxia is associated with antigliadin (and antiendomysium) antibodies and the human leukocyte antigen (HLA) DQB1*0201 haplotype; in some affected patients, biopsy of the small intestine reveals villus atrophy consistent with gluten-sensitive enteropathy **(Chap. 318)**. Finally, subacute progressive ataxia may be caused by a prion disorder, especially when an infectious etiology, such as transmission from contaminated human growth hormone, is responsible **(Chap. 430)**.

Chronic symmetric gait ataxia suggests an inherited ataxia (discussed below), a metabolic disorder, or a chronic infection. Hypothyroidism must always be considered as a readily treatable and reversible form of gait ataxia. Infectious diseases that can present with ataxia are meningovascular syphilis and tabes dorsalis due to degeneration of the posterior columns and spinocerebellar pathways in the spinal cord.

FOCAL ATAXIA

Acute focal ataxia commonly results from cerebrovascular disease, usually ischemic infarction or cerebellar hemorrhage. These lesions typically produce cerebellar symptoms ipsilateral to the injured

TABLE 431-1 Etiology of Cerebellar Ataxia

SYMMETRIC AND PROGRESSIVE SIGNS			FOCAL AND IPSILATERAL CEREBELLAR SIGNS		
ACUTE (HOURS TO DAYS)	**SUBACUTE (DAYS TO WEEKS)**	**CHRONIC (MONTHS TO YEARS)**	**ACUTE (HOURS TO DAYS)**	**SUBACUTE (DAYS TO WEEKS)**	**CHRONIC (MONTHS TO YEARS)**
Intoxication: alcohol, lithium, phenytoin, barbiturates (positive history and toxicology screen)	Intoxication: mercury, solvents, gasoline, glue; cytotoxic chemotherapeutic, hemotherapeutic drugs	Paraneoplastic syndrome	Vascular: cerebellar infarction, hemorrhage, or subdural hematoma	Neoplastic: cerebellar glioma or metastatic tumor (positive for neoplasm on MRI/CT)	Stable gliosis secondary to vascular lesion or demyelinating plaque (stable lesion on MRI/CT older than several months)
Acute viral cerebellitis (CSF supportive of acute viral infection)	Alcoholic-nutritional (vitamin B_1 and B_{12} deficiency)	Antigliadin antibody syndrome	Infectious: cerebellar abscess (mass lesion on MRI/CT, history in support of lesion)	Demyelinating: multiple sclerosis (history, CSF, and MRI are consistent)	Congenital lesion: Chiari or Dandy-Walker malformations (malformation noted on MRI/CT)
Postinfection syndrome	Lyme disease	Hypothyroidism		AIDS-related multifocal leukoencephalopathy (positive HIV test and CD4+ cell count for AIDS)	
		Inherited diseases			
		Tabes dorsalis (tertiary syphilis)			
		Phenytoin toxicity			
		Amiodarone			

Abbreviations: CSF, cerebrospinal fluid; CT, computed tomography; MRI, magnetic resonance imaging.

cerebellum and may be associated with an impaired level of consciousness due to brainstem compression and increased intracranial pressure; ipsilateral pontine signs, including sixth and seventh nerve palsies, may be present. Focal and worsening signs of acute ataxia should also prompt consideration of a posterior fossa subdural hematoma, bacterial abscess, or primary or metastatic cerebellar tumor. Computed tomography (CT) or magnetic resonance imaging (MRI) studies will reveal clinically significant processes of this type. Many of these lesions represent true neurologic emergencies, as sudden herniation, either rostrally through the tentorium or caudal herniation of cerebellar tonsils through the foramen magnum, can occur and is usually devastating. Acute surgical decompression may be required **(Chap. 301)**. Lymphoma or progressive multifocal leukoencephalopathy (PML) in a patient with AIDS may present with an acute or subacute focal cerebellar syndrome. Chronic etiologies of progressive ataxia include multiple sclerosis **(Chap. 436)** and congenital lesions such as a Chiari malformation **(Chap. 434)** or a congenital cyst of the posterior fossa (Dandy-Walker syndrome).

THE INHERITED ATAXIAS

These may show autosomal dominant, autosomal recessive, or maternal (mitochondrial) modes of inheritance. A genomic classification **(Chap. S10)** has now largely superseded previous ones based on clinical expression alone.

Although the clinical manifestations and neuropathologic findings of cerebellar disease dominate the clinical picture, there may also be characteristic changes in the basal ganglia, brainstem, spinal cord, optic nerves, retina, and peripheral nerves. In large families with dominantly inherited ataxias, many gradations are observed from purely cerebellar manifestations to mixed cerebellar and brainstem disorders, cerebellar and basal ganglia syndromes, and spinal cord or peripheral nerve disease. Rarely, dementia is present as well. The clinical picture may be homogeneous within a family with dominantly inherited ataxia, but sometimes most affected family members show one characteristic syndrome, while one or several members have an entirely different phenotype.

AUTOSOMAL DOMINANT ATAXIAS

The autosomal spinocerebellar ataxias (SCAs) include SCA types 1 through 40, dentatorubropallidoluysian atrophy (DRPLA), and episodic ataxia (EA) types 1 to 7 **(Chap. S10)**. SCA1, SCA2, SCA3 (Machado-Joseph disease [MJD]), SCA6, SCA7, and SCA17 are caused by CAG triplet repeat expansions in different genes. SCA8 is due to an untranslated CTG repeat expansion, SCA12 is linked to an untranslated CAG repeat, and SCA10 is caused by an untranslated pentanucleotide repeat. The clinical phenotypes of these SCAs overlap. The genotype has become the gold standard for diagnosis and classification. CAG encodes glutamine, and these expanded CAG triplet repeat expansions result in expanded polyglutamine proteins, termed *ataxins*, that produce a toxic gain of function with autosomal dominant inheritance. Although the phenotype is variable for any given disease gene, a pattern of neuronal loss with gliosis is produced that is relatively unique for each ataxia. Immunohistochemical and biochemical studies have shown cytoplasmic (SCA2), neuronal (SCA1, MJD, SCA7), and nucleolar (SCA7) accumulation of the specific mutant polyglutamine-containing ataxin proteins. Expanded polyglutamine ataxins with more than ~40 glutamines are potentially toxic to neurons for a variety of reasons including: high levels of gene expression for the mutant polyglutamine ataxin in affected neurons; conformational change of the aggregated protein to a β-pleated structure; abnormal transport of the ataxin into the nucleus (SCA1, MJD, SCA7); binding to other polyglutamine proteins, including the TATA-binding transcription protein and the CREB-binding protein, impairing their functions; altering the efficiency of the ubiquitin-proteasome system of protein turnover; and inducing neuronal apoptosis. An earlier age of onset (anticipation) and more aggressive disease in subsequent generations are due to further expansion of the CAG triplet repeat and increased polyglutamine number in the mutant ataxin. The most common disorders are discussed below.

SCA1

SCA1 was previously referred to as *olivopontocerebellar atrophy*, but genomic data have shown that that entity represents several different genotypes with overlapping clinical features.

Symptoms and Signs SCA1 is characterized by the development in early or middle adult life of progressive cerebellar ataxia of the trunk and limbs, impairment of equilibrium and gait, slowness of voluntary movements, scanning speech, nystagmoid eye movements, and oscillatory tremor of the head and trunk. Dysarthria, dysphagia, and oculomotor and facial palsies may also occur. Extrapyramidal symptoms include rigidity, an immobile face, and parkinsonian tremor. The reflexes are usually normal, but knee and ankle jerks may be lost, and extensor plantar responses may occur. Dementia may be noted but is usually mild. Impairment of sphincter function is common, with urinary and sometimes fecal incontinence. Cerebellar and brainstem atrophy are evident on MRI **(Fig. 431-1)**.

Marked shrinkage of the ventral half of the pons, disappearance of the olivary eminence on the ventral surface of the medulla, and atrophy of the cerebellum are evident on gross postmortem inspection of the brain. Variable loss of Purkinje cells, reduced numbers of cells in the molecular and granular layer, demyelination of the middle cerebellar peduncle and the cerebellar hemispheres, and severe loss of cells in the pontine nuclei and olives are found on histologic examination. Degenerative changes in the striatum, especially the putamen, and loss of the pigmented cells of the substantia nigra may be found in cases with extrapyramidal features. More widespread degeneration in the central nervous system (CNS), including involvement of the posterior columns and the spinocerebellar fibers, is often present.

GENETIC CONSIDERATIONS

SCA1 encodes a gene product, called *ataxin-1*, which is a novel protein of unknown function. The mutant allele has 40 CAG repeats located within the coding region, whereas alleles from unaffected individuals have ≤36 repeats. A few patients with 38–40 CAG repeats have been described. There is a direct correlation between a larger number of repeats and a younger age of onset for SCA1. Juvenile patients have higher numbers of repeats, and anticipation is present in subsequent generations. Transgenic mice carrying SCA1 developed ataxia and Purkinje cell pathology. Leucine-rich acidic nuclear protein localization, but not aggregation, of ataxin-1 appears to be required for cell death initiated by the mutant protein.

SCA2

Symptoms and Signs Another clinical phenotype, SCA2, has been described in patients from Cuba and India. Cuban patients

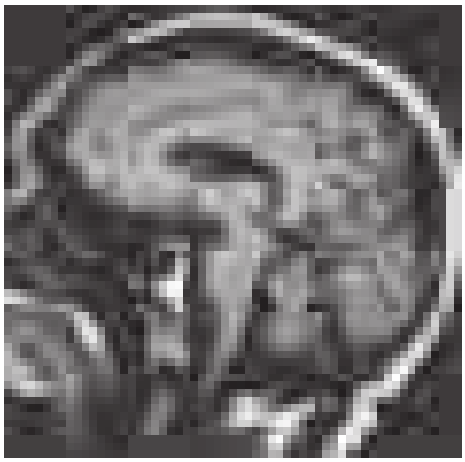


FIGURE 431-1 Sagittal magnetic resonance imaging (MRI) of the brain of a 60-year-old man with gait ataxia and dysarthria due to spinocerebellar ataxia type 1 (SCA1), illustrating cerebellar atrophy (*arrows*). *(Reproduced with permission from RN Rosenberg, P Khemani, in RN Rosenberg, JM Pascual [eds]: Rosenberg's Molecular and Genetic Basis of Neurological and Psychiatric Disease, 5th ed. London, Elsevier, 2015.)*

probably are descendants of a common ancestor, and the population may be the largest homogeneous group of patients with ataxia yet described. The age of onset ranges from 2 to 65 years, and there is considerable clinical variability within families. Although neuropathologic and clinical findings are compatible with a diagnosis of SCA1, including slow saccadic eye movements, ataxia, dysarthria, parkinsonian rigidity, optic disc pallor, mild spasticity, and retinal degeneration, SCA2 is a unique form of cerebellar degenerative disease.

GENETIC CONSIDERATIONS

The gene in SCA2 families also contains CAG repeat expansions coding for a polyglutamine-containing protein, ataxin-2. Normal alleles contain 15–32 repeats; mutant alleles have 35–77 repeats.

MACHADO-JOSEPH DISEASE/SCA3

MJD was first described among the Portuguese and their descendants in New England and California. Subsequently, MJD has been found in families from Portugal, Australia, Brazil, Canada, China, England, France, India, Israel, Italy, Japan, Spain, Taiwan, and the United States. In most populations, it is the most common autosomal dominant ataxia.

Symptoms and Signs MJD has been classified into three clinical types. In type I MJD (amyotrophic lateral sclerosis-parkinsonism-dystonia type), neurologic deficits appear in the first two decades and involve weakness and spasticity of extremities, especially the legs, often with dystonia of the face, neck, trunk, and extremities. Patellar and ankle clonus are common, as are extensor plantar responses. The gait is slow and stiff, with a slightly broadened base and lurching from side to side; this gait results from spasticity, not true ataxia. There is no truncal titubation. Pharyngeal weakness and spasticity cause difficulty with speech and swallowing. Of note is the prominence of horizontal and vertical nystagmus, loss of fast saccadic eye movements, hypermetric and hypometric saccades, and impairment of upward vertical gaze. Facial fasciculations, facial myokymia, lingual fasciculations without atrophy, ophthalmoparesis, and ocular prominence are common early manifestations.

In type II MJD (ataxic type), true cerebellar deficits of dysarthria and gait and extremity ataxia begin in the second to fourth decades along with corticospinal and extrapyramidal deficits of spasticity, rigidity, and dystonia. Type II is the most common form of MJD. Ophthalmoparesis, upward vertical gaze deficits, and facial and lingual fasciculations are also present. Type II MJD can be distinguished from the clinically similar disorders SCA1 and SCA2.

Type III MJD (ataxic-amyotrophic type) presents in the fifth to the seventh decades with a pancerebellar disorder that includes dysarthria and gait and extremity ataxia. Distal sensory loss involving pain, touch, vibration, and position senses and distal atrophy are prominent, indicating the presence of peripheral neuropathy. The deep tendon reflexes are depressed to absent, and there are no corticospinal or extrapyramidal findings.

The mean age of onset of symptoms in MJD is 25 years. Neurologic deficits invariably progress and lead to death from debilitation within 15 years of onset, especially in patients with types I and II disease. Usually, patients retain full intellectual function.

The major pathologic findings are variable loss of neurons and glial replacement in the corpus striatum and severe loss of neurons in the pars compacta of the substantia nigra. A moderate loss of neurons occurs in the dentate nucleus of the cerebellum and in the red nucleus. Purkinje cell loss and granule cell loss occur in the cerebellar cortex. Cell loss also occurs in the dentate nucleus and in the cranial nerve motor nuclei. Sparing of the inferior olives distinguishes MJD from other dominantly inherited ataxias.

GENETIC CONSIDERATIONS

The gene for MJD maps to 14q24.3-q32. Unstable CAG repeat expansions are present in the MJD gene coding for a polyglutamine-containing protein named ataxin-3, or MJD-ataxin. An earlier age of onset is associated with longer repeats. Alleles from normal individuals have between 12 and 37 CAG repeats, whereas MJD alleles have 60–84 CAG repeats. Polyglutamine-containing aggregates of ataxin-3 (MJD-ataxin) have been described in neuronal nuclei undergoing degeneration. MJD-ataxin codes for a ubiquitin protease, which is inactive due to expanded polyglutamines. Proteosome function is impaired, resulting in altered clearance of proteins and cerebellar neuronal loss.

SCA6

Genomic screening for CAG repeats in other families with autosomal dominant ataxia and vibratory and proprioceptive sensory loss have yielded another locus. Of interest is that different mutations in the same gene for the α_{1A} voltage-dependent calcium channel subunit (CACNLIA4; also referred to as the *CACNA1A* gene) at 19p13 result in different clinical disorders. CAG repeat expansions (21–27 in patients; 4–16 triplets in normal individuals) result in late-onset progressive ataxia with cerebellar degeneration. Missense mutations in this gene result in familial hemiplegic migraine. Nonsense mutations resulting in termination of protein synthesis of the gene product yield hereditary paroxysmal cerebellar ataxia or EA. Some patients with familial hemiplegic migraine develop progressive ataxia and also have cerebellar atrophy.

SCA7

This disorder is distinguished from all other SCAs by the presence of retinal pigmentary degeneration. The visual abnormalities first appear as blue-yellow color blindness and proceed to frank visual loss with macular degeneration. In almost all other respects, SCA7 resembles several other SCAs in which ataxia is accompanied by various noncerebellar findings, including ophthalmoparesis and extensor plantar responses. The genetic defect is an expanded CAG repeat in the SCA7 gene at 3p14-p21.1. The expanded repeat size in SCA7 is highly variable. Consistent with this, the severity of clinical findings varies from essentially asymptomatic to mild late-onset symptoms to severe, aggressive disease in childhood with rapid progression. Marked anticipation has been recorded, especially with paternal transmission. The disease protein, ataxin-7, forms aggregates in nuclei of affected neurons, as has also been described for SCA1 and SCA3/MJD. Ataxin 7 is a subunit of GCN5, a histone acetyltransferase-containing complex.

SCA8

This form of ataxia is caused by a CTG repeat expansion in an untranslated region of a gene on chromosome 13q21. There is marked maternal bias in transmission, perhaps reflecting contractions of the repeat during spermatogenesis. The mutation is not fully penetrant. Symptoms include slowly progressive dysarthria and gait ataxia beginning at ~40 years of age with a range between 20 and 65 years. Other features include nystagmus, leg spasticity, and reduced vibratory sensation. Severely affected individuals are nonambulatory by the fourth to sixth decades. MRI shows cerebellar atrophy. The mechanism of disease may involve a dominant "toxic" effect occurring at the RNA level, as occurs in myotonic dystrophy.

DENTATORUBROPALLIDOLUYSIAN ATROPHY

DRPLA has a variable presentation that may include progressive ataxia, choreoathetosis, dystonia, seizures, myoclonus, and dementia. DRPLA is due to unstable CAG triplet repeats in the open reading frame of a gene named *atrophin* located on chromosome 12p12-ter. Larger expansions are found in patients with earlier onset. The number of repeats is 49 in patients with DRPLA and ≤26 in normal individuals. Anticipation occurs in successive generations, with earlier onset of disease in association with an increasing CAG repeat number in children who inherit the disease from their father. One well-characterized family in North Carolina has a phenotypic variant known as the *Haw River syndrome*, now recognized to be due to the DRPLA mutation.

EPISODIC ATAXIA

EA types 1 and 2 are two rare dominantly inherited disorders that have been mapped to chromosomes 12p (a potassium channel gene, KCNA1, Phe249Leu mutation) for type 1 and 19p for type 2. Patients with EA-1 have brief episodes of ataxia with myokymia and nystagmus that last only minutes. Startle, sudden change in posture, and exercise can induce episodes. Acetazolamide or anticonvulsants may be therapeutic. Patients with EA-2 have episodes of ataxia with nystagmus that can last for hours or days. Stress, exercise, or excessive fatigue may be precipitants. Acetazolamide may be therapeutic and can reverse the relative intracellular alkalosis detected by magnetic resonance spectroscopy. Stop codon, nonsense mutations causing EA-2 have been found

in the *CACNA1A* gene, encoding the α_{1A} voltage-dependent calcium channel subunit (see "SCA6," above).

AUTOSOMAL RECESSIVE ATAXIAS

Friedreich's Ataxia This is the most common form of inherited ataxia, comprising one-half of all hereditary ataxias. It can occur in a classic form or in association with a genetically determined vitamin E deficiency syndrome; the two forms are clinically indistinguishable.

SYMPTOMS AND SIGNS Friedreich's ataxia presents before 25 years of age with progressive staggering gait, frequent falling, and titubation. The lower extremities are more severely involved than the upper ones. Dysarthria occasionally is the presenting symptom; rarely, progressive scoliosis, foot deformity, nystagmus, or cardiopathy is the initial sign.

The neurologic examination reveals nystagmus, loss of fast saccadic eye movements, truncal titubation, dysarthria, dysmetria, and ataxia of trunk and limb movements. Extensor plantar responses (with normal tone in trunk and extremities), absence of deep tendon reflexes, and weakness (greater distally than proximally) are usually found. Loss of vibratory and proprioceptive sensation occurs. The median age of death is 35 years. Women have a significantly better prognosis than men.

Cardiac involvement occurs in 90% of patients. Cardiomegaly, symmetric hypertrophy, murmurs, and conduction defects are reported. Moderate mental retardation or psychiatric syndromes are present in a small percentage of patients. A high incidence of diabetes mellitus (20%) is found and is associated with insulin resistance and pancreatic β-cell dysfunction. Musculoskeletal deformities are common and include pes cavus, pes equinovarus, and scoliosis. MRI of the spinal cord shows atrophy **(Fig. 431-2)**.

The primary sites of pathology are the spinal cord, dorsal root ganglion cells, and the peripheral nerves. Slight atrophy of the cerebellum and cerebral gyri may occur. Sclerosis and degeneration occur predominantly in the spinocerebellar tracts, lateral corticospinal tracts, and posterior columns. Degeneration of the glossopharyngeal, vagus, hypoglossal, and deep cerebellar nuclei is described. The cerebral cortex is histologically normal except for loss of Betz cells in the precentral gyri. The peripheral nerves are extensively involved, with a loss of large myelinated fibers. Cardiac pathology consists of myocytic hypertrophy and fibrosis, focal vascular fibromuscular dysplasia with subintimal or medial deposition of periodic acid-Schiff (PAS)-positive material, myocytopathy with unusual pleomorphic nuclei, and focal degeneration of nerves and cardiac ganglia.

GENETIC CONSIDERATIONS

The classic form of Friedreich's ataxia has been mapped to 9q13-q21.1, and the mutant gene, *frataxin*, contains expanded GAA triplet repeats in the first intron. There is homozygosity for expanded GAA repeats in >95% of patients. Normal persons have 7–22 GAA repeats, and patients have 200–900 GAA repeats. A more varied clinical syndrome has been described in compound heterozygotes who have one copy of the GAA expansion and the other copy a point mutation in the *frataxin* gene. When the point mutation is located in the region of the gene that encodes the amino-terminal half of frataxin, the phenotype is milder, often consisting of a spastic gait, retained or exaggerated reflexes, no dysarthria, and mild or absent ataxia.

Patients with Friedreich's ataxia have undetectable or extremely low levels of *frataxin* mRNA, as compared with carriers and unrelated individuals; thus, disease appears to be caused by a loss of expression of the frataxin protein. Frataxin is a mitochondrial protein involved in iron homeostasis. Mitochondrial iron accumulation due to loss of the iron transporter coded by the mutant *frataxin* gene results in oxidized intramitochondrial iron. Excess oxidized iron results in turn in the oxidation of cellular components and irreversible cell injury.

Two forms of hereditary ataxia associated with abnormalities in the interactions of vitamin E (α-tocopherol) with very-low-density lipoprotein (VLDL) have been delineated. These are abetalipoproteinemia (Bassen-Kornzweig syndrome) and ataxia with vitamin E deficiency (AVED). Abetalipoproteinemia is caused by mutations in the gene coding for the larger subunit of the microsomal triglyceride transfer protein (MTP). Defects in MTP result in impairment of formation and secretion of VLDL in liver. This defect results in a deficiency of delivery of vitamin E to tissues, including the central and peripheral nervous system, as VLDL is the transport molecule for vitamin E and other fat-soluble substitutes. AVED is due to mutations in the gene for α-tocopherol transfer protein (α-TTP). These patients have an impaired ability to bind vitamin E into the VLDL produced and secreted by the liver, resulting in a deficiency of vitamin E in peripheral tissues. Hence, either absence of VLDL (abetalipoproteinemia) or impaired binding of vitamin E to VLDL (AVED) causes an ataxic syndrome. Once again, a genotype classification has proved to be essential in sorting out the various forms of the Friedreich's disease syndrome, which may be clinically indistinguishable.

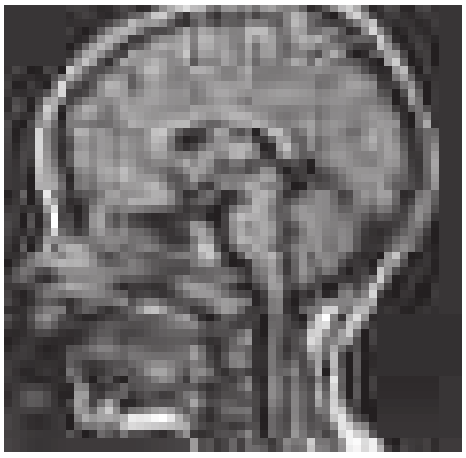


FIGURE 431-2 Sagittal magnetic resonance imaging (MRI) of the brain and spinal cord of a patient with Friedreich's ataxia, demonstrating spinal cord atrophy. *(Reproduced with permission from RN Rosenberg, P Khemani, in RN Rosenberg, JM Pascual [eds]: Rosenberg's Molecular and Genetic Basis of Neurological and Psychiatric Disease, 5th ed. London, Elsevier, 2015.)*

Ataxia Telangiectasia • **SYMPTOMS AND SIGNS** Patients with ataxia telangiectasia (AT) present in the first decade of life with progressive telangiectatic lesions associated with deficits in cerebellar function and nystagmus. The neurologic manifestations correspond to those in Friedreich's disease, which should be included in the differential diagnosis. Truncal and limb ataxia, dysarthria, extensor plantar responses, myoclonic jerks, areflexia, and distal sensory deficits may develop. There is a high incidence of recurrent pulmonary infections and neoplasms of the lymphatic and reticuloendothelial system in patients with AT. Thymic hypoplasia with cellular and humoral (IgA and IgG2) immunodeficiencies, premature aging, and endocrine disorders such as type 1 diabetes mellitus are described. There is an increased incidence of lymphomas, Hodgkin's disease, acute T cell leukemias, and breast cancer.

The most striking neuropathologic changes include loss of Purkinje, granule, and basket cells in the cerebellar cortex as well as of neurons in the deep cerebellar nuclei. The inferior olives of the medulla may also have neuronal loss. There is a loss of anterior horn neurons in the spinal cord and of dorsal root ganglion cells associated with posterior column spinal cord demyelination. A poorly developed or absent thymus gland is the most consistent defect of the lymphoid system.

GENETIC CONSIDERATIONS

The gene for AT (the *ATM* gene) at 11q22-23 encodes a protein that is similar to several yeast and mammalian phosphatidylinositol-3′ kinases involved in mitogenic signal transduction, meiotic recombination, and cell cycle control. Defective DNA repair in AT fibroblasts exposed to ultraviolet light has been demonstrated. The discovery of *ATM* permits early diagnosis and identification of heterozygotes who are at risk for cancer (e.g., breast cancer). Elevated serum alpha-fetoprotein and immunoglobulin deficiency are noted.

MITOCHONDRIAL ATAXIAS

Spinocerebellar syndromes have been identified with mutations in mitochondrial DNA (mtDNA). Thirty pathogenic mtDNA point

 mutations and 60 different types of mtDNA deletions are known, several of which cause or are associated with ataxia **(Chap. 441)**.

TREATMENT

Ataxic Disorders

The most important goal in management of patients with ataxia is to identify treatable disease entities. Mass lesions must be recognized promptly and treated appropriately. Paraneoplastic disorders can often be identified by the clinical patterns of disease that they produce, measurement of specific autoantibodies, and uncovering the primary cancer; these disorders are often refractory to therapy, but some patients improve following removal of the tumor or immunotherapy **(Chap. 90)**. Ataxia with antigliadin antibodies and gluten-sensitive enteropathy may improve with a gluten-free diet. Malabsorption syndromes leading to vitamin E deficiency may lead to ataxia. The vitamin E deficiency form of Friedreich's ataxia must be considered, and serum vitamin E levels measured. Vitamin E therapy is indicated for these rare patients. Vitamin B_1 and B_{12} levels in serum should be measured, and the vitamins administered to patients having deficient levels. Hypothyroidism is easily treated. The cerebrospinal fluid should be tested for a syphilitic infection in patients with progressive ataxia and other features of tabes dorsalis. Similarly, antibody titers for Lyme disease and *Legionella* should be measured and appropriate antibiotic therapy should be instituted in antibody-positive patients. Aminoacidopathies, leukodystrophies, urea-cycle abnormalities, and mitochondrial encephalomyopathies may produce ataxia, and some dietary or metabolic therapies are available for these disorders. The deleterious effects of phenytoin and alcohol on the cerebellum are well known, and these exposures should be avoided in patients with ataxia of any cause.

There is no proven therapy for any of the autosomal dominant ataxias (SCA1 to SCA40). There is preliminary evidence that idebenone, a free-radical scavenger, can improve myocardial hypertrophy in patients with classic Friedreich's ataxia; there is no current evidence, however, that it improves neurologic function. A small preliminary study in a mixed population of patients with different inherited ataxias raised the possibility that the glutamate antagonist riluzole may offer modest benefit. Iron chelators and antioxidant drugs are potentially harmful in Friedreich's patients because they may increase heart muscle injury. Acetazolamide can reduce the duration of symptoms of EA. At present, identification of an at-risk person's genotype, together with appropriate family and genetic counseling, can reduce the incidence of these cerebellar syndromes in future generations **(Chap. 457)**.

GENETIC DIAGNOSTIC LABORATORIES

1. Baylor College of Medicine; Houston, Texas, 1-713-798-6522
   http://www.bcm.edu/genetics/index.cfm?pmid=21387
2. GeneDx
   http://www.genedx.com
3. Transgenomic, 1-877-274-9432
   http://www.transgenomic.com/labs/neurology

GLOBAL FEATURES

Ataxias with autosomal dominant, autosomal recessive, X-linked, or mitochondrial forms of inheritance are present on a worldwide basis. Machado-Joseph disease (SCA3) (autosomal dominant) and Friedreich's ataxia (autosomal recessive) are the most common types in most populations. Genetic markers are now commercially available to precisely identify the genetic mutation for correct diagnosis and also for family planning. Early detection of asymptomatic preclinical disease can reduce or eliminate the inherited form of ataxia in some families on a global, worldwide basis.

FURTHER READING

Anheim M, Tranchant C, Koenig M: The autosomal recessive cerebellar ataxias. N Engl J Med 16:636, 2012.

Jacobi H et al: Long-term disease progression in spinocerebellar ataxia types 1, 2, 3, and 6: A longitudinal cohort study. Lancet Neurol 14:1101, 2015.

Romano S et al: Riluzole in patients with hereditary cerebellar ataxia: A randomised, double-blind, placebo-controlled trial. Lancet Neurol 14:985, 2015.

432 Disorders of the Autonomic Nervous System

Christopher H. Gibbons, John W. Engstrom

The autonomic nervous system (ANS) innervates the entire neuraxis and influences all organ systems. It regulates blood pressure (BP), heart rate, sleep, glandular, pupillary, bladder and bowel function. It maintains organ homeostasis and operates automatically; its full importance becomes recognized only when ANS function is compromised, resulting in dysautonomia. **Hypothalamic disorders that cause disturbances in homeostasis are discussed in Chaps. 15 and 371.**

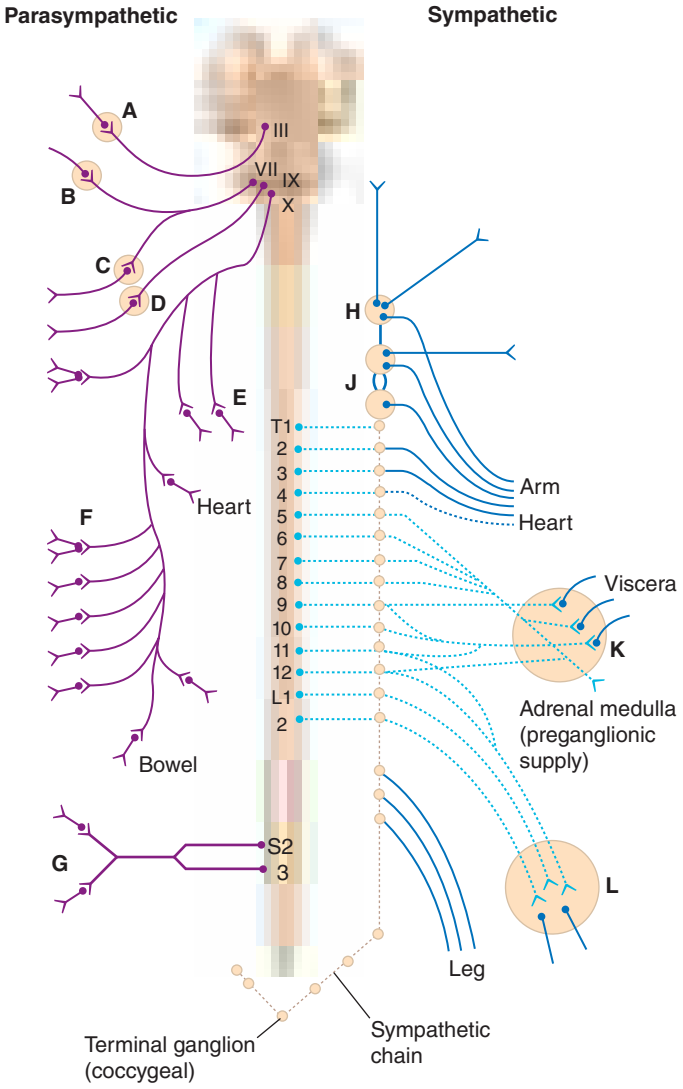
ANATOMIC ORGANIZATION

The activity of the ANS is regulated by central neurons responsive to diverse afferent inputs. After central integration of afferent information, autonomic outflow is adjusted to permit the functioning of the major organ systems in accordance with the needs of the whole organism. Connections between the cerebral cortex and the autonomic centers in the brainstem coordinate autonomic outflow with higher mental functions.

The preganglionic neurons of the parasympathetic nervous system leave the central nervous system (CNS) in the third, seventh, ninth, and tenth cranial nerves as well as the second and third sacral nerves, while the preganglionic neurons of the sympathetic nervous system exit the spinal cord between the first thoracic and the second lumbar segments **(Fig. 432-1)**. The autonomic preganglionic fibers are thinly myelinated. The postganglionic neurons, located in ganglia outside the CNS, give rise to the postganglionic unmyelinated autonomic nerves that innervate organs and tissues throughout the body. Responses to sympathetic and parasympathetic stimulation are frequently antagonistic **(Table 432-1)**, reflecting highly coordinated interactions within the CNS; the resultant changes in parasympathetic and sympathetic activity provide more precise control of autonomic responses than could be achieved by the modulation of a single system.

Acetylcholine (ACh) is the preganglionic neurotransmitter for both the sympathetic and parasympathetic divisions of the ANS as well as the postganglionic neurotransmitter of the parasympathetic neurons; the preganglionic receptors are nicotinic, and the postganglionic are muscarinic in type. Norepinephrine (NE) is the neurotransmitter of the postganglionic sympathetic neurons, except for cholinergic neurons innervating the eccrine sweat glands.

The gastrointestinal (GI) tract has long been described as part of the sympathetic and parasympathetic nervous systems. However, it is has many unique characteristics such that it is now considered separately as the enteric, or intrinsic, nervous system. Parasympathetic control of the GI system is through the craniospinal nerves (vagus and S2-S4 nerves) while sympathetic control is through the thoracolumbar region. The enteric nervous system itself is made up of a series of ganglia that form a network of plexuses with several hundred million cells (the equivalent of the number of cells in the spinal cord). Meissner's (submucosal) plexus, Auerbach's (myenteric), Cajal's (deep muscular), mucosal and submucosal plexuses comprise the majority of nerves within the enteric nervous system. Numerous neurotransmitters have now been identified within the enteric nervous system, with many neurons containing both primary and co-transmitter neurotransmitters.



Parasympathetic system
from cranial nerves III, VII, IX, X and from sacral nerves 2 and 3

Sympathetic system
from T1-L2
Preganglionic fibers
Postganglionic fibers

- **A** Ciliary ganglion
- **B** Sphenopalatine (pterygopalatine) ganglion
- **C** Submandibular ganglion
- **D** Otic ganglion
- **E** Vagal ganglion cells in the heart wall
- **F** Vagal ganglion cells in bowel wall
- **G** Pelvic ganglia
- **H** Superior cervical ganglion
- **J** Middle cervical ganglion and inferior cervical (stellate) ganglion including T1 ganglion
- **K** Coeliac and other abdominal ganglia
- **L** Lower abdominal sympathetic ganglia

FIGURE 432-1 Schematic representation of the autonomic nervous system. *(From M Moskowitz: Clin Endocrinol Metab 6:77, 1977.)*

TABLE 432-1 Functional Consequences of Normal ANS Activation

	SYMPATHETIC	PARASYMPATHETIC
Heart rate	Increased	Decreased
Blood pressure	Increased	Mildly decreased
Bladder	Increased sphincter tone	Voiding (decreased tone)
Bowel motility	Decreased motility	Increased
Lung	Bronchodilation	Bronchoconstriction
Sweat glands	Sweating	—
Pupils	Dilation	Constriction
Adrenal glands	Catecholamine release	—
Sexual function	Ejaculation, orgasm	Erection
Lacrimal glands	—	Tearing
Parotid glands	—	Salivation

CLINICAL EVALUATION

CLASSIFICATION

Disorders of the ANS may result from pathology of either the CNS or the peripheral nervous system (PNS) (Table 432-2). Signs and symptoms may result from interruption of the afferent limb, CNS processing centers, or efferent limb of reflex arcs controlling autonomic responses. For example, a lesion of the medulla produced by a posterior fossa tumor can impair BP responses to postural changes and result in orthostatic hypotension (OH). OH can also be caused by lesions of the afferent limb of the baroreflex arc (e.g., radiation or congenital disease), spinal cord or peripheral vasomotor nerve fibers (e.g., diabetic autonomic neuropathy). Lesions of the efferent limb cause the most consistent and severe OH. The site of reflex interruption is usually established by the clinical context in which the dysautonomia arises, combined with judicious use of ANS testing and neuroimaging studies. The presence or absence of CNS signs, association with sensory or motor polyneuropathy, medical illnesses, medication use, and family history are often important considerations. Some syndromes do not fit easily into any classification scheme.

SYMPTOMS OF AUTONOMIC DYSFUNCTION

Clinical manifestations can result from loss of function, overactivity, or dysregulation of autonomic circuits. Disorders of autonomic function should be considered in patients with unexplained OH, syncope, sleep dysfunction, altered sweating (hyperhidrosis or hypohidrosis), impotence, constipation or other GI symptoms (bloating, nausea, vomiting of old food, diarrhea), or bladder disorders (urinary frequency, hesitancy, or incontinence). Symptoms may be widespread or regional in distribution. An autonomic history focuses on systemic functions (orthostatic symptoms, BP, heart rate, sleep, fever, sweating) and involvement of individual organ systems (pupils, bowel, bladder, sexual function). The autonomic symptom profile is a self-report questionnaire that can be used for formal assessment. A recent consensus guideline has provided some useful screening questions for detection of OH, and there is a recently developed OH questionnaire for more detailed symptom assessment. It is important to recognize the modulating effects of age and duration of disease. For example, OH typically produces lightheadedness when of acute onset, but may present with subtle cognitive manifestations in chronic disease. Specific symptoms of orthostatic intolerance are diverse (Table 432-3). Autonomic symptoms may vary dramatically, reflecting the dynamic nature of autonomic control over homeostatic function. For example, OH might be manifest only in the early morning, following a meal, with exercise, or with raised ambient temperature, depending on the regional vascular bed affected by the dysautonomia.

Early autonomic symptoms may be overlooked. Impotence, although not specific for autonomic failure, often heralds autonomic failure in men and may precede other symptoms by years (Chap. 390). A decrease in the frequency of spontaneous early morning erections may occur months before loss of nocturnal penile tumescence and development of total impotence. Bladder dysfunction may appear early in men and women, particularly in those with a CNS etiology. Cold feet may indicate increased peripheral vasomotor constriction, although this symptom is a very common complaint among healthy individuals as well. Brain and spinal cord disease above the level of the lumbar spine results first in urinary frequency and small bladder volumes and eventually in incontinence (upper motor neuron or spastic bladder). By contrast, PNS disease of autonomic nerve fibers results in large bladder volumes, urinary frequency, and overflow incontinence (lower motor neuron flaccid bladder). Measurements of bladder volume (postvoid residual), or urodynamic studies, are useful tests for distinguishing between upper and lower motor neuron bladder dysfunction in the early stages of dysautonomia. GI autonomic dysfunction typically

TABLE 432-2 Classification of Clinical Autonomic Disorders

I. Autonomic Disorders with Brain Involvement

A. Associated with multisystem degeneration
  1. Multisystem degeneration: autonomic failure clinically prominent
    a. Multiple system atrophy (MSA)
    b. Parkinson's disease with autonomic failure
    c. Diffuse Lewy body disease with autonomic failure
  2. Multisystem degeneration: autonomic failure clinically not usually prominent
    a. Parkinson's disease without autonomic failure
    b. Other extrapyramidal disorders (inherited spinocerebellar atrophies, progressive supranuclear palsy, corticobasal degeneration, Machado-Joseph disease, fragile X syndrome [FXTAS])

B. Unassociated with multisystem degeneration (focal CNS disorders)
  1. Disorders mainly due to cerebral cortex involvement
    a. Frontal cortex lesions causing urinary/bowel incontinence
    b. Focal seizures (temporal lobe or anterior cingulate)
    c. Cerebral infarction of the insula
  2. Disorders of the limbic and paralimbic circuits
    a. Shapiro's syndrome (agenesis of corpus callosum, hyperhidrosis, hypothermia)
    b. Autonomic seizures
    c. Limbic encephalitis
  3. Disorders of the hypothalamus
    a. Thiamine deficiency (Wernicke-Korsakoff syndrome)
    b. Diencephalic syndrome
    c. Neuroleptic malignant syndrome
    d. Serotonin syndrome
    e. Fatal familial insomnia
    f. Antidiuretic hormone (ADH) syndromes (diabetes insipidus, inappropriate ADH secretion)
    g. Disturbances of temperature regulation (hyperthermia, hypothermia)
    h. Disturbances of sexual function
    i. Disturbances of appetite
    j. Disturbances of BP/HR and gastric function
    k. Horner's syndrome
  4. Disorders of the brainstem and cerebellum
    a. Posterior fossa tumors
    b. Syringobulbia and Arnold-Chiari malformation
    c. Disorders of BP control (hypertension, hypotension)
    d. Cardiac arrhythmias
    e. Central sleep apnea
    f. Baroreflex failure
    g. Horner's syndrome
    h. Vertebrobasilar and lateral medullary (Wallenberg's) syndromes
    i. Brainstem encephalitis

II. Autonomic Disorders with Spinal Cord Involvement

A. Traumatic quadriplegia
B. Syringomyelia
C. Subacute combined degeneration
D. Multiple sclerosis and neuromyelitis optica
E. Amyotrophic lateral sclerosis
F. Tetanus
G. Stiff-person syndrome
H. Spinal cord tumors

III. Autonomic Neuropathies

A. Acute/subacute autonomic neuropathies
  a. Subacute autoimmune autonomic ganglionopathy (AAG)
  b. Subacute paraneoplastic autonomic neuropathy
  c. Guillain-Barré syndrome
  d. Botulism
  e. Porphyria
  f. Drug induced autonomic neuropathies-stimulants, drug withdrawal, vasoconstrictor, vasodilators, beta-receptor antagonists, beta-agonists
  g. Toxin-induced autonomic neuropathies
  h. Subacute cholinergic neuropathy

B. Chronic peripheral autonomic neuropathies
  1. Distal small fiber neuropathy
  2. Combined sympathetic and parasympathetic failure
    a. Amyloid
    b. Diabetic autonomic neuropathy
    c. AAG (paraneoplastic and idiopathic)
    d. Sensory neuronopathy with autonomic failure
    e. Familial dysautonomia (Riley-Day syndrome)
    f. Diabetic, uremic, or nutritional deficiency
    g. Geriatric dysautonomia (age >80 years)
  3. Disorders of orthostatic intolerance: reflex syncope; POTS; prolonged bed rest; space flight; chronic fatigue

Abbreviations: BP, blood pressure; CNS, central nervous system; HR, heart rate; POTS, postural orthostatic tachycardia syndrome.

presents as progressively severe constipation. Diarrhea may develop (typically in diabetes mellitus) due to many reasons including rapid transit of contents, uncoordinated small-bowel motor activity, an osmotic basis from bacterial overgrowth associated with small-bowel stasis, and anorectal dysfunction with diminished sphincter control and increased intestinal secretion. Impaired glandular secretory function may cause difficulty with food intake due to decreased salivation or eye irritation due to decreased lacrimation. Loss of sweat function (anhidrosis), a critical element of thermoregulation, may result in hyperthermia. Patients with a length-dependent neuropathy may present with distal anhidrosis but the primary complaint is proximal hyperhidrosis that occurs to maintain thermoregulation **(Chap. 15)**. Lack of sweating after a hot bath, during exercise, or on a hot day can suggest sudomotor failure.

TABLE 432-3 Symptoms of Orthostatic Intolerance

Lightheadedness (dizziness)	88%
Weakness or tiredness	72%
Cognitive difficulty (thinking/concentrating)	47%
Blurred vision	47%
Tremulousness	38%
Vertigo	37%
Pallor	31%
Anxiety	29%
Palpitations	26%
Clammy feeling	19%
Nausea	18%

Source: PA Low et al: Mayo Clin Proc 70:617, 1995.

OH (also called *orthostatic or postural hypotension*) is perhaps the most disabling feature of autonomic dysfunction. There are numerous causes of OH (e.g., medications, anemia, dehydration or volume depletion), but when the OH is specifically due to dysfunction of the ANS it is referred to as neurogenic OH. The prevalence of OH is relatively high, especially when OH associated with aging and diabetes mellitus is included **(Table 432-4)**. OH can cause a variety of symptoms, including dimming or loss of vision, lightheadedness, diaphoresis, diminished hearing, pallor, weakness, and shortness of breath. Syncope results

TABLE 432-4 Prevalence of Orthostatic Hypotension in Different Situations

DISORDER	PREVALENCE
Aging	14–20%
Diabetic neuropathy	10%
Other autonomic neuropathies	>60%
Multiple system atrophy	>90%
Pure autonomic failure	>95%

when the drop in BP impairs cerebral perfusion. Other manifestations of impaired baroreflexes are supine hypertension, a heart rate that is fixed regardless of posture, postprandial hypotension, and an excessively high nocturnal BP. Many patients with OH have a preexisting diagnosis of hypertension or have concomitant supine hypertension, reflecting the great importance of baroreflexes in maintaining postural and supine normotension. The appearance of OH in patients receiving antihypertensive treatment may indicate overtreatment or the onset of an autonomic disorder. The most common causes of OH are not neurologic in origin **(Table 432-5)**; these must be distinguished from the neurogenic causes. The mortality rates of nonneurogenic OH are similar to that of the general population while neurogenic OH carries a three- to sevenfold higher mortality rate. **Neurocardiogenic and cardiac causes of syncope are considered in Chap. 18.**

TABLE 432-5 Nonneurogenic Causes of Orthostatic Hypotension

Cardiac Pump Failure
Myocardial infarction
Myocarditis
Constrictive pericarditis
Aortic stenosis
Tachyarrhythmias
Bradyarrhythmias
Salt-losing nephropathy
Adrenal insufficiency
Diabetes insipidus
Venous obstruction
Reduced Intravascular Volume
Straining or heavy lifting, urination, defecation
Dehydration
Diarrhea, emesis
Hemorrhage
Burns
Metabolic
Adrenocortical insufficiency
Hypoaldosteronism
Pheochromocytoma
Severe potassium depletion
Venous Pooling
Alcohol
Postprandial dilation of splanchnic vessel beds
Vigorous exercise with dilation of skeletal vessel beds
Heat: hot environment, hot showers and baths, fever
Prolonged recumbency or standing
Sepsis
Medications
Antihypertensives
Diuretics
Vasodilators: nitrates, hydralazine
Alpha- and beta-blocking agents
Central nervous system sedatives: barbiturates, opiates
Tricyclic antidepressants
Phenothiazines

APPROACH TO THE PATIENT

Orthostatic Hypotension and Other ANS Disorders

The first step in the evaluation of symptomatic OH is the exclusion of treatable causes. The history should include a review of medications that may affect the ANS **(Table 432-6)**. The main classes of drugs that may cause OH are diuretics, antihypertensive agents (preload reducers, vasodilators, negative inotropic or chronotropic agents), antidepressants (tricyclic antidepressants and SSRIs), ethanol, opioids, insulin, dopamine agonists, and barbiturates. However, the precipitation of OH by medications may also be the first sign of an underlying autonomic disorder. The history may reveal an underlying cause for symptoms (e.g., diabetes, Parkinson's disease) or specific underlying mechanisms (e.g., cardiac pump failure, reduced intravascular volume). The relationship of symptoms to meals (splanchnic pooling), standing on awakening in the morning (intravascular volume depletion), ambient warming (vasodilatation), or exercise (muscle arteriolar vasodilatation) should be sought. Standing time to first symptom and to presyncope **(Chap. 18)** should be followed for management.

Physical examination includes measurement of supine and standing pulse and BP. OH is defined as a sustained drop in systolic (≥20 mmHg) or diastolic (≥10 mmHg) BP after 3 min of standing. In non-neurogenic causes of OH (such as hypovolemia), the BP drop is accompanied by a compensatory increase in heart rate of >15 beats/min. A clue that the patient has neurogenic OH is the aggravation or precipitation of OH by autonomic stressors (a meal, hot bath, or exercise). Neurologic examination should include mental status (neurodegenerative disorders such as Lewy body dementia can be accompanied by significant dysautonomia), cranial nerves (abnormal pupils with Horner's or Adie's syndrome), motor tone (parkinsonian syndromes), motor strength and sensation (polyneuropathies). In patients without a clear diagnosis initially, follow-up evaluations every few months or whenever symptoms worsen may reveal the underlying cause.

Disorders of autonomic function should be considered in patients with symptoms of altered sweating (hyperhidrosis or hypohidrosis), gastroparesis (bloating, nausea, vomiting of old food), impotence, constipation, or bladder disturbances (urinary frequency, hesitancy, or incontinence).

AUTONOMIC TESTING

Autonomic function tests are helpful to document and localize abnormalities when findings on history and examination are inconclusive; to detect subclinical involvement; or to follow the course of an autonomic disorder.

Heart Rate Variation With Deep Breathing This tests the parasympathetic component of cardiovascular reflexes via the vagus nerve. Results are influenced by multiple factors including the subject's position (recumbent, sitting, or standing), rate and depth

TABLE 432-6 Some Drugs That Affect Autonomic Function

SYMPTOM	DRUG CLASS	SPECIFIC EXAMPLES
Impotence	Opioids	Tylenol #3
	Anabolic steroids	—
	Some antiarrhythmics	Prazosin
	Some antihypertensives	Clonidine
	Some diuretics	Benazepril
	Some SSRIs	Venlafaxine
Urinary retention	Opioids	Fentanyl
	Decongestants	Brompheniramine
		Diphenhydramine
Diaphoresis	Some antihypertensives	Amlodipine
	Some SSRIs	Citalopram
	Opioids	Morphine

Abbreviations: CCBs, calcium channel blockers; HCTZ, hydrochlorothiazide; SSRIs, selective serotonin reuptake inhibitors.

TABLE 432-7 Normal Blood Pressure and Heart Rate Changes During the Valsalva Maneuver

PHASE	MANEUVER	BLOOD PRESSURE	HEART RATE	COMMENTS
I	Forced expiration against a partially closed glottis	Rises; aortic compression from raised intrathoracic pressure	Decreases	Mechanical
II *early*	Continued expiration	Falls; decreased venous return to the heart	Increases (reflex tachycardia)	Reduced vagal tone
II *late*	Continued expiration	Rises; reflex increase in peripheral vascular resistance	Increases at slower rate	Requires intact efferent sympathetic response
III	End of expiration	Falls; increased capacitance of pulmonary bed	Increases further	Mechanical
IV	Recovery	Rises; persistent vasoconstriction and increased cardiac output	Compensatory bradycardia	Requires intact efferent sympathetic response

of respiration (6 breaths per minute and a forced vital capacity [FVC] >1.5 L are optimal), age, medications, weight, and degree of hypocapnia. Interpretation of results requires comparison of test data with results from age-matched controls collected under identical test conditions. For example, the lower limit of normal heart rate variation with deep breathing in persons <20 years is >15–20 beats/min, but for persons aged >60 it is 5–8 beats/min. Heart rate variation with deep breathing (respiratory sinus arrhythmia) is abolished by the muscarinic ACh receptor antagonist atropine but is unaffected by sympathetic postganglionic blockade (e.g., propranolol).

Valsalva Response This response **(Table 432-7)** assesses the integrity of the baroreflex control of heart rate (parasympathetic) and BP (sympathetic adrenergic). Under normal conditions, increases in BP at the carotid bulb trigger a reduction in heart rate (increased vagal tone), and decreases in BP trigger an increase in heart rate (reduced vagal tone). The Valsalva response is tested in the supine position. The subject exhales against a closed glottis (or into a manometer maintaining a constant expiratory pressure of 40 mmHg) for 15 s while measuring changes in heart rate and beat-to-beat BP. Without directly measuring expiratory pressure, heart rate and beat-to-beat blood pressure the Valsalva maneuver cannot be interpreted correctly. There are four phases of the BP and heart rate response to the Valsalva maneuver. Phases I and III are mechanical and related to changes in intrathoracic and intraabdominal pressure. In early phase II, reduced venous return results in a fall in stroke volume and BP, counteracted by a combination of reflex tachycardia and increased total peripheral resistance. Increased total peripheral resistance arrests the BP drop ~5–8 s after the onset of the maneuver. Late phase II begins with a progressive rise in BP toward or above baseline. Venous return and cardiac output return to normal in phase IV. Persistent peripheral arteriolar vasoconstriction and increased cardiac adrenergic tone result in a temporary BP overshoot and phase IV bradycardia (mediated by the baroreceptor reflex). Abnormalities in blood pressure during phase II recovery or phase IV overshoot suggest sympathetic adrenergic dysfunction.

Autonomic parasympathetic function during the Valsalva maneuver is measured using heart rate changes. The *Valsalva ratio* is defined as the maximum phase II tachycardia divided by the minimum phase IV bradycardia **(Table 432-8)** and is predominantly a measure of parasympathetic function.

Sudomotor Function Sweating is induced by release of ACh from sympathetic postganglionic fibers. The quantitative sudomotor axon reflex test (QSART) is a measure of regional autonomic function mediated by ACh-induced sweating. A reduced or absent response indicates a lesion of the postganglionic sudomotor axon. For example, sweating may be reduced in the feet as a result of distal polyneuropathy (e.g., diabetes). The thermoregulatory sweat test (TST) is a qualitative measure of global sweat production in response to an elevation of body temperature under controlled conditions. An indicator powder placed on the anterior surface of the body changes color with sweat production during temperature elevation. The pattern of color change measures the integrity of both the preganglionic and postganglionic sudomotor function. A postganglionic lesion is present if both QSART and TST show absent sweating. In a preganglionic lesion, the QSART is normal but TST shows anhidrosis.

Orthostatic BP Recordings Beat-to-beat BP measurements determined in supine, 70° tilt, and tilt-back positions are useful to quantitate orthostatic failure of BP control. Allow a 20-min period of rest in the supine position before assessing changes in BP during tilting. The BP change combined with heart rate monitoring is useful for the evaluation of patients with suspected OH or unexplained syncope.

Tilt Table Testing For Syncope The great majority of patients with syncope do not have autonomic failure. Tilt table testing can be used to make the diagnosis of vasovagal syncope with sensitivity, specificity, and reproducibility. A standardized protocol is used that specifies the tilt apparatus, tilt angle, and duration of tilt. A passive phase for 30–40 min with a tilt angle at 60–70 degrees can identify reflex syncope, psychogenic syncope, or be nondiagnostic. Pharmacologic provocation of syncope (with intravenous, sublingual, or spray nitroglycerin) is controversial because it increases sensitivity at the cost of specificity. Recommendations for the performance of tilt studies for syncope have been incorporated in consensus guidelines.

SPECIFIC SYNDROMES OF ANS DYSFUNCTION

MULTIPLE SYSTEM ATROPHY

Multiple system atrophy (MSA) is an entity that comprises autonomic failure (OH or a neurogenic bladder) *and* either parkinsonism (MSA-p) or a cerebellar syndrome (MSA-c). MSA-p is the more common form; the parkinsonism is atypical in that there is more symmetric motor involvement than in Parkinson's disease (PD; **Chap. 427**), tremor is not as prominent, and there is a poor or only transient response to levodopa. Symptomatic OH within 1 year of onset of parkinsonism

TABLE 432-8 Neural Pathways Underlying Some Standardized Autonomic Tests

TEST EVALUATED	PROCEDURE	AUTONOMIC FUNCTION
HRDB	6 deep breaths/min	Cardiovagal (parasympathetic) function
Valsalva ratio	Expiratory pressure, 40 mmHg for 10–15 s	Cardiovagal (parasympathetic) function
QSART	Axon-reflex test 4 limb sites	Postganglionic (sympathetic cholinergic) sudomotor function
BP_{BB} to VM	BP_{BB} response to VM	Sympathetic adrenergic function: baroreflex adrenergic control of vagal and vasomotor function
HUT	BP_{BB} and heart rate response to HUT	Sympathetic adrenergic and cardiovagal (parasympathetic) responses to HUT

Abbreviations: BP_{BB}, beat-to-beat blood pressure; HRDB, heart rate response to deep breathing; HUT, head-up tilt; QSART, quantitative sudomotor axon reflex test; VM, Valsalva maneuver.

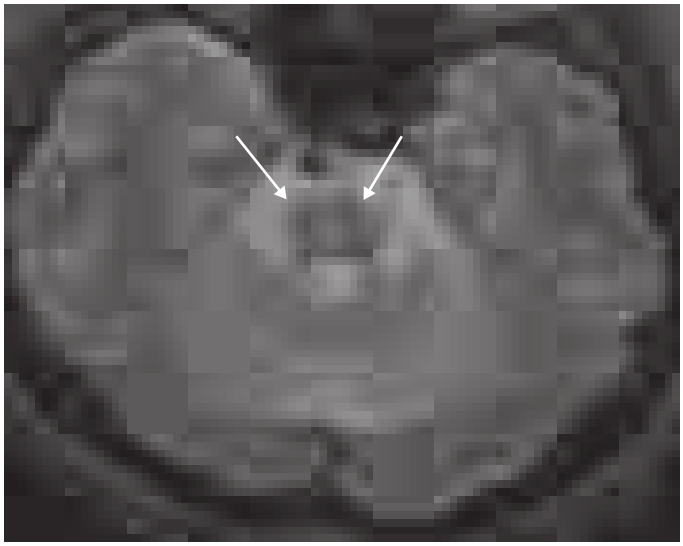


FIGURE 432-2 Multiple system atrophy, cerebellar type (MSA-c). Axial T2-weighted magnetic resonance image at the level of the pons shows a characteristic hyperintense signal, the "hot cross buns" sign (*arrows*). This appearance can also be seen in some spinocerebellar atrophies, as well as other neurodegenerative conditions affecting the brainstem.

is suggestive of MSA-p. There is a very high frequency of impotence in men. Although autonomic abnormalities are common in advanced PD, the severity and distribution of autonomic failure are more severe and generalized in MSA. Brain magnetic resonance imaging (MRI) is a useful diagnostic adjunct: in MSA-p, iron deposition in the striatum may be evident as T2 hypointensity, and in MSA-c, cerebellar atrophy is present with a characteristic T2 hyperintense signal ("hot cross buns sign") in the pons **(Fig. 432-2)**. However, these MRI findings are typically present only with advanced disease. Cardiac postganglionic adrenergic innervation, measured by uptake of fluorodopamine on positron emission tomography, is markedly impaired in the dysautonomia of PD but is usually normal in MSA. Neuropathologic changes include neuronal loss and gliosis in many CNS regions, including the brainstem, cerebellum, striatum, and intermediolateral cell column of the thoracolumbar spinal cord.

MSA is uncommon, with a prevalence estimated at 2–5 per 100,000 individuals. Onset is typically in the mid-fifties, men are slightly more often affected than women, and most cases are sporadic. The diagnosis should be considered in adults aged >30 years who present with OH or urinary incontinence and either parkinsonism that is poorly responsive to dopamine replacement or a cerebellar syndrome. MSA generally progresses relentlessly to death 7–10 years after onset, but survival beyond 15 years has been reported. MSA-p is more prevalent in Western countries, while MSA-c is more common in Japan. Factors that predict a worse prognosis include early autonomic dysfunction, rapid progression of disability, bladder dysfunction, female gender, the MSA-p subtype, and an older age at onset. Attempts to slow the progression of MSA have thus far been unsuccessful, including trials of lithium, growth hormone, riluzole, rasagiline, minocycline, and rifampicin.

Management is symptomatic for neurogenic OH (see below), sleep disorders including laryngeal stridor, GI, and urinary dysfunction. GI management includes frequent small meals, soft diet, stool softeners, and bulk agents. Gastroparesis is difficult to treat; metoclopramide stimulates gastric emptying but worsens parkinsonism by blocking central dopamine receptors. The peripheral dopamine (D_2 and D_3) receptor antagonist domperidone has been used patients with various GI conditions in many countries, and although not available in the United States, it can be obtained through the U.S. Food and Drug Administration's (FDA) Expanded Access to Investigational Drugs program.

Autonomic dysfunction is also a common feature in dementia with Lewy bodies **(Chap. 426)**; with the severity usually intermediate between that found in MSA and PD. In multiple sclerosis (MS; **Chap. 436**), autonomic complications reflect the CNS location of MS involvement and generally worsen with disease duration and disability, but are generally a secondary complaint and not of the severity seen in the synucleinopathies.

SPINAL CORD

Spinal cord lesions from any cause can result in focal autonomic deficits or autonomic hyperreflexia (e.g., spinal cord transection or hemisection) affecting bowel, bladder, sexual, temperature-regulation, or cardiovascular functions. Quadriparetic patients exhibit both supine hypertension and OH after upward tilting. *Autonomic dysreflexia* describes a dramatic increase in BP in patients with traumatic spinal cord lesions above the T6 level, often in response to irritation of the bladder, skin, or muscles. The triggers may be clinically silent because perception of painful sensations arising from structures innervated below the level of a spinal cord lesion is often blunted or absent. A distended bladder, often from an obstructed Foley catheter or a urinary infection, are common triggers of dysreflexia. Associated symptoms can include facial flushing, headache, hypertension, or piloerection. Potential complications include intracranial vasospasm or hemorrhage, cardiac arrhythmia, and death. Awareness of the syndrome, identifying the trigger, and careful monitoring of BP during procedures in patients with acute or chronic spinal cord injury are essential. In patients with supine hypertension, BP can be lowered by tilting the head upward or sitting the patient up. Vasodilator drugs may be used to treat acute elevations in BP. Clonidine can be used prophylactically to reduce the hypertension resulting from bladder stimulation. Dangerous increases or decreases in body temperature may result from an inability to sense heat or cold exposure or control peripheral vasoconstriction or sweating below the level of the spinal cord injury.

PERIPHERAL NERVE AND NEUROMUSCULAR JUNCTION DISORDERS

Peripheral neuropathies **(Chap. 438)** are the most common cause of chronic autonomic insufficiency. Polyneuropathies that affect small myelinated and unmyelinated fibers of the sympathetic and parasympathetic nerves commonly occur in diabetes mellitus, amyloidosis, chronic alcoholism, porphyria, and Guillain-Barré syndrome. Neuromuscular junction disorders with autonomic involvement include botulism and Lambert-Eaton syndrome **(Chap. 440)**.

Diabetes Mellitus The presence of autonomic neuropathy in patients with diabetes increases the mortality rate 1.5- to 3-fold, even after adjusting for other cardiovascular risk factors. Estimates of 5-year mortality risk among these patients range from 15 to 53%. Although many deaths are due to secondary vascular disease, there are patients who specifically suffer cardiac arrest due to autonomic neuropathy. The autonomic involvement is also predictive of other complications including renal disease, stroke, and sleep apnea. Tight glycemic control with insulin significantly reduces the long-term risk of autonomic cardiovascular neuropathy. **Diabetes mellitus is discussed in Chaps. 396–398.**

Amyloidosis Autonomic neuropathy occurs in both sporadic and familial forms of amyloidosis. The AL (immunoglobulin light chain) type is associated with primary amyloidosis or amyloidosis secondary to multiple myeloma. The amyloid transthyretin (ATTR) type, with transthyretin as the primary protein component, is responsible for the most common form of inherited amyloidosis. Although patients usually present with a distal sensorimotor polyneuropathy accompanied by autonomic insufficiency that can precede the development of the polyneuropathy or occur in isolation. The diagnosis can be made by protein electrophoresis of blood and urine, tissue biopsy (abdominal fat pad, rectal mucosa, or sural nerve) to search for amyloid deposits, and genetic testing for transthyretin mutations in familial cases. Death is usually due to cardiac or renal involvement. Postmortem studies reveal amyloid deposition in many organs, including two sites that contribute to autonomic failure: intraneural blood vessels and autonomic ganglia. Pathologic examination reveals a loss of both unmyelinated and myelinated nerve fibers. **Clinical manifestations and treatment of the various forms of amyloidosis are discussed in detail in Chap. 108.**

Alcoholic Neuropathy Abnormalities in parasympathetic vagal and efferent sympathetic function are usually mild in alcoholic polyneuropathy. OH is usually due to brainstem involvement, rather than injury to the PNS. Impotence is a major problem, but concurrent gonadal hormone abnormalities may play a role in this symptom. Clinical symptoms of autonomic failure generally appear only when the stocking-glove polyneuropathy is severe, and there is usually coexisting Wernicke's encephalopathy (Chap. 301). Autonomic involvement may contribute to the high mortality rates associated with alcoholism (Chap. 445).

Porphyria (Chap. 409) Autonomic dysfunction is most extensively documented in acute intermittent porphyria but can also occur with variegate porphyria and hereditary coproporphyria. Autonomic symptoms include tachycardia, sweating, urinary retention, abdominal pain, nausea and vomiting, insomnia, hypertension, and (less commonly) hypotension. Another prominent symptom is anxiety. Abnormal autonomic function can occur both during acute attacks and during remissions. Elevated catecholamine levels during acute attacks correlate with the degree of tachycardia and hypertension that is present.

Guillain-Barré Syndrome (Chap. 439) BP fluctuations and arrhythmias from autonomic instability can be severe. It is estimated that between 2 and 10% of patients with severe Guillain-Barré syndrome suffer fatal cardiovascular collapse. GI autonomic involvement, sphincter disturbances, abnormal sweating, and pupillary dysfunction can also occur. Demyelination has been described in the vagus and glossopharyngeal nerves, the sympathetic chain, and the white rami communicantes. Interestingly, the degree of autonomic involvement appears to be independent of the severity of motor or sensory neuropathy. Acute autonomic and sensory neuropathy is a variant that spares the motor system and presents with neurogenic OH and varying degrees of sensory loss. It is treated similarly to Guillain-Barré syndrome, but prognosis is less favorable, with persistent severe sensory deficits and variable degrees of OH in many patients.

Autoimmune Autonomic Ganglionopathy (AAG) This disorder presents with the subacute development of autonomic disturbances including OH, enteric neuropathy (gastroparesis, ileus, constipation/diarrhea), flaccid bladder, and cholinergic failure (e.g., loss of sweating, sicca complex, and a tonic pupil). A chronic form of AAG resembles pure autonomic failure (PAF) (see below). Autoantibodies against the α3 subunit of the ganglionic Ach receptor, present in approximately half of patients, are considered diagnostic of AAG. Pathology shows preferential involvement of small unmyelinated nerve fibers, with sparing of larger myelinated ones. Onset of the neuropathy follows a viral infection in approximately half of cases. Up to one-third of untreated patients experience significant functional improvement over time. Immunotherapies that have been reported to be helpful include plasmapheresis, intravenous immune globulin, glucocorticoids, azathioprine, rituximab, and mycophenolate mofetil. OH, gastroparesis, and sicca symptoms can be managed symptomatically.

AAG can also occur on a paraneoplastic basis, with adenocarcinoma or small-cell carcinoma of the lung, lymphoma, or thymoma being the most common (Chap. 90). Cerebellar involvement or dementia may coexist (see Tables 90-1–90-3), and the neoplasm can be occult.

Botulism Botulinum toxin binds presynaptically to cholinergic nerve terminals and, after uptake into the cytosol, blocks ACh release. This acute cholinergic neuropathy presents as motor paralysis and autonomic disturbances that include blurred vision, dry mouth, nausea, unreactive or sluggishly reactive pupils, constipation, and urinary retention (Chap. 148).

PURE AUTONOMIC FAILURE (PAF)

This sporadic syndrome consists of postural hypotension, impotence, bladder dysfunction, and impaired sweating. The disorder begins in midlife and occurs in women more often than men. The symptoms can be disabling, but life span is unaffected. The clinical and pharmacologic characteristics suggest primary involvement of postganglionic autonomic neurons. A severe reduction in the density of neurons within sympathetic ganglia results in low supine plasma NE levels and noradrenergic supersensitivity. Some patients who are initially labeled with this diagnosis subsequently go on to develop AAG, but more often a neurodegenerative disease supervenes, typically Lewy body dementia, PD, or MSA. In one recent series, more than one-third of patients initially diagnosed with PAF developed a CNS synucleinopathy within 4 years, and the presence of rapid eye movement sleep behavior disorder (RBD; Chap. 27) was predictive of subsequent CNS disease. Skin biopsies and autopsy studies demonstrate phosphorylated α-synuclein inclusions in postganglionic sympathetic adrenergic and cholinergic nerve fibers, distinguishing PAF from AAG and indicating that PAF is a synucleinopathy; notably, patients with PD also have α-synuclein inclusions in sympathetic nerve biopsies.

POSTURAL ORTHOSTATIC TACHYCARDIA SYNDROME (POTS)

This syndrome is characterized by symptomatic orthostatic intolerance without OH, accompanied by either an increase in heart rate to >120 beats/min or an increase of 30 beats/min with standing that subsides on sitting or lying down. Women are affected approximately five times more often than men, and most develop the syndrome between the ages of 15 and 50. Presyncopal symptoms (lightheadedness, weakness, blurred vision) combined with symptoms of autonomic overactivity (palpitations, tremulousness, nausea) are common. The pathogenesis is typically multifactorial which frequently confounds the clinical picture. A number of potential causes have been reported, including sympathetic denervation distally in the legs with preserved cardiovascular function or reduced cardiac function due to deconditioning. Hypovolemia, venous pooling, impaired brainstem baroreceptor regulation, or increased sympathetic activity may also play a role. No standardized approach to diagnosis has been established, and therapy typically has included symptomatic relief with a focus on cardiovascular rehabilitation, including a sustained exercise program. Expansion of fluid volume with water, salt, and fludrocortisone can be helpful as an initial intervention. In some patients, low-dose propranolol (20 mg) provides a modest improvement in heart rate control and exercise capacity. If these approaches are inadequate, then midodrine, pyridostigmine, or clonidine can be considered.

INHERITED DISORDERS

Five hereditary sensory and autonomic neuropathies (HSANs) exist, designated HSAN I–V. The most important autonomic variants are HSAN I and HSAN III. HSAN I is dominantly inherited and often presents as a distal small-fiber neuropathy (burning feet syndrome) associated with sensory loss and foot ulcers. The most common responsible gene, on chromosome 9q, is *SPTLC1*. SPTLC is a key enzyme in the regulation of ceramide. Cells from HSAN I patients with the mutation produce higher-than-normal levels of glucosyl ceramide, perhaps triggering apoptosis. HSAN III (Riley-Day syndrome; familial dysautonomia) is an autosomal recessive disorder of Ashkenazi Jewish children and adults and is much less prevalent than HSAN I. Decreased tearing, hyperhidrosis, reduced sensitivity to pain, areflexia, absent fungiform papillae on the tongue, and labile BP may be present. Individuals with HSAN III have afferent, but not efferent, baroreflex failure that causes the classic episodic abdominal crises and blood pressure surges in response to emotional stimuli. Pathologic examination of nerves reveals a loss of sympathetic, parasympathetic, and sensory neurons. The defective gene, *IKBKAP*, prevents normal transcription of important molecules in neural development.

PRIMARY HYPERHIDROSIS

This syndrome presents with excess sweating of the palms of the hands and soles of the feet beginning in childhood or early adulthood. The condition tends to improve with age. The disorder affects 0.6–1.0% of the population. The etiology is unclear, but there may be a genetic component because 25% of patients have a positive family history. The condition can be socially embarrassing (e.g., shaking hands) or even disabling (e.g., inability to write without soiling the paper). Topical

antiperspirants are occasionally helpful. More useful are potent anticholinergic drugs such as glycopyrrolate 1–2 mg PO tid or oxybutynin 5 mg po bid. T2 ganglionectomy or sympathectomy is successful in >90% of patients with palmar hyperhidrosis. The advent of endoscopic transaxillary T2 sympathectomy has lowered the complication rate of the procedure. The most common postoperative complication is compensatory hyperhidrosis, which improves spontaneously over months. Other potential complications include recurrent hyperhidrosis (16%), Horner's syndrome (<2%), gustatory sweating, wound infection, hemothorax, and intercostal neuralgia. Local injection of botulinum toxin has also been used to block cholinergic, postganglionic sympathetic fibers to sweat glands. This approach is effective but limited by the need for repetitive injections (the effect usually lasts 4 months before waning).

ACUTE SYMPATHETIC OVERACTIVITY SYNDROMES

An *autonomic storm* is an acute state of sustained sympathetic surge that results in variable combinations of alterations in BP and heart rate, body temperature, respiration, and sweating. Causes of autonomic storm include brain and spinal cord injury, toxins and drugs, autonomic neuropathy, and chemodectomas (e.g., pheochromocytoma). Brain injury is the most common cause of autonomic storm and typically follows severe head trauma and postresuscitation anoxic-ischemic brain injury. Autonomic storm can also occur with other acute intracranial lesions such as hemorrhage, cerebral infarction, rapidly expanding tumors, subarachnoid hemorrhage, hydrocephalus, or (less commonly) an acute spinal cord lesion. The most consistent setting is that of an acute intracranial catastrophe of sufficient size and rapidity to produce a massive catecholaminergic surge. The surge can cause seizures, neurogenic pulmonary edema, and myocardial injury. Manifestations include fever, tachycardia, hypertension, tachypnea, hyperhidrosis, pupillary dilatation, and flushing. Lesions of the afferent limb of the baroreflex can result in milder recurrent autonomic storms; these can be associated with tumors or follow neck irradiation or surgery that damages the vagus and glossopharyngeal nerves.

Drugs and toxins may also be responsible, including sympathomimetics such as phenylpropanolamine, cocaine, amphetamines, and tricyclic antidepressants; tetanus; and, less often, botulinum toxin. Cocaine, including "crack," can cause a hypertensive state with CNS hyperstimulation. An overdose of tricyclic antidepressants, such as amitriptyline, can cause flushing, hypertension, tachycardia, fever, mydriasis, anhidrosis, and a toxic psychosis. The hyperadrenergic state associated with Guillain-Barré syndrome can produce a moderate autonomic storm. Pheochromocytoma presents with a paroxysmal or sustained hyperadrenergic state, headache, hyperhidrosis, palpitations, anxiety, tremulousness, and hypertension.

Neuroleptic malignant syndrome refers to a syndrome of muscle rigidity, hyperthermia, and hypertension in patients treated with neuroleptic agents (including lower potency and atypical antipsychotic agents, and even antiemetic drugs such as metoclopramide, promethazine) **(Chap. 429)**. Management of autonomic storm includes ruling out other causes of autonomic instability, including malignant hyperthermia, porphyria, and seizures. Sepsis and encephalitis need to be excluded with appropriate studies. An electroencephalogram (EEG) should be done to search for seizure activity; MRI of the brain and spine is often necessary. The patient should be managed in an intensive care unit. Management with morphine sulphate (10 mg every 4 h) and labetalol (100–200 mg twice daily) may be helpful. Supportive treatment may need to be maintained for several weeks. For chronic and milder autonomic storm, propranolol and/or clonidine can be effective.

MISCELLANEOUS

Other conditions associated with autonomic failure include infections, malignancy, and poisoning (organophosphates). Disorders of the hypothalamus can affect autonomic function and produce abnormalities in temperature control, satiety, sexual function, and circadian rhythms **(Chap. 373)**.

COMPLEX REGIONAL PAIN SYNDROMES (CRPS)

The failure to identify a primary role of the ANS in the pathogenesis of these disorders has resulted in a change of nomenclature. The terms CRPS types I and II are now used in place of reflex sympathetic dystrophy (RSD) and causalgia.

CRPS type I is a regional pain syndrome that often develops after tissue injury and most commonly affects one limb. Examples of associated injury include minor shoulder or limb trauma, fractures, myocardial infarction, or stroke. *Allodynia* (the perception of a nonpainful stimulus as painful), *hyperpathia* (an exaggerated pain response to a painful stimulus), and spontaneous pain occur. The symptoms are unrelated to the severity of the initial trauma and are not confined to the distribution of a single peripheral nerve. CRPS type II is a regional pain syndrome that develops after injury to a specific peripheral nerve, often a major nerve trunk. Spontaneous pain initially develops within the territory of the affected nerve but eventually may spread outside the nerve distribution. Although CRPS type I (RSD) has been classically divided into three clinical phases, there is little evidence that CRPS "progresses" from one stage to another. Currently, the Budapest consensus criteria for clinical diagnosis of CRPS delete staging and require at least three symptoms and two signs in the following four categories: (1) sensory, (2) vasomotor, (3) sudomotor/edema, and (4) motor/trophic. Pain (usually burning or electrical in quality) is the primary clinical feature of CRPS. Limb pain syndromes that do not meet these criteria are best classified as "limb pain—not otherwise specified." In CRPS, localized sweating (increased resting sweat output) and changes in blood flow may produce temperature differences between affected and unaffected limbs.

The natural history of typical CRPS may be more benign and more variable than previously recognized. A variety of surgical and medical treatments have been developed, with conflicting reports of efficacy. Clinical trials suggest that early mobilization with physical therapy or a brief course of glucocorticoids may be helpful for early CRPS type I or II. Chronic glucocorticoid treatment is not recommended. Current treatment paradigms are multidisciplinary with a focus on early mobilization, physical therapy, pain management, patient education, and psychological support.

TREATMENT

Autonomic Failure

Management of autonomic failure is aimed at specific treatment of the cause and alleviation of symptoms. Of particular importance is the removal of drugs or amelioration of underlying conditions that cause or aggravate the autonomic symptoms, especially in the elderly. For example, OH can be caused or aggravated by antihypertensive agents, antidepressants, levodopa or dopaminergic agonists, ethanol, opioids, insulin, and barbiturates. A summary of drugs that can cause impotence, urinary retention, or diaphoresis by class and putative mechanism is shown in Table 432-6.

PATIENT EDUCATION

Only a minority of patients with OH require drug treatment. All patients should be taught the mechanisms of postural normotension (volume status, resistance and capacitance bed, autoregulation) and the nature of orthostatic stressors (time of day and the influence of meals, heat, standing, and exercise). Patients should learn to recognize orthostatic symptoms early (especially subtle cognitive symptoms, weakness, and fatigue) and to modify or avoid activities that provoke episodes. Other helpful measures may include keeping a BP log and dietary education (salt/fluids). Learning physical counter-maneuvers that reduce standing OH and practicing postural and resistance training and cardiovascular reconditioning are helpful measures.

SYMPTOMATIC TREATMENT

Nonpharmacologic approaches are summarized in **Table 432-9**. Adequate intake of salt and fluids to produce a voiding volume between 1.5 and 2.5 L of urine (containing >170 meq/L of Na^+) each 24 h is essential. Sleeping with the head of the bed elevated will minimize the effects of supine nocturnal hypertension. Prolonged recumbency should be avoided when possible. Patients are advised

TABLE 432-9 Initial Treatment of Orthostatic Hypotension (OH)

Patient education: mechanisms and stressors of OH
High-salt diet (10–20 g/d)
High-fluid intake (2 L/d)
Elevate head of bed 10 cm (4 in.) to minimize supine hypertension
Maintain postural stimuli
Learn physical counter-maneuvers
Compression garments
Correct anemia

to sit with legs dangling over the edge of the bed for several minutes before attempting to stand in the morning; other postural stressors should be similarly approached in a gradual manner. One maneuver that can reduce OH is leg-crossing with maintained contraction of leg muscles for 30 s; this compresses leg veins and increases systemic resistance. Compressive garments, such as compression stockings or abdominal binders, are helpful on occasion but are uncomfortable for many patients. For transient worsening of OH, drinking two 250-mL (8-oz) glasses of water within 5 min can raise standing BP 20–30 mmHg for about 2 h, beginning ~5 min after the fluid load. The patient can increase intake of salt and fluids (bouillon treatment), increase use of physical counter-maneuvers (elevate the legs when supine), or temporarily resort to a full-body stocking (compression pressure 30–40 mmHg).

Anemia can be corrected with erythropoietin, administered subcutaneously at doses of 25–75 U/kg three times per week. The hematocrit increases after 2–6 weeks. A weekly maintenance dose is usually necessary. However, the increased intravascular volume that accompanies the rise in hematocrit can exacerbate supine hypertension and requires monitoring.

If these measures are not sufficient, additional pharmacologic treatment may be necessary. Midodrine, a directly acting α_1-agonist that does not cross the blood-brain barrier, is effective. It has a duration of action of 2–4 h. The usual dose is 5–10 mg orally tid, but some patients respond best to a decremental dose (e.g., 15 mg on awakening, 10 mg at noon, and 5 mg in the afternoon). Midodrine should not be taken after 6:00 P.M. Side effects include pruritus, uncomfortable piloerection, and supine hypertension, especially at higher doses. Droxidopa (Northera) for treatment of neurogenic OH associated with PAF, PD, or MSA is effective in decreasing symptoms of OH. Pyridostigmine appears to improve OH without aggravating supine hypertension by enhancing ganglionic transmission (maximal when orthostatic, minimal when supine), but with only modest clinical effects on BP. Fludrocortisone will reduce OH but aggravates supine hypertension. At doses between 0.1 mg/d and 0.3 mg bid orally, it enhances renal sodium conservation and increases the sensitivity of arterioles to NE. Susceptible patients may develop fluid overload, congestive heart failure, supine hypertension, or hypokalemia. Potassium supplements are often necessary with chronic administration of fludrocortisone. Sustained elevations of supine BP >180/110 mmHg should be avoided. Supine hypertension (>180/110 mmHg) can be self-treated by avoiding the supine position (e.g., sleeping in a recumbent chair or elevating the head of the bed) and reducing fludrocortisone. If these simple measures are not adequate, drugs to be considered include oral hydralazine (25 mg qhs), oral nifedipine (Procardia; 10 mg qhs), or a nitroglycerin patch.

A promising drug combination (atomoxetine and yohimbine) has been studied for use in human subjects with severe OH not responsive to other agents, as can occur is some patients with diabetes and severe autonomic neuropathy not responsive to other medications. The atomoxetine blocks the NE reuptake transporter, and yohimbine blocks α_2 receptors that mediate the sympathetic feedback loop for downregulation of BP in response to atomoxetine. The result is a dramatic increase in BP and standing tolerance. Yohimbine is no longer produced commercially and must be obtained from a compounding pharmacy. This combination is not FDA approved for this purpose.

Postprandial OH may respond to several measures. Frequent, small, low-carbohydrate meals may diminish splanchnic shunting of blood after meals and reduce postprandial OH. Prostaglandin inhibitors (ibuprofen or indomethacin) taken with meals or midodrine (10 mg with the meal) can be helpful. The somatostatin analogue octreotide can be useful in the treatment of postprandial syncope by inhibiting the release of GI peptides that have vasodilator and hypotensive effects. The subcutaneous dose ranges from 25 µg bid to 200 µg tid.

ACKNOWLEDGMENT

The authors want to thank Phillip A. Low for his contributions to previous editions of this chapter.

FURTHER READING

GIBBONS CH et al: The recommendations of a consensus panel for the screening, diagnosis, and treatment of neurogenic orthostatic hypotension and associated supine hypertension. J Neurol 264:1587, 2017.

KAUFMANN H et al: Natural history of pure autonomic failure: A United States prospective cohort. Ann Neurol 81:287, 2017.

MCKEON A, BENARROCH EE: Autoimmune autonomic disorders. Handb Clin Neurol 133:405, 2016.

SINGER W et al: Pure autonomic failure: Predictors of conversion to clinical CNS involvement. Neurology 88:1129, 2017.

VINIK AI, ERBAS T: Diabetic autonomic neuropathy. Handb Clin Neurol 117:279, 2013.

433 Trigeminal Neuralgia, Bell's Palsy, and Other Cranial Nerve Disorders

M. Flint Beal, Stephen L. Hauser

Symptoms and signs of cranial nerve pathology are common in internal medicine. They often develop in the context of a widespread neurologic disturbance, and in such situations, cranial nerve involvement may represent the initial manifestation of the illness. In other disorders, involvement is largely restricted to one or several cranial nerves; these distinctive disorders are reviewed in this chapter. **Disorders of ocular movement are discussed in Chap. 28, disorders of hearing in Chap. 30, and vertigo and disorders of vestibular function in Chap. 19.**

FACIAL PAIN OR NUMBNESS

ANATOMIC CONSIDERATIONS

The trigeminal (fifth cranial) nerve supplies sensation to the skin of the face and anterior half of the head **(Fig. 433-1)**. The motor part innervates the muscles involved in chewing (including masseters and pterygoids) as well as the tensor tympani of the middle ear (hearing especially for high-pitched tones). It is the largest of the cranial nerves. It exits in the lateral midpons and traverses the middle cranial fossa to the semilunar (gasserian, trigeminal) ganglion in Meckel's cave, where the nerve divides into three divisions (ophthalmic [V1], maxillary [V2], and mandibular [V3]). V1 and V2 traverse the cavernous sinus to exit in the superior orbital fissure and foramen rotundum, located above and below the eye socket respectively; V3 exits through the foramen ovale. The trigeminal nerve is predominantly sensory, and motor innervation is exclusively carried in V3. The cornea is primarily innervated by V1, although an inferior crescent may be V2. Upon entering the pons, pain and temperature fibers descend ipsilaterally to the upper cervical spinal cord as the spinal tract of V, before synapsing with the spinal nucleus of V; this accounts for the facial numbness that can occur with spinal cord

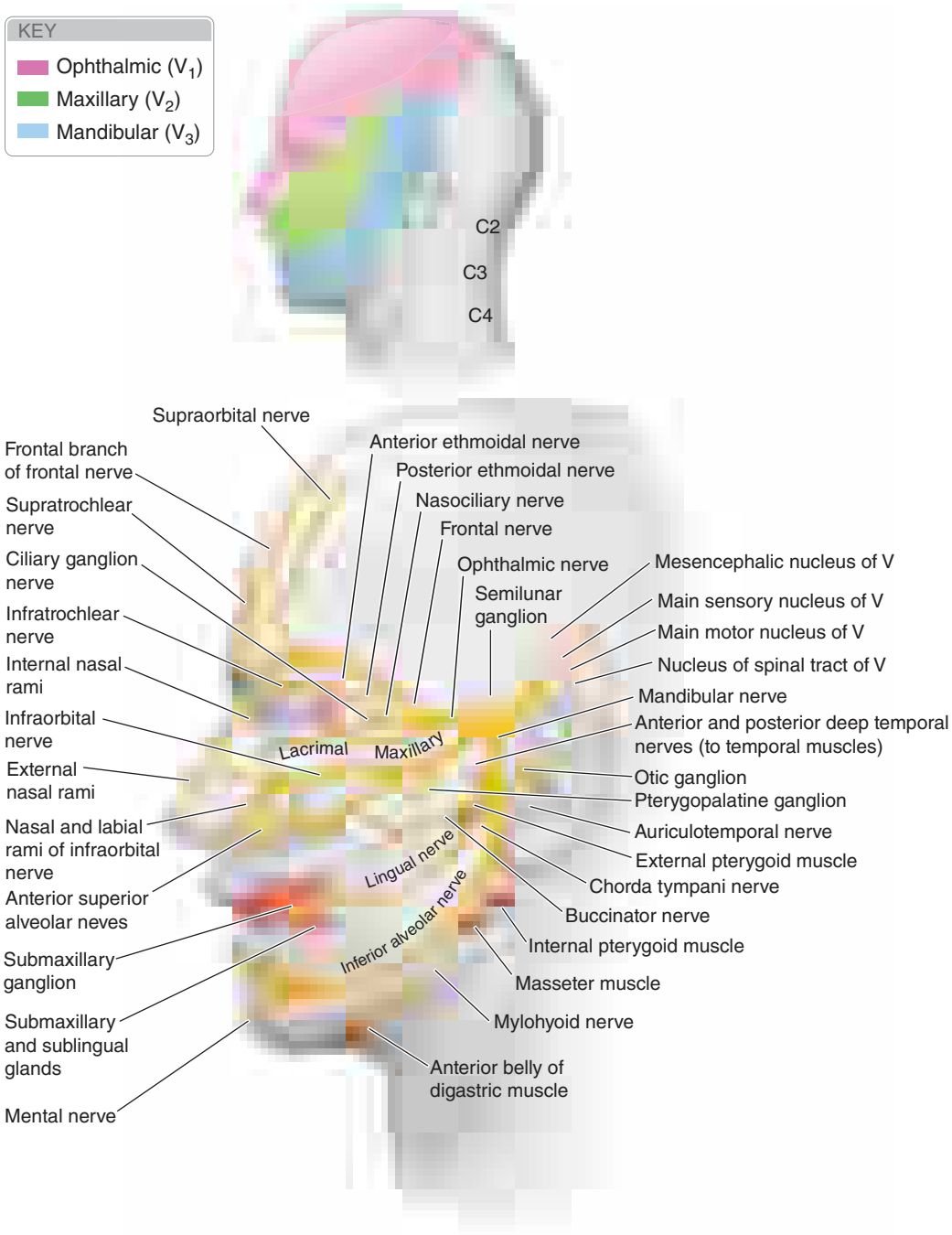


FIGURE 433-1 The trigeminal nerve and its branches and sensory distribution on the face. The three major sensory divisions of the trigeminal nerve consist of the ophthalmic, maxillary, and mandibular nerves. *(Adapted from Waxman SG: Clinical Neuroanatomy, 26th ed. http://www.accessmedicine.com. Copyright © The McGraw-Hill Companies, Inc. All rights reserved.)*

lesions above C2. In the brainstem, the spinal tract of V is also located adjacent to crossed ascending fibers of the spinothalamic tract, producing a "crossed" sensory loss for pain and temperature (ipsilateral face, contralateral arm/trunk/leg) with lesions of the lateral lower brainstem. CN V is also ensheathed by oligodendrocyte-derived, rather than Schwann cell–derived, myelin for up to 7 mm after it leaves the brainstem, unlike just a few millimeters for other cranial and spinal nerves; this may explain the high frequency of trigeminal neuralgia in multiple sclerosis (MS) **(Chap. 436)**, a disorder of oligodendrocyte myelin.

TRIGEMINAL NEURALGIA (TIC DOULOUREUX)

Clinical Manifestations Trigeminal neuralgia is characterized by excruciating paroxysms of pain in the lips, gums, cheek, or chin and, very rarely, in the distribution of the ophthalmic division of the fifth nerve. The pain seldom lasts more than a few seconds or a minute or two but may be so intense that the patient winces, hence the term tic. The paroxysms, experienced as single jabs or clusters, tend to recur frequently, both day and night, for several weeks at a time. They may occur spontaneously or with movements of affected areas evoked by speaking, chewing, or smiling. Another characteristic feature is the presence of trigger zones, typically on the face, lips, or tongue, that provoke attacks; patients may report that tactile stimuli—e.g., washing the face, brushing the teeth, or exposure to a draft of air—generate excruciating pain. An essential feature of trigeminal neuralgia is that objective signs of sensory loss cannot be demonstrated on examination.

Trigeminal neuralgia is relatively common, with an estimated annual incidence of 4–8 per 100,000 individuals. Middle-aged and elderly persons are affected primarily, and ~60% of cases occur in women. Onset is typically sudden, and bouts tend to persist for weeks or months before remitting spontaneously. Remissions may be long-lasting, but in most patients, the disorder ultimately recurs.

Pathophysiology Symptoms result from ectopic generation of action potentials in pain-sensitive afferent fibers of the fifth cranial nerve root just before it enters the lateral surface of the pons. Compression or other pathology in the nerve leads to demyelination of large myelinated fibers that do not themselves carry pain sensation but become hyperexcitable and electrically coupled with smaller unmyelinated or poorly myelinated pain fibers in close proximity; this may explain why tactile stimuli, conveyed via the large myelinated fibers, can stimulate paroxysms of pain. Compression of the trigeminal nerve root by a blood vessel, most often the superior cerebellar artery or on occasion a tortuous vein, is now believed to be the source of trigeminal neuralgia in most patients. In cases of vascular compression, age-related brain sagging and increased vascular thickness and tortuosity may explain the prevalence of trigeminal neuralgia in later life.

Differential Diagnosis Trigeminal neuralgia must be distinguished from other causes of face and head pain (Chap. 13) and from pain arising from diseases of the jaw, teeth, or sinuses. Pain from migraine or cluster headache tends to be deep-seated and steady, unlike the superficial stabbing quality of trigeminal neuralgia; rarely, cluster headache is associated with trigeminal neuralgia, a syndrome known as *cluster-tic*. Other rare headaches include short-lasting unilateral headache attacks with conjunctival injection and tearing (SUNCT; Chap. 422). In temporal arteritis, superficial facial pain is present but is not typically shock-like, the patient frequently complains of myalgias and other systemic symptoms, and an elevated erythrocyte sedimentation rate (ESR) or C-reactive protein (CRP) is usually present (Chap. 356). When trigeminal neuralgia develops in a young adult or is bilateral, MS is a key consideration, and in such cases, the cause is a demyelinating plaque near the root entry zone of the fifth nerve in the pons; often, evidence of facial sensory loss can be found on careful examination. Cases that are secondary to mass lesions—such as aneurysms, neurofibromas, acoustic schwannomas, or meningiomas—usually produce objective signs of sensory loss in the trigeminal nerve distribution (trigeminal neuropathy, see below).

Laboratory Evaluation An ESR or CRP is indicated if temporal arteritis is suspected. In typical cases of trigeminal neuralgia, neuroimaging studies are usually unnecessary but may be valuable if MS is a consideration or in assessing overlying vascular lesions in order to plan for decompression surgery.

TREATMENT

Trigeminal Neuralgia

Drug therapy with carbamazepine is effective in ~50–75% of patients. Carbamazepine should be started as a single daily dose of 100 mg taken with food and increased gradually (by 100 mg daily in divided doses every 1–2 days) until substantial (>50%) pain relief is achieved. Most patients require a maintenance dose of 200 mg qid. Doses >1200 mg daily provide no additional benefit. Dizziness, imbalance, sedation, and rare cases of agranulocytosis are the most important side effects of carbamazepine. If treatment is effective, it is usually continued for 1 month and then tapered as tolerated. Oxcarbazepine (300–1200 mg bid) is an alternative to carbamazepine that has less bone marrow toxicity and probably is equally efficacious. If these agents are not well tolerated or are ineffective, lamotrigine, 400 mg daily, and phenytoin, 300–400 mg daily, are other options. Baclofen may also be tried, either alone or in combination with an anticonvulsant. The initial dose is 5–10 mg tid, gradually increasing as needed to 20 mg qid.

If drug treatment fails, surgical therapy should be offered. The most widely used method is currently microvascular decompression to relieve pressure on the trigeminal nerve as it exits the pons. This procedure requires a suboccipital craniotomy. This procedure appears to have a >70% efficacy rate and a low rate of pain recurrence in responders; the response is better for classic tic-like symptoms than for nonlancinating facial pains. In a small number of cases, there is perioperative damage to the eighth or seventh cranial nerves or to the cerebellum or a postoperative cerebrospinal fluid leak syndrome. High-resolution magnetic resonance angiography is useful preoperatively to visualize the relationships between the fifth cranial nerve root and nearby blood vessels.

Gamma knife radiosurgery of the trigeminal nerve root is also used for treatment and results in complete pain relief, sometimes delayed in onset, in more than two-thirds of patients and a low risk of persistent facial numbness; the response is sometimes long-lasting, but recurrent pain develops over 2–3 years in half of patients. Compared with surgical decompression, gamma knife surgery appears to be somewhat less effective but has few serious complications.

Another procedure, radiofrequency thermal rhizotomy, creates a heat lesion of the trigeminal ganglion or nerve. It is used less often now than in the past. Short-term relief is experienced by >95% of patients; however, long-term studies indicate that pain recurs in up to one-third of treated patients. Postoperatively, partial numbness of the face is common, masseter (jaw) weakness may occur especially following bilateral procedures, and corneal denervation with secondary keratitis can follow rhizotomy for first-division trigeminal neuralgia.

TRIGEMINAL NEUROPATHY

A variety of diseases can affect the trigeminal nerve (Table 433-1). Most present with sensory loss on the face or with weakness of the jaw muscles. Deviation of the jaw on opening indicates weakness of the pterygoids on the side to which the jaw deviates. Some cases are due to Sjögren's syndrome or a collagen-vascular disease such as systemic lupus erythematosus, scleroderma, or mixed connective tissue disease. Among infectious causes, herpes zoster (acute or postherpetic) and leprosy should be considered. Tumors of the middle cranial fossa (meningiomas), of the trigeminal nerve (schwannomas), or of the base of the skull (metastatic tumors) may cause a combination of motor and sensory signs. Lesions in the cavernous sinus can affect the first and second divisions of the trigeminal nerve, and lesions of the superior orbital fissure can affect the first (ophthalmic) division; the accompanying corneal anesthesia increases the risk of ulceration (neurokeratitis).

TABLE 433-1 Trigeminal Nerve Disorders

Nuclear (Brainstem) Lesions
Multiple sclerosis
Stroke
Syringobulbia
Glioma
Lymphoma
Preganglionic Lesions
Acoustic neuroma
Meningioma
Metastasis
Chronic meningitis
Cavernous carotid aneurysm
Gasserian Ganglion Lesions
Trigeminal neuroma
Herpes zoster
Infection (spread from otitis media or mastoiditis)
Peripheral Nerve Lesions
Nasopharyngeal carcinoma
Trauma
Guillain-Barré syndrome
Sjögren's syndrome
Collagen-vascular diseases
Sarcoidosis
Leprosy
Drugs (stilbamidine, trichloroethylene)
Idiopathic trigeminal neuropathy

Isolated sensory loss over the chin (mental neuropathy) can be the only manifestation of systemic malignancy. Rarely, an idiopathic form of trigeminal neuropathy is observed. It is characterized by numbness and paresthesias, sometimes bilaterally, with loss of sensation in the territory of the trigeminal nerve but without weakness of the jaw. Gradual recovery is the rule. Tonic spasm of the masticatory muscles, known as trismus, is symptomatic of tetanus **(Chap. 147)** or may occur in patients treated with phenothiazines.

FACIAL WEAKNESS

ANATOMIC CONSIDERATIONS

(Fig. 433-2) The seventh cranial nerve supplies all the muscles concerned with facial expression. The sensory component is small (the nervus intermedius); it conveys taste sensation from the anterior two-thirds of the tongue and probably cutaneous impulses from the anterior wall of the external auditory canal. The motor nucleus of the seventh nerve lies anterior and lateral to the abducens nucleus. After leaving the pons, the seventh nerve enters the internal auditory meatus with the acoustic nerve. The nerve continues its course in its own bony channel, the facial canal, and exits from the skull via the stylomastoid foramen. It then passes through the parotid gland and subdivides to supply the facial muscles.

A complete interruption of the facial nerve at the stylomastoid foramen paralyzes all muscles of facial expression. The corner of the mouth droops, the creases and skinfolds are effaced, the forehead is unfurrowed, and the eyelids will not close. Upon attempted closure of the lids, the eye on the paralyzed side rolls upward (Bell's phenomenon). The lower lid sags and falls away from the conjunctiva, permitting tears to spill over the cheek. Food collects between the teeth and lips, and saliva may dribble from the corner of the mouth. The patient complains of a heaviness or numbness in the face, but sensory loss is rarely demonstrable and taste is intact.

If the lesion is in the middle-ear portion, taste is lost over the anterior two-thirds of the tongue on the same side. If the nerve to the stapedius is interrupted, there is hyperacusis (sensitivity to loud sounds). Lesions in the internal auditory meatus may affect the adjacent auditory and vestibular nerves, causing deafness, tinnitus, or dizziness. Intrapontine lesions that paralyze the face usually affect the abducens nucleus as well, and often the corticospinal and sensory tracts.

If the peripheral facial paralysis has existed for some time and recovery of motor function is incomplete, a continuous diffuse contraction of facial muscles may appear. The palpebral fissure becomes narrowed, and the nasolabial fold deepens. Attempts to move one group of facial muscles may result in contraction of all (associated movements, or synkinesis). Facial spasms, initiated by movements of the face, may develop (hemifacial spasm). Anomalous regeneration of seventh nerve fibers may result in other troublesome phenomena. If fibers originally connected with the orbicularis oculi come to innervate the orbicularis oris, closure of the lids may cause a retraction of the mouth, or if fibers originally connected with muscles of the face later innervate the lacrimal gland, anomalous tearing ("crocodile tears") may occur with any activity of the facial muscles, such as eating. Another facial synkinesia is triggered by jaw opening, causing closure of the eyelids on the side of the facial palsy (jaw-winking).

BELL'S PALSY

The most common form of facial paralysis is Bell's palsy. The annual incidence of this idiopathic disorder is ~25 per 100,000 annually, or about 1 in 60 persons in a lifetime. Risk factors include pregnancy and diabetes mellitus.

Clinical Manifestations The onset of Bell's palsy is fairly abrupt, with maximal weakness being attained by 48 h as a general rule. Pain behind the ear may precede the paralysis for a day or two. Taste sensation may be lost unilaterally, and hyperacusis may be present. In some cases, there is mild cerebrospinal fluid lymphocytosis. Magnetic resonance imaging (MRI) may reveal swelling and uniform enhancement of the geniculate ganglion and facial nerve and, in some cases, entrapment of the swollen nerve in the temporal bone. Approximately 80% of patients recover within a few weeks or months. Electromyography may be of some prognostic value; evidence of denervation after 10 days indicates there has been axonal degeneration, that there will be a long delay (3 months as a rule) before regeneration occurs, and that it may be incomplete. The presence of incomplete paralysis in the first week is the most favorable prognostic sign. Recurrences are reported in ~7% of cases.

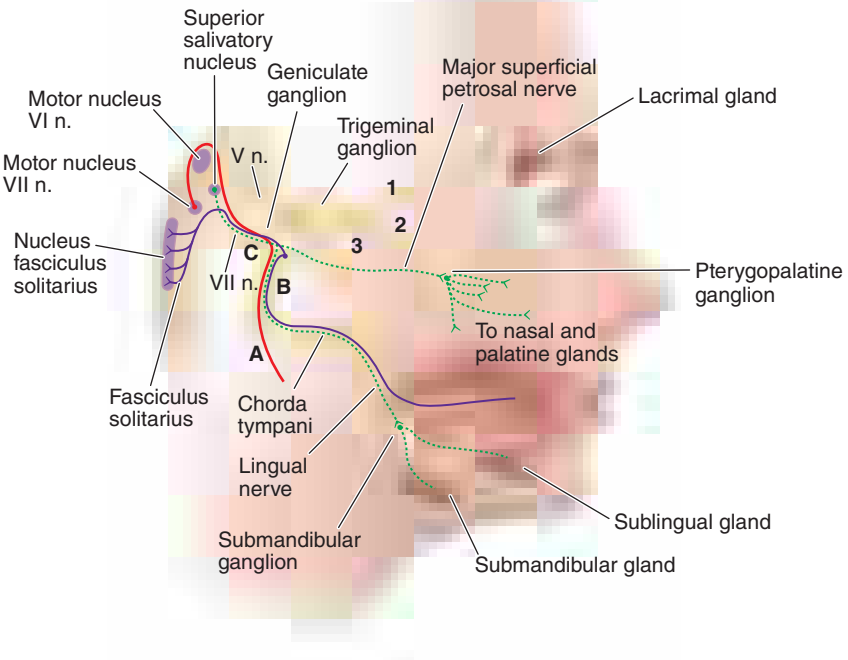


FIGURE 433-2 The facial nerve. A, B, and C denote lesions of the facial nerve at the stylomastoid foramen, distal and proximal to the geniculate ganglion, respectively. *Green lines* indicate the parasympathetic fibers, *red line* indicates motor fibers, and *purple lines* indicate visceral afferent fibers (taste). *(Adapted from MB Carpenter: Core Text of Neuroanatomy, 2nd ed. Williams & Wilkins, 1978.)*

Pathophysiology In acute Bell's palsy, there is inflammation of the facial nerve with mononuclear cells, consistent with an infectious or immune cause. Herpes simplex virus (HSV) type 1 DNA was frequently detected in endoneurial fluid and posterior auricular muscle, suggesting that a reactivation of this virus in the geniculate ganglion may be responsible for most cases. Reactivation of varicella-zoster virus is associated with Bell's palsy in up to one-third of cases and may represent the second most frequent cause. A variety of other viruses have also been implicated less commonly. An increased incidence of Bell's palsy was also reported among recipients of inactivated intranasal influenza vaccine, and it was hypothesized that this could have resulted from the *Escherichia coli* enterotoxin used as adjuvant or reactivation of latent virus.

Differential Diagnosis There are many other causes of acute facial palsy that must be considered in the differential diagnosis of Bell's palsy. Lyme disease can cause unilateral or bilateral facial palsies; in endemic areas, ≥10% of cases of facial palsy are likely due to infection with *Borrelia burgdorferi* **(Chap. 181)**. The Ramsay Hunt

syndrome, caused by reactivation of herpes zoster in the geniculate ganglion, consists of a severe facial palsy associated with a vesicular eruption in the external auditory canal and sometimes in the pharynx and other parts of the cranial integument; often the eighth cranial nerve is affected as well. Facial palsy that is often bilateral occurs in sarcoidosis **(Chap. 360)** and in Guillain-Barré syndrome **(Chap. 439)**. Leprosy frequently involves the facial nerve, and facial neuropathy may also occur in diabetes mellitus, connective tissue diseases including Sjögren's syndrome, and amyloidosis. The rare Melkersson-Rosenthal syndrome consists of recurrent facial paralysis; recurrent—and eventually permanent—facial (particularly labial) edema; and, less constantly, plication of the tongue. Its cause is unknown. Acoustic neuromas frequently involve the facial nerve by local compression. Infarcts, demyelinating lesions of MS, and tumors are the common pontine lesions that interrupt the facial nerve fibers; other signs of brainstem involvement are usually present. Tumors that invade the temporal bone (carotid body, cholesteatoma, dermoid) may produce a facial palsy, but the onset is insidious and the course progressive.

All these forms of nuclear or peripheral facial palsy must be distinguished from the supranuclear type. In the latter, the frontalis and orbicularis oculi muscles of the forehead are involved less than those of the lower part of the face, since the upper facial muscles are innervated by corticobulbar pathways from both motor cortices, whereas the lower facial muscles are innervated only by the opposite hemisphere. In supranuclear lesions, there may be a dissociation of emotional and voluntary facial movements, and often some degree of paralysis of the arm and leg or an aphasia (in dominant hemisphere lesions) is present.

Laboratory Evaluation The diagnosis of Bell's palsy can usually be made clinically in patients with (1) a typical presentation, (2) no risk factors or preexisting symptoms for other causes of facial paralysis, (3) absence of cutaneous lesions of herpes zoster in the external ear canal, and (4) a normal neurologic examination with the exception of the facial nerve. Particular attention to the eighth cranial nerve, which courses near to the facial nerve in the pontomedullary junction and in the temporal bone, and to other cranial nerves is essential. In atypical or uncertain cases, an ESR or CRP, testing for diabetes mellitus, a Lyme titer, angiotensin-converting enzyme and chest imaging studies for possible sarcoidosis, a lumbar puncture for possible Guillain-Barré syndrome, or MRI scanning may be indicated. MRI often shows swelling and enhancement of the facial nerve in idiopathic Bell's palsy **(Fig. 433-3)**.

TREATMENT

Bell's Palsy

Symptomatic measures include (1) the use of paper tape to depress the upper eyelid during sleep and prevent corneal drying, (2) artificial tears; and (3) massage of the weakened muscles. A course of glucocorticoids, given as prednisone 60–80 mg daily during the first 5 days and then tapered over the next 5 days, modestly shortens the recovery period and improves the functional outcome. Although large and well-controlled randomized trials found no added benefit of the antiviral agents valacyclovir (1000 mg daily for 5–7 days) or acyclovir (400 mg five times daily for 10 days) compared to glucocorticoids alone, some earlier data suggested that combination therapy with prednisone plus valacyclovir might be marginally better than prednisone alone, especially in patients with severe clinical presentations. For patients with permanent paralysis from Bell's palsy, a number of cosmetic surgical procedures have been used to restore a relatively symmetric appearance to the face.

OTHER MOTOR DISORDERS OF THE FACE

Hemifacial spasm consists of painless irregular involuntary contractions on one side of the face. Most cases appear related to vascular compression of the exiting facial nerve in the pons. Other cases develop as a sequela to Bell's palsy or are secondary to compression and/or demyelination of the nerve by tumor, infection, or MS. Mild cases can be treated with carbamazepine, gabapentin, or, if these drugs fail, baclofen. Local injections of botulinum toxin into affected muscles can relieve spasms for 3–4 months, and the injections can be repeated. Refractory cases due to vascular compression usually respond to surgical decompression of the facial nerve. Blepharospasm is an involuntary recurrent spasm of both eyelids that usually occurs in elderly persons as an isolated phenomenon or with varying degrees of spasm of other facial muscles. Severe, persistent cases of blepharospasm can be treated by local injection of botulinum toxin into the orbicularis oculi. Facial myokymia refers to a fine rippling activity of the facial muscles; it may be caused by MS or follow Guillain-Barré syndrome **(Chap. 439)**.

Facial hemiatrophy occurs mainly in women and is characterized by a disappearance of fat in the dermal and subcutaneous tissues on one side of the face. It usually begins in adolescence or the early adult years and is slowly progressive. In its advanced form, the affected side of the face is gaunt, and the skin is thin, wrinkled, and brown. The facial hair may turn white and fall out, and the sebaceous glands become atrophic. Bilateral involvement may occur. A limited form of systemic sclerosis (scleroderma) may be the cause of some cases. Treatment is cosmetic, consisting of transplantation of skin and subcutaneous fat.

OTHER CRANIAL NERVE DISORDERS

GLOSSOPHARYNGEAL NEURALGIA

This form of neuralgia involves the ninth (glossopharyngeal) and sometimes portions of the tenth (vagus) cranial nerves. It resembles trigeminal neuralgia in many respects but is much less common. The pain is intense and paroxysmal; it originates on one side of the throat,

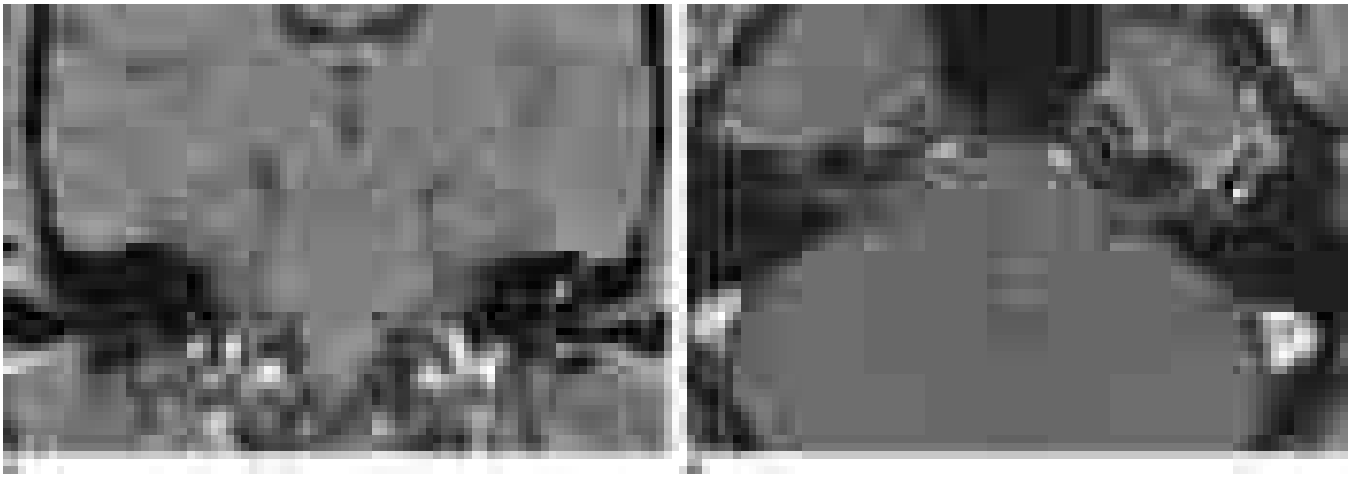


FIGURE 433-3 Axial and coronal T1-weighted images after gadolinium with fat suppression demonstrate diffuse smooth linear enhancement of the left facial nerve, involving the genu, tympanic, and mastoid segments within the temporal bone (*arrows*), without evidence of mass lesion. Although highly suggestive of Bell's palsy, similar findings may be seen with other etiologies such as Lyme disease, sarcoidosis, and perineural malignant spread.

TABLE 433-2 Cranial Nerve Syndromes

SITE	CRANIAL NERVES	USUAL CAUSE
Sphenoid fissure (superior orbital)	III, IV, first division V, VI	Invasive tumors of sphenoid bone; aneurysms
Lateral wall of cavernous sinus	III, IV, first division V, VI, often with proptosis	Infection, thrombosis, aneurysm, or fistula of cavernous sinus; invasive tumors from sinuses and sella turcica; benign granuloma responsive to glucocorticoids
Retrosphenoid space	II, III, IV, V, VI	Large tumors of middle cranial fossa
Apex of petrous bone	V, VI	Petrositis; tumors of petrous bone
Internal auditory meatus	VII, VIII	Tumors of petrous bone (dermoids, etc.); infectious processes; acoustic neuroma
Pontocerebellar angle	V, VII, VIII, and sometimes IX	Acoustic neuroma; meningioma
Jugular foramen	IX, X, XI	Tumors and aneurysms
Posterior laterocondylar space	IX, X, XI, XII	Tumors of parotid gland and carotid body and metastatic tumors
Posterior retroparotid space	IX, X, XI, XII, and Horner's syndrome	Tumors of parotid gland, carotid body, lymph nodes; metastatic tumor; tuberculous adenitis

approximately in the tonsillar fossa. In some cases, the pain is localized in the ear or may radiate from the throat to the ear because of involvement of the tympanic branch of the glossopharyngeal nerve. Spasms of pain may be initiated by swallowing or coughing. There is no demonstrable motor or sensory deficit; the glossopharyngeal nerve supplies taste sensation to the posterior third of the tongue and, together with the vagus nerve, sensation to the posterior pharynx. Cardiac symptoms—bradycardia or asystole, hypotension, and fainting—have been reported. Glossopharyngeal neuralgia can result from vascular compression, MS, or tumors, but many cases are idiopathic. Medical therapy is similar to that for trigeminal neuralgia, and carbamazepine is generally the first choice. If drug therapy is unsuccessful, surgical procedures—including microvascular decompression if vascular compression is evident—or rhizotomy of glossopharyngeal and vagal fibers in the jugular bulb is frequently successful.

Glossopharyngeal neuropathy in conjunction with vagus and accessory nerve palsies may occur with herpes zoster infection or with a tumor or aneurysm in the posterior fossa or in the jugular foramen. Hoarseness due to vocal cord paralysis, some difficulty in swallowing, deviation of the soft palate to the intact side, anesthesia of the posterior wall of the pharynx, and weakness of the upper part of the trapezius and sternocleidomastoid muscles make up the jugular foramen syndrome **(Table 433-2)**.

DYSPHAGIA AND DYSPHONIA

When the intracranial portion of one vagus (tenth cranial) nerve is interrupted, the soft palate droops ipsilaterally and does not rise in phonation. There is loss of the gag reflex on the affected side, as well as of the "curtain movement" of the lateral wall of the pharynx, whereby the faucial pillars move medially as the palate rises in saying "ah." The voice is hoarse and slightly nasal, and the vocal cord lies immobile midway between abduction and adduction. Loss of sensation at the external auditory meatus and the posterior pinna may also be present.

The pharyngeal branches of both vagal nerves may be affected in diphtheria; the voice has a nasal quality, and regurgitation of liquids through the nose occurs during swallowing.

Injury to the vagus nerve in the carotid sheath can also occur with carotid dissection or following endarterectomy. The vagus nerve may be involved at the meningeal level by neoplastic and infectious processes and within the medulla by tumors, vascular lesions (e.g., the lateral medullary syndrome), and motor neuron disease. This nerve may be involved by infection with varicella zoster virus. Polymyositis and dermatomyositis, which cause hoarseness and dysphagia by direct involvement of laryngeal and pharyngeal muscles, may be confused with diseases of the vagus nerves. Dysphagia is also a symptom in some patients with myotonic dystrophy. **Nonneurologic causes of dysphagia are discussed in Chap. 40**.

The recurrent laryngeal nerves, especially the left, are most often damaged as a result of intrathoracic disease. Aneurysm of the aortic arch, an enlarged left atrium, and tumors of the mediastinum and bronchi are much more frequent causes of an isolated vocal cord palsy than are intracranial disorders. However, a substantial number of cases of recurrent laryngeal palsy remain idiopathic.

When confronted with a case of laryngeal palsy, the physician must attempt to determine the site of the lesion. If it is intramedullary, there are usually other signs, such as ipsilateral cerebellar dysfunction, loss of pain and temperature sensation over the ipsilateral face and contralateral arm and leg, and an ipsilateral Horner's syndrome. If the lesion is extramedullary, the glossopharyngeal and spinal accessory nerves are frequently involved (jugular foramen syndrome). If it is extracranial in the posterior laterocondylar or retroparotid space, there may be a combination of ninth, tenth, eleventh, and twelfth cranial nerve palsies and a Horner's syndrome (Table 433-2). If there is no sensory loss over the palate and pharynx and no palatal weakness or dysphagia, the lesion is below the origin of the pharyngeal branches, which leave the vagus nerve high in the cervical region; the usual site of disease is then the mediastinum.

NECK WEAKNESS

Isolated involvement of the accessory (eleventh cranial) nerve can occur anywhere along its route, resulting in partial or complete paralysis of the sternocleidomastoid and trapezius muscles. More commonly, involvement occurs in combination with deficits of the ninth and tenth cranial nerves in the jugular foramen or after exit from the skull (Table 433-2). An idiopathic form of accessory neuropathy, akin to Bell's palsy, has been described, and it may be recurrent in some cases. Most but not all patients recover.

TONGUE PARALYSIS

The hypoglossal (twelfth cranial) nerve supplies the ipsilateral muscles of the tongue. The nucleus of the nerve or its fibers of exit may be involved by intramedullary lesions such as tumor, poliomyelitis, or most often motor neuron disease. Lesions of the basal meninges and the occipital bones (platybasia, invagination of occipital condyles, Paget's disease) may compress the nerve in its extramedullary course or in the hypoglossal canal. Isolated lesions of unknown cause can occur. Atrophy and fasciculation of the tongue develop weeks to months after interruption of the nerve.

MULTIPLE CRANIAL NERVE PALSIES

Several cranial nerves may be affected by the same disease process. In this situation, the main clinical problem is to determine whether the lesion lies within the brainstem or outside it. Lesions that lie on the surface of the brainstem are characterized by involvement of adjacent cranial nerves (often occurring in succession) and late and rather slight involvement of the long sensory and motor pathways and segmental structures lying within the brainstem. The opposite is true of primary lesions within the brainstem. The extramedullary lesion is more likely to cause bone erosion or enlargement of the foramens of exit of cranial nerves. The intramedullary lesion involving cranial nerves often produces a crossed sensory or motor paralysis (cranial nerve signs on one side of the body and tract signs on the opposite side).

Involvement of multiple cranial nerves outside the brainstem is frequently the result of trauma, localized infections including varicella-zoster virus, infectious and noninfectious (especially carcinomatous) causes of meningitis **(Chaps. 133 and 134)**, granulomatous diseases such as granulomatosis with polyangiitis **(Chap. 356)**, Behçet's disease, vascular disorders including those associated with diabetes, enlarging aneurysms, or locally infiltrating tumors. Among the tumors, nasopharyngeal cancers, lymphomas, neurofibromas, meningiomas, chordomas, cholesteatomas, carcinomas, and sarcomas have all been observed to involve a succession of lower cranial nerves. Owing to their anatomic relationships, the

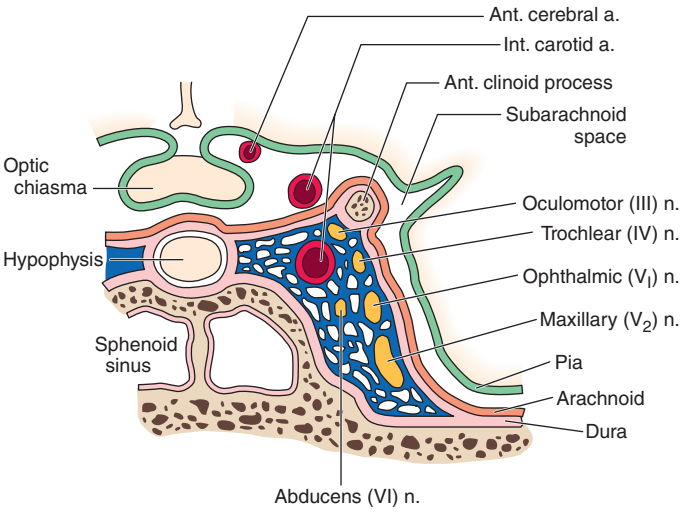


FIGURE 433-4 Anatomy of the cavernous sinus in coronal section, illustrating the location of the cranial nerves in relation to the vascular sinus, internal carotid artery (which loops anteriorly to the section), and surrounding structures.

multiple cranial nerve palsies form a number of distinctive syndromes, listed in Table 433-2. Sarcoidosis is the cause of some cases of multiple cranial neuropathy; tuberculosis, the Chiari malformation, platybasia, and basilar invagination of the skull are additional causes. A purely motor disorder without atrophy always raises the question of myasthenia gravis **(Chap. 440)**. As noted above, Guillain-Barré syndrome commonly affects the facial nerves bilaterally. In the Fisher variant of the Guillain-Barré syndrome, oculomotor paresis occurs with ataxia and areflexia in the limbs **(Chap. 439)**. Wernicke's encephalopathy can cause a severe ophthalmoplegia combined with other brainstem signs **(Chap. 301)**.

The cavernous sinus syndrome **(Fig. 433-4)** is a distinctive and frequently life-threatening disorder. It often presents as orbital or facial pain; orbital swelling and chemosis due to occlusion of the ophthalmic veins; fever; oculomotor neuropathy affecting the third, fourth, and sixth cranial nerves; and trigeminal neuropathy affecting the ophthalmic (V1) and occasionally the maxillary (V2) divisions of the trigeminal nerve. Cavernous sinus thrombosis, often secondary to infection from orbital cellulitis (frequently *Staphylococcus aureus*), a cutaneous source on the face, or sinusitis (especially with mucormycosis in diabetic patients), is the most frequent cause; other etiologies include aneurysm of the carotid artery, a carotid-cavernous fistula (orbital bruit may be present), meningioma, nasopharyngeal carcinoma, other tumors, or an idiopathic granulomatous disorder (Tolosa-Hunt syndrome). The two cavernous sinuses directly communicate via intercavernous channels; thus, involvement on one side may extend to become bilateral. Early diagnosis is essential, especially when due to infection, and treatment depends on the underlying etiology.

In infectious cases, prompt administration of broad-spectrum antibiotics, drainage of any abscess cavities, and identification of the offending organism are essential. Anticoagulant therapy may benefit cases of primary thrombosis. Repair or occlusion of the carotid artery may be required for treatment of fistulas or aneurysms. The Tolosa-Hunt syndrome generally responds to glucocorticoids. A dramatic improvement in pain is usually evident within a few days; oral prednisone (60 mg daily) is usually continued for 2 weeks and then gradually tapered over a month, or longer if pain recurs. Occasionally an immunosuppressive medication, such as azathioprine or methotrexate, needs to be added to maintain an initial response to glucocorticoids.

An idiopathic form of multiple cranial nerve involvement on one or both sides of the face is occasionally seen. The syndrome consists of a subacute onset of boring facial pain, followed by paralysis of motor cranial nerves. The clinical features overlap those of the Tolosa-Hunt syndrome and appear to be due to idiopathic inflammation of the dura mater, which may be visualized by MRI. The syndrome is usually responsive to glucocorticoids.

FURTHER READING

Cruccu G et al: Trigeminal neuralgia: New classification and diagnostic grading for practice and research. Neurology 87:220, 2016.

Gagyor I et al: Antiviral treatment for Bell's palsy (idiopathic facial paralysis). Cochrane Database Syst Rev (11):CD001869, 2015.

Kelly HR, Curtin HD: Imaging of skull base lesions. Handb Clin Neurol 135:637, 2016.

Madhok VB et al: Corticosteroids for Bell's palsy (idiopathic facial paralysis). Cochrane Database Syst Rev 7:CD001942, 2016.

434 Diseases of the Spinal Cord

Stephen L. Hauser

Diseases of the spinal cord are frequently devastating. They produce quadriplegia, paraplegia, and sensory deficits far beyond the damage they would inflict elsewhere in the nervous system because the spinal cord contains, in a small cross-sectional area, almost the entire motor output and sensory input of the trunk and limbs. Many spinal cord diseases are reversible if recognized and treated at an early stage **(Table 434-1)**; thus, they are among the most critical of neurologic emergencies. The efficient use of diagnostic procedures, guided by knowledge of the anatomy and the clinical features of spinal cord diseases, is required to maximize the likelihood of a successful outcome.

TABLE 434-1 Treatable Spinal Cord Disorders

Compressive
Epidural, intradural, or intramedullary neoplasm
Epidural abscess
Epidural hemorrhage
Cervical spondylosis
Herniated disk
Posttraumatic compression by fractured or displaced vertebra or hemorrhage
Vascular
Arteriovenous malformation and dural fistula
Antiphospholipid syndrome and other hypercoagulable states
Inflammatory
Multiple sclerosis
Neuromyelitis optica
Transverse myelitis
Sarcoidosis
Sjögren-related myelopathy
Systemic lupus erythematosus-related myelopathy
Vasculitis
Infectious
Viral: VZV, HSV-1 and 2, CMV, HIV, HTLV-1, others
Bacterial and mycobacterial: *Borrelia*, *Listeria*, syphilis, others
Mycoplasma pneumoniae
Parasitic: schistosomiasis, toxoplasmosis, cystercercosis
Developmental
Syringomyelia
Meningomyelocele
Tethered cord syndrome
Metabolic
Vitamin B_{12} deficiency (subacute combined degeneration)
Copper deficiency

Abbreviations: CMV, cytomegalovirus; HSV, herpes simplex virus; HTLV, human T cell lymphotropic virus; VZV, varicella-zoster virus.

APPROACH TO THE PATIENT

Spinal Cord Disease

SPINAL CORD ANATOMY RELEVANT TO CLINICAL SIGNS

The spinal cord is a thin, tubular extension of the central nervous system contained within the bony spinal canal. It originates at the medulla and continues caudally to the conus medullaris at the lumbar level; its fibrous extension, the filum terminale, terminates at the coccyx. The adult spinal cord is ~46 cm (18 in.) long, oval in shape, and enlarged in the cervical and lumbar regions, where neurons that innervate the upper and lower extremities, respectively, are located. The white matter tracts containing ascending sensory and descending motor pathways are located peripherally, whereas nerve cell bodies are clustered in an inner region of gray matter shaped like a four-leaf clover that surrounds the central canal (anatomically an extension of the fourth ventricle). The membranes that cover the spinal cord—the pia, arachnoid, and dura—are continuous with those of the brain, and the cerebrospinal fluid is contained within the subarachnoid space between the pia and arachnoid.

The spinal cord has 31 segments, each defined by an exiting ventral motor root and entering dorsal sensory root. During embryologic development, growth of the cord lags behind that of the vertebral column, and the mature spinal cord ends at approximately the first lumbar vertebral body. The lower spinal nerves take an increasingly downward course to exit via intervertebral foramina. The first seven pairs of cervical spinal nerves exit above the same-numbered vertebral bodies, whereas all the subsequent nerves exit below the same-numbered vertebral bodies because of the presence of eight cervical spinal cord segments but only seven cervical vertebrae. The relationship between spinal cord segments and the corresponding vertebral bodies is shown in **Table 434-2**. These relationships assume particular importance for localization of lesions that cause spinal cord compression. Sensory loss below the circumferential level of the umbilicus, for example, corresponds to the T10 cord segment but indicates involvement of the cord adjacent to the seventh or eighth thoracic vertebral body **(see Figs. 22-2 and 22-3)**. In addition, at every level, the main ascending and descending tracts are somatotopically organized with a laminated distribution that reflects the origin or destination of nerve fibers.

Determining the Level of the Lesion The presence of a horizontally defined level below which sensory, motor, and autonomic function is impaired is a hallmark of a lesion of the spinal cord. This *sensory level* is sought by asking the patient to identify a pinprick or cold stimulus applied to the proximal legs and lower trunk and successively moved up toward the neck on each side. Sensory loss below this level is the result of damage to the spinothalamic tract on the opposite side, one to two segments higher in the case of a unilateral spinal cord lesion, and at the level of a bilateral lesion. The discrepancy in the level of a unilateral lesion is the result of the course of the second-order sensory fibers, which originate in the dorsal horn, and ascend for one or two levels as they cross anterior to the central canal to join the opposite spinothalamic tract. Lesions that transect the descending corticospinal and other motor tracts cause paraplegia or quadriplegia with heightened deep tendon reflexes, Babinski signs, and eventual spasticity (the upper motor neuron syndrome). Transverse damage to the cord also produces autonomic disturbances consisting of absent sweating below the implicated cord level and bladder, bowel, and sexual dysfunction.

The uppermost level of a spinal cord lesion can also be localized by attention to the *segmental signs* corresponding to disturbed motor or sensory innervation by an individual cord segment. A band of altered sensation (hyperalgesia or hyperpathia) at the upper end of the sensory disturbance, fasciculations or atrophy in muscles innervated by one or several segments, or a muted or absent deep tendon reflex may be noted at this level. These signs also can occur with focal root or peripheral nerve disorders; thus, they are most useful when they occur together with signs of long tract damage. With severe and acute transverse lesions, the limbs initially may be flaccid rather than spastic. This state of "spinal shock" lasts for several days, rarely for weeks, and may be mistaken for extensive damage to the anterior horn cells over many segments of the cord or for an acute polyneuropathy.

The main features of transverse damage at each level of the spinal cord are summarized below.

Cervical Cord Upper cervical cord lesions produce quadriplegia and weakness of the diaphragm. The uppermost level of weakness and reflex loss with lesions at C5-C6 is in the biceps; at C7, in finger and wrist extensors and triceps; and at C8, finger, and wrist flexion. Horner's syndrome (miosis, ptosis, and facial hypohidrosis) may accompany a cervical cord lesion at any level.

Thoracic Cord Lesions here are localized by the sensory level on the trunk and, if present, by the site of midline back pain. Useful markers of the sensory level on the trunk are the nipples (T4) and umbilicus (T10). Leg weakness and disturbances of bladder and bowel function accompany the paralysis. Lesions at T9-T10 paralyze the lower—but not the upper—abdominal muscles, resulting in upward movement of the umbilicus when the abdominal wall contracts (*Beevor's sign*).

Lumbar Cord Lesions at the L2-L4 spinal cord levels paralyze flexion and adduction of the thigh, weaken leg extension at the knee, and abolish the patellar reflex. Lesions at L5-S1 paralyze only movements of the foot and ankle, flexion at the knee, and extension of the thigh, and abolish the ankle jerks (S1).

Sacral Cord/Conus Medullaris The conus medullaris is the tapered caudal termination of the spinal cord, comprising the sacral and single coccygeal segments. The distinctive conus syndrome consists of bilateral saddle anesthesia (S3-S5), prominent bladder and bowel dysfunction (urinary retention and incontinence with lax anal tone), and impotence. The bulbocavernosus (S2-S4) and anal (S4-S5) reflexes are absent **(Chap. 415)**. Muscle strength is largely preserved. By contrast, lesions of the cauda equina, the nerve roots derived from the lower cord, are characterized by low back and radicular pain, asymmetric leg weakness and sensory loss, variable areflexia in the lower extremities, and relative sparing of bowel and bladder function. Mass lesions in the lower spinal canal often produce a mixed clinical picture with elements of both cauda equina and conus medullaris syndromes. **Cauda equina syndromes are also discussed in Chap. 14.**

Special Patterns of Spinal Cord Disease The location of the major ascending and descending pathways of the spinal cord are shown in **Fig. 434-1**. Most fiber tracts—including the posterior columns and the spinocerebellar and pyramidal tracts—are situated on the side of the body they innervate. However, afferent fibers mediating pain and temperature sensation ascend in the spinothalamic tract contralateral to the side they supply. The anatomic configurations of these tracts produce characteristic syndromes that provide clues to the underlying disease process.

Brown-Sequard Hemicord Syndrome This consists of ipsilateral weakness (corticospinal tract) and loss of joint position and vibratory sense (posterior column), with contralateral loss of pain and temperature sense (spinothalamic tract) one or two levels below the

TABLE 434-2 Spinal Cord Levels Relative to the Vertebral Bodies

SPINAL CORD LEVEL	CORRESPONDING VERTEBRAL BODY
Upper cervical	Same as cord level
Lower cervical	1 level higher
Upper thoracic	2 levels higher
Lower thoracic	2–3 levels higher
Lumbar	T10-T12
Sacral	T12-L1

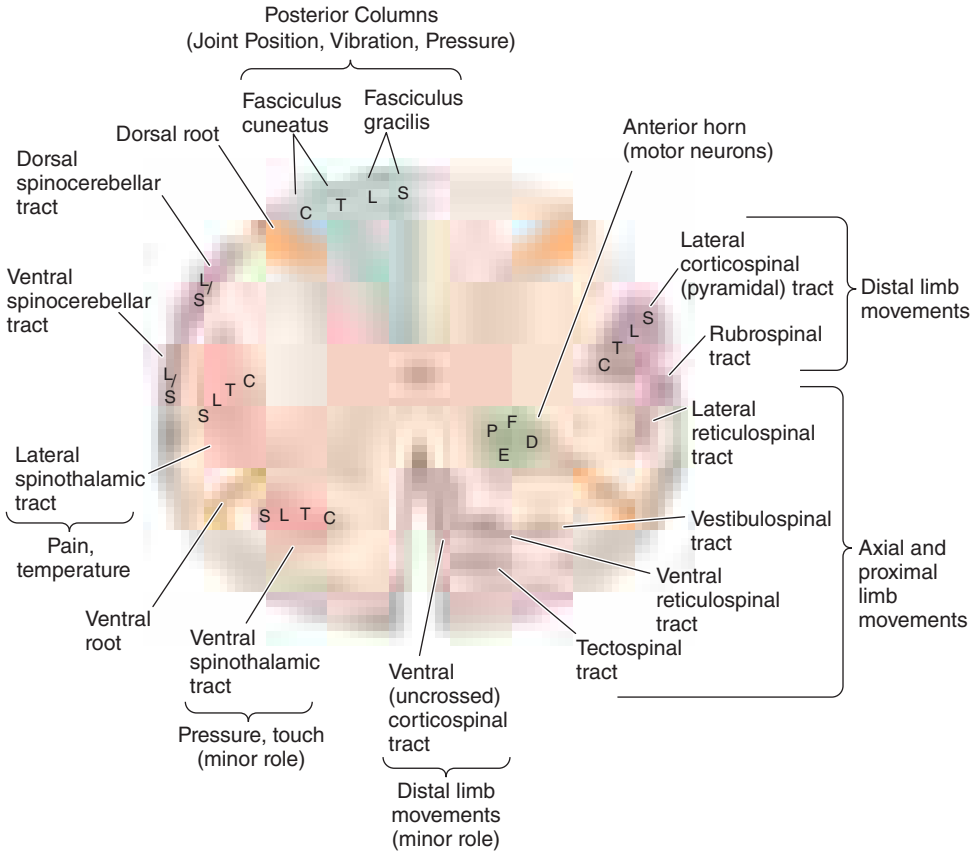


FIGURE 434-1 Transverse section through the spinal cord, composite representation, illustrating the principal ascending (*left*) and descending (*right*) pathways. The lateral and ventral spinothalamic tracts ascend contralateral to the side of the body that is innervated. C, cervical; D, distal; E, extensors; F, flexors; L, lumbar; P, proximal; S, sacral; T, thoracic.

lesion. Segmental signs, such as radicular pain, muscle atrophy, or loss of a deep tendon reflex, are unilateral. Partial forms are more common than the fully developed syndrome.

Central Cord Syndrome This syndrome results from selective damage to the gray matter nerve cells and crossing spinothalamic tracts surrounding the central canal. In the cervical cord, the central cord syndrome produces arm weakness out of proportion to leg weakness and a "dissociated" sensory loss, meaning loss of pain and temperature sensations over the shoulders, lower neck, and upper trunk (cape distribution), in contrast to preservation of light touch, joint position, and vibration sense in these regions. Spinal trauma, syringomyelia, and intrinsic cord tumors are the main causes.

Anterior Spinal Artery Syndrome Infarction of the cord is generally the result of occlusion or diminished flow in this artery. The result is bilateral tissue destruction at several contiguous levels that spares the posterior columns. All spinal cord functions—motor, sensory, and autonomic—are lost below the level of the lesion, with the striking exception of retained vibration and position sensation.

Foramen Magnum Syndrome Lesions in this area interrupt decussating pyramidal tract fibers destined for the legs, which cross caudal to those of the arms, resulting in weakness of the legs (*crural paresis*). Compressive lesions near the foramen magnum may produce weakness of the ipsilateral shoulder and arm followed by weakness of the ipsilateral leg, then the contralateral leg, and finally the contralateral arm, an "around the clock" pattern that may begin in any of the four limbs. There is typically suboccipital pain spreading to the neck and shoulders.

Intramedullary and Extramedullary Syndromes It is useful to differentiate *intramedullary* processes, arising within the substance of the cord, from *extramedullary* ones that lie outside the cord and compress the spinal cord or its vascular supply. The differentiating features are only relative and serve as clinical guides. With extramedullary lesions, radicular pain is often prominent, and there is early sacral sensory loss and spastic weakness in the legs with incontinence due to the superficial location of the corresponding sensory and motor fibers in the spinothalamic and corticospinal tracts (Fig. 434-1). Intramedullary lesions tend to produce poorly localized burning pain rather than radicular pain and to spare sensation in the perineal and sacral areas ("sacral sparing"), reflecting the laminated configuration of the spinothalamic tract with sacral fibers outermost; corticospinal tract signs appear later. Regarding extramedullary lesions, a further distinction is made between extradural and intradural masses, as the former are generally malignant and the latter benign (neurofibroma being a common cause). Consequently, a long duration of symptoms favors an intradural origin.

ACUTE AND SUBACUTE SPINAL CORD DISEASES

Symptoms of the cord diseases that evolve over days or weeks are focal neck or back pain, followed by various combinations of paresthesias, sensory loss, motor weakness, and sphincter disturbance. There may be mild sensory symptoms only or a devastating functional transection of the cord. When paresthesias begin in the feet and then ascend a polyneuropathy is often considered, and in such cases the presence of bladder disturbances and a sharply demarcated spinal cord level provide important clues to the spinal cord origin of the disease.

In severe and abrupt cases, areflexia reflecting spinal shock may be present, but hyperreflexia supervenes over days or weeks; persistent areflexic paralysis with a sensory level usually indicates necrosis over multiple segments of the spinal cord.

APPROACH TO THE PATIENT

Compressive and Noncompressive Myelopathy

DISTINGUISHING COMPRESSIVE FROM NONCOMPRESSIVE MYELOPATHY

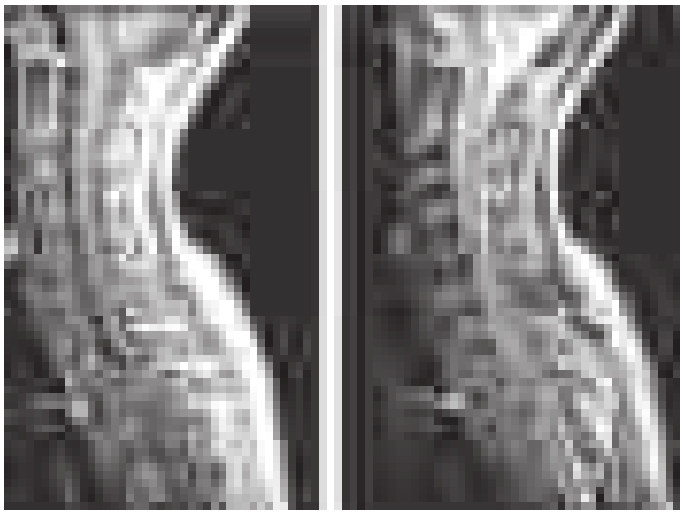
The first priority is to exclude treatable compression of the cord by a mass lesion. The common causes are tumor, epidural abscess or hematoma, herniated disk, and spondylitic vertebral pathology. Epidural compression due to malignancy or abscess often causes warning signs of neck or back pain, bladder disturbances, and sensory symptoms that precede the development of paralysis. Spinal subluxation, hemorrhage, and noncompressive etiologies such as infarction are more likely to produce myelopathy without antecedent symptoms. Magnetic resonance imaging (MRI) with gadolinium, centered on the clinically suspected level, is the initial diagnostic procedure if it is available; it is often appropriate to image the entire spine (cervical through sacral regions) to search for additional clinically silent lesions. Once compressive lesions have been excluded, noncompressive causes of acute myelopathy that are intrinsic to the cord are considered, primarily vascular, inflammatory, and infectious etiologies.

COMPRESSIVE MYELOPATHIES

Neoplastic Spinal Cord Compression In adults, most neoplasms are epidural in origin, resulting from metastases to the adjacent vertebral column. The propensity of solid tumors to metastasize to the vertebral column probably reflects the high proportion of bone marrow located in the axial skeleton. Almost any malignant tumor can metastasize to the spinal column, with breast, lung, prostate, kidney, lymphoma, and myeloma being particularly frequent. The thoracic spinal column is most commonly involved; exceptions are metastases from prostate and ovarian cancer, which occur disproportionately in the sacral and lumbar vertebrae, probably from spread through Batson's plexus, a network of veins along the anterior epidural space. Retroperitoneal neoplasms (especially lymphomas or sarcomas) enter the spinal canal laterally through the intervertebral foramina and produce radicular pain with signs of weakness that corresponds to the level of involved nerve roots.

Pain is usually the initial symptom of spinal metastasis; it may be aching and localized or sharp and radiating in quality and typically worsens with movement, coughing, or sneezing and characteristically awakens patients at night. A recent onset of persistent back pain, particularly if in the thoracic spine (which is uncommonly involved by spondylosis), should prompt consideration of vertebral metastasis. Rarely, pain is mild or absent. Plain radiographs of the spine and radionuclide bone scans have a limited role in diagnosis because they do not identify 15–20% of metastatic vertebral lesions and fail to detect paravertebral masses that reach the epidural space through the intervertebral foramina. MRI provides excellent anatomic resolution of the extent of spinal tumors **(Fig. 434-2)** and is able to distinguish between malignant lesions and other masses—epidural abscess, tuberculoma, lipoma, or epidural hemorrhage, among others—that present in a similar fashion. Vertebral metastases are usually hypointense relative to a normal bone marrow signal on T1-weighted MRI; after the administration of gadolinium, contrast enhancement may deceptively "normalize" the appearance of the tumor by increasing its intensity to that of normal bone marrow. Infections of the spinal column (osteomyelitis and related disorders) are distinctive in that, unlike tumor, they often cross the disk space to involve the adjacent vertebral body.

If spinal cord compression is suspected, imaging should be obtained promptly. If there are radicular symptoms but no evidence of myelopathy, it may be safe to defer imaging for 24–48 h. Up to 40% of patients who present with cord compression at one level are found to have asymptomatic epidural metastases elsewhere; thus, imaging of the entire length of the spine is important to define the extent of disease.



A ***B***

FIGURE 434-2 Epidural spinal cord compression due to breast carcinoma. Sagittal T1-weighted (***A***) and T2-weighted (***B***) magnetic resonance imaging scans through the cervicothoracic junction reveal an infiltrated and collapsed second thoracic vertebral body with posterior displacement and compression of the upper thoracic spinal cord. The low-intensity bone marrow signal in ***A*** signifies replacement by tumor.

TREATMENT

Neoplastic Spinal Cord Compression

Management of cord compression includes glucocorticoids to reduce cord edema, local radiotherapy (initiated as early as possible) to the symptomatic lesion, and specific therapy for the underlying tumor type. Glucocorticoids (typically dexamethasone, 10 mg intravenously) can be administered before an imaging study if there is clinical suspicion of cord compression and continued at a lower dose (4 mg every 6 h orally) until definitive treatment with radiotherapy (generally 30–40 Gy administered in 8–10 fractions) and/or surgical decompression is completed. In one trial, initial management with surgery followed by radiotherapy was more effective than radiotherapy alone for patients with a single area of spinal cord compression by extradural tumor; however, patients with recurrent cord compression, brain metastases, radiosensitive tumors, or severe motor symptoms of >48 h in duration were excluded from this study. Radiotherapy alone may be effective even for some typically radioresistant metastases. A good response to therapy can be expected in individuals who are ambulatory at presentation. Treatment usually prevents new weakness, and some recovery of motor function occurs in up to one-third of patients. Motor deficits (paraplegia or quadriplegia), once established for >12 h, do not usually improve, and beyond 48 h the prognosis for substantial motor recovery is poor. Although most patients do not experience recurrences in the months following radiotherapy, with survival beyond 2 years recurrence becomes increasingly likely and can be managed with additional radiotherapy. Newer techniques such as stereotactic radiosurgery can deliver high doses of focused radiation with similar rates of response compared to traditional radiotherapy, and these are increasingly being used, particularly for patients with traditionally radioresistant tumors or those requiring re-irradiation. Biopsy of the epidural mass is unnecessary in patients with known primary cancer, but it is indicated if a history of underlying cancer is lacking. Surgery, either decompression by laminectomy or vertebral body resection, is also indicated when signs of cord compression worsen despite radiotherapy, when the maximum-tolerated dose of radiotherapy has been delivered previously to the site, or when a vertebral compression fracture or spinal instability contributes to cord compression.

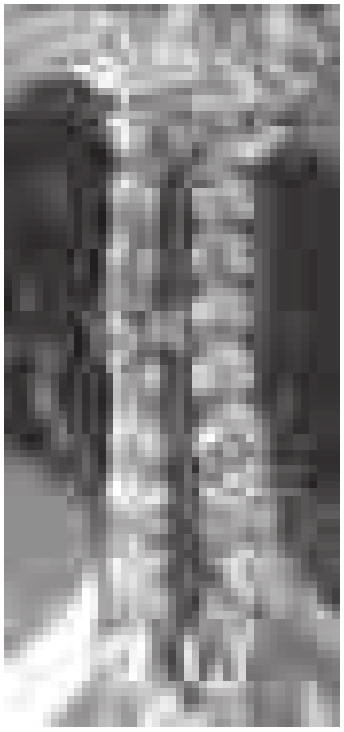


FIGURE 434-3 Magnetic resonance imaging of a thoracic meningioma. Coronal T1-weighted postcontrast image through the thoracic spinal cord demonstrates intense and uniform enhancement of a well-circumscribed extramedullary mass (*arrows*) that displays the spinal cord to the left.

In contrast to tumors of the epidural space, most intradural mass lesions are slow-growing and benign. Meningiomas and neurofibromas account for most of these, with occasional cases caused by chordoma, lipoma, dermoid, or sarcoma. Meningiomas **(Fig. 434-3)** are often located posterior to the thoracic cord or near the foramen magnum, although they can arise from the meninges anywhere along the spinal canal. Neurofibromas are benign tumors of the nerve sheath that typically arise from the posterior root; when multiple, neurofibromatosis is the likely etiology. Symptoms usually begin with radicular sensory symptoms followed by an asymmetric, progressive spinal cord syndrome. Therapy is surgical resection.

Primary intramedullary tumors of the spinal cord are uncommon. They present as central cord or hemicord syndromes, often in the cervical region. There may be poorly localized burning pain in the extremities and sparing of sacral sensation. In adults, these lesions are ependymomas, hemangioblastomas, or low-grade astrocytomas **(Fig. 434-4)**. Complete resection of an intramedullary ependymoma is often possible with microsurgical techniques. Debulking of an intramedullary astrocytoma can also be helpful, as these are often slowly growing lesions; the value of adjunctive radiotherapy and chemotherapy is uncertain. Secondary (metastatic) intramedullary tumors also occur, especially in patients with advanced metastatic disease **(Chap. 86)**, although these are not nearly as frequent as brain metastases.

Spinal Epidural Abscess Spinal epidural abscess presents with midline back or neck pain, fever, and progressive limb weakness. Prompt recognition of this distinctive process may prevent permanent sequelae. Aching pain is almost always present, either over the spine or in a radicular pattern. The duration of pain prior to presentation is generally ≤2 weeks but may on occasion be several months or longer. Fever is typically but not invariably present, accompanied by elevated white blood cell count, sedimentation rate, and C-reactive protein. As the abscess expands, further spinal cord damage results from venous congestion and thrombosis. Once weakness and other signs of myelopathy appear, progression may be rapid and irreversible. A more chronic sterile granulomatous form of abscess is also known, usually after treatment of an acute epidural infection.

FIGURE 434-4 Magnetic resonance imaging of an intramedullary astrocytoma. Sagittal T1-weighted postcontrast image through the cervical spine demonstrates expansion of the upper cervical spine by a mass lesion emanating from within the spinal cord at the cervicomedullary junction. Irregular peripheral enhancement occurs within the mass (*arrows*).

Risk factors include an impaired immune status (HIV, diabetes mellitus, renal failure, alcoholism, malignancy), intravenous drug abuse, and infections of the skin or other tissues. Two-thirds of epidural infections result from hematogenous spread of bacteria from the skin (furunculosis), soft tissue (pharyngeal or dental abscesses; sinusitis), or deep viscera (bacterial endocarditis). The remainder arises from direct extension of a local infection to the subdural space; examples of local predisposing conditions are vertebral osteomyelitis, decubitus ulcers, lumbar puncture, epidural anesthesia, or spinal surgery. Most cases are due to *Staphylococcus aureus*; gram-negative bacilli, *Streptococcus*, anaerobes, and fungi can also cause epidural abscesses. Methicillin resistant *Staphylococcus aureus* (MRSA) is an important consideration, and therapy should be tailored to this possibility. Tuberculosis from an adjacent vertebral source (Pott's disease) remains an important cause in the developing world.

MRI **(Fig. 434-5)** localizes the abscess and excludes other causes of myelopathy. Blood cultures are positive in more than half of cases, but direct aspiration of the abscess at surgery is often required for a microbiologic diagnosis. Lumbar puncture is only required if encephalopathy or other clinical signs raise the question of associated meningitis, a feature that is found in <25% of cases. The level of the puncture should be planned to minimize the risk of meningitis due to passage of the needle through infected tissue. A high cervical tap is sometimes the safest approach. Cerebrospinal fluid (CSF) abnormalities in epidural and subdural abscess consist of pleocytosis with a preponderance of polymorphonuclear cells, an elevated protein level, and a reduced glucose level, but the responsible organism is not cultured unless there is associated meningitis.

TREATMENT

Spinal Epidural Abscess

Treatment is by decompressive laminectomy with debridement combined with long-term antibiotic treatment. Surgical evacuation prevents development of paralysis and may improve or reverse paralysis in evolution, but it is unlikely to improve deficits of more than several days in duration. Broad-spectrum antibiotics (typically vancomycin 15–20 mg/kg q12h (staphylococcus including MRSA, streptococcus), ceftriaxone 2 gm q24h (gram-negative bacilli), and when indicated metronidazole 30 mg/kg/d divided into q6h intervals (anaerobes) should be started empirically before surgery and then modified on the basis of culture results; medication is generally

A *B*

FIGURE 434-5 Magnetic resonance (MR) imaging of a spinal epidural abscess due to tuberculosis. ***A.*** Sagittal T2-weighted free spin-echo MR sequence. A hypointense mass replaces the posterior elements of C3 and extends epidurally to compress the spinal cord (*arrows*). ***B.*** Sagittal T1-weighted image after contrast administration reveals a diffuse enhancement of the epidural process (*arrows*) with extension into the epidural space.

continued for 6–8 weeks. If surgery is contraindicated or if there is a fixed paraplegia or quadriplegia that is unlikely to improve following surgery, long-term administration of systemic and oral antibiotics can be used; in such cases, the choice of antibiotics may be guided by results of blood cultures. Surgical management remains the treatment of choice unless the abscess is limited in size and causes few or no neurologic signs.

With prompt diagnosis and treatment of spinal epidural abscess, up to two-thirds of patients experience significant recovery.

Spinal Epidural Hematoma Hemorrhage into the epidural (or subdural) space causes acute focal or radicular pain followed by variable signs of a spinal cord or conus medullaris disorder. Therapeutic anticoagulation, trauma, tumor, or blood dyscrasias are predisposing conditions. Rare cases complicate lumbar puncture or epidural anesthesia. MRI and computed tomography (CT) confirm the clinical suspicion and can delineate the extent of the bleeding. Treatment consists of prompt reversal of any underlying clotting disorder and surgical decompression. Surgery may be followed by substantial recovery, especially in patients with some preservation of motor function preoperatively. Because of the risk of hemorrhage, lumbar puncture should be avoided whenever possible in patients with severe thrombocytopenia or other coagulopathies.

Hematomyelia Hemorrhage into the substance of the spinal cord is a rare result of trauma, intraparenchymal vascular malformation (see below), vasculitis due to polyarteritis nodosa or systemic lupus erythematosus (SLE), bleeding disorders, or a spinal cord neoplasm. Hematomyelia presents as an acute painful transverse myelopathy. With large lesions, extension into the subarachnoid space results in subarachnoid hemorrhage **(Chap. 302)**. Diagnosis is by MRI or CT. Therapy is supportive, and surgical intervention is generally not useful. An exception is hematomyelia due to an underlying vascular malformation, for which spinal angiography and endovascular occlusion may be indicated, or surgery to evacuate the clot and remove the underlying vascular lesion.

NONCOMPRESSIVE MYELOPATHIES

The most frequent causes of noncompressive acute transverse myelopathy are spinal cord infarction; systemic inflammatory disorders, including SLE and sarcoidosis; demyelinating diseases, including multiple sclerosis (MS); neuromyelitis optica (NMO); postinfectious or idiopathic transverse myelitis, which is presumed to be an immune condition related to acute disseminated encephalomyelitis **(Chap. 436)**; and infectious (primarily viral) causes. After spinal cord compression is excluded, the evaluation generally requires a lumbar puncture and a search for underlying systemic disease **(Table 434-3)**.

TABLE 434-3 Evaluation of Acute Transverse Myelopathy

1. MRI of spinal cord with and without contrast (exclude compressive causes).
2. CSF studies: Cell count, protein, glucose, IgG index/synthesis rate, oligoclonal bands, VDRL; Gram's stain, acid-fast bacilli, and India ink stains; PCR for VZV, HSV-2, HSV-1, EBV, CMV, HHV-6, enteroviruses, HIV; antibody for HTLV-1, *Borrelia burgdorferi*, *Mycoplasma pneumoniae*, and *Chlamydia pneumoniae*; viral, bacterial, mycobacterial, and fungal cultures.
3. Blood studies for infection: HIV; RPR; IgG and IgM enterovirus antibody; IgM mumps, measles, rubella, group B arbovirus, *Brucella melitensis*, *Chlamydia psittaci*, *Bartonella henselae*, schistosomal antibody; cultures for *B. melitensis*. Also consider nasal/pharyngeal/anal cultures for enteroviruses; stool O&P for *Schistosoma* ova.
4. Systemic immune-mediated disorders: ESR; ANA; ENA; dsDNA; rheumatoid factor; anti-SSA; anti-SSB, complement levels; antiphospholipid and anticardiolipin antibodies; p-ANCA; antimicrosomal and antithyroglobulin antibodies; if Sjögren's syndrome suspected, Schirmer test, salivary gland scintigraphy, and salivary/lacrimal gland biopsy.
5. Neuromyelitis Optica: Serum anti-aquaporin-4 antibody, anti-MOG antibody.
6. Demyelinating disease: Brain MRI scan; evoked potentials.
7. Sarcoidosis: Serum angiotensin-converting enzyme; serum Ca; 24-h urine Ca; chest x-ray; chest CT; total-body gallium scan; lymph node biopsy.
8. Vascular causes: MRI, CT myelogram; spinal angiogram.

Abbreviations: ANA, antinuclear antibodies; CMV, cytomegalovirus; CSF, cerebrospinal fluid; CT, computed tomography; EBV, Epstein-Barr virus; ENA, epithelial neutrophil-activating peptide; ESR, erythrocyte sedimentation rate; HHV, human herpes virus; HSV, herpes simplex virus; HTLV, human T cell leukemia/lymphoma virus; MOG, myelin oligodendrocyte glycoprotein; MRI, magnetic resonance imaging; O&P, ova and parasites; p-ANCA, perinuclear antineutrophilic cytoplasmic antibodies; PCR, polymerase chain reaction; RPR, rapid plasma reagin (test); VDRL, Venereal Disease Research Laboratory; VZV, varicella-zoster virus.

Spinal Cord Infarction The cord is supplied by three arteries that course vertically over its surface: a single anterior spinal artery and paired posterior spinal arteries. The anterior spinal artery originates in paired branches of the vertebral arteries at the cranciocervical junction and is fed by additional radicular vessels that arise at C6, at an upper thoracic level, and, most consistently, at T11-L2 (artery of Adamkiewicz). At each spinal cord segment, paired penetrating vessels branch from the anterior spinal artery to supply the anterior two-thirds of the cord; the posterior spinal arteries, which often become less distinct below the midthoracic level, supply the posterior columns.

Spinal cord ischemia can occur at any level; however, the presence of the artery of Adamkiewicz below, and the anterior spinal artery circulation above, creates a region of marginal blood flow in the upper thoracic segments. With hypotension or cross-clamping of the aorta, cord infarction typically occurs at the level of T3-T4, and also at boundary zones between the anterior and posterior spinal artery territories. The latter may result in a rapidly progressive syndrome over hours of weakness and spasticity with little sensory change.

Acute infarction in the territory of the *anterior spinal artery* produces paraplegia or quadriplegia, dissociated sensory loss affecting pain and temperature sense but sparing vibration and position sense, and loss of sphincter control ("anterior cord syndrome"). Onset may be sudden but more typically is progressive over minutes or a few hours, unlike stroke in the cerebral hemispheres. Sharp midline or radiating back pain localized to the area of ischemia is frequent. Areflexia due to spinal shock is often present initially; with time, hyperreflexia and spasticity appear. Less common is infarction in the territory of the *posterior spinal arteries*, resulting in loss of posterior column function either on one side or bilaterally.

Causes of spinal cord infarction include aortic atherosclerosis, dissecting aortic aneurysm, vertebral artery occlusion or dissection in the neck, aortic surgery, or profound hypotension from any cause. A *surfer's myelopathy* usually in the thoracic region, has been associated with prolonged back extension due to lifting the upper body off the board while waiting for waves; it, typically manifests as back pain followed

by an anterior cord syndrome with progressive paralysis and loss of sphincter control, and is likely vascular in origin. Cardiogenic emboli, vasculitis **(Chap. 356)**, and collagen vascular disease (particularly SLE **[Chap. 349]**, Sjögren's syndrome **[Chap. 354]**, and the antiphospholipid antibody syndrome **[Chap. 350]**) are other etiologies. Occasional cases develop from *embolism of nucleus pulposus* material into spinal vessels, usually from local spine trauma. In a substantial number of cases, no cause can be found, and thromboembolism in arterial feeders is suspected. MRI may fail to demonstrate infarctions of the cord, especially in the first day, but often the imaging becomes abnormal at the affected level.

In cord infarction due to presumed thromboembolism, acute anticoagulation is not indicated, with the possible exception of the unusual transient ischemic attack or incomplete infarction with a stuttering or progressive course. The antiphospholipid antibody syndrome is treated with anticoagulation **(Chap. 350)**. Increasing systemic blood pressure to a mean arterial pressure of >90 mmHg, or lumbar drainage of spinal fluid, was reportedly helpful in a few published cases of cord infarction, but neither of these approaches has been studied systematically. Prognosis following spinal cord infarction is influenced by the severity of the deficits at presentation; patients with severe motor weakness and those with persistent areflexia usually do poorly, but in one recent large series some improvement over time occurred in many patients, with more than half ultimately regaining some ambulation.

Inflammatory and Immune Myelopathies (Myelitis) This broad category includes the demyelinating conditions MS, NMO, and postinfectious myelitis, as well as sarcoidosis and systemic autoimmune disease. In approximately one-quarter of cases of myelitis, no underlying cause can be identified. Some will later manifest additional symptoms of an immune-mediated disease. *Recurrent episodes of myelitis* are usually due to one of the immune-mediated diseases or to infection with herpes simplex virus (HSV) type 2 (below).

MULTIPLE SCLEROSIS MS may present with acute myelitis, particularly in individuals of Asian or African ancestry. In Caucasians, MS attacks rarely cause a transverse myelopathy (i.e., attacks of bilateral sensory disturbances, unilateral or bilateral weakness, and bladder or bowel symptoms), but it is among the most common causes of a partial cord syndrome. MRI findings in MS-associated myelitis typically consist of mild swelling of the cord and diffuse or multifocal "shoddy" areas of abnormal signal on T2-weighted sequences. Contrast enhancement, indicating disruption in the blood-brain barrier associated with inflammation, is present in many acute cases. In one study 68% of patients presenting with partial myelitis developed MS after a mean follow-up of 4 years; risk factors for conversion to MS included age <40 years; inflammatory CSF, and >3 periventricular lesions on brain MRI.

Treatment of acute episodes of MS-associated myelitis consists of intravenous methylprednisolone (500 mg qd for 3 days) followed by oral prednisone (1 mg/kg/d for several weeks, then gradual taper). A course of plasma exchange may be indicated for severe cases if glucocorticoids are ineffective. **MS is discussed in Chap. 436.**

NEUROMYELITIS OPTICA NMO is an immune-mediated demyelinating disorder consisting of a severe myelopathy that is typically longitudinally extensive, meaning that the lesion spans three or more vertebral segments. NMO is associated with optic neuritis that is often bilateral and may precede or follow myelitis by weeks or months, and also by brainstem and, in some cases, hypothalamic or focal cerebral white matter involvement. Recurrent myelitis without optic nerve or other involvement can also occur in NMO. CSF studies reveal a variable mononuclear pleocytosis of up to several hundred cells per microliter; unlike MS, oligoclonal bands are generally absent. Diagnostic serum autoantibodies against the water channel protein aquaporin-4 are present in 60–70% of patients with NMO, and less commonly autoantibodies against the CNS myelin protein myelin oligodendrocyte glycoprotein (MOG) are found. NMO has also been associated with SLE (see below) as well as with other systemic autoimmune diseases; rare cases are paraneoplastic in origin. There have been no definitive trials of therapy for NMO. Recommended treatment of acute relapses is with glucocorticoids and, for refractory cases, plasma exchange. Prophylactic treatment with azathioprine, mycophenylate, or rituximab may protect against subsequent relapses; treatment for 5 years or longer is generally recommended. **NMO is discussed in Chap. 437.**

SYSTEMIC IMMUNE-MEDIATED DISORDERS Myelitis occurs in a small number of patients with SLE, many cases of which are associated with antibodies to aquaporin-4 and satisfy diagnostic criteria for NMO-spectrum disorder **(Chap. 437)**. These patients are at high risk of developing future episodes of myelitis and/or optic neuritis. In others the etiology of SLE-associated myelitis is uncertain; anti-phospholipid antibodies have been suggested to play a role, however the presence of these antibodies appears to be no more frequent in SLE patients with and without myelitis. The CSF in NMO-associated myelitis typically shows a pleocytosis with polymorphonuclear leukocytes and no oligoclonal bands; in cases not due to NMO a mild lymphocytic pleocytosis and oligoclonal bands are variable findings. Although there are no systematic trials of therapy for SLE myelitis, based on limited data, high-dose glucocorticoids followed by cyclophosphamide have been recommended. Acute severe episodes of transverse myelitis that do not initially respond to glucocorticoids are often treated with a course of plasma exchange. Sjögren's syndrome **(Chap. 354)** can also be associated with NMO-spectrum disorder and also with cases of chronic progressive myelopathy. Other immune-mediated myelitides include antiphospholipid antibody syndrome **(Chap. 350)**, mixed connective tissue disease **(Chap. 353)**, Behçet's syndrome **(Chap. 357)**, and vasculitis related to polyarteritis nodosa, perinuclear antineutrophilic cytoplasmic (p-ANCA) antibodies, or primary central nervous system vasculitis **(Chap. 356)**.

Another important consideration in this group is sarcoid myelopathy that may present as a slowly progressive or relapsing disorder. MRI reveals an edematous swelling of the spinal cord that may mimic tumor; there is almost always gadolinium enhancement of active lesions and in some cases nodular enhancement of the adjacent surface of the cord; lesions may be single or multiple, and on axial images, enhancement of the central cord is often present. The typical CSF profile consists of a mild lymphocytic pleocytosis and mildly elevated protein level; in a minority of cases, reduced glucose and oligoclonal bands are found. The diagnosis is particularly difficult when systemic manifestations of sarcoid are minor or absent (nearly 50% of cases) or when other typical neurologic manifestations of the disease—such as cranial neuropathy, hypothalamic involvement, or meningeal enhancement visualized by MRI—are lacking. A slit-lamp examination of the eye to search for uveitis, chest x-ray and CT to assess pulmonary involvement and mediastinal lymphadenopathy, serum or CSF angiotensin-converting enzyme (ACE; CSF values elevated in only a minority of cases), serum calcium, and a gallium scan may assist in the diagnosis. The usefulness of spinal fluid ACE is uncertain. Initial treatment is with oral glucocorticoids; immunosuppressant drugs, including the tumor necrosis factor α inhibitor infliximab, have been used for resistant cases. **Sarcoidosis is discussed in Chap. 360.**

POSTINFECTIOUS MYELITIS Many cases of myelitis, termed *postinfectious* or *postvaccinal*, follow an infection or vaccination. Numerous organisms have been implicated, including Epstein-Barr virus (EBV), cytomegalovirus (CMV), mycoplasma, influenza, measles, varicella, mumps, and yellow fever. As in the related disorder acute disseminated encephalomyelitis **(Chap. 436)**, postinfectious myelitis often begins as the patient appears to be recovering from an acute febrile infection, or in the subsequent days or weeks, but an infectious agent cannot be isolated from the nervous system or CSF. The presumption is that the myelitis represents an autoimmune disorder triggered by infection and is not due to direct infection of the spinal cord. No randomized controlled trials of therapy exist; treatment is usually with glucocorticoids or, in fulminant cases, plasma exchange.

ACUTE INFECTIOUS MYELITIS Many viruses have been associated with an acute myelitis that is infectious in nature rather than postinfectious. Nonetheless, the two processes are often difficult to distinguish.

Herpes zoster is the best characterized viral myelitis, but HSV types 1 and 2, EBV, CMV, and rabies virus are other well-described causes and Zika virus has also been recognized as a cause of infectious myelitis. HSV-2 (and less commonly HSV-1) produces a distinctive syndrome of recurrent sacral cauda equina neuritis in association with outbreaks of genital herpes (Elsberg's syndrome). Poliomyelitis is the prototypic viral myelitis, but it is more or less restricted to the anterior gray matter of the cord containing the spinal motoneurons. A polio-like syndrome can also be caused by a large number of enteroviruses (including enterovirus 71 and coxsackie), and with Japanese encephalitis and other flaviviruses such as West Nile virus. Recently, cases of paralysis in children and adolescents were associated with enterovirus D-68 infection but a causal role for this virus has not been established. Chronic viral myelitic infections, such as those due to HIV or human T cell lymphotropic virus type 1 (HTLV-1), are discussed below.

Bacterial and mycobacterial myelitis (most are essentially abscesses) are less common than viral causes and much less frequent than cerebral bacterial abscess. Almost any pathogenic species may be responsible, including *Borrelia burgdorferi* (Lyme disease), *Listeria monocytogenes, Mycobacterium tuberculosis,* and *Treponema pallidum* (syphilis). *Mycoplasma pneumoniae* may be a cause of myelitis, but its status is uncertain because many cases are more properly classified as postinfectious.

Schistosomiasis **(Chap. 229)** is an important cause of parasitic myelitis in endemic areas. The process is intensely inflammatory and granulomatous, caused by a local response to tissue-digesting enzymes from the ova of the parasite, typically *Schistosoma hematobium* or *Schistosoma mansoni*. Toxoplasmosis **(Chap. 223)** can occasionally cause a focal myelopathy, and this diagnosis should especially be considered in patients with AIDS **(Chap. 197)**. *Cysticercosis* **(Chap. 230)** is another consideration, although myelitis from this helminth is far less common than parenchymal brain or meningeal involvement.

In cases of suspected viral myelitis, it may be appropriate to begin specific therapy pending laboratory confirmation. Herpes zoster, HSV, and EBV myelitis are treated with intravenous acyclovir (10 mg/kg q8h) or oral valacyclovir (2 g tid) for 10–14 days; CMV is treated with ganciclovir (5 mg/kg IV bid) plus foscarnet (60 mg/kg IV tid) or cidofovir (5 mg/kg per week for 2 weeks).

High-Voltage Electrical Injury Spinal cord injuries are prominent following electrocution from lightning strikes or other accidental electrical exposures. The syndrome consists of transient weakness acutely (often with an altered sensorium and focal cerebral disturbances), sometimes followed several days or even weeks later by a myelopathy that can be severe and permanent. This is a rare injury type, and limited data incriminate a vascular pathology involving the anterior spinal artery and its branches in some cases. Therapy is supportive.

CHRONIC MYELOPATHIES

SPONDYLOTIC MYELOPATHY

Spondylotic myelopathy is one of the most common causes of chronic cord compression and of gait difficulty in the elderly. Neck and shoulder pain with stiffness are early symptoms; impingement of bone and soft tissue overgrowth on nerve roots results in radicular arm pain, most often in a C5 or C6 distribution. Compression of the cervical cord, which occurs in fewer than one-third of cases, produces a slowly progressive spastic paraparesis, at times asymmetric and often accompanied by paresthesias in the feet and hands. Vibratory sense is diminished in the legs, there is a Romberg sign, and occasionally there is a sensory level for vibration or pinprick on the upper thorax. In some cases, coughing or straining produces leg weakness or radiating arm or shoulder pain. Dermatomal sensory loss in the arms, atrophy of intrinsic hand muscles, increased deep-tendon reflexes in the legs, and extensor plantar responses are common. Urinary urgency or incontinence occurs in advanced cases, but there are many alternative causes of these problems in older individuals. A tendon reflex in the arms is often diminished at some level; most often at the biceps (C5-C6). In individual cases, radicular, myelopathic, or combined signs may predominate. The diagnosis should be considered in appropriate cases of progressive cervical myelopathy, paresthesias of the feet and hands, or wasting of the hands.

Diagnosis is usually made by MRI and may be suspected from CT images; plain x-rays are less helpful. Extrinsic cord compression and deformation are appreciated on axial MRI views, and T2-weighted sequences may reveal areas of high signal intensity within the cord adjacent to the site of compression. A cervical collar may be helpful in milder cases, but the likelihood of progression of medically-treated myelopathy is high, estimated at 8% over 1 year. Definitive therapy consists of surgical decompression, either posterior laminectomy or an anterior approach with resection of the protruded disk and bony material. **Cervical spondylosis and related degenerative diseases of the spine are discussed in Chap. 14.**

VASCULAR MALFORMATIONS OF THE CORD AND DURA

Vascular malformations, comprising ~4% of all mass lesions of the cord and overlying dura, are treatable causes of progressive myelopathy. Most common are fistulas located within the dura or posteriorly along the surface of the cord. Most dural arteriovenous (AV) fistulas are located at or below the midthoracic level, usually consisting of a direct connection between a radicular feeding artery in the nerve root sleeve with dural veins. The typical presentation is a middle-aged man with a progressive myelopathy that worsens slowly or intermittently and may have periods of remission, sometimes mimicking MS. Acute deterioration due to hemorrhage into the spinal cord (hematomyelia) or subarachnoid space may also occur but is rare. In many cases, progression results from local ischemia and edema due to venous congestion. Most patients have incomplete sensory, motor, and bladder disturbances. The motor disorder may predominate and produce a mixture of upper and restricted lower motor neuron signs, simulating amyotrophic lateral sclerosis (ALS). Pain over the dorsal spine, dysesthesias, or radicular pain may be present. Other symptoms suggestive of AV malformation (AVM) or dural fistula include intermittent claudication; symptoms that change with posture, exertion, Valsalva maneuver, or menses; and fever.

Less commonly, AVM disorders are intramedullary rather than dural. One unusual disorder is a progressive thoracic myelopathy with paraparesis developing over weeks or months, characterized pathologically by abnormally thick, hyalinized vessels within the cord (subacute necrotic myelopathy, or Foix-Alajouanine syndrome).

Spinal bruits are infrequent but may be sought at rest and after exercise in suspected cases. A vascular nevus on the overlying skin may indicate an underlying vascular malformation as occurs with Klippel-Trenaunay-Weber syndrome. MR angiography and CT angiography can detect the draining vessels of many AVMs **(Fig. 434-6)**. Definitive diagnosis requires selective spinal angiography, which defines the feeding vessels and the extent of the malformation. Treatment is tailored to the anatomy and location of the lesion, and generally consists of microsurgical resection, endovascular embolization of the major feeding vessels, or a combination of the two approaches.

RETROVIRUS-ASSOCIATED MYELOPATHIES

The myelopathy associated with HTLV-1, formerly called tropical spastic paraparesis, is a slowly progressive spastic syndrome with variable sensory and bladder disturbance. Approximately half of patients have mild back or leg pain. The neurologic signs may be asymmetric, often lacking a well-defined sensory level; the only sign in the arms may be hyperreflexia after several years of illness. The onset is usually insidious, and the tempo of progression of the illness occurs at a variable rate; in one study, median time for progression to cane, walker, or wheelchair dependent state was 6, 13, and 21 years, respectively. Progression appears to be more rapid in older patients and those with higher viral loads. Diagnosis is made by demonstration of HTLV-1-specific antibody in serum by enzyme-linked immunosorbent assay (ELISA), confirmed by radioimmunoprecipitation or Western blot analysis. Especially in endemic areas, a finding of HTLV-1 seropositivity in a patient with myelopathy does not necessarily prove that HTLV-1

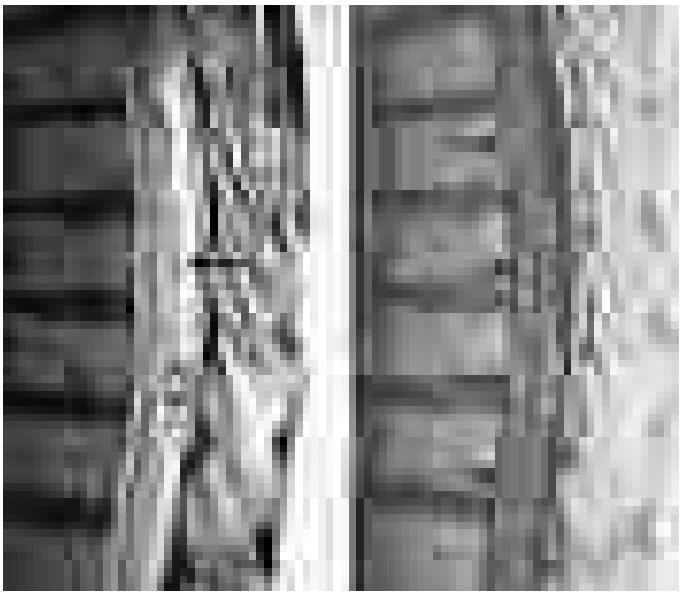


FIGURE 434-6 Arteriovenous malformation. Sagittal magnetic resonance scans of the thoracic spinal cord: T2 fast spin-echo technique (*left*) and T1 postcontrast image (*right*). On the T2-weighted image (*left*), abnormally high signal intensity is noted in the central aspect of the spinal cord (*arrowheads*). Numerous punctate flow voids indent the dorsal and ventral spinal cord (*arrow*). These represent the abnormally dilated venous plexus supplied by a dural arteriovenous fistula. After contrast administration (*right*), multiple, serpentine, enhancing veins (*arrows*) on the ventral and dorsal aspect of the thoracic spinal cord are visualized, diagnostic of arteriovenous malformation. This patient was a 54-year-old man with a 4-year history of progressive paraparesis.

FIGURE 434-7 Magnetic resonance imaging of syringomyelia associated with a Chiari malformation. Sagittal T1-weighted image through the cervical and upper thoracic spine demonstrates descent of the cerebellar tonsils below the level of the foramen magnum (*black arrows*). Within the substance of the cervical and thoracic spinal cord, a cerebrospinal fluid collection dilates the central canal (*white arrows*).

is causative. The CSF/serum antibody index may provide support by establishing intrathecal synthesis of antibodies, including oligoclonal antibodies, favoring HTVL-1 myelopathy over asymptomatic carriage. Measuring proviral DNA by polymerase chain reaction (PCR) in serum and CSF cells can be useful as an ancillary part of diagnosis. The pathogenesis of the myelopathy is uncertain. It could result from an immune response directed against HTLV-1 antigens in the nervous system, or alternatively to secondary autoimmunity triggered by the viral infection. There is no proven effective treatment. Based on limited evidence, the use of chronic low dose oral glucocorticoids can be tried; interferon is of uncertain value, and antiviral treatment is ineffective. Symptomatic therapy for spasticity and bladder symptoms may be helpful.

A progressive myelopathy can also result from HIV infection **(Chap. 197)**. It is characterized by vacuolar degeneration of the posterior and lateral tracts, resembling subacute combined degeneration (see below).

SYRINGOMYELIA

Syringomyelia is a developmental cavity of the cervical cord that may enlarge and produce progressive myelopathy or may remain asymptomatic. Symptoms begin insidiously in adolescence or early adulthood, progress irregularly, and may undergo spontaneous arrest for several years. Many young patients acquire a cervical-thoracic scoliosis. More than half of all cases are associated with Chiari type 1 malformations in which the cerebellar tonsils protrude through the foramen magnum and into the cervical spinal canal. The pathophysiology of syrinx expansion is controversial, but some interference with the normal flow of CSF seems likely, perhaps by the Chiari malformation. Acquired cavitations of the cord in areas of necrosis are also termed *syrinx cavities*; these follow trauma, myelitis, necrotic spinal cord tumors, and chronic arachnoiditis due to tuberculosis and other etiologies.

The presentation is a central cord syndrome consisting of a regional dissociated sensory loss (loss of pain and temperature sensation with sparing of touch and vibration) and areflexic weakness in the upper limbs. The sensory deficit has a distribution that is "suspended" over the nape of the neck, shoulders, and upper arms (cape distribution) or in the hands. Most cases begin asymmetrically with unilateral sensory loss in the hands that leads to injuries and burns that are not appreciated by the patient. Muscle wasting in the lower neck, shoulders, arms, and hands with asymmetric or absent reflexes in the arms reflects expansion of the cavity in the gray matter of the cord. As the cavity enlarges and compresses the long tracts, spasticity and weakness of the legs, bladder and bowel dysfunction, and a Horner's syndrome appear. Some patients develop facial numbness and sensory loss from damage to the descending tract of the trigeminal nerve (C2 level or above). In cases with Chiari malformations, cough-induced headache and neck, arm, or facial pain may be reported. Extension of the syrinx into the medulla, syringobulbia, causes palatal or vocal cord paralysis, dysarthria, horizontal or vertical nystagmus, episodic dizziness or vertigo, and tongue weakness with atrophy.

MRI accurately identifies developmental and acquired syrinx cavities and their associated spinal cord enlargement **(Fig. 434-7)**. Images of the brain and the entire spinal cord should be obtained to delineate the full longitudinal extent of the syrinx, assess posterior fossa structures for the Chiari malformation, and determine whether hydrocephalus is present.

TREATMENT

Syringomyelia

Treatment of syringomyelia is generally unsatisfactory. The Chiari tonsillar herniation may be decompressed, generally by suboccipital craniectomy, upper cervical laminectomy, and placement of a dural graft. Fourth ventricular outflow is reestablished by this procedure. If the syrinx cavity is large, some surgeons recommend direct decompression or drainage, but the added benefit of this procedure is uncertain, and complications are common. With Chiari malformations, shunting of hydrocephalus generally precedes any attempt to correct the syrinx. Surgery may stabilize the neurologic deficit, and some patients improve. Patients with few symptoms and signs from the syrinx do not require surgery and are followed by serial clinical and imaging examinations.

Syrinx cavities secondary to trauma or infection, if symptomatic, are treated with a decompression and drainage procedure in which a small shunt is inserted between the cavity and subarachnoid space;

alternatively, the cavity can be fenestrated. Cases due to intramedullary spinal cord tumor are generally managed by resection of the tumor.

CHRONIC MYELOPATHY OF MULTIPLE SCLEROSIS

A chronic progressive myelopathy is the most frequent cause of disability in both primary progressive and secondary progressive forms of MS. Involvement is typically bilateral but asymmetric and produces motor, sensory, and bladder/bowel disturbances. Fixed motor disability appears to result from extensive loss of axons in the corticospinal tracts. Diagnosis is facilitated by identification of earlier attacks such as optic neuritis. MRI, CSF, and evoked response testing are confirmatory. Treatment with ocrelizumab, an anti-CD20 B-cell monoclonal antibody, is effective in patients with primary progressive MS, and disease modifying therapy is also indicated in patients with secondary progressive MS who have coexisting MS relapses. **MS is discussed in Chap. 436.**

SUBACUTE COMBINED DEGENERATION (VITAMIN B_{12} DEFICIENCY)

This treatable myelopathy presents with subacute paresthesias in the hands and feet, loss of vibration and position sensation, and a progressive spastic and ataxic weakness. Loss of reflexes due to an associated peripheral neuropathy in a patient who also has Babinski signs is an important diagnostic clue. Optic atrophy and irritability or other cognitive changes may be prominent in advanced cases and are occasionally the presenting symptoms. The myelopathy of subacute combined degeneration tends to be diffuse rather than focal; signs are generally symmetric and reflect predominant involvement of the posterior and lateral tracts, including Romberg's sign. Causes include dietary deficiency, especially in vegans, and gastric malabsorption syndromes including pernicious anemia **(Chap. 95)**. The diagnosis is confirmed by the finding of macrocytic red blood cells, a low serum B_{12} concentration, and elevated serum levels of homocysteine and methylmalonic acid. Treatment is by replacement therapy, beginning with 1000 µg of intramuscular vitamin B_{12} repeated at regular intervals or by subsequent oral treatment.

HYPOCUPRIC MYELOPATHY

This myelopathy is similar to subacute combined degeneration (described above), except there is no neuropathy, and explains cases with normal serum levels of B_{12}. Low levels of serum copper are found, and often there is also a low level of serum ceruloplasmin. Some cases follow gastrointestinal procedures, particularly bariatric surgery, that result in impaired copper absorption; others have been associated with excess zinc from health food supplements or in the past zinc-containing denture creams, all of which impair copper absorption via induction of metallothionein, a copper-binding protein. Many cases are idiopathic. Improvement or at least stabilization may be expected with reconstitution of copper stores by oral supplementation. There is microcytic or macrocytic anemia. The pathophysiology and pathology of the idiopathic form are not known.

TABES DORSALIS

The classic syphilitic syndromes of tabes dorsalis and meningovascular inflammation of the spinal cord are now less frequent than in the past but must be considered in the differential diagnosis of spinal cord disorders. The characteristic symptoms of tabes are fleeting and repetitive lancinating pains, primarily in the legs or less often in the back, thorax, abdomen, arms, and face. Ataxia of the legs and gait due to loss of position sense occurs in half of patients. Paresthesias, bladder disturbances, and acute abdominal pain with vomiting (visceral crisis) occur in 15–30% of patients. The cardinal signs of tabes are loss of reflexes in the legs; impaired position and vibratory sense; Romberg sign; and, in almost all cases, bilateral Argyll Robertson pupils, which fail to constrict to light but accommodate. Diabetic polyradiculopathy may simulate this condition. Treatment of tabes dorsalis and other forms of neurosyphilis consists of penicillin G administered intravenously, or intramuscularly in combination with oral probenecid **(Chap. 177)**.

HEREDITARY SPASTIC PARAPLEGIA

Many cases of slowly progressive myelopathy are genetic in origin **(Chap. 429)**. More than 60 different causative loci have been identified, including autosomal dominant, autosomal recessive, and X-linked forms. Especially for the recessive and X-linked forms, a family history of myelopathy may be lacking. Most patients present with almost imperceptibly progressive spasticity and weakness in the legs, usually but not always symmetrical. Sensory symptoms and signs are absent or mild, but sphincter disturbances may be present. In some families, additional neurologic signs are prominent, including nystagmus, ataxia, or optic atrophy. The onset may be as early as the first year of life or as late as middle adulthood. Only symptomatic therapies are available.

ADRENOMYELONEUROPATHY

This X-linked disorder is a variant of adrenoleukodystrophy (ALD). Most affected males have a history of adrenal insufficiency and then develop a progressive spastic (or ataxic) paraparesis beginning in early or sometimes middle adulthood; some patients also have a mild peripheral neuropathy. Female heterozygotes may develop a slower, insidiously progressive spastic myelopathy beginning later in adulthood and without adrenal insufficiency. Diagnosis is usually made by demonstration of elevated levels of very-long-chain fatty acids in plasma and in cultured fibroblasts. The responsible gene encodes the adrenoleukodystrophy protein (ALDP), a peroxisomal membrane transporter involved in carrying long-chain fatty acids to peroxisomes for degradation. Corticosteroid replacement is indicated if hypoadrenalism is present. Allogeneic bone marrow transplantation has been successful in slowing progression of cognitive decline in ALD, but appears to be ineffective for the myelopathy of ALD. Nutritional supplements (Lorenzo's oil) have also been attempted for this condition without evidence of efficacy.

OTHER CHRONIC MYELOPATHIES

Primary lateral sclerosis **(Chap. 429)** is a mid to late life onset degenerative disorder characterized by progressive spasticity with weakness, eventually accompanied by dysarthria and dysphonia; bladder symptoms occur in approximately half of patients. Sensory function is spared. The disorder resembles ALS and is considered a variant of the motor neuron degenerations, but without the characteristic lower motor neuron disturbance and with typically a slower progression. Some cases may represent late-onset cases of familial spastic paraplegia, particularly autosomal recessive or X-linked varieties in which a family history may be absent.

Tethered cord syndrome is a developmental disorder of the lower spinal cord and nerve roots that rarely presents in adulthood as low back pain accompanied by a progressive lower spinal cord and/or nerve root syndrome. Some patients have a small leg or foot deformity indicating a long-standing process, and in others, a dimple, patch of hair, or sinus tract on the skin overlying the lower back is the clue to a congenital lesion. Diagnosis is made by MRI, which demonstrates a low-lying conus medullaris and thickened filum terminale. The MRI may also reveal diastematomyelia (division of the lower spinal cord into two halves), lipomas, cysts, or other congenital abnormalities of the lower spine coexisting with the tethered cord. Treatment is with surgical release.

There are a number of rare toxic causes of spastic myelopathy, including lathyrism due to ingestion of chickpeas containing the excitotoxin β-*N*-oxalylamino-L-alanine (BOAA), seen primarily in the developing world, and nitrous oxide inhalation producing a myelopathy identical to subacute combined degeneration. SLE, Sjögren's syndrome, and sarcoidosis may each cause a myelopathy without overt evidence of systemic disease. Cancer-related causes of chronic myelopathy, besides the common neoplastic compressive myelopathy discussed earlier, include radiation injury **(Chap. 86)** and rare paraneoplastic myelopathies. The last of these are most often associated with lung cancer and anti-Hu or anti-CV2/CRMP5 antibodies **(Chap. 90)** or with lymphoma that causes a syndrome of destruction of anterior horn cells; NMO with aquaporin-4 antibodies **(Chap. 437)** can also rarely be paraneoplastic in

origin. Metastases to the cord are probably more common than either of these in patients with cancer. Often, a cause of intrinsic myelopathy can be identified only through periodic reassessment.

REHABILITATION OF SPINAL CORD DISORDERS

The prospects for recovery from an acute destructive spinal cord lesion fade after ~6 months. There are currently no effective means to promote repair of injured spinal cord tissue; promising but entirely experimental approaches include the use of factors that influence reinnervation by axons of the corticospinal tract, nerve and neural sheath graft bridges, forms of electrical stimulation at the site of injury, and the local introduction of stem cells. The disability associated with irreversible spinal cord damage is determined primarily by the level of the lesion and by whether the disturbance in function is complete or incomplete **(Table 434-4)**. Even a complete high cervical cord lesion may be compatible with a productive life. The primary goals are development of a rehabilitation plan framed by realistic expectations and attention to the neurologic, medical, and psychological complications that commonly arise.

Many of the usual symptoms associated with medical illnesses, especially somatic and visceral pain, may be lacking because of the destruction of afferent pain pathways. Unexplained fever, worsening of spasticity, or deterioration in neurologic function should prompt a search for infection, thrombophlebitis, or an intraabdominal pathology. The loss of normal thermoregulation and inability to maintain normal body temperature can produce recurrent fever (*quadriplegic fever*), although most episodes of fever are due to infection of the urinary tract, lung, skin, or bone.

Bladder dysfunction generally results from loss of supraspinal innervation of the detrusor muscle of the bladder wall and the sphincter musculature. Detrusor spasticity is treated with anticholinergic drugs (oxybutynin, 2.5–5 mg qid) or tricyclic antidepressants with anticholinergic properties (imipramine, 25–200 mg/d). Failure of the sphincter muscle to relax during bladder emptying (urinary dyssynergia) may be managed with the α-adrenergic blocking agent terazosin hydrochloride (1–2 mg tid or qid), with intermittent catheterization, or, if that is not feasible, by use of a condom catheter in men or a permanent indwelling catheter. Surgical options include the creation of an artificial bladder by isolating a segment of intestine that can be catheterized intermittently (enterocystoplasty) or can drain continuously to an external appliance (urinary conduit). Bladder areflexia due to acute spinal shock or conus lesions is best treated by catheterization. Bowel regimens and disimpaction are necessary in most patients to ensure at least biweekly evacuation and avoid colonic distention or obstruction.

Patients with acute cord injury are at risk for venous thrombosis and pulmonary embolism. Use of calf-compression devices and anticoagulation with low-molecular-weight heparin is recommended. In cases of persistent paralysis, anticoagulation should probably be continued for 3 months.

Prophylaxis against decubitus ulcers should involve frequent changes in position in a chair or bed, the use of special mattresses, and cushioning of areas where pressure sores often develop, such as the sacral prominence and heels. Early treatment of ulcers with careful cleansing, surgical or enzyme debridement of necrotic tissue, and appropriate dressing and drainage may prevent infection of adjacent soft tissue or bone.

Spasticity is aided by stretching exercises to maintain mobility of joints. Drug treatment is effective but may result in reduced function, as some patients depend on spasticity as an aid to stand, transfer, or walk. Baclofen (up to 240 mg/d in divided doses) is effective; it acts by facilitating γ-aminobutyric acid–mediated inhibition of motor reflex arcs. Diazepam acts by a similar mechanism and is useful for leg spasms that interrupt sleep (2–4 mg at bedtime). Tizanidine (2–8 mg tid), an α_2 adrenergic agonist that increases presynaptic inhibition of motor neurons, is another option. For nonambulatory patients, the direct muscle inhibitor dantrolene (25–100 mg qid) may be used, but it is potentially hepatotoxic. In refractory cases, intrathecal baclofen administered via an implanted pump, botulinum toxin injections, or dorsal rhizotomy may be required to control spasticity.

Despite the loss of sensory function, many patients with spinal cord injury experience chronic pain sufficient to diminish their quality of life. Randomized controlled studies indicate that gabapentin or pregabalin is useful in this setting. Epidural electrical stimulation and intrathecal infusion of pain medications have been tried with some success. **Management of chronic pain is discussed in Chap. 10.**

A paroxysmal autonomic hyperreflexia may occur following lesions above the major splanchnic sympathetic outflow at T6. Headache, flushing, and diaphoresis above the level of the lesion, as well as hypertension with bradycardia or tachycardia, are the major symptoms. The trigger is typically a noxious stimulus—for example, bladder or bowel distention, a urinary tract infection, or a decubitus ulcer—below the level of the cord lesion. Treatment consists of removal of offending stimuli; ganglionic blocking agents (mecamylamine, 2.5–5 mg) or other short-acting antihypertensive drugs are useful in some patients.

Attention to these details allows longevity and a productive life for patients with complete transverse myelopathies.

TABLE 434-4 Expected Neurologic Function Following Complete Cord Lesions

LEVEL	SELF-CARE	TRANSFERS	MAXIMUM MOBILITY
High quadriplegia (C1-C4)	Dependent on others; requires respiratory support	Dependent on others	Motorized wheelchair
Low quadriplegia (C5-C8)	Partially independent with adaptive equipment	May be dependent or independent	May use manual wheelchair, drive an automobile with adaptive equipment
Paraplegia (below T1)	Independent	Independent	Ambulates short distances with aids

Source: Adapted from JF Ditunno, CS Formal: Chronic spinal cord injury. N Engl J Med 330:550, 1994; with permission.

FURTHER READING

Kato S et al: Comparison of anterior and posterior surgery for degenerative cervical myelopathy: An MRI-based propensity-score-matched analysis using data from the prospective multicenter AOSpine CSM North America and international studies. J Bone Joint Surg Am 99:1013, 2017.

Kühl JS et al: Long-term outcomes of allogeneic haematopoietic stem cell transplantation for adult cerebral X-linked adrenoleukodystrophy. Brain 140:953, 2017.

Loblaw DA et al: A 2011 updated systematic review and clinical practice guideline for the management of malignant extradural spinal cord compression. Int J Radiat Oncol Biol Phys 84:312, 2012.

Ozpinar A et al: Epidemiology, clinical presentation, diagnostic evaluation, and prognosis of spinal arteriovenous malformations. Handb Clin Neurol 143:145, 2017.

Patchell RA et al: Direct decompressive surgical resection in the treatment of spinal cord compression caused by metastatic cancer: A randomised trial. Lancet 366:643, 2005.

Robertson CE et al: Recovery after spinal cord infarcts: Long-term outcome in 115 patients. Neurology 78:114, 2012.

Ropper AE, Ropper AH: Acute spinal cord compression. N Engl J Med 376:1358, 2017.

Ruet A et al: Predictive factors for multiple sclerosis in patients with clinically isolated spinal cord syndrome. Mult Scler 17:312, 2011.

Yáñez ML et al: Diagnosis and treatment of epidural metastases. Cancer 123:1106, 2017.

435 Concussion and Other Traumatic Brain Injuries

Geoffrey T. Manley, Stephen L. Hauser, Michael McCrea

INTRODUCTION

Traumatic brain injury (TBI) represents a significant global public health problem facing the United States and other countries around the world. In the United States, estimates of the frequency of TBI range between 2.5 and 4 million cases per year, depending on the study and methods used to define and include cases. Age-specific rates show a bimodal distribution, with highest risk in younger individuals and older adults. The most common mechanism of injury in the young is motor vehicle accidents and is more common in men, while in older adults falls are the major cause of injury and are more likely to occur in women.

TBI imposes substantial demands on health care systems. Worldwide, at least 10 million TBIs are serious enough to result in death or hospitalization. In the United States, the estimated annual cost is >$76 billion. Due to advances in medical care and other factors, more people are surviving traumatic brain injury than ever before. Brain injury accounts for more lost productivity at work among Americans than any other form of injury. An estimated 5.3 million Americans are living with significant disabilities resulting from TBI that complicate their return to a full and productive life. Increased media attention to military and sports-related TBI has highlighted the growing concern that injuries that were previously dismissed can have life-long consequences for some individuals.

Head injuries are so common that almost all physicians will be called upon to provide some aspect of immediate care or to see patients who are suffering from various sequelae. Patients initially need education regarding the natural history of TBI along with treatment of acute symptoms such as headache. Follow-up of TBI patients is important to make sure that the sequelae that some patients experience—such as postconcussive disorder (PCD), depression, or sleep disorders—can be identified and treated by a coordinated multidisciplinary team.

DEFINITION AND CLASSIFICATION

TBI is commonly defined as *an alteration in brain function, or other evidence of brain pathology, caused by an external force, and characterized by the following: (1) any period of loss or decreased level of consciousness (LOC), (2) any loss of memory for events immediately before (retrograde) or after (posttraumatic) the injury, (3) any neurological deficits, and/or (4) any alteration in mental state at the time of injury.*

Evidence of TBI can include visual, neuroradiologic, or laboratory confirmation of damage to the brain, but TBI is more often diagnosed on the basis of acute clinical criteria. In addition to standard computed tomography (CT) imaging, modern structural magnetic resonance imaging (MRI), and functional imaging (resting state functional MRI) techniques show increasing sensitivity, and it is likely that sensitive blood-based biomarkers will be developed in the near future.

Mechanisms of TBI: Common mechanisms of TBI include the head being struck by an object, the head striking an object, the brain undergoing an acceleration/deceleration movement, a foreign body penetrating the brain, or forces generated from events such as a blast or explosion. Motor vehicle crashes have historically been cited as the most common cause of TBI. All forms of transportation, however, are common causes of TBI, including motorcycle crashes, bicycle accidents, skateboard and pedestrian injuries. The other leading causes of TBI are falls, assaults, and sports, with varied frequency across the lifespan. Certainly, there has been an increased focus on the high frequency of mild TBI (mTBI), often referred to as concussion, encountered by athletes participating in contact and collision sports at all competitive levels, as well as the potential short-term effects and long-term risks associated with sport-related concussion.

Classification of TBI Severity: Numerous systems have been developed over the years to define and classify TBI severity along a continuum from mild to moderate to severe. These systems are usually most applicable to closed head injuries. In nearly all classification systems, traumatic brain injury severity is graded based on acute injury characteristics rather than postacute injury status, as other factors can intervene to influence functional outcome. Historically, the presence and duration of unconsciousness and amnesia have been the main points of distinction along the gradient of TBI severity. Current TBI classification systems are symptom-based and do not incorporate patho-anatomical or molecular features.

The Glasgow Coma Scale (GCS) is the most recognized and widely used method for grading TBI severity. The GCS provides a practical indicator of gross neurologic status by assessing motor function, verbal responses, and the patient's ability to open his or her eyes voluntarily or in response to external commands and stimuli. The grading is applied to the best response that can be elicited from the patient at the time of assessment, preferably before any paralyzing or sedating medication is administered or the patient is intubated, as these interventions confound interpretation of the score. The GCS assessment produces scores ranging from 3 to 15 **(Table 435-1)**.

Upon the 40th anniversary of the GCS, the wording for responses was revised, and recommendations were made to improve its utility. Importantly, individual patients are best described by the three components of the coma scale (eye, verbal, motor, e.g., E3V4M6); the derived total coma score (e.g., 13) is less informative and should only be used to characterize groups of patients.

Several injury classification systems have been developed to go beyond GCS score or acute injury characteristics and incorporate chief signs and symptoms in defining mTBI. The use of multiple severity indicators is intended to improve sensitivity in the detection of mTBI (GCS 13–15), while also taking into consideration traditional acute injury characteristics that have been presumed to predict outcome following mild and moderate brain injury. Loss of consciousness (LOC) and posttraumatic amnesia (PTA) remain the most common injury characteristics referenced in these classification systems. In the case of moderate (GCS 9–12) and severe (GCS 3–8) TBI, GCS score and the duration of LOC and PTA can be robust predictors of long-term outcome and morbidity. In cases of mTBI, however, while PTA and LOC are important indicators of acute injury severity, they are less predictive of eventual recovery time and outcome.

TBI TYPES AND PATHOLOGIES

Mild TBI (Concussion): It is estimated that between 70 and 90% of all treated traumatic brain injuries are mild in severity based on traditional case definitions and acute injury characteristics, with most reported estimates in the order of 85%. The published figures likely under-represent the true incidence of mTBI because of variable case definitions and heterogeneous methods. Moreover, because a subgroup

TABLE 435-1 Glasgow Coma Scale

Eye Opening (E)		Verbal Response (V)	
Spontaneous	4	Oriented	5
To speech	3	Confused	4
To pressure	2	Words	3
None	1	Sounds	2
		None	1
Best Motor Response (M)			
Obeying commands	6		
Localizing	5		
Normal flexion	4		
Abnormal flexion	3		
Extension	2		
None	1		

Note: Revised GCS (2014).

Source: From G Teasdale et al: The Glasgow Coma Scale at 40 years: Standing the test of time. Lancet Neurol 13:844, 2014.

of individuals with milder brain injuries do not seek medical attention, epidemiological studies that depend on hospital-based data also underestimate the true incidence.

The term concussion, while popular, is vague and is not based on widely accepted objective criteria, resulting in multiple definitions from various groups. There has been debate as to whether concussion is part of the TBI spectrum or a separate entity. In 2017, the Concussion in Sports Group issued a consensus statement that "concussion is a traumatic brain injury" (McCrory et al, 2017). By firmly placing concussion in the spectrum of TBI, the underlying pathophysiological processes common to all TBI presentations can now be considered together.

SKULL FRACTURE, EXTRA-AXIAL HEMATOMA, CONTUSION, AND AXONAL INJURY

Skull Fracture A blow to the skull that exceeds the elastic tolerance of the bone causes a fracture. Intracranial lesions accompany roughly two-thirds of skull fractures, and the presence of a fracture increases many-fold the chances of an underlying subdural or epidural hematoma. Consequently, fractures are primarily markers of the site and severity of injury. If the underlying arachnoid membrane has been torn, fractures also provide potential pathways for entry of bacteria to the cerebrospinal fluid (CSF) with a risk of meningitis and for leakage of CSF outward through the dura. If there is leakage of CSF, severe orthostatic headache results from lowered pressure in the spinal fluid compartment.

Most fractures are linear and extend from the point of impact toward the base of the skull. Basilar skull fractures are often extensions of adjacent linear fractures over the convexity of the skull but may occur independently owing to stresses on the floor of the middle cranial fossa or occiput. Basilar fractures are usually parallel to the petrous bone or along the sphenoid bone and directed toward the sella turcica and ethmoidal groove. Although most basilar fractures are uncomplicated, they can cause CSF leakage, pneumocephalus, and delayed cavernous-carotid fistulas. Hemotympanum (blood behind the tympanic membrane), ecchymosis over the mastoid process (Battle sign), and periorbital ecchymosis ("raccoon sign") are associated with basilar fractures and should be suspected if these clinical signs are present.

EPIDURAL AND SUBDURAL HEMATOMAS

Hemorrhages between the dura and skull (epidural) or beneath the dura (subdural) have characteristic clinical and imaging features. They are sometimes associated with underlying brain contusions and other injuries, often making it difficult to determine the relative contribution of each component to the clinical state. The mass effect of raised intracranial pressure (ICP) caused by these hematomas can be life threatening, making it imperative to identify them rapidly by CT or MRI scan and to surgically remove them when appropriate.

Epidural Hematoma (Fig. 435-1) These highly dangerous lesions usually arise from an injury to a meningeal arterial vessel and evolve rapidly. They are often accompanied by a "lucid interval" of several minutes to hours prior to neurological deterioration. They occur in up to 10% of cases of severe head injury, but are less often associated with underlying cortical damage compared to subdural hematomas. Rapid surgical evacuation and ligation or cautery of the damaged vessel, usually the middle meningeal artery that has been lacerated by an overlying skull fracture, is indicated. If recognized and treated rapidly, patients often have a favorable outcome.

Acute Subdural Hematoma (Fig. 435-2) Direct cranial trauma may be minor and is not always required for acute subdural hemorrhage to occur, especially in the elderly and those taking anticoagulant medications. Acceleration forces alone, as from whiplash, are sometimes sufficient to produce subdural hematoma. Up to one-third of patients have a lucid interval lasting minutes to hours before coma supervenes, but most are drowsy or comatose from the moment of injury. A unilateral headache and slightly enlarged pupil on the side of the hematoma are frequently, but not invariably, present.

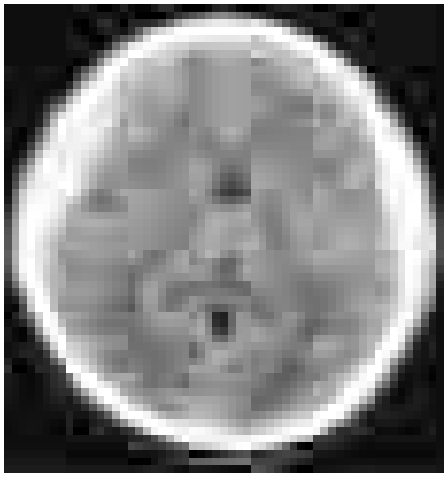


FIGURE 435-1 Acute epidural hematoma. The tightly attached dura is stripped from the inner table of the skull, producing a characteristic lenticular-shaped hemorrhage on noncontrast computed tomography scan. Epidural hematomas are usually caused by tearing of the middle meningeal artery following fracture of the temporal bone.

Small subdural hematomas may be asymptomatic and usually do not require surgical evacuation if they do not enlarge. Stupor or coma, hemiparesis, and unilateral pupillary enlargement are signs of larger hematomas. The bleeding that causes larger subdural hematomas is primarily venous in origin, although arterial bleeding sites are sometimes found at operation, and a few large hematomas have a purely arterial origin. In an acutely deteriorating patient, an emergency craniotomy is required. In contrast to epidural hematomas, there is significant morbidity and mortality associated with acute subdural hematomas that require surgery.

Chronic Subdural Hematoma A subacutely evolving syndrome due to subdural hematoma occurs days or weeks after injury with drowsiness, headache, confusion, or mild hemiparesis, usually in the elderly with age-related atrophy and often after only minor or unnoticed trauma. On imaging studies, chronic subdural hematomas appear as crescentic clots over the convexity of one or both hemispheres, most commonly in the frontotemporal region (Fig. 435-3). A history of trauma may or may not be elicited in relation to chronic subdural hematoma; the injury may have been trivial and forgotten, particularly in the elderly and those with clotting disorders. Headache is common but not invariable. Additional features that may appear weeks later include slowed thinking, vague change in personality,

FIGURE 435-2 Acute subdural hematoma. Noncontrast computed tomography scan reveals a hyperdense clot that has an irregular border with the brain and causes more horizontal displacement (mass effect) than might be expected from its thickness. The disproportionate mass effect is the result of the large rostral-caudal extent of these hematomas. Compare to Fig. 435-1.

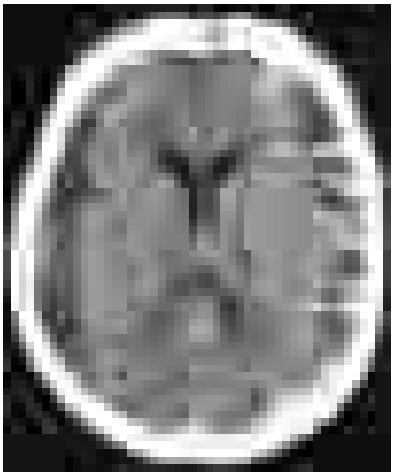


FIGURE 435-3 Computed tomography scan of chronic bilateral subdural hematomas of different ages. The collections began as acute hematomas and have become hypodense in comparison to the adjacent brain after a period during which they were isodense and difficult to appreciate. Some areas of resolving blood are contained on the more recently formed collection on the left (*arrows*).

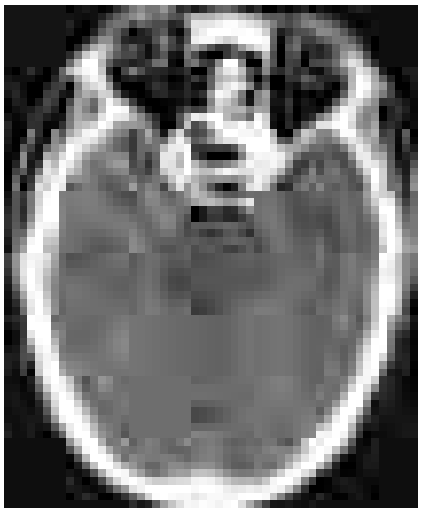


FIGURE 435-4 Traumatic cerebral contusion. Noncontrast computed tomography scan demonstrating a hyperdense hemorrhagic region in the anterior temporal lobe.

seizure, or a mild hemiparesis. The headache typically fluctuates in severity, sometimes with changes in head position. Drowsiness, inattentiveness, and incoherence of thought are generally more prominent than focal signs such as hemiparesis. Rarely, chronic hematomas cause brief episodes of hemiparesis or aphasia that are indistinguishable from transient ischemic attacks.

CT without contrast initially shows a low-density mass over the convexity of the hemisphere. Between 2 and 6 weeks after the initial bleeding, the clot becomes isodense compared to adjacent brain and may be inapparent. Many subdural hematomas that are several weeks in age contain areas of blood and intermixed serous fluid. Infusion of contrast material demonstrates enhancement of the vascular fibrous capsule surrounding the collection. MRI reliably identifies both subacute and chronic hematomas.

Clinical observation coupled with serial imaging is a reasonable approach to patients with few symptoms and small chronic subdural collections that do not cause mass effect. Treatment with surgical evacuation through burr holes is usually successful, if a cranial drain is used postoperatively. The fibrous membranes that grow from the dura and encapsulate the collection may require removal with a craniotomy to prevent recurrent fluid accumulation.

TRAUMATIC SUBARACHNOID HEMORRHAGE

Subarachnoid hemorrhage (SAH) is common in TBI. Rupture of small cortical arteries or veins can cause bleeding into the subarachnoid space. Traumatic SAH is often seen in the sulci and is frequently the only radiographic finding on CT following mild TBI. SAH occurs diffusely after severe TBI and confers an increase in mortality. In mild TBI, SAH provides an objective imaging biomarker for TBI, and in some patients is associated with unfavorable outcomes.

Contusion (Fig. 435-4) A surface bruise of the brain, or contusion, consists of varying degrees of petechial hemorrhage, edema, and tissue destruction. Contusions and deeper hemorrhages result from mechanical forces that displace and compress the hemispheres forcefully and by deceleration of the brain against the inner skull, either under a point of impact (coup lesion) or, as the brain swings back, in the antipolar area (contrecoup lesion). Trauma sufficient to cause prolonged unconsciousness usually produces some degree of contusion. Blunt deceleration impact, as occurs against an automobile dashboard or from falling forward onto a hard surface, causes contusions on the orbital surfaces of the frontal lobes and the anterior and basal portions of the temporal lobes. With lateral forces, as from impact on an automobile door frame, contusions are situated on the lateral convexity of the hemisphere. The clinical signs of contusion are determined by the location and size of the lesion; often, there are no focal abnormalities with a routine neurological exam, but these injured regions are later the sites of gliotic scars that may produce seizures. A hemiparesis or gaze preference is fairly typical of moderately sized contusions. Large bilateral contusions produce stupor with extensor posturing, while those limited to the frontal lobes cause a taciturn state. Contusions in the temporal lobe may cause delirium or an aggressive, combative syndrome. Torsional or shearing forces within the brain can cause hemorrhages of the basal ganglia and other deep regions. Large contusions and hemorrhages after minor trauma should raise concerns for coagulopathy due to an underlying disease or more commonly anticoagulant therapy.

Acute contusions are easily visible on CT and MRI scans, appearing as inhomogeneous hyperdensities on CT and as hyperintensities on T2 and fluid-attenuated inversion recovery (FLAIR) MRI sequences; there is usually surrounding localized brain edema and some subarachnoid bleeding. Blood in the CSF due to trauma may provoke a mild inflammatory reaction. Over a few days, contusions acquire a surrounding contrast enhancement and edema that may be mistaken for tumor or abscess.

Axonal Injury (Fig. 435-5) Traumatic axonal injury (TAI) is one of the most common injuries after TBI. There is disruption, or shearing, of axons at the time of impact and this is associated with microhemorrhages. It occurs following high-speed deceleration injuries, such as motor vehicle collisions (Johnson et al, 2013). The presence of ≥4 areas

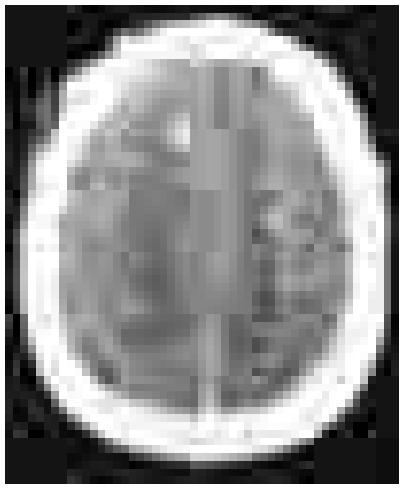


FIGURE 435-5 Multiple small areas of hemorrhage and tissue disruption in the white matter of the frontal lobes on noncontrast computed tomography scan. These appear to reflect an extreme type of the diffuse axonal shearing lesions that occur with closed head injury.

of TAI is called diffuse axonal injury (DAI), and when widespread, has been proposed to explain persistent coma and the vegetative state after TBI (Chap. 300). Only severe TAI lesions that contain substantial blood are visualized by CT, usually in the corpus callosum and centrum semiovale. More commonly, the CT will be negative for TAI, but subsequent MRI, particularly gradient-echo or susceptibility weighted imaging, will show hemosiderin deposits reflective of microhemorrhages in addition to the axonal damage on diffusion sequences.

CRANIAL NERVE INJURIES

The cranial nerves most often injured with TBI are the olfactory, optic, oculomotor, and trochlear; the first and second branches of the trigeminal nerve; and the facial and auditory nerves. Anosmia and an apparent loss of taste (actually a loss of perception of aromatic flavors, with retained elementary taste perception) occur in ~10% of persons with serious head injuries, particularly from falls on the back of the head. This is the result of displacement of the brain and shearing of the fine olfactory nerve filaments that course through the cribriform bone. At least partial recovery of olfactory and gustatory function is expected, but if bilateral anosmia persists for several months, the prognosis is poor. Partial optic nerve injuries from closed trauma result in blurring of vision, central or paracentral scotomas, or sector defects. Direct orbital injury may cause short-lived blurred vision for close objects due to reversible iridoplegia. Diplopia limited to downward gaze and corrected when the head is tilted away from the side of the affected eye indicates trochlear (fourth nerve) nerve damage. It occurs frequently as an isolated problem after minor head injury or may develop for unknown reasons after a delay of several days. Facial nerve injury caused by a basilar fracture is present immediately in up to 3% of severe injuries; it may also be delayed for 5–7 days. Fractures through the petrous bone, particularly the less common transverse type, are liable to produce facial palsy. Delayed facial palsy occurring up to a week after injury, the mechanism of which is unknown, has a good prognosis. Injury to the eighth cranial nerve from a fracture of the petrous bone causes loss of hearing, vertigo, and nystagmus immediately after injury. Deafness from eighth nerve injury is rare and must be distinguished from blood in the middle ear or disruption of the middle ear ossicles. Dizziness, tinnitus, and high-tone hearing loss occur from cochlear concussion.

SEIZURES

Convulsions are surprisingly uncommon immediately after TBI, but a brief period of tonic extensor posturing or a few clonic movements of the limbs just after the moment of impact can occur. However, the cortical scars that evolve from contusions are highly epileptogenic and may later manifest as seizures, even after many months or years (Chap. 418). The severity of injury roughly determines the risk of future seizures. It has been estimated that 17% of individuals with brain contusion, subdural hematoma, or prolonged LOC will develop a seizure disorder and that this risk extends for an indefinite period of time, whereas the risk is ≤2% after mild injury. The majority of convulsions in the latter group occur within 5 years of injury but may be delayed for decades. Penetrating injuries have a much higher rate of subsequent epilepsy.

CLINICAL SYNDROMES AND TREATMENT OF HEAD INJURY

CONCUSSION/MILD TBI

The patient who has briefly lost consciousness or been stunned after a minor head injury usually becomes fully alert and attentive within minutes but may complain of headache, dizziness, faintness, nausea, a single episode of emesis, difficulty with concentration, a brief amnestic period, or slight blurring of vision. This typical concussion syndrome has a good prognosis with little risk of subsequent deterioration. Children are particularly prone to drowsiness, vomiting, and irritability, symptoms that are sometimes delayed for several hours after apparently minor injuries. Vasovagal syncope that follows injury may cause undue concern. Generalized or frontal headache is common in the following days. It may be migrainous (throbbing and hemicranial) in nature or aching and bilateral. After several hours of observation, patients with minor injury may be accompanied home and observed for a day by a family member or friend, with written instructions to return if symptoms worsen.

Persistent severe headache and repeated vomiting in the context of normal alertness and no focal neurologic signs is usually benign, but CT should be obtained and a longer period of observation is appropriate. The decision to perform imaging tests also depends on clinical signs that indicate that the impact was severe (e.g., persistent confusion, repeated vomiting, palpable skull fracture); the presence of other serious bodily injuries, an underlying coagulopathy, or age >65 years; and on the degree of surveillance that can be anticipated after discharge. Guidelines have also indicated that older age (>65), two or more episodes of vomiting, >30 min of retrograde or persistent anterograde amnesia, seizure, and concurrent drug or alcohol intoxication are sensitive (but not specific) indicators of intracranial hemorrhage that justify CT scanning.

SPORT-RELATED CONCUSSION

Based on its reported prevalence and acute effects, and fears over potential long-term neurological consequences, sport-related concussion has become the focus of increasing concern from clinicians, researchers, sporting organizations, and athletes themselves. Concussion is a frequent injury in contact and collision sports (e.g., football, hockey, wrestling) at all levels of participation, including youth sports. One study indicated that from 1997 to 2007 emergency department visits for 8- to 13-year-old children affected by concussion in organized team sports doubled, and increased by >200% in the 14- to 19-year-old group; these increases could represent improvements in identification in addition to actual changes in incidence rates.

Research over the last decade has advanced our understanding of the true natural history of clinical recovery following sport-related concussion. In general, the findings on acute recovery are favorable. A 2003 report was the first to chart the continuous time course of acute recovery within several days after concussion, indicating that >90% of athletes reported symptom recovery within 1 week. Several other prospective studies have since demonstrated that the overwhelming majority of athletes achieve a complete recovery in symptoms, cognitive functioning, postural stability, and other functional impairments over a period of 1–3 weeks following concussion.

There are frequent anecdotal reports, however, of athletes who remain symptomatic or impaired on functional testing well beyond the window of recovery commonly reported in group studies. The greatest challenge arguably still facing sport medicine clinicians and public health experts is how to most effectively manage and reduce risk in this subset of athletes who do not follow the "typical" course of recovery. The precise frequency of athletes who do not follow the typical course of rapid, spontaneous recovery and instead exhibit prolonged postconcussive symptoms or other functional impairments after concussion remains unclear. There is little empirical evidence regarding which risk factors may be associated with prolonged recovery time or poor outcome in athletes and how these risks can be modified in a clinical setting.

In the current absence of adequate data, a common sense approach to athletic concussion has been to remove the individual from play immediately and avoid contact sports for at least several days after a mild injury, and for a longer period if there are more severe injuries or if there are protracted neurologic symptoms such as headache and difficulty concentrating. No individual should return to play unless all symptoms have resolved and an assessment has been made by a health care professional who has experience with treatment of concussion. Once cleared, the individual can then begin a graduated program of increasing activity. Younger athletes are particularly likely to experience protracted concussive symptoms, and a slower return to play in this age group may be reasonable. These guidelines are designed in part to avoid a perpetuation of symptoms but also to prevent the rare *second impact syndrome*, in which diffuse and fatal cerebral swelling follows a second minor head injury.

In the past, mental decline in boxers late in their careers had been called *dementia pugilistica*. There is some evidence that repeated concussions from other sports, and especially in professional American football players, are associated with a similar delayed and progressive cognitive disorder, sometimes with prominent behavioral symptoms that can include depression, insomnia, violent behaviors, and suicidality. The brains of these patients display a characteristic deposition of tau protein in neurons located in the superficial cortical layers and perivascular regions, and particularly in the depths of sulci, a pattern named *chronic traumatic encephalopathy* (CTE). CTE is an intensively studied and provocative entity. A recent neuropathologic study of athletes who had donated their brains for research reported that changes of CTE were extremely common findings. However, the majority of former football players do not complain of cognitive symptoms, and at this time the true prevalence of CTE is unknown. Its contribution, if any, to late-life dementia and parkinsonism in former athletes, soldiers, or others who have sustained repeated concussive injuries is unknown. **CTE is also discussed in Chap. 417.**

POSTCONCUSSIVE STATES

The *postconcussion syndrome* (PCS) refers to a state following minor TBI consisting of combinations of fatigue, dizziness, headache, and difficulty in concentration. Management is difficult and generally requires the identification and management of the specific problem or problems that are most troubling to the individual. A clear explanation of the symptoms that may follow concussion has been shown to reduce subsequent complaints. Care is taken to avoid prolonged use of drugs that produce dependence. Headache may initially be treated with acetaminophen and small doses of amitriptyline. Vestibular exercises **(Chap. 19)** and small doses of vestibular suppressants such as promethazine (Phenergan) may be helpful when dizziness is the main problem. Patients who after minor or moderate injury have difficulty with memory or with complex cognitive tasks at work may be reassured that these problems usually improve over several months, and workload may be reduced in the interim.

For the vast majority of individuals with mTBI, the symptoms of PCS subside and resolve within a few weeks of injury. For a subset of individuals with mTBI, complaints of postconcussion symptoms persist beyond the expectation derived from TBI severity markers. The term PCD has been proposed for diagnostic use when symptoms following mTBI such as neurologic, cognitive, behavioral or somatic complaints persist beyond the acute and subacute periods and become chronic, often operationalized as persisting beyond 3 months. Although the overall risk of developing PCD following mTBI is low, the frequency of mTBI patients who meet criteria for a diagnosis of PCD and present in a clinical setting is believed to be higher.

mTBI patients with PCD frequently present to the outpatient clinics of primary care physicians, physiatrists or neurologists seeking relief for lingering PCD-related symptoms. While some patients will have already received an initial medical work-up to rule out a more serious brain injury during the acute phase, many patients will have had no prior contact with health care specialists. A medical work-up ordered in the outpatient setting for PCD-related complaints is typically unremarkable for any identifiable neurologic cause to account for the persisting symptoms reported by the patient. The development of uniform decision trees or "standard of care" treatment regimens for PCD-related symptoms has been limited by the diversity of symptoms that patients experience, even within mTBI subgroups that have sustained very similar injury patterns. While some patients experience somatic symptoms, others complain of subjective cognitive or behavioral changes.

PCD is not a unidimensional condition but rather an outcome influenced by diverse cognitive, emotional, medical, psychosocial, and motivational factors. Because of this complexity, treatments targeting persistent and refractory PCD-related symptoms should be tailored to the needs and expectations of the individual patient, with referrals to specialists as needed for assistance with management of headache, neck and back pain, dizziness and vertigo, and other symptoms reported within the context of PCD. In addition, patients are frequently referred to behavioral health providers such as neuropsychologists, rehabilitation psychologists, health psychologists, and/or psychiatrists for a variety of reasons, but particularly when they are experiencing cognitive, emotional, or behavioral changes that accompany PCD. Patients with mood disorders (e.g., depression), anxiety disorders (e.g., posttraumatic stress disorder), or adjustment reactions may benefit from psychiatric consultation for appropriate medication trials or from time-limited psychotherapy such as cognitive behavioral therapy.

INJURY OF INTERMEDIATE SEVERITY

Patients who are not fully alert or have persistent confusion, behavioral changes, extreme dizziness, or focal neurologic signs such as hemiparesis should be admitted to the hospital and undergo a cerebral imaging study. A cerebral contusion or hematoma will usually be found. Common syndromes include: (1) delirium with a disinclination to be examined or moved, expletive speech, and resistance if disturbed (anterior temporal lobe contusions); (2) a quiet, disinterested, slowed mental state (abulia) alternating with irascibility (inferior frontal and frontopolar contusions); (3) a focal deficit such as aphasia or mild hemiparesis (due to subdural hematoma or convexity contusion or, less often, carotid artery dissection); (4) confusion and inattention, poor performance on simple mental tasks, and fluctuating orientation (associated with several types of injuries, including those described above, and with medial frontal contusions and interhemispheric subdural hematoma); (5) repetitive vomiting, nystagmus, drowsiness, and unsteadiness (labyrinthine concussion, but occasionally due to a posterior fossa subdural hematoma or vertebral artery dissection); and (6) diabetes insipidus (damage to the median eminence or pituitary stalk). Injuries of this degree are often complicated by drug or alcohol intoxication, and clinically inapparent cervical spine injury may be present. Blast injuries are often accompanied by rupture of the tympanic membranes.

After surgical removal of hematomas, patients in this category improve over weeks to months. During the first week, the state of alertness, memory, and other cognitive functions often fluctuate, and agitation and somnolence are common. Behavioral changes tend to be worse at night, as with many other encephalopathies, and may be treated with small doses of antipsychotic medications. Subtle abnormalities of attention, intellect, spontaneity, and memory return toward normal weeks or months after the injury, sometimes abruptly. However, the full extent of recovery may not be realized for several years. Persistent cognitive problems are discussed below.

SEVERE INJURY

Patients who are comatose from the moment of injury require immediate neurologic attention and resuscitation. After intubation, with care taken to immobilize the cervical spine, the depth of coma, pupillary size and reactivity, limb movements, and Babinski responses are assessed. As soon as vital functions permit and cervical spine x-rays and a CT scan have been obtained, the patient should be transported to a critical care unit. Hypoxia should be reversed, and normal saline used as the resuscitation fluid in preference to albumin. The finding of an epidural or subdural hematoma or large intracerebral hemorrhage is usually an indication for prompt surgery and intracranial decompression in an otherwise salvageable patient. Measurement of ICP with a ventricular catheter or fiberoptic device in order to guide treatment has been favored by many units but has not improved outcome. Hyperosmolar intravenous solutions are used in various regimens to limit intracranial pressure. Prophylactic antiepileptic medications are recommended for 7 days and should be discontinued unless there are multiple seizures postinjury. **Management of raised ICP, a frequent feature of severe head injury, is discussed in Chap. 301.**

Despite the improvement in mortality for severe TBI over the past few decades, a great deal of therapeutic nihilism persists in TBI. The common use of a 6-month outcome for TBI clinical studies reinforces this misconception. The recovery from severe TBI can take years. Furthermore, the ability to predict long-term outcome is limited and frequently incorrect. Recent best practice guidelines recommend, in the absence of brain death, that aggressive therapy be instituted for at least 72 h in the acute injury period.

Acknowledgment

The authors wish to acknowledge the contributions of Dr. Allan Ropper to earlier editions of this chapter.

FURTHER READING

Jagoda AS et al: American College of Emergency Physicians and Centers for Disease Control and Prevention. Clinical policy: Neuroimaging and decisionmaking in adult mild traumatic brain injury in the acute setting. Ann Emerg Med 52:714, 2008.

Johnson VE et al: Axonal pathology in traumatic brain injury. Exp Neurol 246:35, 2013.

McCrory P et al: Consensus statement on concussion in sport—the 5th international conference on concussion in sport held in Berlin, October 2016. Br J Sports Med 51:838, 2017.

Mez J et al: Clinicopathological evaluation of chronic traumatic encephalopathy in players of American football. JAMA 25;318:360, 2017.

Taylor CA et al: Traumatic brain injury-related emergency department visits, hospitalizations, and deaths—United States, 2007 and 2013. MMWR Surveill Summ 66:1, 2017.

436 Multiple Sclerosis

Bruce A. C. Cree, Stephen L. Hauser

MULTIPLE SCLEROSIS

Multiple sclerosis (MS) is an autoimmune disease of the central nervous system (CNS) characterized by chronic inflammation, demyelination, gliosis (plaques or scarring), and neuronal loss; the course can be relapsing or progressive. MS plaques typically develop at different times and in different CNS locations (i.e., MS is said to be disseminated in time and space). Approximately more than 900,000 individuals in the United States and millions of individuals worldwide are affected. The clinical course is extremely variable, ranging from a relatively benign condition to a rapidly evolving and incapacitating disease requiring profound lifestyle adjustments.

CLINICAL MANIFESTATIONS

The onset of MS may be abrupt or insidious. Symptoms may be severe or seem so trivial that a patient may not seek medical attention for months or years. Indeed, at autopsy, ~0.1% of individuals who were asymptomatic during life will be found, unexpectedly, to have pathologic evidence of MS. Similarly, an magnetic resonance imaging (MRI) scan obtained for an unrelated reason may show evidence of asymptomatic MS. Symptoms of MS are extremely varied and depend on the location and severity of lesions within the CNS **(Table 436-1)**. Examination often reveals evidence of neurologic dysfunction, often in asymptomatic locations. For example, a patient may present with symptoms in one leg but signs in both.

TABLE 436-1 Initial Symptoms of Multiple Sclerosis (MS)

SYMPTOM	PERCENTAGE OF CASES	SYMPTOM	PERCENTAGE OF CASES
Sensory loss	37	Lhermitte	3
Optic neuritis	36	Pain	3
Weakness	35	Dementia	2
Paresthesias	24	Visual loss	2
Diplopia	15	Facial palsy	1
Ataxia	11	Impotence	1
Vertigo	6	Myokymia	1
Paroxysmal attacks	4	Epilepsy	1
Bladder	4	Falling	1

Source: After WB Matthews et al: *McAlpine's Multiple Sclerosis.* New York, Churchill Livingstone, 1991.

Sensory symptoms are varied and include both paresthesias (e.g., tingling, prickling sensations, formications, "pins and needles," or painful burning) and hypesthesia (e.g., reduced sensation, numbness, or a "dead" feeling). Unpleasant sensations (e.g., feelings that body parts are swollen, wet, raw, or tightly wrapped) are also common. Sensory impairment of the trunk and legs below a horizontal line on the torso (a sensory level) indicates that the spinal cord is the origin of the sensory disturbance. It is often accompanied by a bandlike sensation of tightness around the torso. Pain is a common symptom of MS, experienced by >50% of patients. Pain can occur anywhere on the body and can change locations over time.

Optic neuritis (ON) presents as diminished visual acuity, dimness, or decreased color perception (desaturation) in the central field of vision. These symptoms can be mild or may progress to severe visual loss. Rarely, there is complete loss of light perception. Visual symptoms are generally monocular but may be bilateral. Periorbital pain (aggravated by eye movement) often precedes or accompanies the visual loss. An afferent pupillary defect **(Chap. 28)** is usually present. Funduscopic examination may be normal or reveal optic disc swelling (papillitis). Pallor of the optic disc (optic atrophy) commonly follows ON. Uveitis is uncommon and should raise the possibility of alternative diagnoses such as sarcoid or lymphoma.

Weakness of the limbs may manifest as loss of strength, speed, or dexterity, as fatigue, or as a disturbance of gait. Exercise-induced weakness is a characteristic symptom of MS. The weakness is of the upper motor neuron type **(Chap. 21)** and is usually accompanied by other pyramidal signs such as spasticity, hyperreflexia, and Babinski signs. Occasionally, a tendon reflex may be lost (simulating a lower motor neuron lesion) if an MS lesion disrupts the afferent reflex fibers in the spinal cord **(see Fig. 21-2)**.

Facial weakness due to a lesion in the pons may resemble idiopathic Bell's palsy **(Chap. 433)**. Unlike Bell's palsy, facial weakness in MS is usually not associated with ipsilateral loss of taste sensation or retroauricular pain.

Spasticity **(Chap. 21)** is commonly associated with spontaneous and movement-induced muscle spasms. More than 30% of MS patients have moderate to severe spasticity, especially in the legs. This is often accompanied by painful spasms interfering with ambulation, work, or self-care. Occasionally, spasticity provides support for the body weight during ambulation, and in these cases, treatment of spasticity may actually do more harm than good.

Visual blurring in MS may result from ON or diplopia (double vision); if the symptom resolves when either eye is covered, the cause is diplopia. *Diplopia* may result from internuclear ophthalmoplegia (INO) or from palsy of the sixth cranial nerve (rarely the third or fourth). An INO consists of impaired adduction of one eye due to a lesion in the ipsilateral medial longitudinal fasciculus **(Chaps. 28 and V3)**. Prominent nystagmus is often observed in the abducting eye, along with a small skew deviation. A bilateral INO is particularly suggestive of MS. Other common gaze disturbances in MS include (1) a horizontal gaze palsy, (2) a "one and a half" syndrome (horizontal gaze palsy plus an INO), and (3) acquired pendular nystagmus.

Ataxia usually manifests as cerebellar tremors **(Chap. 431)**. Ataxia may also involve the head and trunk or the voice, producing a characteristic cerebellar dysarthria (scanning speech).

Vertigo may appear suddenly from a brainstem lesion, superficially resembling acute labyrinthitis **(Chap. 19)**. *Hearing loss* **(Chap. 30)** may also occur in MS but is uncommon.

Ancillary Symptoms *Paroxysmal symptoms* are distinguished by their brief duration (10 s to 2 min), high frequency (5–40 episodes per day), lack of any alteration of consciousness or change in background electroencephalogram during episodes, and a self-limited course (generally lasting weeks to months). They may be precipitated by hyperventilation or movement. These syndromes may include Lhermitte's symptom; tonic contractions of a limb, face, or trunk (tonic seizures); paroxysmal dysarthria and ataxia; paroxysmal sensory disturbances; and several other less well-characterized syndromes. Paroxysmal

symptoms probably result from spontaneous discharges, arising at the edges of demyelinated plaques and spreading to adjacent white matter tracts.

Lhermitte's symptom is an electric shock–like sensation (typically induced by flexion or other movements of the neck) that radiates down the back into the legs. Rarely, it radiates into the arms. It is generally self-limited but may persist for years. Lhermitte's symptom can also occur with other disorders of the cervical spinal cord (e.g., cervical spondylosis).

Trigeminal neuralgia, hemifacial spasm, and glossopharyngeal neuralgia **(Chap. 433)** can occur when the demyelinating lesion involves the root entry (or exit) zone of the fifth, seventh, and ninth cranial nerve, respectively. Trigeminal neuralgia (tic douloureux) is a very brief lancinating facial pain often triggered by an afferent input from the face or teeth. Most cases of trigeminal neuralgia are not MS related; however, atypical features such as onset before age 50 years, bilateral symptoms, objective sensory loss, or nonparoxysmal pain should raise the possibility that MS could be responsible.

Facial myokymia consists of either persistent rapid flickering contractions of the facial musculature (especially the lower portion of the orbicularis oculus) or a contraction that slowly spreads across the face. It results from lesions of the corticobulbar tracts or brainstem course of the facial nerve.

Heat sensitivity refers to neurologic symptoms produced by an elevation of the body's core temperature. For example, unilateral visual blurring may occur during a hot shower or with physical exercise (*Uhthoff's symptom*). It is also common for MS symptoms to worsen transiently, sometimes dramatically, during febrile illnesses (see "Acute Attacks or Initial Demyelinating Episodes," below). Such heat-related symptoms probably result from transient conduction block (see above).

Bladder dysfunction is present in >90% of MS patients, and in a third of patients, dysfunction results in weekly or more frequent episodes of incontinence. During normal reflex voiding, relaxation of the bladder sphincter (α-adrenergic innervation) is coordinated with contraction of the detrusor muscle in the bladder wall (muscarinic cholinergic innervation). *Detrusor hyperreflexia*, due to impairment of suprasegmental inhibition, causes urinary frequency, urgency, nocturia, and uncontrolled bladder emptying. *Detrusor sphincter dyssynergia*, due to loss of synchronization between detrusor and sphincter muscles, causes difficulty in initiating and/or stopping the urinary stream, producing hesitancy, urinary retention, overflow incontinence, and recurrent infection.

Constipation occurs in >30% of patients. Fecal urgency or *bowel incontinence* is less common (<15%) but can be socially debilitating.

Sexual dysfunction may manifest as decreased libido, impaired genital sensation, impotence in men, and diminished vaginal lubrication or adductor spasms in women.

Cognitive dysfunction can include memory loss; impaired attention; difficulties in executive functioning, memory, and problem solving; slowed information processing; and problems shifting between cognitive tasks. Euphoria (elevated mood) was once thought to be characteristic of MS but is actually uncommon, occurring in <20% of patients. Cognitive dysfunction sufficient to impair activities of daily living is rare.

Depression, experienced by approximately half of patients, can be reactive, endogenous, or part of the illness itself and can contribute to fatigue.

Fatigue **(Chap. 20)** is experienced by 90% of patients; this symptom is the most common reason for work-related disability in MS. Fatigue can be exacerbated by elevated temperatures, depression, expending exceptional effort to accomplish basic activities of daily living, or sleep disturbances (e.g., from frequent nocturnal awakenings to urinate).

DISEASE COURSE

Three clinical types of MS exist **(Fig. 436-1)**:

1. *Relapsing or bout onset MS* (RMS) accounts for 90% of MS cases and is characterized by discrete attacks of neurological dysfunction that generally evolve over days to weeks (rarely over hours). With initial attacks, there is often substantial or complete recovery over the ensuing weeks to months. However, as attacks continue recovery may be less evident **(Fig. 436-1*A*)**. Between attacks, patients are neurologically stable.
2. *Secondary progressive MS* (SPMS) always begins as RMS **(Fig. 436-1*B*)**. At some point, however, the clinical course changes so that the patient experiences deterioration in function unassociated with acute attacks. SPMS produces a greater amount of fixed neurologic disability than RMS. For a patient with RMS, the risk of developing SPMS is ~2% each year, meaning that the great majority of RMS ultimately evolves into SPMS. As such, SPMS appears to represent a late stage of the same underlying illness as RMS.
3. *Primary progressive MS* (PPMS) accounts for ~10% of cases. These patients do not experience attacks but rather steadily decline in function from disease onset **(Fig. 436-1*C*)**. Compared to RMS, the sex distribution is more even, the disease begins later in life (mean age ~40 years), and disability develops faster (relative to the onset of the first clinical symptom). Despite these differences, PPMS appears to represent the same underlying illness as RMS.

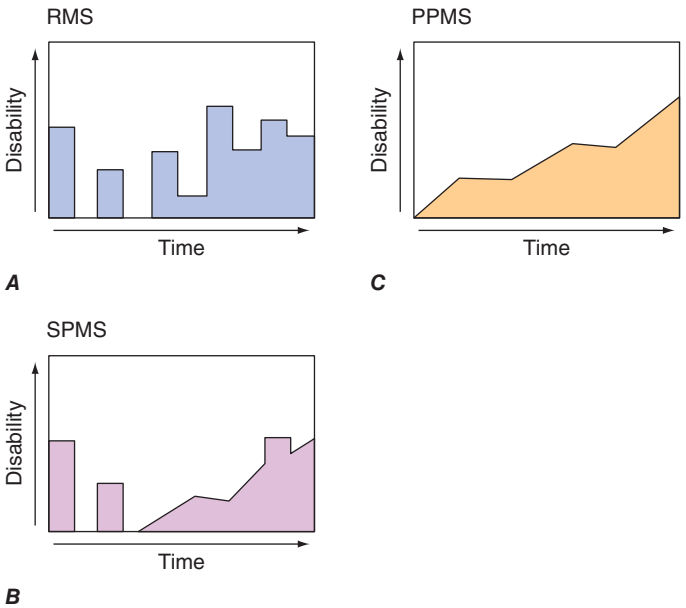


FIGURE 436-1 Clinical course of multiple sclerosis (MS). *A.* Relapsing MS (RMS). ***B.*** Secondary progressive MS (SPMS). ***C.*** Primary progressive MS (PPMS).

Progressive MS and disease activity. Patients with SPMS or even PPMS will occasionally experience relapses, albeit far less often than in RMS. Progressive MS patients experiencing relapses or who are found to have acute new lesions on MRI are considered to have "active" MS. In contrast, the term "progression" is reserved to describe neurological worsening that accumulates independently from disease activity.

Epidemiology MS is approximately threefold more common in women than men. The age of onset is typically between 20 and 40 years (slightly later in men than in women), but the disease can present across the lifespan. Approximately 10% of cases begin before the age of 18 years, and a small percentage of cases begin before the age of 10 years.

Geographical gradients are observed in MS, with the highest known prevalence for MS (250 per 100,000) in the Orkney Islands, located north of Scotland. In other temperate zone areas (e.g., northern North America, northern Europe, southern Australia, and southern New Zealand), the prevalence of MS is 0.1–0.2%. By contrast, in the tropics (e.g., Asia, equatorial Africa, and the Middle East), the prevalence is often tenfold to twentyfold less.

The prevalence of MS has increased steadily (and dramatically) in several regions around the world over the past half-century, presumably reflecting the impact of some environmental shift. Moreover, the fact that this increase has occurred primarily (or exclusively) in women

indicates that women are more responsive to this environmental change.

Well-established risk factors for MS include a genetic predisposition, vitamin D deficiency, Epstein-Barr virus (EBV) exposure after early childhood, and cigarette smoking.

Vitamin D deficiency is associated with an increase in MS risk, and data suggest that ongoing deficiency also increases disease activity after MS begins. Immunoregulatory effects of vitamin D could explain these apparent relationships. Exposure of the skin to ultraviolet-B (UVB) radiation from the sun is essential for the biosynthesis of vitamin D, and this endogenous production is the most important source of vitamin D in most individuals. A diet rich in fatty fish represents another source of vitamin D. At high latitudes, the amount of UVB radiation reaching the earth's surface is often insufficient, particularly during winter months, and consequently, low serum levels of vitamin D are common in temperate zones. The common practice to avoid direct sun exposure and the widespread use of sun block would be expected to exacerbate any population-wide vitamin D deficiency (sun protection factor [SPF] 15 blocks 94% of incoming UVB radiation).

Evidence of a remote EBV infection playing some role in MS is supported by numerous epidemiologic and laboratory studies. A higher risk of infectious mononucleosis (associated with relatively late EBV infection) and higher antibody titers to latency-associated EBV nuclear antigen have been repeatedly associated with MS risk, although a causal role for EBV is not established.

A history of cigarette smoking also is associated with MS risk. Interestingly, in an animal model of MS, the lung was identified as a critical site for activation of pathogenic T lymphocytes responsible for autoimmune demyelination.

GENETIC CONSIDERATIONS

Whites are inherently at higher risk for MS than Africans or Asians, even when residing in a similar environment. MS also aggregates within some families, and adoption, half-sibling, twin, and spousal studies indicate that familial aggregation is due to genetic, and not environmental, factors **(Table 436-2)**.

Susceptibility to MS is polygenic, with each gene contributing a relatively small amount to the overall risk. The strongest susceptibility signal genome-wide maps to the HLA-DRB1 gene in the class II region of the major histocompatibility complex (MHC), and this association accounts for ~10% of the disease risk. This HLA association, first described in the early 1970s, suggests that MS, at its core, is an autoimmune disease. Whole-genome association studies have now identified ~200 other MS susceptibility variants, each of which individually has only a very small effect on MS risk. Many of these MS-associated genes have known roles in the adaptive immune system, for example the genes for the interleukin (IL) 7 receptor (CD127), IL-2 receptor (CD25), and T cell costimulatory molecule LFA-3 (CD58); some variants also influence susceptibility to other autoimmune diseases in addition to MS. The variants identified so far all lack specificity and sensitivity for MS; thus, at present, they are not useful for diagnosis or prediction of the future disease course.

PATHOGENESIS

Pathology New MS lesions begin with perivenular cuffing by inflammatory mononuclear cells, predominantly T cells and macrophages, which also infiltrate the surrounding white matter. At sites of inflammation, the blood-brain barrier (BBB) is disrupted, but unlike vasculitis, the vessel wall is preserved. Involvement of the humoral immune system is also evident; small numbers of B lymphocytes infiltrate the nervous system, myelin-specific autoantibodies are present on degenerating myelin sheaths, and complement is activated. Demyelination is the pathological hallmark and evidence of myelin degeneration is found at the earliest time points of tissue injury. Although relative sparing of axons is typical of MS, partial or total axonal destruction can also occur, especially within highly inflammatory lesions. In some lesions, surviving oligodendrocytes or those that differentiate from precursor cells partially remyelinate the surviving axons, producing so-called *shadow plaques*. However, in many lesions, although oligodendrocyte precursor cells are present, they fail to differentiate into mature myelin-producing cells. As lesions evolve, there is prominent astrocytic proliferation (gliosis) and the term *sclerosis* refers to these gliotic plaques that have a rubbery or hardened texture at autopsy.

MS is not solely a disease of myelin, and neuronal pathology is increasingly recognized as a major contributor to irreversible neurologic disability. Inflammation, demyelination, and plaque formation are also present in the cerebral cortex, and significant axon loss indicating death of neurons is widespread, especially in advanced cases (see "Neurodegeneration," below). Cortical plaques may extend upward from demyelinated white matter, or may be restricted to the cortex itself, or located underneath the pia. A recently recognized feature of MS pathology is the presence of ectopic clusters of lymphocytes, termed lymphoid follicles, consisting of aggregates of T, B, and plasma cells resembling secondary lymphoid tissue located in the meninges, especially overlying deep cortical sulci; they are also present in perivascular spaces and less commonly within brain parenchyma. These structures appear to be more prevalent in progressive MS and are located in proximity to cortical plaques suggesting that perhaps diffused factors from these ectopic follicles contribute to subpial cortical demyelination and neurodegeneration.

Inflammation is always present when active demyelination or axonal injury occurs, and the presence of T-cell and B-cell infiltration is related to the extent of demyelination and axonal injury. However, the nature of the inflammatory response appears to be somewhat different between early and later stages of MS. In relapsing MS, inflammation is associated with focal perivenular parenchymal infiltration of lymphocytes and monocytes associated with BBB disruption and active demyelination. In contrast, inflammation in progressive MS is more diffuse and is characterized by widespread microglial activation. Acute perivascular infiltrates are fewer in number, and lymphocytes and monocytes in chronic MS plaques aggregate at the lesion border suggesting ongoing inflammatory injury at the lesion edge. In addition, a diffuse low-grade inflammation with microglial proliferation is observed across large areas of white matter, associated with reduced myelin staining and axonal injury ("dirty white matter"). Activated astrocytes induced by microglia may also contribute to tissue damage **(Chap. 417)**. These observations imply that ongoing inflammation occurs behind a partially repaired BBB in many patients with progressive MS, and this feature could explain the failure of immunotherapies not capable of crossing the BBB to benefit patients with progressive MS.

Physiology Nerve conduction in myelinated axons occurs in a saltatory manner, with the nerve impulse jumping from one node of Ranvier to the next without depolarization of the axonal membrane underlying the myelin sheath between nodes **(Fig. 436-2)**. This produces considerably faster conduction velocities (~70 m/s) than the slow velocities (~1 m/s) produced by continuous propagation in unmyelinated nerves. Conduction block occurs when the nerve impulse is unable to traverse the demyelinated segment. This can happen when the resting axon membrane becomes hyperpolarized due to the exposure of voltage-dependent potassium channels that are normally buried underneath the myelin sheath. A temporary conduction block often follows a demyelinating event before sodium channels (originally concentrated at the nodes) redistribute along the naked axon (Fig. 436-2). This redistribution ultimately allows continuous propagation of nerve action potentials through the demyelinated segment. Conduction block may be incomplete, affecting high- but

TABLE 436-2 Risk of Developing Multiple Sclerosis (MS)

1 in 3	If an identical twin has MS
1 in 15	If a fraternal twin has MS
1 in 25	If a sibling has MS
1 in 50	If a parent or half-sibling has MS
1 in 100	If a first cousin has MS
1 in 1000	If a spouse has MS
1 in 1000	If no one in the family has MS

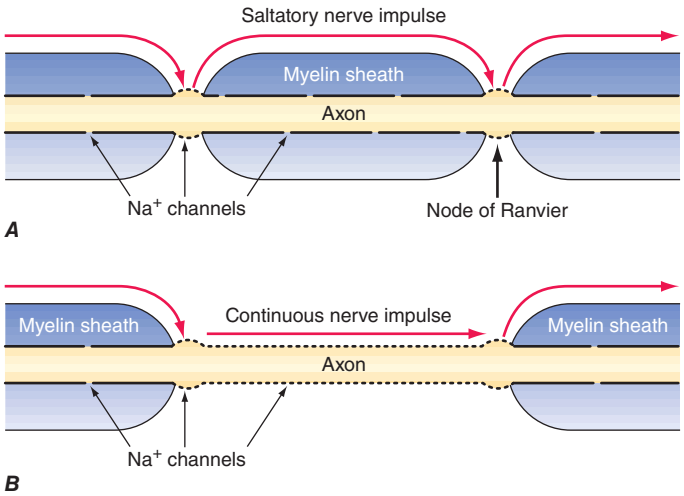


FIGURE 436-2 Nerve conduction in myelinated and demyelinated axons. ***A.*** Saltatory nerve conduction in myelinated axons occurs with the nerve impulse jumping from one node of Ranvier to the next. Sodium channels (shown as breaks in the solid black line) are concentrated at the nodes where axonal depolarization occurs. ***B.*** Following demyelination, additional sodium channels are redistributed along the axon itself, thereby allowing continuous propagation of the nerve action potential despite the absence of myelin.

not low-frequency volleys of impulses. Variable conduction block can occur with raised body temperature or metabolic alterations and may explain clinical fluctuations that vary from hour to hour or appear with fever or exercise. Conduction slowing occurs when the demyelinated segments of the axonal membrane are reorganized to support continuous (slow) nerve impulse propagation.

IMMUNOLOGY A proinflammatory autoimmune response directed against a component of CNS myelin, and perhaps other neural elements as well, remains the cornerstone of current concepts of MS pathogenesis.

AUTOREACTIVE T LYMPHOCYTES Myelin basic protein (MBP), an intracellular protein involved in myelin compaction, is an important T cell antigen in experimental allergic encephalomyelitis (EAE), a laboratory model, and possibly also in human MS. Activated MBP-reactive T cells have been identified in the blood, in cerebrospinal fluid (CSF), and within MS lesions. Moreover, *DRB1*15:01* may influence the autoimmune response because it binds with high affinity to a fragment of MBP (spanning amino acids 89–96), stimulating T cell responses to this self-protein. Two different populations of proinflammatory T cells are likely to mediate autoimmunity in MS. T-helper type 1 (T_H1) cells producing interferon γ (IFN-γ) are one key effector population, and a role for highly proinflammatory T_H17 T cells has also been established, at least in some patients. T_H17 cells are induced by transforming growth factor β (TGF-β) and IL-6 and are amplified by IL-21 and IL-23. T_H17 cells and levels of their corresponding cytokine IL-17 are increased in MS lesions. T_H1 cytokines, including IL-2, tumor necrosis factor (TNF)-α, and IFN-γ, also play key roles in activating and maintaining autoimmune responses, and TNF-α and IFN-γ may directly injure oligodendrocytes or the myelin membrane.

HUMORAL AUTOIMMUNITY B cell activation and antibody responses are also involved in the development of demyelinating lesions. Clonally restricted populations of activated, antigen-experienced, memory B cells and plasma cells are present in MS lesions, in meningeal lymphoid follicle-like structures overlying the cerebral cortex, and in the CSF. Similar populations are found in each compartment, indicating that a highly focused B cell response occurs locally within the CNS. Myelin-specific autoantibodies, some directed against an extracellular myelin protein, myelin oligodendrocyte glycoprotein (MOG), have been detected bound to vesiculated myelin debris in MS plaques. In the CSF, elevated levels of locally synthesized immunoglobulins and oligoclonal antibodies, derived from clonally restricted CNS B cells and plasma cells, are also characteristic of MS. The pattern of oligoclonal banding is unique to each individual, and attempts to identify the targets of these antibodies have been largely unsuccessful. Moreover, when proteins recognized by CSF restricted oligoclonal bands (OCBs) have been found, they appear to recognize a variety of antigens including intracellular ubiquitous proteins. Therefore, although intrathecal OCBs and elevated intrathecal synthesis of immunoglobulins are characteristic of MS, their role in disease pathogenesis remains uncertain.

NEURODEGENERATION

Axonal damage occurs in every newly formed MS lesion, and cumulative axonal and neuronal loss is considered to be the most important contributor to irreversible neurologic disability. As many as 70% of axons are lost from the lateral corticospinal (e.g., motor) tracts in patients with advanced paraparesis from MS, and longitudinal MRI studies suggest that there is progressive axonal loss over time within established lesions. Demyelination can result in reduced trophic support for axons, redistribution of ion channels, and destabilization of action potential membrane potentials. Axons can adapt initially to these injuries, but over time distal and retrograde degeneration ("dying–back" axonopathy) occurs. Therefore, promoting remyelination to protect axons remains an important therapeutic goal.

In addition to white matter plaques and axonopathy, as noted above (see Pathology), recent studies in progressive MS have highlighted an important role for a primary injury to the cerebral cortex, perhaps related to overlying meningeal inflammation.

Data also support a role for one, or more likely several, of the following mechanisms in progressive MS. Axonal and neuronal death may result from glutamate-mediated excitotoxicity, oxidative injury, iron accumulation, and/or mitochondrial failure either occurring as a consequence of free-radical damage or due to accumulation of deletions in mitochondrial DNA.

DIAGNOSIS

There is no single diagnostic test for MS. Diagnostic criteria for clinically definite MS require documentation of two or more episodes of symptoms and two or more signs that reflect pathology in anatomically noncontiguous white matter tracts of the CNS **(Table 436-3)**. Symptoms must last for >24 h and occur as distinct episodes that are separated by a month or more. In patients who have only one of the two required signs on neurologic examination, the second may be documented by abnormal tests such as MRI or evoked potentials (EPs). Similarly, in the most recent diagnostic scheme, the second clinical event (in time) may be supported solely by MRI findings, consisting of either the development of new focal white matter lesions on MRI or the simultaneous presence of both an enhancing lesion and a nonenhancing lesion in an asymptomatic location. In patients whose course is progressive from onset for ≥6 months without superimposed relapses, documentation of intrathecal IgG synthesis may be used to support a diagnosis of PPMS.

DIAGNOSTIC TESTS

Magnetic Resonance Imaging MRI has revolutionized the diagnosis and management of MS **(Fig. 436-3)**; characteristic abnormalities are found in >95% of patients, although >90% of the lesions visualized by MRI are asymptomatic. An increase in vascular permeability from a breakdown of the BBB is detected by leakage of intravenous gadolinium (Gd) into the parenchyma. Such leakage occurs early in the development of an MS lesion and serves as a useful marker of inflammation. Gd enhancement typically persists for <1 month, and the residual MS plaque remains visible indefinitely as a focal area of hyperintensity (a lesion) on T2-weighted images. Lesions are frequently oriented perpendicular to the ventricular surface, corresponding to the pathologic pattern of perivenous demyelination (Dawson's fingers). Lesions are multifocal within the brain, brainstem, and spinal cord. Lesions >6 mm located in the corpus callosum, periventricular white matter, brainstem, cerebellum, or spinal cord are particularly helpful diagnostically. Current criteria for the use of MRI in the diagnosis of MS are shown in Table 436-3.

TABLE 436-3 Diagnostic Criteria for Multiple Sclerosis (MS)

CLINICAL PRESENTATION	ADDITIONAL DATA NEEDED FOR MS DIAGNOSIS
2 or more attacks; objective clinical evidence of 2 or more lesions or objective clinical evidence of 1 lesion with reasonable historical evidence of a prior attack	None
2 or more attacks; objective clinical evidence of 1 lesion	Dissemination in space, demonstrated by • ≥1 T2 lesion on MRI in at least 2 out of 4 MS-typical regions of the CNS (periventricular, juxtacortical, infratentorial, or spinal cord) OR • Await a further clinical attack implicating a different CNS site
1 attack; objective clinical evidence of 2 or more lesions	Dissemination in time, demonstrated by • Simultaneous presence of asymptomatic gadolinium-enhancing and nonenhancing lesions at any time OR • A new T2 and/or gadolinium-enhancing lesion(s) on follow-up MRI, irrespective of its timing with reference to a baseline scan OR • Await a second clinical attack
1 attack; objective clinical evidence of 1 lesion (clinically isolated syndrome)	Dissemination in space and time, demonstrated by: For dissemination in space • ≥1 T2 lesion in at least 2 out of 4 MS-typical regions of the CNS (periventricular, juxtacortical, infratentorial, or spinal cord) OR • Await a second clinical attack implicating a different CNS site AND For dissemination in time • Simultaneous presence of asymptomatic gadolinium-enhancing and nonenhancing lesions at any time OR • A new T2 and/or gadolinium-enhancing lesion(s) on follow-up MRI, irrespective of its timing with reference to a baseline scan OR • Await a second clinical attack
Insidious neurologic progression suggestive of MS (PPMS)	1 year of disease progression (retrospectively or prospectively determined) PLUS 2 out of the 3 following criteria Evidence for dissemination in space in the brain based on ≥1 T2+ lesions in the MS-characteristic periventricular, juxtacortical, or infratentorial regions Evidence for dissemination in space in the spinal cord based on ≥2 T2+ lesions in the cord Positive CSF (isoelectric focusing evidence of oligoclonal bands and/or elevated IgG index)

Abbreviations: CNS, central nervous system; CSF, cerebrospinal fluid; MRI, magnetic resonance imaging; PPMS, primary progressive multiple sclerosis.

Source: From CH Polman et al: Diagnostic criteria for multiple sclerosis: 2010 Revisions to the "McDonald Criteria." Ann Neurol 69:292, 2011.

Serial MRI studies in early relapsing-remitting MS reveal that bursts of focal inflammatory disease activity occur far more frequently than would have been predicted by the frequency of relapses. Thus, early in MS, most disease activity is clinically silent.

The total volume of T2-weighted signal abnormality (the "burden of disease") shows a significant (albeit weak) correlation with clinical disability. Quantitative measures of brain and spinal cord atrophy are evidence of diffuse tissue injury and correlate more strongly with measures of disability or progressive MS. Serial MRI studies also indicate that progressive whole brain atrophy occurs even in very early MS and continues throughout the disease course. Approximately one-third of T2-weighted lesions appear as hypointense lesions (black holes) on T1-weighted imaging. Black holes may be a marker of irreversible demyelination and axonal loss, although even this measure depends on the timing of the image acquisition (e.g., most acute Gd-enhancing T2 lesions are T1 dark).

Evoked Potentials EP testing assesses function in afferent (visual, auditory, and somatosensory) or efferent (motor) CNS pathways. EPs use computer averaging to measure CNS electric potentials evoked by repetitive stimulation of selected peripheral nerves or of the brain. These tests provide the most information when the pathways studied are clinically uninvolved. For example, in a patient with a relapsing spinal cord syndrome with sensory deficits in the legs, an abnormal somatosensory EP following posterior tibial nerve stimulation provides little new information. By contrast, an abnormal visual EP in this circumstance would permit a diagnosis of clinically definite MS (Table 436-3). Abnormalities on one or more EP modalities occur in 80–90% of MS patients. EP abnormalities are not specific to MS, although a marked delay in the latency of a specific EP component (as opposed to a reduced amplitude or distorted wave-shape) is suggestive of demyelination.

Cerebrospinal Fluid CSF abnormalities found in MS include a mononuclear cell pleocytosis and an increased level of intrathecally synthesized IgG. The total CSF protein is usually normal or mildly elevated. Various formulas distinguish intrathecally synthesized IgG from IgG that entered the CNS passively from the serum. One formula, the CSF IgG index, expresses the ratio of IgG to albumin in the CSF divided by the same ratio in the serum. The IgG synthesis rate uses serum and CSF IgG and albumin measurements to calculate the rate of CNS IgG synthesis. The measurement of OCBs by agarose gel electrophoresis in the CSF also assesses intrathecal production of IgG. Two or more discrete OCBs, not present in a paired serum sample, are found in >75% of patients with MS. OCBs may be absent at the onset of MS, and in individual patients, the number of bands may increase with time.

A mild CSF pleocytosis (>5 cells/μL) is present in ~25% of cases, usually in young patients with RMS. A pleocytosis of >75 cells/μL, the presence of polymorphonuclear leukocytes, or a protein concentration >1 g/L (>100 mg/dL) in CSF should raise concern that the patient may not have MS.

DIFFERENTIAL DIAGNOSIS

The possibility of an alternative diagnosis should always be considered (Table 436-4), particularly when (1) symptoms are localized exclusively to the posterior fossa, craniocervical junction, or spinal cord; (2) the patient is <15 or >60 years of age; (3) the clinical course is progressive from onset; (4) the patient has never experienced visual, sensory, or bladder symptoms; or (5) laboratory findings (e.g., MRI, CSF, or EPs) are atypical. Similarly, uncommon or rare symptoms in MS (e.g., aphasia, parkinsonism, chorea, isolated dementia, severe muscular atrophy, peripheral neuropathy, episodic loss of consciousness, fever, headache, seizures, or coma) should increase concern about an alternative diagnosis. Diagnosis is also difficult in patients with a rapid or explosive (stroke-like) onset or with mild symptoms and a normal neurologic examination. Rarely, intense inflammation and swelling may produce a mass lesion that mimics a primary or metastatic tumor. Disorders possibly mistaken for MS include: neuromyelitis optica (Chap. 437), sarcoidosis, vascular disorders (antiphospholipid syndrome and vasculitis), rarely CNS lymphoma and still more rarely infections such as syphilis or Lyme disease. The specific tests required to exclude alternative diagnoses will vary with each clinical situation; however, an erythrocyte sedimentation rate, serum B_{12} level, anti-nuclear antibodies, and treponemal antibody should probably be obtained in all patients with suspected MS.

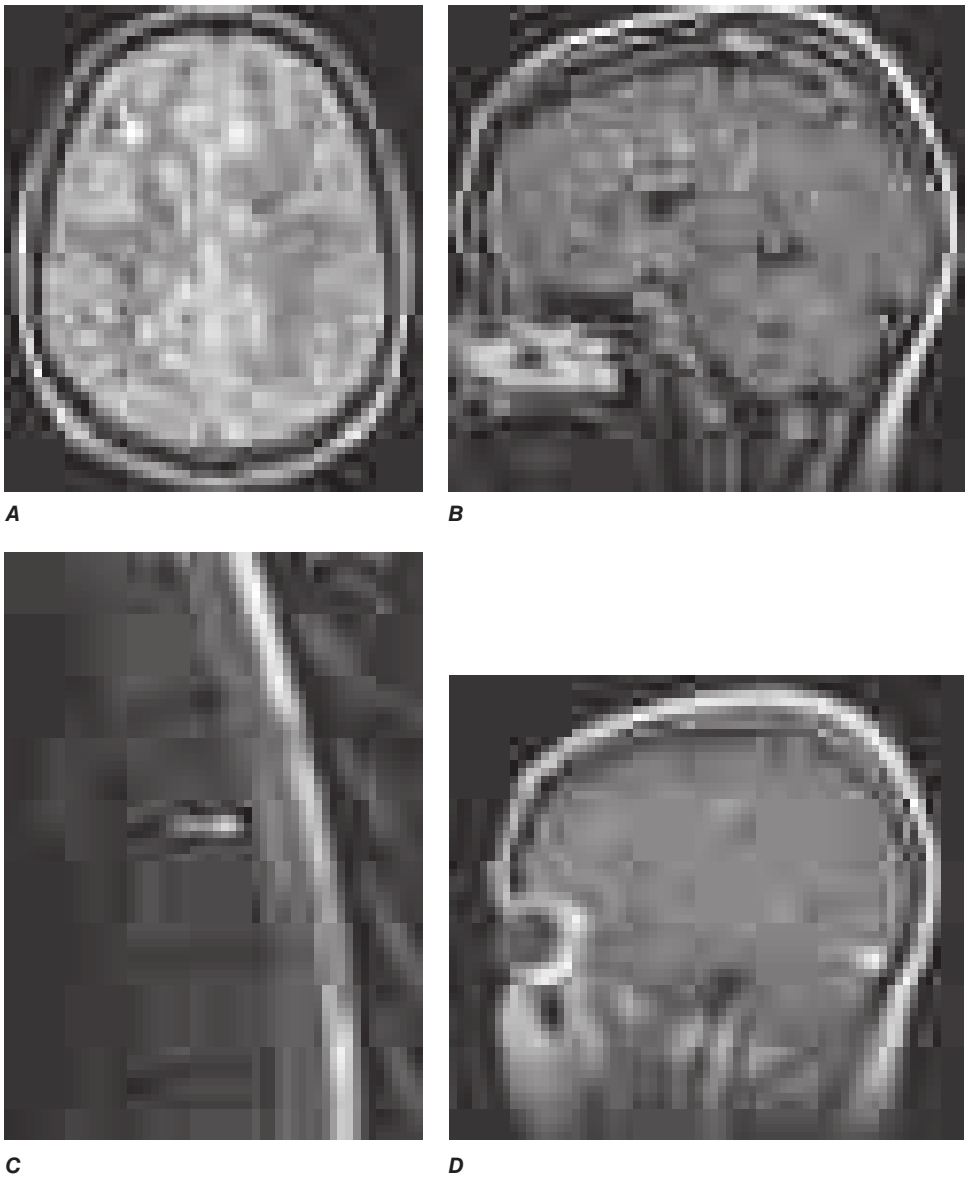


FIGURE 436-3 Magnetic resonance imaging findings in multiple sclerosis (MS). ***A.*** Axial first-echo image from T2-weighted sequence demonstrates multiple bright signal abnormalities in white matter, typical for MS. ***B.*** Sagittal T2-weighted fluid-attenuated inversion recovery (FLAIR) image in which the high signal of cerebrospinal fluid (CSF) has been suppressed. CSF appears dark, whereas areas of brain edema or demyelination appear high in signal, as shown here in the corpus callosum (*arrows*). Lesions in the anterior corpus callosum are frequent in MS and rare in vascular disease. ***C.*** Sagittal T2-weighted fast spin echo image of the thoracic spine demonstrates a fusiform high-signal-intensity lesion in the midthoracic spinal cord. ***D.*** Sagittal T1-weighted image obtained after the intravenous administration of gadolinium DTPA reveals focal areas of blood-brain barrier disruption, identified as high-signal-intensity regions (*arrows*).

PROGNOSIS

Most patients with clinically evident MS ultimately experience progressive neurologic disability. In older studies conducted before disease-modifying therapies for MS were available, 15 years after onset, only 20% of patients had no functional limitation, and between one-third and one-half of RMS patients progressed to SPMS and required assistance with ambulation; furthermore, 25 years after onset, ~80% of MS patients reached this level of disability. The long-term prognosis for MS has improved substantially in recent years, and transition from RMS to SPMS now occurs at approximately a 1% annual rate compared with 2–3% in the pretreatment era. This improvement is almost certainly due, at least in part, to widespread use of disease modifying therapies for RMS. Although the prognosis in an individual is difficult to establish, certain clinical features suggest a more favorable prognosis. These include ON or sensory symptoms at onset; fewer than two relapses in the first year of illness; and minimal impairment after 5 years. By contrast, patients with truncal ataxia, action tremor, pyramidal symptoms, or a progressive disease course are more likely to become disabled. Patients with a long-term favorable course are likely to have developed fewer MRI lesions and have less brain atrophy during the early years of disease, and vice versa. Importantly, some MS patients have a benign variant of MS and never develop neurologic disability. The likelihood of having benign MS is thought to be <10%. Patients with benign MS 15 years after onset who have entirely normal neurologic examinations are likely to maintain their benign course.

In patients with their first demyelinating event (i.e., a clinically isolated syndrome), the brain MRI provides prognostic information. With three or more typical T2-weighted lesions, the risk of developing MS after 20 years is ~80%. Conversely, with a normal brain MRI, the likelihood of developing MS is <20%. Similarly, the presence of two or more Gd-enhancing lesions at baseline is highly predictive of future MS, as is the appearance of either new T2-weighted lesions or new Gd enhancement ≥3 months after the initial episode.

Effect of Pregnancy Pregnant MS patients experience fewer attacks than expected during gestation (especially in the last trimester), but more attacks than expected in the first 3 months postpartum. When considering the pregnancy year as a whole (i.e., 9 months of pregnancy plus 3 months postpartum), the overall disease course is unaffected. Decisions about childbearing should thus be made based on (1) the

mother's physical state, (2) her ability to care for the child, and (3) the availability of social support. Disease-modifying therapy is generally discontinued during pregnancy, although the actual risk from the interferons and glatiramer acetate (see below) appears to be low.

TREATMENT

Multiple Sclerosis

Therapy for MS can be divided into several categories: (1) treatment of acute attacks, (2) treatment with disease-modifying agents that reduce the biologic activity of MS, and (3) symptomatic therapy. Treatments that promote remyelination or neural repair do not currently exist, but several promising approaches are being actively investigated.

The Expanded Disability Status Score (EDSS) is a widely used measure of neurologic impairment in MS **(Table 436-5)**. Most patients with EDSS scores <3.5 walk normally, and are generally not disabled; by contrast, patients with EDSS scores >4.0 have progressive MS (SPMS or PPMS), are gait-impaired, and often are occupationally disabled.

ACUTE ATTACKS OR INITIAL DEMYELINATING EPISODES

When patients experience acute deterioration, it is important to consider whether this change reflects new disease activity or a "pseudoexacerbation" resulting from an increase in ambient temperature, fever, or an infection. When the clinical change is thought to reflect a pseudoexacerbation, glucocorticoid treatment is inappropriate. Glucocorticoids are used to manage either first attacks or acute exacerbations. They provide short-term clinical benefit by reducing the severity and shortening the duration of attacks. Whether treatment provides any long-term benefit on the course of the illness is less clear. Therefore, mild attacks are often not treated. Physical and occupational therapy can help with mobility and manual dexterity.

Glucocorticoid treatment is usually administered as intravenous methylprednisolone, 500–1000 mg/d for 3–5 days, either without a taper or followed by a course of oral prednisone beginning at a dose of 60–80 mg/d and gradually tapered over 2 weeks. Orally administered methylprednisolone, prednisone, or dexamethasone (in equivalent dosages) can be substituted for the intravenous portion of the therapy. Outpatient treatment is almost always possible.

Side effects of short-term glucocorticoid therapy include fluid retention, potassium loss, weight gain, gastric disturbances, acne, and emotional lability. Concurrent use of a low-salt, potassium-rich diet and avoidance of potassium-wasting diuretics are advisable. Lithium carbonate (300 mg orally bid) may help manage emotional lability and insomnia associated with glucocorticoid therapy. Patients with a history of peptic ulcer disease may require cimetidine (400 mg bid) or ranitidine (150 mg bid). Proton pump inhibitors such as pantoprazole (40 mg orally bid) may reduce the likelihood of gastritis, especially when large doses are administered orally. Plasma exchange (five to seven exchanges: 40–60 mL/kg per exchange, every other day for 14 days) may benefit patients with fulminant attacks of demyelination that are unresponsive to glucocorticoids. However, the cost is high, and conclusive evidence of efficacy is lacking.

TABLE 436-4 Disorders That Can Mimic Multiple Sclerosis (MS)

Acute disseminated encephalomyelitis (ADEM)
Antiphospholipid antibody syndrome
Behçet's disease
Cerebral autosomal dominant arteriopathy, subcortical infarcts, and leukoencephalopathy (CADASIL)
Congenital leukodystrophies (e.g., adrenoleukodystrophy, metachromatic leukodystrophy)
Human immunodeficiency virus (HIV) infection
Ischemic optic neuropathy (arteritic and nonarteritic)
Lyme disease
Mitochondrial encephalopathy with lactic acidosis and stroke (MELAS)
Neoplasms (e.g., lymphoma, glioma, meningioma)
Sarcoid
Sjögren's syndrome
Stroke and ischemic cerebrovascular disease
Syphilis
Systemic lupus erythematosus and related collagen vascular disorders
Tropical spastic paraparesis (HTLV-1/2 infection)
Vascular malformations (especially spinal dural AV fistulas)
Vasculitis (primary CNS or other)
Vitamin B_{12} deficiency

Abbreviations: AV, arteriovenous; CNS, central nervous system; HTLV, human T cell lymphotropic virus.

DISEASE-MODIFYING THERAPIES FOR RELAPSING FORMS OF MS (RMS, SPMS WITH EXACERBATIONS)

More than a dozen immunomodulatory and immunosuppressive agents are approved by regulatory bodies for treatment of RMS. In phase 3 clinical trials, each was shown to reduce the frequency of clinical relapses and evolution of new brain MRI lesions in relapsing forms of MS **(Table 436-6)**. Each can also be used in SPMS patients who continue to experience attacks, both because SPMS can be difficult to distinguish from relapsing MS and because the available clinical trials, although not definitive, suggest that such patients may sometimes derive therapeutic benefit. When considering the data in Table 436-6, however, it is important to note that the relative efficacy of the different agents has not been directly tested in head-to head studies and that cross-trial comparisons are inaccurate. However, given the increasingly complex landscape of therapeutics for MS, for convenience the discussion of these agents has been divided into those used more and less frequently; and also by an estimate of their relative (modest, moderate or high) perceived level of efficacy. These are meant to serve as only very rough guides, and considerable variance exists in practice patterns, as well as availability of these agents, in different parts of the world.

FREQUENTLY USED AGENTS FOR RMS

Interferon β (Modestly Effective) Interferon β (IFN-β) is a class I interferon originally identified by its antiviral properties. Efficacy in MS probably results from immunomodulatory properties including (1) downregulating expression of MHC molecules on antigen-presenting cells, (2) reducing proinflammatory and increasing regulatory cytokine levels, (3) inhibiting T-cell proliferation, and (4) limiting the trafficking of inflammatory cells in the CNS. IFN-β reduces the attack rate and slows accumulation of disability and MRI-documented disease burden. IFN-β should be considered in patients with either relapsing forms of MS (either RMS or SPMS with superimposed relapses). Head-to-head trials suggest that dosing IFN-β more frequently and at higher doses has better efficacy but is also more likely to induce neutralizing antibodies (see below). IFN-β-1a (Avonex), 30 μg, is administered by intramuscular injection once every week. IFN-β-1a (Rebif), 44 μg, is administered by subcutaneous injection three times per week. IFN-β-1b (Betaseron or Extavia), 250 μg, is administered by subcutaneous injection every other day. Pegylated IFN-β-1a (Plegridy), 125 μg, is administered by subcutaneous injection once every 14 days. Pegylated IFN-β-1a is an interferon to which a single, linear 20,000 dalton methoxy poly(ethyleneglycol)-O-2-methylpopionaldehyde molecule is covalently attached; the pegylated molecule contributes to reduced in vivo clearance allowing less frequent administration.

Common side effects of IFN-β therapy include flulike symptoms (e.g., fevers, chills, and myalgias) and mild abnormalities on routine laboratory evaluation (e.g., elevated liver function tests or lymphopenia). Rarely, more severe hepatotoxicity may occur. Subcutaneous IFN-β also causes reactions at the injection site (e.g., pain, redness, induration, or, rarely, skin necrosis). Side effects can usually be managed with concomitant nonsteroidal anti-inflammatory medications. Depression, increased spasticity, and cognitive changes have been reported, although these symptoms can also be due to

TABLE 436-5 Scoring Systems for Multiple Sclerosis (MS)

Kurtzke Expanded Disability Status Score (EDSS)

0.0 = Normal neurologic examination (all grade 0 in functional status [FS])
1.0 = No disability, minimal signs in one FS (i.e., grade 1)
1.5 = No disability, minimal signs in more than one FS (more than one grade 1)
2.0 = Minimal disability in one FS (one FS grade 2, others 0 or 1)
2.5 = Minimal disability in two FS (two FS grade 2, others 0 or 1)
3.0 = Moderate disability in one FS (one FS grade 3, others 0 or 1) or mild disability in three or four FS (three/four FS grade 2, others 0 or 1) although fully ambulatory
3.5 = Fully ambulatory but with moderate disability in one FS (one grade 3) and one or two FS grade 2; or two FS grade 3; or five FS grade 2 (others 0 or 1)
4.0 = Ambulatory without aid or rest for ~500 m
4.5 = Ambulatory without aid or rest for ~300 m
5.0 = Ambulatory without aid or rest for ~200 m
5.5 = Ambulatory without aid or rest for ~100 m
6.0 = Unilateral assistance required to walk about 100 m with or without resting
6.5 = Constant bilateral assistance required to walk about 20 m without resting
7.0 = Unable to walk beyond about 5 m even with aid; essentially restricted to wheelchair; wheels self and transfers alone
7.5 = Unable to take more than a few steps; restricted to wheelchair; may need aid to transfer
8.0 = Essentially restricted to bed or chair or perambulated in wheelchair, but out of bed most of day; retains many self-care functions; generally has effective use of arms
8.5 = Essentially restricted to bed much of the day; has some effective use of arm(s); retains some self-care functions
9.0 = Helpless bed patient; can communicate and eat
9.5 = Totally helpless bed patient; unable to communicate or eat
10.0 = Death due to MS

Functional Status (FS) Score

A. Pyramidal functions
0 = Normal
1 = Abnormal signs without disability
2 = Minimal disability
3 = Mild or moderate paraparesis or hemiparesis, or severe monoparesis
4 = Marked paraparesis or hemiparesis, moderate quadriparesis, or monoplegia
5 = Paraplegia, hemiplegia, or marked quadriparesis
6 = Quadriplegia

B. Cerebellar functions
0 = Normal
1 = Abnormal signs without disability
2 = Mild ataxia
3 = Moderate truncal or limb ataxia
4 = Severe ataxia all limbs
5 = Unable to perform coordinated movements due to ataxia

C. Brainstem functions
0 = Normal
1 = Signs only
2 = Moderate nystagmus or other mild disability
3 = Severe nystagmus, marked extraocular weakness, or moderate disability of other cranial nerves
4 = Marked dysarthria or other marked disability
5 = Inability to swallow or speak

D. Sensory functions
0 = Normal
1 = Vibration or figure-writing decrease only, in 1 or 2 limbs
2 = Mild decrease in touch or pain or position sense, and/or moderate decrease in vibration in 1 or 2 limbs, or vibratory decrease alone in 3 or 4 limbs
3 = Moderate decrease in touch or pain or position sense, and/or essentially lost vibration in 1 or 2 limbs, or mild decrease in touch or pain, and/or moderate decrease in all proprioceptive tests in 3 or 4 limbs
4 = Marked decrease in touch or pain or loss of proprioception, alone or combined, in 1 or 2 limbs or moderate decrease in touch or pain and/or severe proprioceptive decrease in >2 limbs
5 = Loss (essentially) of sensation in 1 or 2 limbs or moderate decrease in touch or pain and/or loss of proprioception for most of the body below the head
6 = Sensation essentially lost below the head

E. Bowel and bladder functions
0 = Normal
1 = Mild urinary hesitancy, urgency, or retention
2 = Moderate hesitancy, urgency, retention of bowel or bladder, or rare urinary incontinence
3 = Frequent urinary incontinence
4 = In need of almost constant catheterization
5 = Loss of bladder function
6 = Loss of bowel and bladder function

F. Visual (or optic) functions
0 = Normal
1 = Scotoma with visual acuity (corrected) better than 20/30
2 = Worse eye with scotoma with maximal visual acuity (corrected) of 20/30 to 20/59
3 = Worse eye with large scotoma, or moderate decrease in fields, but with maximal visual acuity (corrected) of 20/60 to 20/99
4 = Worse eye with marked decrease of fields and maximal acuity (corrected) of 20/100 to 20/200; grade 3 plus maximal acuity of better eye of 20/60 or less
5 = Worse eye with maximal visual acuity (corrected) <20/200; grade 4 plus maximal acuity of better eye of ≤20/60
6 = Grade 5 plus maximal visual acuity of better eye of ≤20/60

G. Cerebral (or mental) functions
0 = Normal
1 = Mood alteration only (does not affect EDSS score)
2 = Mild decrease in mentation
3 = Moderate decrease in mentation
4 = Marked decrease in mentation
5 = Chronic brain syndrome—severe or incompetent

Source: Adapted from JF Kurtzke: Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS). Neurology 33:1444, 1983.

the underlying disease. Side effects due to IFN-β therapy usually subside over time.

Approximately 2–10% of IFN-β-1a (Avonex) recipients, 15–25% of IFN-β-1a (Rebif) recipients, and 30–40% of IFN-β-1b (Betaseron/Extavia) recipients develop neutralizing antibodies to IFN-β, which may disappear over time. Less than 1% of patients treated with pegylated IFN-β-1a develop neutralizing antibodies. For a patient doing well on therapy, the presence of antibodies should not affect treatment. Conversely, for a patient doing poorly on therapy, alternative treatment should be considered, even if there are no detectable antibodies.

Glatiramer Acetate (Modestly Effective) Glatiramer acetate is a synthetic, random polypeptide composed of four amino acids (L-glutamic acid, L-lysine, L-alanine, and L-tyrosine). Its mechanism of action may include (1) induction of antigen-specific suppressor T cells; (2) binding to MHC molecules, thereby displacing bound MBP; or (3) altering the balance between proinflammatory and regulatory cytokines. Glatiramer acetate reduces the attack rate

TABLE 436-6 Outcomes for FDA-Approved Therapies for Multiple Sclerosis (MS)[a]

DOSE, ROUTE, AND SCHEDULE	STUDY DURATION WEEKS	COMPARATOR	CLINICAL OUTCOMES[b]		MRI OUTCOMES[c]	
			ATTACK RATE, MEAN	CHANGE IN DISEASE SEVERITY	NEW T2 LESIONS[d]	TOTAL BURDEN OF DISEASE
RELAPSING MS						
IFN-β-1b, 250 μg SC qod	96	PBO	–34%[e]	–29% (NS)	–83%[f]	–17%[e]
IFN-β-1a, 30 μg IM qw	96	PBO	–18%[g]	–37%[g]	–36%[f]	NS
IFN-β-1a, 44 μg SC tiw	96	PBO	–32%[e]	–30%[g]	–78%[e]	–15%[e]
Peg- IFN-β-1a, 125 μg SC q2w	48	PBO	–36%[e]	–38%[g]	–67%[e]	–2%[e]
GA, 20 mg SC qd	96	PBO	–29%[f]	–12% (NS)	–38%[f]	–8%[f]
MTX, 12 mg/m^2 IV q3mo	96	PBO	–66%[e]	–75%[g]	–79%[g]	NR
NTZ, 300 mg IV qmo	96	PBO	–68%[e]	–42%[e]	–83%[e]	–18%[e]
FNG, 0.5 mg PO qd	96	PBO	–55%[e]	–34%[f]	–74%[e]	–23%[e]
DMF, 240 mg PO bid	96	PBO	–52%[e]	–40%[f]	–71%[e]	NR
TF, 14 mg PO qd	96	PBO	–31%[e]	–26%[g]	–70%[e]	–20%[g]
FNG, 0.5 mg PO qd	48	IFN-β-1a, 30 μg IM qw	–52%[e]	NS	–35%[e]	NS
ALEM, 12 mg/m^2 IV/5 d	104	IFN-β-1a, 44 μg SC tiw	–49%[e]	–42%[f]	–32%[e]	NS
OCR, 600 mg IV, Q6 mo	96	IFN-β-1a, 44 μg SC tiw	–46%[e,h]	–33%[e,h]	–80%[e,h]	NR
Primary Progressive MS						
OCR, 600 mg IV, Q6 mo	96	PBO	–NR	–24%	–92%	–11%

[a]Percentage reductions (or increases) have been calculated by dividing the reported rates in the treated group by the comparable rates in the placebo group, except for magnetic resonance imaging (MRI) disease burden, which was calculated as the difference in the median percent change between the treated and placebo groups. [b]Severity = 1 point Expanded Disability Status Score progression, sustained for 3 months (in the IFN-β-1a 30 μg qw trial, this change was sustained for 6 months; in the IFN-β-1b trial, this was over 3 years). [c]Different studies measured these MRI measures differently, making comparisons difficult (numbers for new T2 represent the best case scenario for each trial). [d]New lesions seen on T2-weighted MRI. [e]$p = .001$. [f]$p = .01$. [g]$p = .05$. [h]Pooled analysis from OPERA 1 and 2 studies.

Abbreviations: DMF, dimethyl fumarate; FDA, U.S. Food and Drug Administration; FNG, fingolimod; GA, glatiramer acetate; IFN-β, interferon β; IM, intramuscular; IV, intravenous; MTX, mitoxantrone; NR, not reported; NS, not significant; NTZ, natalizumab; PO, oral; q3mo, once every 3 months; qd, daily; qmo, once per month; qod, every other day; qw, once per week; qyr, once per year; SC, subcutaneous; TF, teriflunomide; tiw, three times per week.

(whether measured clinically or by MRI) in RMS. Glatiramer acetate also benefits disease severity measures, although, for clinical disability, this is less well established than for IFN-β. Nevertheless, two head-to-head trials demonstrated that the impact of glatiramer acetate on clinical relapse rates and disability was comparable to high-dose, high-frequency IFN-β. Therefore, glatiramer acetate should be considered as an equally effective alternative to IFN-β in RMS patients. Its usefulness in progressive disease is unknown. Glatiramer acetate is administered by subcutaneous injection of either 20 mg every day or 40 mg thrice weekly. Injection-site reactions also occur with glatiramer acetate. In addition, ~15% of patients experience one or more episodes of flushing, chest tightness, dyspnea, palpitations, and anxiety after injection. This systemic reaction is unpredictable, brief (duration <1 h), and tends not to recur. Finally, some patients experience lipoatrophy, which, on occasion, can be disfiguring and require cessation of treatment. Recently, glatiramer acetate was U.S. Food and Drug Administration (FDA) approved as a biosimilar medication (Glatopa) and is dosed at 20 mg every day. Although clinical trials were not performed with biosimilar glatiramer acetate, the efficacy and safety are presumed to be similar to the branded product.

Fingolimod (Moderately Effective) Fingolimod is a sphingosine-1-phosphate (S1P) inhibitor that prevents the egress of lymphocytes from secondary lymphoid organs such as the lymph nodes and spleen. Its mechanism of action is probably due to sequestration of lymphocytes in the periphery, thereby inhibiting their trafficking to the CNS. Fingolimod reduces the attack rate and significantly improves all measures of disease severity in MS. It is well tolerated, and the daily oral dosing schedule makes it convenient for patients. A head-to-head phase 3 randomized study demonstrated the superiority of fingolimod over low-dose (weekly) IFN-β-1a.

Fingolimod, 0.5 mg, is administered orally each day. Treatment with fingolimod is well tolerated. Mild abnormalities on routine laboratory evaluation (e.g., elevated liver function tests or lymphopenia) are more common than in controls, sometimes requiring discontinuation of the medication. First- and second-degree heart block and bradycardia can also occur when fingolimod therapy is initiated. A 6-h period of observation (including electrocardiogram monitoring) is recommended for all patients receiving their first dose. Other side effects include macular edema and, rarely, disseminated varicella-zoster virus (VZV) and cryptococcal infections; prior to initiating therapy with fingolimod, an ophthalmic examination and VZV vaccination for seronegative individuals are indicated. Fingolimod can also cause QT prolongation with the potential for drug-drug interactions with other medications that also prolong the QT interval.

Dimethyl Fumarate (DMF) (Moderately Effective) DMF is a small molecule and is a Krebs cycle metabolite with anti-inflammatory effects in psoriasis. Although the precise mechanisms of action of DMF are not fully understood, it seems to modulate the expression of proinflammatory and anti-inflammatory cytokines. Also, DMF inhibits the ubiquitylation and degradation of nuclear factor E2-related factor 2 (Nrf2)—a transcription factor that binds to the antioxidant response elements (AREs) located on the DNA and thereby induces the transcription of several antioxidant proteins. DMF reduces the attack rate and significantly improves all measures of disease severity in MS patients. However, its twice-daily oral dosing schedule makes it somewhat less convenient for patients than daily oral therapies. In addition, compliance is likely to be less with a twice-daily dosing regimen—a factor that could be of concern given the observation (in a small clinical trial) that once-daily DMF lacks efficacy. A head-to-head trial provided evidence that DMF was superior to glatiramer acetate on some outcome measures.

DMF, 240 mg, is administered orally twice each day. Gastrointestinal side effects (abdominal discomfort, nausea, vomiting, flushing, and diarrhea) are common at the start of therapy but generally subside with continued administration. Other adverse events include flushing mild decreases in neutrophil and lymphocyte counts and elevations in liver enzymes. Nevertheless, in general, treatment with DMF is well tolerated after an initial period of adjustment. Following the release of DMF, several cases of progressive multifocal leukoencephalopathy (PML) were reported in patients receiving products that contained DMF. Most of these patients were lymphopenic and monitoring for lymphopenia every 6 months is recommended. Patients who are persistently lympopenic (lymphocyte count <500 cells/mL) are recommended to consider alternate treatments

due to the PML risk. Clinically significant liver injury has been reported with DMF treatment. Liver function tests should be assessed before treatment and when clinically indicated. Elevations in liver function tests resolve the following treatment discontinuation.

Natalizumab (Highly Effective) Natalizumab is a humanized monoclonal antibody directed against the α_4 subunit of $\alpha_4\beta_1$ integrin, a cellular adhesion molecule expressed on the surface of lymphocytes. It prevents lymphocytes from binding to endothelial cells, thereby preventing lymphocytes from penetrating the BBB and entering the CNS. Natalizumab is highly effective in reducing the attack rate and significantly improves all measures of disease severity in MS (both clinical and MRI). Moreover, it is well-tolerated, and the dosing schedule of monthly intravenous infusions makes it very convenient for patients. Natalizumab, 300 mg, is administered by IV infusion each month. Treatment with natalizumab is, in general, well tolerated. A small percentage (<10%) of patients experience hypersensitivity reactions (including anaphylaxis), and ~6% develop neutralizing antibodies to the molecule (only half of which persist).

The major concern with long-term treatment is the risk of PML, a life-threatening condition resulting from infection by the John Cunningham (JC) virus. PML has occurred in ~0.4% of patients treated with natalizumab. The incidence of PML is very low in the first year of treatment but then rises in subsequent years of treatment to reach a level of about 2 cases per 1000 patients per year. Nevertheless, the measurement of antibodies against the JC virus in the serum can be used to stratify this risk. Approximately half of the adult population is JC antibody positive, indicating that they experienced an asymptomatic infection with the JC virus at some time in the past. Thus, in patients who do not have these antibodies, the risk of PML is minimal (<1:10,000 as long as they remain JC antibody free). Conversely, in patients who have these antibodies (especially those who have them in high titer), the risk may be as high as ≥1.1%. Up to 2% of seronegative MS patients undergoing treatment with natalizumab seroconvert annually; thus, it is recommended that JC antibody status be assessed at 6-month intervals in all patients receiving natalizumab treatment. In antibody-positive patients, a change to another disease-modifying therapy should be strongly considered. The risk of PML is also high in patients who previously received immunosuppressive therapy. Natalizumab is generally recommended only for JC antibody–negative patients, unless they have failed alternative therapies or if they have a particularly aggressive disease course.

Ocrelizumab (Highly Effective) Ocrelizumab is a humanized monoclonal antibody directed against the CD20 molecule present on the surface of mature B cells. CD20 is not expressed on early B-cell precursors or on antibody-producing plasma cells, thus treatment with ocrelizumab selectively depletes mature B cells while preserving preexisting humoral immunity and the capacity for B-cell reconstitution by lymphoid stem cells. Ocrelizumab rapidly depletes circulating B cells through antibody-dependent cellular toxicity and complement-dependent cytotoxicity. The beneficial effects of B-cell depletion in MS are not fully understood but may involve interruption in trafficking of B cells from the periphery to the CNS and through reduction in antigen presentation and/or modulation of cytokine secretion by B cells. Ocrelizumab targets the same molecule as rituximab, a monoclonal antibody indicated for non-Hodgkin's lymphoma and rheumatoid arthritis, and ofatumumab, indicated for treatment of chronic lymphocytic leukemia. In two phase 3 trials, ocrelizumab demonstrated a high degree of efficacy against RMS, reducing annualized relapse rates by 47%, reducing new MRI lesions by 95%, and improving other measures of inflammatory and degenerative disease activity, compared with three times per week interferon β-1a (Rebif). Ocrelizumab 600 mg is administered by intravenous infusion every 24 weeks (administered as two 300-mg infusions spaced 2 weeks apart for the first dose, and as a single 600-mg infusion thereafter); intravenous methylprednisolone 100 mg is given prior to each infusion and optional prophylaxis with analgesics/antipyretics and antihistamines are recommended, along with adjustment of the infusion rate to manage infusion-related reactions.

Ocrelizumab is generally well tolerated with infusion-related reactions occurring in a minority of patients; these are most often observed with the first infusion and are usually mild in degree. Rituximab is associated with a very small risk of PML (estimated at <1:25,000/year), thus it is possible that ocrelizumab will also carry a nonzero risk. Ocrelizumab may also carry some risk of increased malignancies including breast cancer, although rituximab is not associated with an increased risk of malignancy. Additional study of ocrelizumab in the postmarketing setting is needed to determine whether there is in fact a fractional increased malignancy risk associated with ocrelizumab.

LESS COMMONLY USED AGENTS FOR RMS

Teriflunomide (Modestly Effective) Teriflunomide inhibits the mitochondrial enzyme dihydro-orotate dehydrogenase, which is a key part of the pathway for de novo pyrimidine biosynthesis from carbamoyl phosphate and aspartate. It is the active metabolite of the drug leflunomide (FDA-approved for rheumatoid arthritis), and it exerts its anti-inflammatory effects by limiting the proliferation of rapidly dividing T and B cells. This enzyme is not involved in the so-called "salvage pathway," by which existing pyrimidine pools are recycled for DNA and RNA synthesis in resting and homeostatically proliferating cells. Consequently, teriflunomide is considered to be cytostatic rather than cytotoxic. Teriflunomide reduces the attack rate and significantly improves all measures of disease severity in MS patients. It is well tolerated, and its daily oral dosing schedule makes it very convenient for patients. A head-to-head trial suggested the equivalence, but not superiority, of teriflunomide and thrice-weekly IFN-β-1a. Teriflunomide, either 7 or 14 mg, is administered orally each day. In the pivotal clinical trials, mild hair thinning and gastrointestinal symptoms (nausea and diarrhea) were more common than in controls, but in general, treatment with teriflunomide was well tolerated. Teriflunomide rarely causes toxic epidermal necrolysis or Stevens-Johnson syndrome. A major limitation, especially in women of childbearing age, is its possible teratogenicity (pregnancy category X); teriflunomide can remain in the bloodstream for 2 years due to hepatobiliary reabsorption. Therefore, it is recommended that exposed men and women who wish to conceive receive cholestyramine or activated charcoal to eliminate residual drug.

Alemtuzumab (Highly Effective) Alemtuzumab is a humanized monoclonal antibody directed against the CD52 antigen that is expressed on both monocytes and lymphocytes. It causes lymphocyte depletion (of both B and T cells) and a change in the composition of lymphocyte subsets. Both of these changes, particularly the impact on lymphocyte subsets, are long lasting. In two phase 3 trials, which used the active comparator of thrice-weekly, high-dose IFN-β-1a, alemtuzumab markedly reduced the attack rate and significantly improved measures of disease severity in MS patients although its impact on clinical disability was found in only one of the two trials. The European and Canadian drug agencies were the first to approve this agent for use in RMS; the FDA has also approved alemtuzumab, but only after an appeal following initial disapproval. The reasons for the initial disapproval were based on a perceived lack of a convincing disability effect and concerns over potential toxicity. The toxicities of concern were the occurrence of (1) autoimmune diseases including thyroiditis, Graves' disease, thrombocytopenia, hemolytic anemia, pancytopenia, antiglomerular basement membrane disease, and membranous glomerulonephritis; (2) malignancies including thyroid cancer, melanoma, breast cancer, human papillomavirus (HPV)–related cancers, and lymphoproliferative disorders including lymphoma; (3) serious infections; and (4) infusion reactions. Because of its toxicity profile, the FDA indicated alemtuzumab only in patients who have tried and failed at least two other DMTs.

Mitoxantrone Hydrochloride (Highly Effective) Mitoxantrone, an anthracenedione, exerts its antineoplastic action by (1) intercalating into DNA and producing both strand breaks and interstrand cross-links, (2) interfering with RNA synthesis, and (3) inhibiting topoisomerase II (involved in DNA repair). The FDA approved mitoxantrone on the basis of a single phase 3 clinical trial in Europe, in addition to an even smaller phase 2 studies. Mitoxantrone is indicated for use in patients with rapidly worsening MS (defined as patients whose neurologic status remains significantly abnormal between MS attacks). Despite this broad indication, however, data supporting its efficacy are less robust compared to other approved therapies.

Mitoxantrone is cardiotoxic (e.g., cardiomyopathy, reduced left ventricular ejection fraction, and irreversible congestive heart failure). As a result, a cumulative dose >140 mg/m^2 is not recommended. At currently approved doses (12 mg/m^2 every 3 months), the maximum duration of therapy can be only 2–3 years. Furthermore, >40% of women will experience amenorrhea, which may be permanent. Finally, there is risk of acute leukemia from mitoxantrone, estimated as at least a 1.4% lifetime risk. Because of these risks, and the availability of alternative therapies, mitoxantrone is now rarely used for MS.

Daclizumab, a monoclonal antibody against CD25, the α subunit of the interleukin 2 receptor, was removed from the market in 2018 because of reports of brain inflammation; it was previously approved for patients who had failed at least two other therapies.

Initiating and Changing Treatment Previously, most patients with relapsing forms of MS received injectable agents (IFN-β or glatiramer acetate) as first-line therapy. However, with the introduction of effective oral agents that include dimethyl fumarate, fingolimod, and teriflunomide, this has begun to change. In addition, the monthly infusion therapy natalizumab, which is highly effective, well tolerated, and apparently safe in JC antibody–negative patients, provides an attractive option for many patients. Ocrelizumab can also be used first-line; the combination of high efficacy, infrequently administered infusions, and a favorable safety profile make its use an attractive option. With the exception of the first-generation injectable agents, mitoxantrone and natalizumab long-term (>10 year) safety data are not available for the newer therapies, and in many cases comparative data are lacking. The value of combination therapy is also largely unknown, although a clinical trial demonstrated no added benefit to the combination of glatiramer acetate with once-weekly IFN-β-1a.

Despite these unknowns, clinicians need to make decisions based on the best available evidence, coupled with practical considerations. One reasonable approach stratifies initial decision-making based on two levels of disease aggressiveness **(Fig. 436-4)**.

Mild Initial Course In the case of recent onset, normal examination or minimal impairment (EDSS ≤2.5 or less), or low disease activity, either an injectable (IFN-β or glatiramer acetate) or an oral (DMF, fingolimod, or teriflunomide) agent is reasonable. Although head-to-head comparisons are not available, natalizumab is thought to be more effective than these other agents, and, therefore, this therapy can be considered even in minimally affected, JCV antibody–seronegative patients. Ocrelizumab is more effective than IFN-β-1a TIW and appears to be safe although long-term data are not available. Therefore, this therapy can also be considered in recent onset MS patients regardless of JCV serology status. The injectable agents (IFN-β and glatiramer acetate) have a superb track record for safety but have a high nuisance factor due to the need for frequent injections, as well as bothersome side effects that contribute to noncompliance. Some of the oral agents (DMF and fingolimod) are probably more effective than the injectables, but long-term risks are still unknown; DMF produces flushing and bothersome gastrointestinal symptoms in many patients at least initially (can be mitigated by beginning at one-quarter strength and gradually advancing to full dose), and fingolimod can lead to bradycardia and other cardiac conduction disturbances of unclear clinical significance. Teriflunomide may be less effective than the other oral agents, and there are concerns about its possible long-lasting pregnancy risks as well as its association with toxic epidermal necrolysis and Stevens-Johnson syndrome. Nevertheless, its long-term safety is likely known because teriflunomide

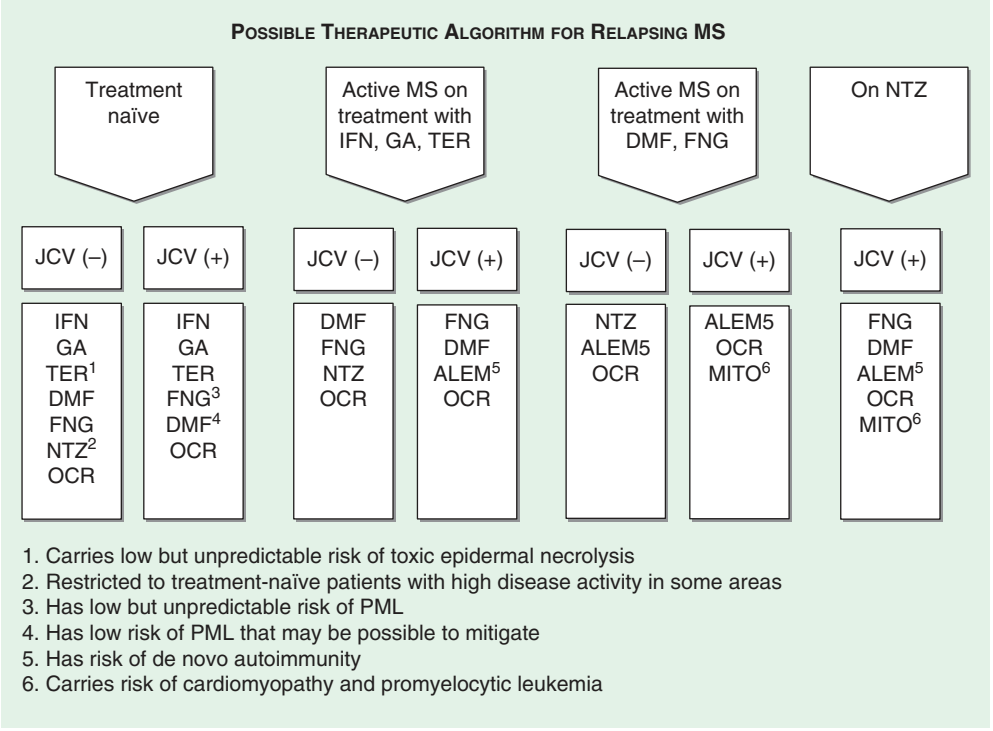


FIGURE 436-4 Therapeutic decision-making for relapsing MS. Options are shown for different clinical scenarios and based on JCV status. Active MS is defined by clinical relapses or the development of new focal MRI white matter lesions. Treatment options can also include trials of different preparations of interferon β (IFN-β), particularly advancing from once-weekly (Avonex) to a more frequent (e.g., Rebif, Betaseron/Extavia) dosing regimen, as well as use of natalizumab in JC virus–positive patients.

is the active metabolite of leflunomide—a drug long approved by the FDA for treatment of rheumatoid arthritis.

Moderate or Severe Initial Course In highly active disease or moderate impairment (EDSS >2.5), either a highly effective oral agent (DMF or fingolimod) or ocrelizumab or, if the patient is JC virus antibody seronegative, infusion therapy with natalizumab is recommended.

Regardless of which agent is chosen first, treatment should probably be changed in patients who continue to have relapses, progressive neurologic impairment or, arguably, ongoing evidence of subclinical MRI activity (Fig. 436-4).

The long-term impact of these treatments on the disease course remains controversial, although several recent observational studies showed that these agents improve the long-term outcome of MS including a prolongation of the time to reach certain disability outcomes (e.g., SPMS and requiring assistance to ambulate) and reduction in MS-related mortality. These benefits seem most conspicuous when treatment begins early in the relapsing stage of the illness. Unfortunately, however, already established progressive symptoms do not respond well to treatment with these disease-modifying therapies. Because progressive symptoms are likely to result from accumulated axonal and neuronal loss, many experts now believe that very early treatment with a disease-modifying drug is appropriate for most MS patients. It may also be reasonable to delay initiating treatment in patients with (1) normal neurologic examinations, (2) a single attack or a low attack frequency, and (3) a low burden of disease as assessed by brain MRI. Untreated patients, however, should be followed closely with periodic brain MRI scans; the need for therapy is reassessed if scans reveal evidence of ongoing, subclinical disease. Finally, vitamin D deficiency should be corrected in all patients with MS, and generally this requires oral supplementation with vitamin D_3, 4000–5000 IU daily. Several clinical trials showed that supplementation with vitamin D in relapsing MS patients reduces MRI measures of disease activity and may also reduce the relapse frequency in patients actively treated with either interferon or glatiramer acetate.

DISEASE-MODIFYING THERAPIES FOR PROGRESSIVE MS

SPMS High-dose IFN-β probably has a beneficial effect in patients with SPMS with active disease (see above). IFN-β is probably ineffective in patients with SPMS who do not have active disease. All of the other agents have not yet been studied in this patient population. Although mitoxantrone was approved for patients with progressive MS, this is not the population studied in the pivotal trial. Therefore, no evidence-based recommendation can be made with regard to its use in this setting.

PPMS Ocrelizumab (see above), a humanized monoclonal antibody that targets CD20 B-cells, was shown in a single phase 3 trial to reduce progression of clinical disability in PPMS by 24%, and also to improve other clinical and MRI markers of inflammatory and degenerative disease activity, compared with placebo treatment. Ocrelizumab represents the first agent to convincingly modify the course of PPMS. The dosing of ocrelizumab for PPMS is identical as for RMS (above).

OFF-LABEL TREATMENT OPTIONS FOR RMS AND SPMS

Azathioprine (2–3 mg/kg per day) has been used primarily in relapsing MS. Meta-analysis of published trials suggests that azathioprine is marginally effective at lowering relapse rates, although a benefit on disability progression has not been demonstrated.

Methotrexate (7.5–20 mg/week) was shown in one study to slow the progression of upper extremity dysfunction in SPMS. Because of the possibility of developing irreversible liver damage, some experts recommend a blind liver biopsy after 2 years of therapy.

Cyclophosphamide (700 mg/m^2, every other month) may be helpful for treatment-refractory patients who are (1) otherwise in good health, (2) ambulatory, and (3) <40 years of age. Because cyclophosphamide can be used for periods in excess of 3 years, it may be preferable to mitoxantrone in these circumstances.

Intravenous immunoglobulin (IVIg), administered in monthly pulses (up to 1 g/kg) for up to 2 years, appears to reduce annual exacerbation rates. However, its use is limited because of its high cost, questions about optimal dose, and uncertainty about its having any impact on long-term disability.

Methylprednisolone in one study, administered as monthly high-dose intravenous pulses, reduced disability progression (see above).

Hematopoietic stem cell transplantation appears to be highly effective in reducing the occurrence of relapses and may improve disability in relapsing MS. However, this procedure carries a significant mortality risk. Randomized trials with appropriate comparators are needed in order to position this procedure with respect to available pharmacological interventions.

OTHER THERAPEUTIC CLAIMS

Many purported treatments for MS have never been subjected to scientific scrutiny. These include dietary therapies (e.g., the Swank diet, the Paleo Diet, the Wahls diet), megadose vitamins, calcium orotate, bee stings, cow colostrum, hyperbaric oxygen, procarin (a combination of histamine and caffeine), chelation, acupuncture, acupressure, various Chinese herbal remedies, and removal of mercury-amalgam tooth fillings, among many others. Patients should avoid costly or potentially hazardous unproven treatments. Many such treatments lack biologic plausibility. For example, no reliable case of mercury poisoning resembling typical MS has ever been described, therefore challenging the notion that removal of mercury-amalgam tooth fillings would be beneficial.

Although potential roles for EBV, human herpesvirus (HHV) 6, or chlamydia have been suggested for MS, these reports are unconfirmed, and treatment with antiviral agents or antibiotics is not recommended.

Recently, chronic cerebrospinal insufficiency (CCSVI) was proposed as a cause of MS with vascular-surgical intervention recommended. However, multiple independent studies have subsequently failed to even approximate the initial claims, and patients should be strongly advised to avoid diagnostic procedures and potentially dangerous surgery for this condition.

SYMPTOMATIC THERAPY

For all patients, it is useful to encourage attention to a healthy lifestyle, including maintaining an optimistic outlook, a healthy diet, and regular exercise as tolerated (swimming is often well-tolerated because of the cooling effect of cold water). It is reasonable also to correct vitamin D deficiency with oral vitamin D.

Ataxia/tremor is often intractable. Clonazepam, 1.5–20 mg/d; primidone, 50–250 mg/d; propranolol, 40–200 mg/d; or ondansetron, 8–16 mg/d, may help. Wrist weights occasionally reduce tremor in the arm or hand. Thalamotomy and deep-brain stimulation have been tried with mixed success.

Spasticity and *spasms* may improve with physical therapy, regular exercise, and stretching. Avoidance of triggers (e.g., infections, fecal impactions, bed sores) is extremely important. Effective medications include baclofen (20–120 mg/d), diazepam (2–40 mg/d), tizanidine (8–32 mg/d), dantrolene (25–400 mg/d), and cyclobenzaprine hydrochloride (10–60 mg/d). For severe spasticity, a baclofen pump (delivering medication directly into the CSF) can provide substantial relief.

Weakness can sometimes be improved with the use of potassium channel blockers such as 4-aminopyridine (20 mg/d) and 3,4-di-aminopyridine (40–80 mg/d), particularly in the setting where lower extremity weakness interferes with the patient's ability to ambulate. The FDA approved extended release 4-aminopyridine (at 10 mg twice daily), and this can be obtained either as dalfampridine (Ampyra) or through a compounding pharmacy. The principal concern with the use of these agents is the possibility of inducing seizures at high doses.

Pain is treated with anticonvulsants (carbamazepine, 100–1000 mg/d; phenytoin, 300–600 mg/d; gabapentin, 300–3600 mg/d; or pregabalin, 50–300 mg/d), antidepressants (amitriptyline, 25–150 mg/d; nortriptyline, 25–150 mg/d; desipramine, 100–300 mg/d; or

venlafaxine, 75–225 mg/d), or antiarrhythmics (mexiletine, 300–900 mg/d). If these approaches fail, patients should be referred to a comprehensive pain management program.

Bladder dysfunction management is best guided by urodynamic testing. Evening fluid restriction or frequent voluntary voiding may help *detrusor hyperreflexia*. If these methods fail, propantheline bromide (10–15 mg/d), oxybutynin (5–15 mg/d), hyoscyamine sulfate (0.5–0.75 mg/d), tolterodine tartrate (2–4 mg/d), or solifenacin (5–10 mg/d) may help. Coadministration of pseudoephedrine (30–60 mg) is sometimes beneficial.

Detrusor/sphincter dyssynergia may respond to phenoxybenzamine (10–20 mg/d) or terazosin hydrochloride (1–20 mg/d). Loss of reflex bladder wall contraction may respond to bethanechol (30–150 mg/d). However, both conditions often require catheterization.

Urinary tract infections should be treated promptly. Patients with postvoid residual urine volumes >200 mL are predisposed to infections. Prevention by urine acidification (with cranberry juice or vitamin C) inhibits some bacteria. Prophylactic administration of antibiotics is sometimes necessary but may lead to colonization by resistant organisms. Intermittent catheterization may help to prevent recurrent infections and reduce overflow incontinence.

Treatment of *constipation* includes high-fiber diets and fluids. Natural or other laxatives may help. *Fecal incontinence* may respond to a reduction in dietary fiber.

Depression should be treated. Useful drugs include the selective serotonin reuptake inhibitors (fluoxetine, 20–80 mg/d, or sertraline, 50–200 mg/d), the tricyclic antidepressants (amitriptyline, 25–150 mg/d; nortriptyline, 25–150 mg/d; or desipramine, 100–300 mg/d), and the nontricyclic antidepressants (venlafaxine, 75–225 mg/d).

Fatigue may improve with assistive devices, help in the home, or successful management of spasticity. Patients with frequent nocturia may benefit from anticholinergic medication at bedtime. Excessive daytime somnolence caused by MS may respond to amantadine (200 mg/d), methylphenidate (5–25 mg/d), modafinil (100–400 mg/d), or armodafinil (150–250 mg/d).

Cognitive problems may respond marginally to lisdexamfetamine (40 mg/d).

Paroxysmal symptoms respond dramatically to low-dose anticonvulsants (acetazolamide, 200–600 mg/d; carbamazepine, 50–400 mg/d; phenytoin, 50–300 mg/d; or gabapentin, 600–1800 mg/d).

Heat sensitivity may respond to heat avoidance, air-conditioning, or cooling garments.

Sexual dysfunction may be helped by lubricants to aid in genital stimulation and sexual arousal. Management of pain, spasticity, fatigue, and bladder/bowel dysfunction may also help. Sildenafil (50–100 mg), tadalafil (5–20 mg), or vardenafil (5–20 mg), taken 1–2 h before sex, are standard treatments for erectile dysfunction.

PROMISING EXPERIMENTAL THERAPIES

Numerous clinical trials are currently under way. These include studies on (1) selective oral sphingosine-1-phosphate receptor antagonists to sequester lymphocytes in secondary lymphoid organs; (2) high dose biotin to improve disability in progressive forms of MS; (3) molecules to promote remyelination; and (4) bone marrow transplantation.

CLINICAL VARIANTS OF MS

Acute MS (*Marburg's variant*) is a fulminant demyelinating process that in some cases progresses inexorably to death within 1–2 years. Typically, there are no remissions. When acute MS presents as a solitary, usually cavitary, lesion, a brain tumor is often suspected **(Fig. 436-5)**. In such cases, a brain biopsy is usually required to establish the diagnosis. Marburg's variant does not seem to follow infection or vaccination, and it is unclear whether this syndrome represents an extreme form of MS or another disease altogether.

Balo's concentric sclerosis is another fulminant demyelinating syndrome characterized by concentric brain or spinal cord lesions with alternating spheres of demyelination and remyelination (Fig. 436-5). For these fulminant demyelinating states, no controlled trials of therapy exist; high-dose glucocorticoids, plasma exchange, and cyclophosphamide have been tried, with uncertain benefit.

ACUTE DISSEMINATED ENCEPHALOMYELITIS (ADEM)

ADEM has a monophasic course and is most frequently associated with an antecedent infection (postinfectious encephalomyelitis); ~5% of ADEM cases follow immunization (postvaccinal encephalomyelitis). ADEM is far more common in children than adults, and many adult cases initially thought to represent ADEM subsequently experience late relapses qualifying as either MS or another chronic inflammatory disorder such as vasculitis, sarcoid, or lymphoma. The hallmark of ADEM is the presence of widely scattered foci of perivenular inflammation and demyelination that can involve both white matter and grey matter structures, in contrast to larger confluent white matter lesions typical of MS. In the most explosive form of ADEM, acute hemorrhagic leukoencephalitis, the lesions are vasculitic and hemorrhagic, and the clinical course is devastating.

Postinfectious encephalomyelitis is most frequently associated with the viral exanthems of childhood. Infection with measles virus is the most common antecedent (1 in 1000 cases). Worldwide, measles encephalomyelitis is still common, although use of the live measles vaccine has dramatically reduced its incidence in developed countries. An ADEM-like illness rarely follows vaccination with live measles vaccine (1–2 in 10^6 immunizations). ADEM is now most frequently associated with varicella (chickenpox) infections (1 in 4000–10,000 cases). It may also follow infection with rubella, mumps, influenza, parainfluenza, EBV, HHV-6, HIV, dengue, Zika, other viruses, and *Mycoplasma pneumoniae*. Some patients may have a nonspecific upper respiratory infection or no known antecedent illness. In addition to measles, postvaccinal encephalomyelitis may also follow the administration of vaccines for smallpox (5 cases per million), the Semple rabies, and Japanese encephalitis. Modern vaccines that do not require viral culture in CNS tissue have reduced the ADEM risk.

All forms of ADEM presumably result from a cross-reactive immune response to the infectious agent or vaccine that then triggers an inflammatory demyelinating response. Autoantibodies to MBP and to other myelin antigens have been detected in the CSF from some patients with ADEM. Attempts to demonstrate direct viral invasion of the CNS have been unsuccessful.

CLINICAL MANIFESTATIONS

In severe cases, onset is abrupt and progression rapid (hours to days). In postinfectious ADEM, the neurologic syndrome generally begins late in the course of the viral illness as the exanthem is fading. Fever reappears, and headache, meningismus, and lethargy progressing to coma may develop. Seizures are common. Signs of disseminated neurologic disease are consistently present (e.g., hemiparesis or quadriparesis, extensor plantar responses, lost or hyperactive tendon reflexes, sensory loss, and brainstem involvement). In ADEM due to chickenpox, cerebellar involvement is often conspicuous. CSF protein is modestly elevated (0.5–1.5 g/L [50–150 mg/dL]). Lymphocytic pleocytosis, generally ≥200 cells/μL, occurs in 80% of patients. Occasional patients have higher counts or a mixed polymorphonuclear-lymphocytic pattern during the initial days of the illness. Transient CSF oligoclonal banding has been reported. MRI usually reveals extensive changes in the brain and spinal cord, consisting of white matter hyperintensities on T2 and fluid-attenuated inversion recovery (FLAIR) sequences with Gd enhancement on T1-weighted sequences.

DIAGNOSIS

The diagnosis is most reliably established when there is a history of recent vaccination or viral exanthematous illness. In severe cases with predominantly cerebral involvement, acute encephalitis due to infection with herpes simplex or other viruses including HIV may be difficult to exclude **(Chap. 132)**; other considerations include hypercoagulable states including the antiphospholipid antibody syndrome, vasculitis, neurosarcoid, primary CNS lymphoma, or metastatic cancer. An explosive presentation of MS can mimic ADEM, and, especially

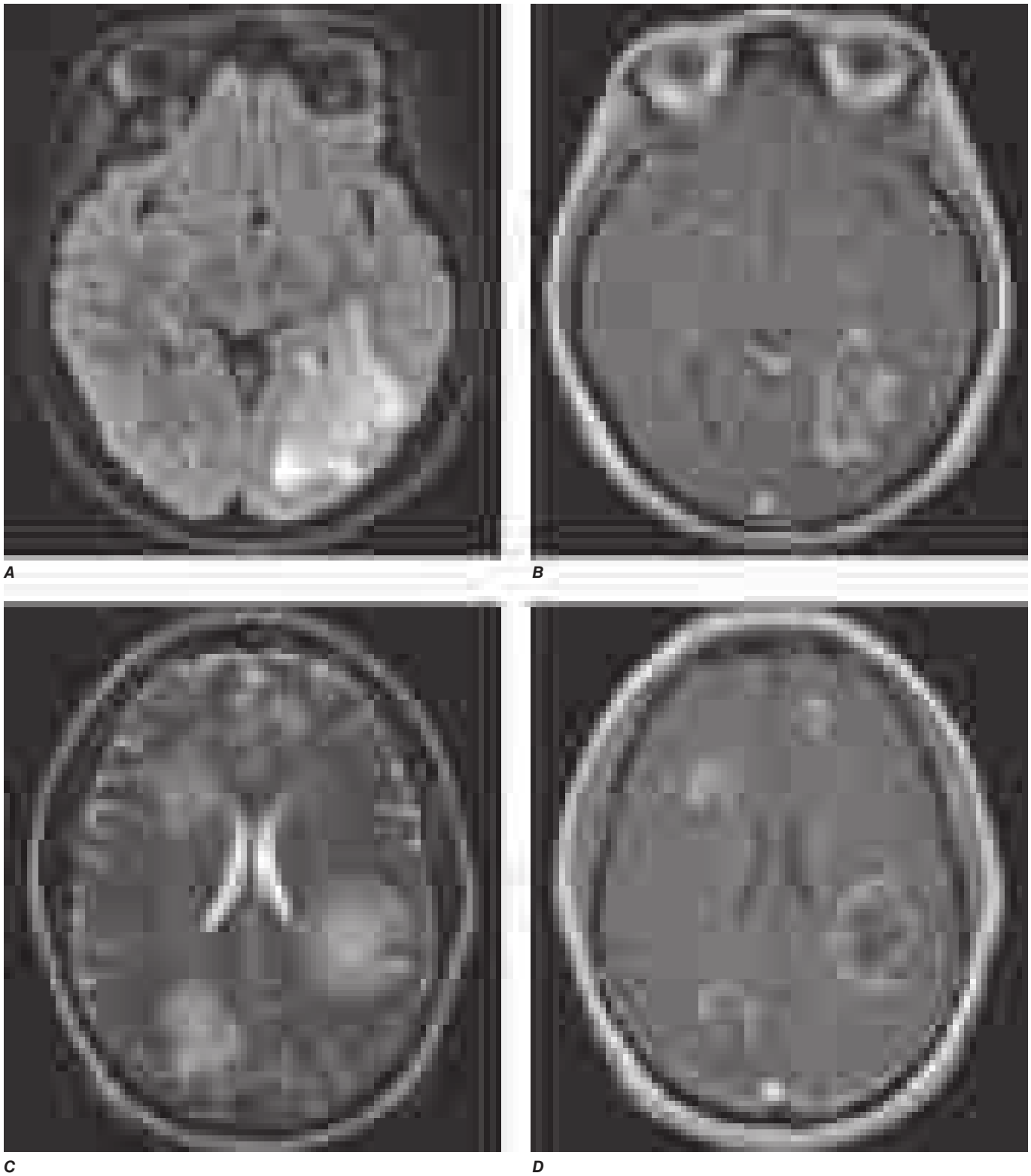


FIGURE 436-5 Magnetic resonance imaging findings in variants of MS. ***A*** and ***B.*** Acute tumefactive MS. In ***A***, a sagittal T2-weighted fluid-attenuated inversion recovery (FLAIR) image of a large solitary right parieto-occipital white matter lesion is shown, with effacement of overlying cortical sulci consistent with mass effect. In ***B***, T1-weighted image obtained after the intravenous administration of gadolinium DTPA reveals a large serpiginous area of blood-brain barrier disruption consistent with acute inflammation. ***C*** and ***D***. Balo's concentric sclerosis. In ***C***, an axial T2-weighted sequence shows multiple areas of abnormal ovoid bright signal in the supratentorial white matter bilaterally; some lesions reveal concentric layers, typical of Balo's concentric sclerosis. In ***D***, T1-weighted MR images postgadolinium demonstrate abnormal enhancement of all lesions with some lesions demonstrating concentric ring enhancement.

in adults, it may not be possible to distinguish these conditions at onset. The simultaneous onset of disseminated symptoms and signs is common in ADEM and rare in MS. Similarly, meningismus, drowsiness, coma, and seizures suggest ADEM rather than MS. Unlike MS, in ADEM, optic nerve involvement is generally bilateral and transverse myelopathy complete. MRI findings that favor ADEM include extensive and relatively symmetric white matter abnormalities, basal ganglia or cortical gray matter lesions, and Gd enhancement of all abnormal areas. By contrast, OCBs in the CSF are more common in MS. In one study of adult patients initially thought to have ADEM, 30% experienced additional relapses over a follow-up period of 3 years, and they were reclassified as having MS. Other patients initially classified

as ADEM are subsequently found to have neuromyelitis optica spectrum disorder (Chap. 437). Occasional patients with "recurrent ADEM" have also been reported, especially children; however, it is not possible to distinguish this entity from atypical MS. Because of the clinical overlap at presentation between ADEM and MS, it is crucial that routine surveillance imaging be performed following recovery from ADEM so that subclinical disease activity due to MS can be recognized and treatment for MS initiated.

TREATMENT

Acute Disseminated Encephalomyelitis

Initial treatment is with high-dose glucocorticoids; depending on the response, treatment may need to be continued for 8 weeks. Patients who fail to respond within a few days may benefit from a course of plasma exchange or intravenous immunoglobulin. The prognosis reflects the severity of the underlying acute illness. In recent case series of presumptive ADEM in adults, mortality rates of 5–20% are reported, and many survivors have permanent neurologic sequelae.

Acknowledgment

The authors want to thank Douglas E. Goodin for his contributions to previous editions of this chapter.

FURTHER READING

Arnold DL et al: Peginterferon beta-1a improves MRI measures and increases the proportion of patients with no evidence of disease activity in relapsing-remitting multiple sclerosis: 2-Year results from the advance randomized controlled trial. BMC Neurol 17:29, 2017.

Correale J et al: Progressive multiple sclerosis: From pathogenic mechanisms to treatment. Brain 140:527, 2017.

Giovannoni G et al: Alemtuzumab improves preexisting disability in active relapsing-remitting MS patients. Neurology 87:1985, 2016.

Hauser SL et al: Ocrelizumab versus interferon beta-1a in relapsing multiple sclerosis. N Engl J Med 376:221, 2017.

Mahad DH et al: Pathological mechanisms in progressive multiple sclerosis. Lancet Neurol 14:183, 2015.

Montalban X et al: Ocrelizumab versus placebo in primary progressive multiple sclerosis. N Engl J Med 376:209, 2017.

Pohl D et al: Acute disseminated encephalomyelitis: Updates on an inflammatory CNS syndrome. Neurology 87(9 Suppl 2):S38, 2016.

437 Neuromyelitis Optica

Bruce A. C. Cree, Stephen L. Hauser

Neuromyelitis optica (NMO; Devic's disease) is an aggressive inflammatory disorder characterized by recurrent attacks of ON and myelitis; the more inclusive term *NMO Spectrum Disorder* (NMOSD) has been proposed to incorporate individuals with partial forms, and also those with involvement of additional structures in the central nervous system (Table 437-1). NMO is more frequent in women than men (>3:1), and typically begins in adulthood but can arise at any age. An important consideration, especially early in its presentation, is distinguishing between NMO and multiple sclerosis (MS; Chap. 436). In patients with NMO, attacks of ON can be bilateral and produce severe visual loss (uncommon in MS); myelitis can be severe and transverse (rare in MS) and is typically longitudinally extensive (Fig. 437-1) involving three or more contiguous vertebral segments. Also in contrast to MS, progressive symptoms typically do not occur in NMO. The brain MRI was earlier thought to be normal in NMO, but it is now recognized that in many cases brain lesions are present, including areas of nonspecific signal change as well as lesions associated with specific syndromes such as the hypothalamus causing an endocrinopathy; the area postrema in the lower medulla presenting as intractable hiccoughs or vomiting; or the cerebral hemispheres producing focal symptoms, encephalopathy, or seizures. Large MRI lesions in the cerebral hemispheres can be asymptomatic, sometimes have a "cloud-like" appearance and, unlike MS lesions, are often not destructive, and can resolve completely. Spinal cord MRI lesions typically consist of focal enhancing areas of swelling and tissue destruction, extending over three or more spinal cord segments, and on axial sequences, these are centered on the gray matter of the cord. Cerebrospinal fluid (CSF) findings include pleocytosis greater than that observed in MS, with neutrophils and eosinophils present in many acute cases; OCBs are uncommon, occurring in <20% of NMO patients. The pathology of NMO is a distinctive astrocytopathy with inflammation, loss of astrocytes, and an absence of staining of the water channel protein AQP4 by immunohistochemistry, plus thickened blood vessel walls, demyelination, and deposition of antibody and complement.

TABLE 437-1 Diagnostic Criteria for Neuromyelitis Optica Spectrum Disorder

Diagnostic Criteria for NMOSA with AQP4-IgG
1. At least 1 core clinical characteristic
2. Positive test for AQP4-IgG using best available detection method (cell-based assay strongly recommended)
3. Exclusion of alternative diagnoses
Diagnostic Criteria for NMOSA Without AQP4-IgG or NMOSD with Unknown AQP4-IgG Status
1. At least 2 core clinical characteristics occurring as a result of one or more clinical attacks and meeting all of the following requirements:
a. At least 1 core clinical characteristic must be optic neuritis, acute myelitis with LETM, or area postrema syndrome
b. Dissemination in space (2 or more different clinical characteristics)
c. Fulfillment of additional MRI requirements, as applicable
2. Negative test for AQP4-IgG using best available detection method or testing unavailable
3. Exclusion of alternative diagnoses
Core Clinical Characteristics
1. Optic neuritis
2. Acute myelitis
3. Area postrema syndrome: episode of otherwise unexplained hiccups or nausea or vomiting
4. Acute brainstem syndrome
5. Symptomatic narcolepsy or acute diencephalic clinical syndrome with NMOSD-typical diencephalic MRI lesions
6. Symptomatic cerebral syndrome with NMOSD-typical brain lesions
Additional MRI Requirements for NMOSD Without AQP4-IgG and NMOSD with Unknown AQP4-IgG Status
1. Acute optic neuritis: requires brain MRI showing (a) normal findings or only nonspecific white matter lesions, OR (b) optic nerve MRI with T2-hyperintense lesion of T1-weighted gadolinium-enhancing lesion extending over >1/2 optic nerve length or involving optic chiasm
2. Acute myelitis: requires associated intramedullary NMRI lesion extending ≥3 contiguous segments (LETM) OR ≥3 contiguous segments of focal spinal cord atrophy in patients with history compatible with acute myelitis
3. Area postrema syndrome requires associated dorsal medulla/area postrema lesions
4. Acute brainstem syndrome requires periependymal brainstem lesions

Source: Adapted from DM Wingerchuk et al: Neurology 85:177–189, 2015.

IMMUNOLOGY

NMO is an autoimmune disease associated with a highly specific autoantibody directed against aquaporin-4 (AQP4) that is present in the sera of ~70% of patients with a clinical diagnosis of NMO. AQP4 is localized to the foot processes of astrocytes in close apposition to endothelial surfaces, as well as at paranodal regions near nodes of Ranvier. It is likely that AQP4 antibodies are pathogenic because passive transfer of AQP4 antibodies into laboratory animals can reproduce histologic features of the disease; complement fixation is thought to mediate astrocyte injury. During acute attacks of myelitis,

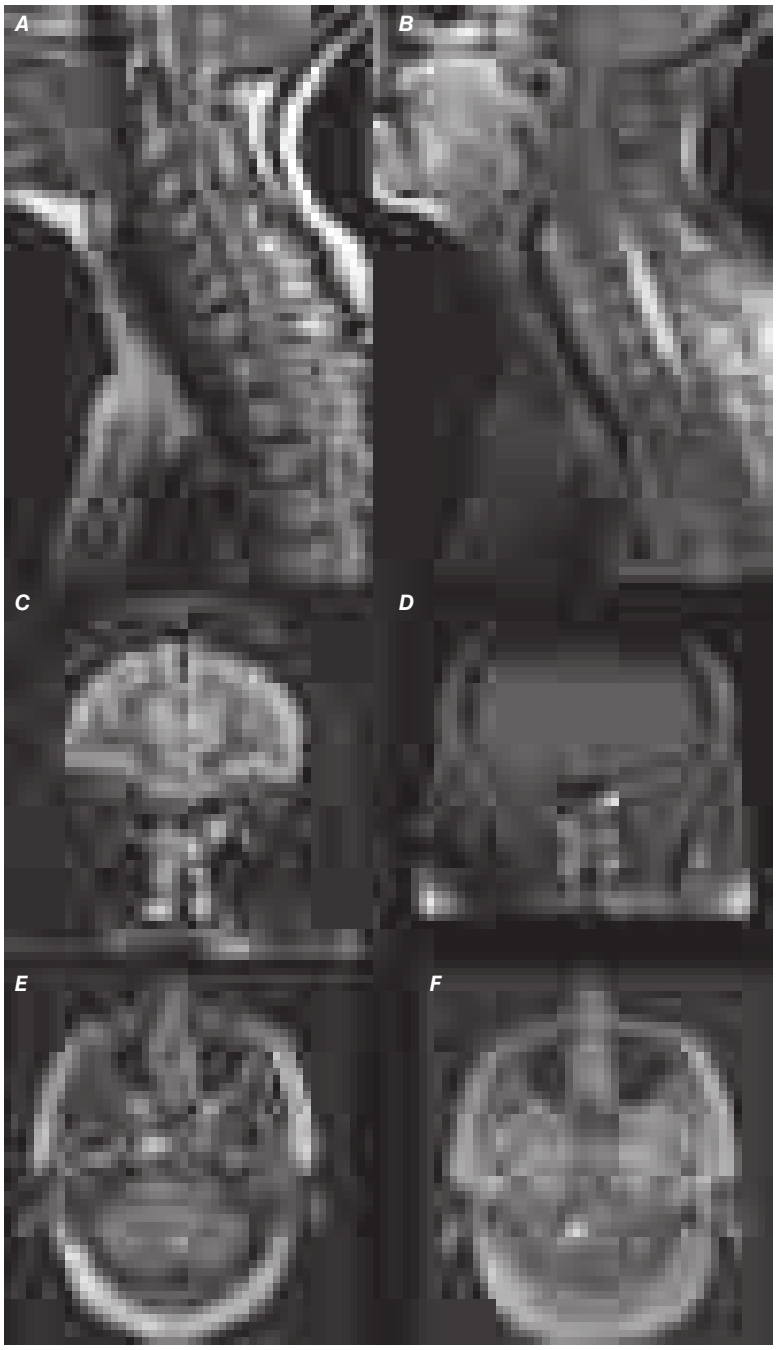


FIGURE 437-1 Imaging findings in neuromyelitis optica: longitudinally extensive transverse myelitis, optic neuritis, and brainstem involvement. ***A.*** Sagittal fluid attenuation inversion recovery (FLAIR) cervical-spine MRI showing an area of increased signal change on T2-weighted imaging spanning >3 vertebral segments in length. ***B.*** Sagittal T1-weighted cervical-spine MRI following gadolinium-diethylene triamine pentaacetic acid (DPTA) infusion showing enhancement. ***C.*** Coronal brain MRI shows hyperintense signal on FLAIR imaging within the left optic nerve. ***D.*** Coronal T1-weighted brain MRI following gadolinium-DPTA infusion shows enhancement of the left optic nerve. ***E.*** Axial brain MRI shows an area of hyperintense signal on T2-weighted imaging within the area postrema (arrow). ***F.*** Axial T1-weighted brain MRI following gadolinium-DPTA infusion shows punctate enhancement of the area postrema (arrow).

CSF levels of interleukin-6 (IL-6; a proinflammatory cytokine) and glial fibrillary acidic protein (GFAP) levels are markedly elevated, consistent with active inflammation and astrocyte injury. Proinflammatory T-lymphocytes of the Th17 type recognize an immunodominant epitope of AQP4 and may also contribute to pathogenesis. Because of the high specificity of the antibody, its presence is considered to be diagnostic when found in conjunction with a typical clinical presentation. Anti-AQP4 seropositive patients have a high risk for future relapses; more than half will relapse within 1 year if untreated.

CLINICAL COURSE

NMO is typically a recurrent disease; the course is monophasic in fewer than 10% of patients. Individuals who test negative for AQP-4 antibodies are somewhat more likely to have a monophasic course. Untreated NMO is usually quite disabling over time; in one series, respiratory failure from cervical myelitis was present in one-third of patients, and 8 years after onset, 60% of patients were blind and more than half had permanent paralysis of one or more limbs. There is limited data indicating that the long-term course of NMO has been substantially improved with the development of therapies to treat acute attacks and prevent relapses.

GLOBAL CONSIDERATIONS

The incidence and prevalence of NMO shows considerable variation between populations and geographic regions, with prevalence estimates that range from <1 to > 4 per 100,000. Although NMO can occur in people of any ethnic background, individuals of Asian and African origin are disproportionately affected. The highest reported prevalence is from Martinique. Among white populations, MS (Chap. 436) is far more common than NMO.

Interestingly, when MS affects individuals of African or Asian ancestry, there is a propensity for demyelinating lesions to involve predominantly the optic nerve and spinal cord, an MS subtype termed *opticospinal MS*. Some individuals with opticospinal MS are seropositive for AQP-4 antibodies, indicating that such cases represent NMOSD.

ASSOCIATED CONDITIONS

Up to 40% of NMO patients have a systemic autoimmune disorder, such as systemic lupus erythematosus, Sjögren's syndrome, perinuclear antineutrophil cytoplasmic antibody (p-ANCA)–associated vasculitis, myasthenia gravis, Hashimoto's thyroiditis, or mixed connective tissue disease. In others, onset may be associated with acute infection with varicella zoster virus, Epstein-Barr virus, HIV, or tuberculosis. Rare cases appear to be paraneoplastic and associated with breast, lung, or other cancers.

TREATMENT

Neuromyelitis Optica

Disease-modifying therapies have not been rigorously studied in NMO. Acute attacks are usually treated with high-dose glucocorticoids (e.g., methylprednisolone 1 g/d for 5–10 days followed by a prednisone taper). Plasma exchange (typically 5–7 exchanges of 1.5 plasma volumes/exchange) is used empirically for acute episodes that do not respond to glucocorticoids. Given the unfavorable natural history of untreated NMO, prophylaxis against relapses is recommended for most patients using one of the following regimens: mycophenolate mofetil (1000 mg bid); rituximab a B cell depleting anti-CD20 monoclonal antibody (2 g IV Q 6 months); or a combination of glucocorticoids (500 mg IV methylprednisolone daily for 5 days; then oral prednisone 1 mg/kg per day for 2 months, followed by slow taper) plus azathioprine (2 mg/kg per day started on week 3). Some therapies with proven efficacy in MS do not appear to be useful for NMO. Available evidence suggests that interferon beta is ineffective and paradoxically may increase the risk of NMO relapses, and based on limited data glatiramer acetate, fingolimod, natalizumab, and alemtuzumab also appear to be ineffective. That therapies not commonly used in MS are empirically used in NMOSD highlight the need for efficient diagnosis of this disorder. Clinical trials with the B-cell depleting anti-CD19 monoclonal antibody (inebilizumab), the terminal complement inhibitor (eculizumab), and an IL-6 receptor blocking antibody (SA-237) are ongoing.

DEMYELINATION ASSOCIATED WITH ANTI-MOG ANTIBODIES

Although long considered to be a likely target for antibody-mediated demyelination, anti-MOG antibodies detected by a cell-based assay that enables recognition of myelin oligodendrocyte glycoprotein (MOG) epitopes in a lipid bilayer were only recently found to be associated with cases of acute disseminated encephalomyelitis (ADEM) **(Chap. 436)** in children, and then with cases of AQP4 seronegative NMO. Further studies showed that patients who are seropositive for anti-MOG antibodies are at risk for bilateral, synchronous optic neuritis and myelitis. A clinical feature that can help distinguish ON associated with anti-MOG antibodies from NMO or MS is the presence of papillitis seen by funduscopy or orbital MRI. ON associated with anti-MOG antibodies is typically longitudinally extensive on MRI, and brain MRI can be normal or show fluffy areas of increased signal change in white or grey matter structures, similar to NMO. MRI lesions that are typical for MS, including finger-like lesions oriented perpendicular to the ventricular surface (Dawson fingers) and T1 hypointense lesions **(Chap. 436)**, are uncommon. Spinal cord lesions can be longitudinally extensive or short. Demyelination associated with anti-MOG antibodies is sometimes monophasic, as in ADEM, but can also be recurrent. Acute episodes are managed with high dose glucocorticoids followed by a prednisone taper and sometimes by plasmapheresis, as with NMO. Brain lesions associated with anti-MOG antibodies often respond rapidly to treatment with glucocorticoids and may resolve entirely. Some patients experience disease recurrence following discontinuation of prednisone and can become glucocorticoid dependent. Clinical trials have not been undertaken and there is limited data on use of other immune-suppressing medications typically used in NMO.

FURTHER READING

Cree BA: Placebo controlled trials in neuromyelitis optica are needed and ethical. Mult Scler Relat Disord 4:536, 2015.

Hinson SR et al: Autoimmune AQP4 channelopathies and neuromyelitis optica spectrum disorders. Handb Clin Neurol 133:377, 2016.

Peschl P et al: Myelin oligodendrocyte glycoprotein: Deciphering a target in inflammatory demyelinating diseases. Front Immunol 8:529, 2017.

Wingerchuk DM et al: International consensus diagnostic criteria for neuromyelitis optica spectrum disorders. Neurology 85:177, 2015.

Section 3 Nerve and Muscle Disorders

438 Peripheral Neuropathy

Anthony A. Amato, Richard J. Barohn

Peripheral nerves are composed of sensory, motor, and autonomic elements. Diseases can affect the cell body of a neuron or its peripheral processes, namely the axons or the encasing myelin sheaths. Most peripheral nerves are mixed and contain sensory and motor as well as autonomic fibers. Nerves can be subdivided into three major classes: large myelinated, small myelinated, and small unmyelinated. Motor axons are usually large myelinated fibers that conduct rapidly (~50 m/s). Sensory fibers may be any of the three types. Large-diameter sensory fibers conduct proprioception and vibratory sensation to the brain, while the smaller-diameter myelinated and unmyelinated fibers transmit pain and temperature sensation. Autonomic nerves are also small in diameter. Thus, peripheral neuropathies can impair sensory, motor, or autonomic function, either singly or in combination. Peripheral neuropathies are further classified into those that primarily affect the cell body (e.g., neuronopathy or ganglionopathy), myelin (myelinopathy), and the axon (axonopathy). These different classes of peripheral neuropathies have distinct clinical and electrophysiologic features. This chapter discusses the clinical approach to a patient suspected of having a peripheral neuropathy, as well as specific neuropathies, including hereditary and acquired neuropathies. **The inflammatory neuropathies are discussed in Chap. 439.**

GENERAL APPROACH

In approaching a patient with a neuropathy, the clinician has three main goals: (1) identify where the lesion is, (2) identify the cause, and (3) determine the proper treatment. The first goal is accomplished by obtaining a thorough history, neurologic examination, and electrodiagnostic and other laboratory studies **(Fig. 438-1)**. While gathering this information, seven key questions are asked **(Table 438-1)**, the answers to which help identify the pattern of involvement and the cause of the neuropathy **(Table 438-2)**. Despite an extensive evaluation, in approximately half of patients no etiology is ever found; these patients typically have a predominately sensory polyneuropathy and have been labeled as having idiopathic or cryptogenic sensory polyneuropathy (CSPN).

INFORMATION FROM THE HISTORY AND PHYSICAL EXAMINATION: SEVEN KEY QUESTIONS (TABLE 438-1)

1. What Systems Are Involved? It is important to determine if the patient's symptoms and signs are motor, sensory, autonomic, or a combination of these. If the patient has only weakness without

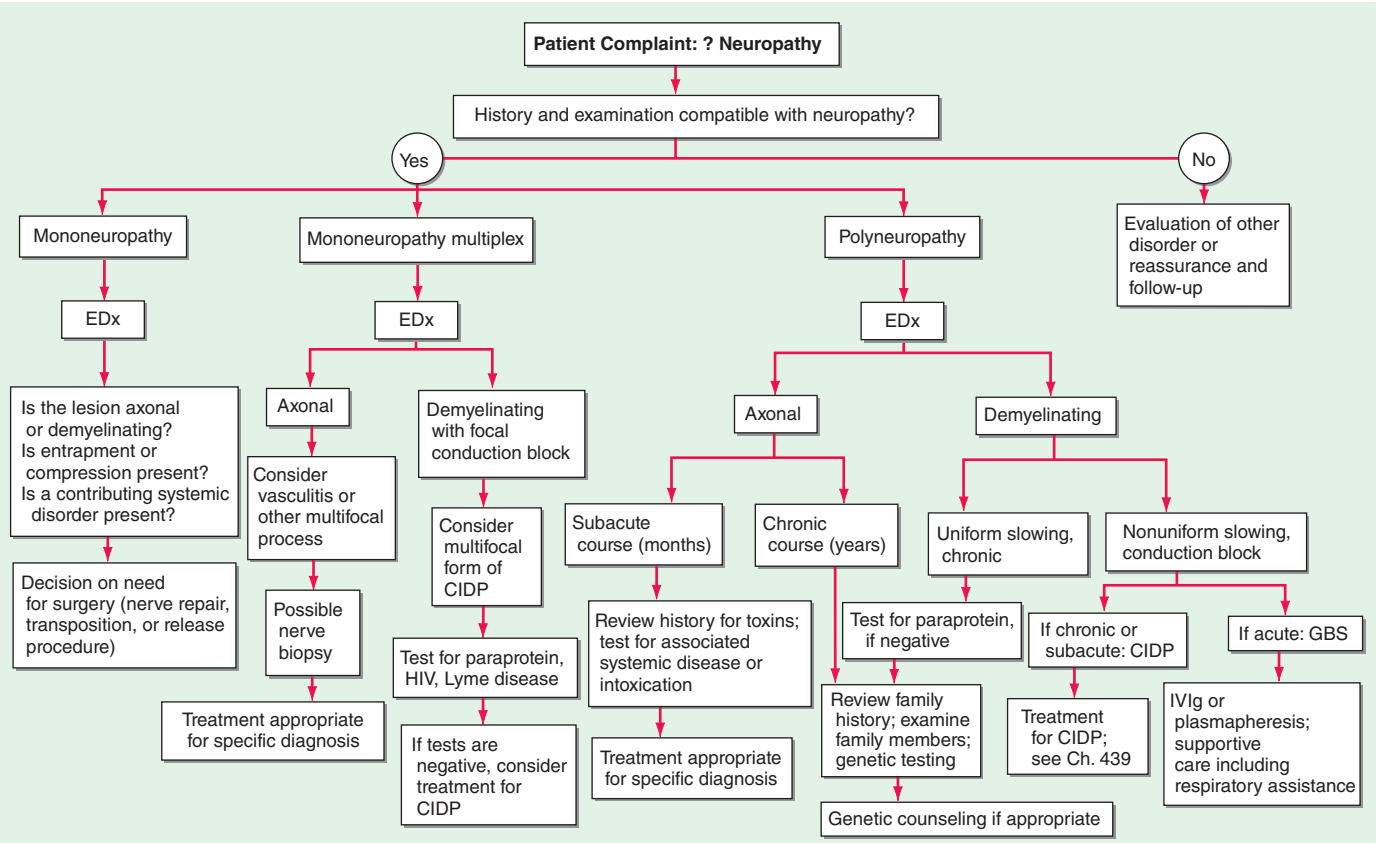


FIGURE 438-1 Approach to the evaluation of peripheral neuropathies. CIDP, chronic inflammatory demyelinating polyradiculoneuropathy; EDx, electrodiagnostic; GBS, Guillain-Barré syndrome; IVIg, intravenous immunoglobulin.

any evidence of sensory or autonomic dysfunction, a motor neuropathy, neuromuscular junction abnormality, or myopathy should be considered. Some peripheral neuropathies are associated with significant autonomic nervous system dysfunction. Symptoms of autonomic involvement include fainting spells or orthostatic lightheadedness; heat intolerance; or any bowel, bladder, or sexual dysfunction (Chap. 432). There will typically be an orthostatic fall in blood pressure without an appropriate increase in heart rate. Autonomic dysfunction in the absence of diabetes should alert the clinician to the possibility of amyloid polyneuropathy. Rarely, a pandysautonomic syndrome can be the only manifestation of a peripheral neuropathy without other motor or sensory findings. The majority of neuropathies are predominantly sensory in nature.

TABLE 438-1 Approach to Neuropathic Disorders: Seven Key Questions

1. **What systems are involved?**
   - Motor, sensory, autonomic, or combinations
2. **What is the distribution of weakness?**
   - Only distal versus proximal and distal
   - Focal/asymmetric versus symmetric
3. **What is the nature of the sensory involvement?**
   - Temperature loss or burning or stabbing pain (e.g., small fiber)
   - Vibratory or proprioceptive loss (e.g., large fiber)
4. **Is there evidence of upper motor neuron involvement?**
   - Without sensory loss
   - With sensory loss
5. **What is the temporal evolution?**
   - Acute (days to 4 weeks)
   - Subacute (4–8 weeks)
   - Chronic (>8 weeks)
   - Monophasic, progressive, or relapsing-remitting
6. **Is there evidence for a hereditary neuropathy?**
   - Family history of neuropathy
   - Lack of sensory symptoms despite sensory signs
7. **Are there any associated medical conditions?**
   - Cancer, diabetes mellitus, connective tissue disease or other autoimmune diseases, infection (e.g., HIV, Lyme disease, leprosy)
   - Medications including over-the-counter drugs that may cause a toxic neuropathy
   - Preceding events, drugs, toxins

2. What Is the Distribution of Weakness? Delineating the pattern of weakness, if present, is essential for diagnosis, and in this regard two additional questions should be answered: (1) Does the weakness only involve the distal extremity, or is it both proximal and distal? and (2) Is the weakness focal and asymmetric, or is it symmetric? Symmetric proximal and distal weakness is the hallmark of acquired immune demyelinating polyneuropathies, both the acute form (Guillain-Barré syndrome [GBS]) and the chronic form (chronic inflammatory demyelinating polyneuropathy [CIDP]) (Chap. 439). The importance of finding symmetric proximal and distal weakness in a patient who presents with both motor and sensory symptoms cannot be overemphasized because this identifies the important subset of patients who may have a treatable acquired demyelinating neuropathic disorder (i.e., GBS or CIDP).

Findings of an asymmetric or multifocal pattern of weakness narrow the differential diagnosis. Some neuropathic disorders may present with unilateral extremity weakness. In the absence of sensory symptoms and signs, such weakness evolving over weeks or months would be worrisome for motor neuron disease (e.g., amyotrophic lateral sclerosis [ALS]), but it would be important to exclude multifocal motor neuropathy that may be treatable (Chap. 439). In a patient presenting with asymmetric subacute or acute sensory and motor symptoms and signs, radiculopathies, plexopathies, compressive mononeuropathies, or multiple mononeuropathies (e.g., mononeuropathy multiplex) must be considered.

TABLE 438-2 Patterns of Neuropathic Disorders

Pattern 1: Symmetric proximal and distal weakness with sensory loss

Consider: inflammatory demyelinating polyneuropathy (GBS and CIDP)

Pattern 2: Symmetric distal sensory loss with or without distal weakness

Consider: cryptogenic or idiopathic sensory polyneuropathy (CSPN), diabetes mellitus and other metabolic disorders, drugs, toxins, familial (HSAN), CMT, amyloidosis, and others

Pattern 3: Asymmetric distal weakness with sensory loss

With involvement of multiple nerves

Consider: multifocal CIDP, vasculitis, cryoglobulinemia, amyloidosis, sarcoid, infectious (leprosy, Lyme, hepatitis B, C, or E, HIV, CMV), HNPP, tumor infiltration

With involvement of single nerves/regions

Consider: may be any of the above but also could be compressive mononeuropathy, plexopathy, or radiculopathy

Pattern 4: Asymmetric proximal and distal weakness with sensory loss

Consider: polyradiculopathy or plexopathy due to diabetes mellitus, meningeal carcinomatosis or lymphomatosis, sarcoid, amyloid, hereditary plexopathy (HNPP, HNA), idiopathic

Pattern 5: Asymmetric distal weakness without sensory loss

With upper motor neuron findings

Consider: motor neuron disease

Without upper motor neuron findings

Consider: progressive muscular atrophy, juvenile monomelic amyotrophy (Hirayama's disease), multifocal motor neuropathy, multifocal acquired motor axonopathy

Pattern 6: Symmetric sensory loss and distal areflexia with upper motor neuron findings

Consider: Vitamin B_{12}, vitamin E, and copper deficiency with combined system degeneration with peripheral neuropathy, chronic liver disease, hereditary leukodystrophies (e.g., adrenomyeloneuropathy) HSP-plus

Pattern 7: Symmetric weakness without sensory loss

With proximal and distal weakness

Consider: SMA

With distal weakness

Consider: hereditary motor neuropathy ("distal" SMA) or atypical CMT

Pattern 8: Focal midline proximal symmetric weakness

Neck extensor weakness

Consider: ALS

Bulbar weakness

Consider: ALS/PLS, isolated bulbar ALS (IBALS), Kennedy's syndrome (X-linked, bulbospinal SMA), bulbar presentation GBS

Diaphragm weakness (SOB)

Consider: ALS

Pattern 9: Asymmetric proprioceptive sensory loss without weakness

Consider causes of a sensory neuronopathy (ganglionopathy):

Cancer (paraneoplastic)

Sjögren's syndrome

Idiopathic sensory neuronopathy (possible GBS variant)

Cisplatin and other chemotherapeutic agents

Vitamin B_6 toxicity

HIV-related sensory neuronopathy

Pattern 10: Autonomic symptoms and signs

Consider neuropathies associated with prominent autonomic dysfunction:

Hereditary sensory and autonomic neuropathy

Amyloidosis (familial and acquired)

Diabetes mellitus

Idiopathic pandysautonomia (may be a variant of Guillain-Barré syndrome)

Porphyria

HIV-related autonomic neuropathy

Vincristine and other chemotherapeutic agents

Abbreviations: CIDP, chronic inflammatory demyelinating polyneuropathy; CMT, Charcot-Marie-Tooth disease; CMV, cytomegalovirus; GBS, Guillain-Barré syndrome; HIV, human immunodeficiency virus; HNA, hereditary neuralgic amyotrophy; HNPP, hereditary neuropathy with liability to pressure palsies; HSAN, hereditary sensory and autonomic neuropathy; HSP-plus, hereditary spastic paraplegia plus neuropathy; SMA, spinal muscular atrophy; SOB, shortness of breath.

ALS can produce prominent neck extensor weakness (head drop), tongue and pharyngeal weakness (dysarthia and dysphagia), or shortness of breath. These focal symmetric weakness patterns can also be seen in neuromuscular junction disorders (myasthenia gravis, Lambert-Eaton myasthenic syndrome [LEMS] **[Chap. 440]**) and some myopathies, particularly isolated neck extensor myopathy **(Chap. 441)**.

3. What Is the Nature of the Sensory Involvement? The patient may have loss of sensation (numbness), altered sensation to touch (hyperpathia or allodynia), or uncomfortable spontaneous sensations (tingling, burning, or aching) **(Chap. 22)**. Neuropathic pain can be burning, dull, and poorly localized (protopathic pain), presumably transmitted by polymodal C nociceptor fibers, or sharp and lancinating (epicritic pain), relayed by A-delta fibers. If pain and temperature perception are lost, while vibratory and position sense are preserved along with muscle strength, deep tendon reflexes, and normal nerve conduction studies (NCS), a small-fiber neuropathy is likely. This is important, because the most likely cause of small-fiber neuropathies, when one is identified, is diabetes mellitus (DM) or glucose intolerance. Amyloid neuropathy should be considered as well in such cases, but most of these small-fiber neuropathies remain idiopathic despite extensive evaluation.

Severe proprioceptive loss also narrows the differential diagnosis. Affected patients will note imbalance, especially in the dark. A neurologic examination revealing a dramatic loss of proprioception with vibration loss and normal strength should alert the clinician to consider a sensory neuronopathy/ganglionopathy (Table 438-2, Pattern 9). In particular, if this loss is asymmetric or affects the arms more than the legs, this pattern suggests a non-length-dependent process as seen in sensory neuronopathies.

4. Is There Evidence of Upper Motor Neuron Involvement? If the patient presents with symmetric distal sensory symptoms and signs suggestive of a distal sensory neuropathy, but there is additional evidence of symmetric upper motor neuron involvement **(Chap. 21)**, the physician should consider a combined system degeneration with neuropathy. The most common cause for this pattern is vitamin B_{12} deficiency, but other etiologies should also be considered (e.g., copper deficiency, human immunodeficiency virus [HIV] infection, severe hepatic disease, adrenomyeloneuropathy [AMN]), and hereditary spastic paraplegia plus a neuropathy.

5. What Is the Temporal Evolution? It is important to determine the onset, duration, and evolution of symptoms and signs. Does the disease have an acute (days to 4 weeks), subacute (4–8 weeks), or chronic (>8 weeks) course? Is the course monophasic, progressive, or relapsing? Most neuropathies are insidious and slowly progressive in nature. Neuropathies with acute and subacute presentations include GBS, vasculitis, and radiculopathies related to diabetes or Lyme disease. A relapsing course can be present in CIDP and porphyria.

6. Is There Evidence for a Hereditary Neuropathy? In patients with slowly progressive distal weakness over many years with few sensory symptoms yet significant sensory deficits on clinical examination, the clinician should consider a hereditary neuropathy (e.g., Charcot-Marie-Tooth disease [CMT]). On examination, the feet may show high or flat arches or hammer toes, and scoliosis may be present. In suspected cases, it may be necessary to perform neurologic and electrophysiologic studies on family members in addition to the patient.

7. Does the Patient Have Any Other Medical Conditions? It is important to inquire about associated medical conditions (e.g., DM, systemic lupus erythematosus [SLE]); preceding or concurrent infections (e.g. diarrheal illness preceding GBS); surgeries (e.g., gastric bypass and nutritional neuropathies); medications (toxic neuropathy), including over-the-counter vitamin preparations (B_6); alcohol; dietary habits; and use of dentures (e.g., fixatives contain zinc that can lead to copper deficiency).

PATTERN RECOGNITION APPROACH TO NEUROPATHIC DISORDERS

Based on the answers to the seven key questions, neuropathic disorders can be classified into several patterns based on the distribution or pattern of sensory, motor, and autonomic involvement (Table 438-2). Each pattern has a limited differential diagnosis, and information from laboratory studies usually permits a final diagnosis to be established.

ELECTRODIAGNOSTIC STUDIES

The electrodiagnostic (EDx) evaluation of patients with a suspected peripheral neuropathy consists of NCS and needle electromyography (EMG). In addition, studies of autonomic function can be valuable. The electrophysiologic data can confirm whether the neuropathic disorder is a mononeuropathy, multiple mononeuropathy (mononeuropathy multiplex), radiculopathy, plexopathy, or generalized polyneuropathy. Similarly, EDx evaluation can ascertain whether the process involves only sensory fibers, motor fibers, autonomic fibers, or a combination of these. Finally, the electrophysiologic data can help distinguish axonopathies from myelinopathies as well as axonal degeneration secondary to ganglionopathies from the more common length-dependent axonopathies.

NCS are most helpful in classifying a neuropathy as due to axonal degeneration or segmental demyelination **(Table 438-3)**. In general, low-amplitude potentials with relatively preserved distal latencies, conduction velocities, and late potentials, along with fibrillations on needle EMG, suggest an axonal neuropathy. On the other hand, slow conduction velocities, prolonged distal latencies and late potentials, relatively preserved amplitudes, and the absence of fibrillations on needle EMG imply a primary demyelinating neuropathy. The presence of nonuniform slowing of conduction velocity, conduction block, or temporal dispersion further suggests an acquired demyelinating neuropathy (e.g., GBS or CIDP) as opposed to a hereditary demyelinating neuropathy (e.g., CMT type 1).

Autonomic studies are used to assess small myelinated (A-delta) or unmyelinated (C) nerve fiber involvement. Such testing includes heart rate response to deep breathing, heart rate, and blood pressure response to both the Valsalva maneuver and tilt-table testing and quantitative sudomotor axon reflex testing **(Chap. 432)**. These studies are particularly useful in patients who have pure small-fiber neuropathy or autonomic neuropathy in which routine NCS are normal.

TABLE 438-3 Electrophysiologic Features: Axonal Degeneration versus Segmental Demyelination

	AXONAL DEGENERATION	SEGMENTAL DEMYELINATION
Motor Nerve Conduction Studies		
CMAP amplitude	Decreased	Normal (except with CB or distal dispersion)
Distal latency	Normal	Prolonged
Conduction velocity	Normal	Slow
Conduction block	Absent	Present
Temporal dispersion	Absent	Present
F wave	Normal or absent	Prolonged or absent
H reflex	Normal or absent	Prolonged or absent
Sensory Nerve Conduction Studies		
SNAP amplitude	Decreased	Normal or decreased
Distal latency	Normal	Prolonged
Conduction velocity	Normal	Slow
Needle EMG		
Spontaneous activity		
Fibrillations	Present	Absent
Fasciculations	Present	Absent
Motor unit potentials		
Recruitment	Decreased	Decreased
Morphology	Long duration, large amplitude, polyphasic (if there is reinnervation)	Normal

Abbreviations: CB, conduction block; CMAP, compound motor action potential; EMG, electromyography; SNAP, sensory nerve action potential.

OTHER IMPORTANT LABORATORY INFORMATION

In patients with generalized symmetric peripheral neuropathy, a standard laboratory evaluation should include a complete blood count, basic chemistries including serum electrolytes and tests of renal and hepatic function, fasting blood glucose (FBS), hemaglobin (Hb) A_{1c}, urinalysis, thyroid function tests, B_{12}, folate, erythrocyte sedimentation rate (ESR), rheumatoid factor, antinuclear antibodies (ANA), serum protein electrophoresis (SPEP) and immunoelectrophoresis or immunofixation, and urine for Bence Jones protein. Quantification of the concentration of serum-free light chains and the kappa/lambda ratio is more sensitive than SPEP, immunoelectrophoresis, or immunofixation to detect a monoclonal gammopathy and therefore should be done if amyloidosis is suspected. A skeletal survey should be performed in patients with acquired demyelinating neuropathies and M-spikes to look for osteosclerotic or lytic lesions. Patients with monoclonal gammopathy should also be referred to a hematologist for consideration of a bone marrow biopsy. An oral glucose tolerance test is indicated in patients with painful sensory neuropathies even if FBS and HbA_{1c} are normal, as the test is abnormal in about one-third of such patients. In addition to the above tests, patients with a mononeuropathy multiplex pattern of involvement should have a vasculitis workup, including antineutrophil cytoplasmic antibodies (ANCA), cryoglobulins, hepatitis serology, Western blot for Lyme disease, HIV, and occasionally a cytomegalovirus (CMV) titer.

There are many autoantibody panels (various antiganglioside antibodies) marketed for screening routine neuropathy patients for a treatable condition. These autoantibodies have no proven clinical utility or added benefit beyond the information obtained from a complete clinical examination and detailed EDx. A heavy metal screen is also not necessary as a screening procedure, unless there is a history of possible exposure or suggestive features on examination (e.g., severe painful sensorimotor and autonomic neuropathy and alopecia—thallium; severe painful sensorimotor neuropathy with or without gastrointestinal [GI] disturbance and Mee's lines—arsenic; wrist or finger extensor weakness and anemia with basophilic stippling of red blood cells—lead).

In patients with suspected GBS or CIDP, a lumbar puncture is indicated to look for an elevated cerebrospinal fluid (CSF) protein. In idiopathic cases of GBS and CIDP, CSF pleocytosis is usually absent. If cells are present, one should consider HIV infection, Lyme disease, sarcoidosis, or lymphomatous or leukemic infiltration of nerve roots. Some patients with GBS and CIDP have abnormal liver function tests. In these cases, it is important to also check for hepatitis B and C, HIV, CMV, and Epstein-Barr virus (EBV) infection. In patients with an axonal GBS (by EMG/NCS) or those with a suspicious coinciding history (e.g., unexplained abdominal pain, psychiatric illness, significant autonomic dysfunction), it is reasonable to screen for porphyria.

In patients with a severe sensory ataxia, a sensory ganglionopathy or neuronopathy should be considered. The most common causes of sensory ganglionopathies are Sjögren's syndrome **(Chap. 354)** and a paraneoplastic neuropathy **(Chap. 90)**. Neuropathy can be the initial manifestation of Sjögren's syndrome. Thus, one should always inquire about dry eyes and mouth in patients with sensory signs and symptoms. Further, some patients can manifest sicca complex without other manifestations of Sjögren's syndrome. Thus, patients with sensory ataxia should be tested for antibodies to SS-A/Ro and SS-B/La, in addition to the routine ANA. To evaluate a possible paraneoplastic sensory ganglionopathy, antineuronal nuclear antibodies (e.g., anti-Hu antibodies) should be obtained. These antibodies are most commonly seen in patients with small-cell carcinoma of the lung but are also present with breast, ovarian, lymphoma, and other cancers. Importantly, the paraneoplastic neuropathy can precede the detection of the cancer, and detection of these autoantibodies should lead to a search for malignancy.

NERVE BIOPSIES

Nerve biopsies are now rarely performed in the evaluation of neuropathies. The primary indication for nerve biopsy is suspicion for amyloid neuropathy or vasculitis. In most instances, the abnormalities present on biopsies do not help distinguish one form of peripheral neuropathy from another (beyond what is already apparent by clinical examination and the NCS). Nerve biopsies should only be performed when the NCS are abnormal. The sural nerve is most commonly biopsied because it is a pure sensory nerve and biopsy will not result in loss of motor function. In suspected vasculitis, a combination biopsy of a superficial peroneal nerve (pure sensory) and the underlying peroneus brevis muscle obtained from a single small incision increases the diagnostic yield. Tissue can be analyzed to assess for evidence of inflammation, vasculitis, or amyloid deposition. Semithin plastic sections, teased fiber preparations, and electron microscopy are used to assess the morphology of the nerve fibers and to distinguish axonopathies from myelinopathies.

SKIN BIOPSIES

Skin biopsies are sometimes used to diagnose a small-fiber neuropathy. Following a punch biopsy of the skin in the distal lower extremity, immunologic staining can be used to measure the density of small unmyelinated fibers. The density of these nerve fibers is reduced in patients with small-fiber neuropathies in whom NCS and routine nerve biopsies are often normal. This technique may allow for an objective measurement in patients with mainly subjective symptoms. However, it often adds little to what one already knows from the clinical examination and EDx.

SPECIFIC DISORDERS

HEREDITARY NEUROPATHIES

CMT disease is the most common type of hereditary neuropathy (Pattern 2, Table 438-2). Rather than one disease, CMT is a syndrome of many genetically distinct disorders **(Table 438-4)**. The various subtypes of CMT are classified according to the nerve conduction velocities (NCVs) and predominant pathology (e.g., demyelination or axonal degeneration), inheritance pattern (autosomal dominant, recessive, or X-linked), and the specific mutated genes. Type 1 CMT (or CMT1) refers to inherited demyelinating sensorimotor neuropathies, whereas the axonal sensory neuropathies are classified as CMT2. By definition, motor conduction velocities in the arms are slowed to <38 m/s in CMT1 and are >38 m/s in CMT2. However, most cases of CMT1 actually have motor NCVs between 20 and 25 m/s. CMT1 and CMT2 usually begin in childhood or early adult life; however, onset later in life can occur, particularly in CMT2. Both are inherited in an autosomal dominant fashion, with a few exceptions. CMT3 is an autosomal dominant neuropathy that appears in infancy and is associated with severe demyelination or hypomyelination. CMT4 is an autosomal recessive neuropathy that typically begins in childhood or early adult life. There are no medical therapies for any of the CMTs, but physical and occupational therapy can be beneficial, as can bracing (e.g., ankle-foot orthotics for footdrop) and other orthotic devices.

CMT1

CMT1 is the most common form of hereditary neuropathy. Affected individuals usually present in the first to third decade of life with distal leg weakness (e.g., foot drop), although patients may remain asymptomatic even late in life. People with CMT generally do not complain of numbness or tingling, which can be helpful in distinguishing CMT from acquired forms of neuropathy in which sensory symptoms usually predominate. Although usually asymptomatic, reduced sensation to all modalities is apparent on examination. Muscle stretch reflexes are unobtainable or reduced throughout. There is often atrophy of the muscles below the knee (particularly the anterior compartment), leading to so-called inverted champagne bottle legs.

Motor NCVs are generally in the 20–25 m/s range. Nerve biopsies usually are not performed on patients suspected of having CMT1, because the diagnosis usually can be made by less invasive testing (e.g., NCS and genetic studies). However, when done, the biopsies reveal reduced numbers of myelinated nerve fibers with a predilection for loss of large-diameter fibers and Schwann cell proliferation around thinly or demyelinated fibers, forming so-called onion bulbs.

CMT1A is the most common subtype of CMT1, representing 70% of cases, and is caused by a 1.5-megabase (Mb) duplication within chromosome 17p11.2-12 encoding the gene for peripheral myelin protein-22 (PMP-22). This results in patients having three copies of the PMP-22 gene rather than two. This protein accounts for 2–5% of myelin protein and is expressed in compact regions of the peripheral myelin sheath. Approximately 20% of patients with CMT1 have CMT1B, caused by mutations in the myelin protein zero (MPZ). CMT1B is for the most part clinically, electrophysiologically, and histologically indistinguishable from CMT1A. MPZ is an integral myelin protein and accounts for more than half of the myelin protein in peripheral nerves. Other forms of CMT1 are much less common and also indistinguishable from one another clinically and electrophysiologically.

CMT2

CMT2 occurs approximately half as frequently as CMT1 and CMT2 tends to present later in life. Affected individuals usually become symptomatic in the second decade; some cases present earlier in childhood, whereas others remain asymptomatic into late adult life. Clinically, CMT2 is for the most part indistinguishable from CMT1. NCS are helpful in this regard; in contrast to CMT1, the velocities are normal or only slightly slowed. The most common cause of CMT2 is a mutation in the gene for mitofusin 2 (MFN2), which accounts for approximately 20–30% of CMT2 cases overall. MFN2 localizes to the outer mitochondrial membrane, where it regulates the mitochondrial network architecture by participating in mitochondrial fusion. The other genes associated with CMT2 are much less common.

CMTDI

In dominant-intermediate CMTs (CMTDIs), the NCVs are faster than usually seen in CMT1 (e.g., >38 m/s) but slower than in CMT2.

CMT3

CMT3 was originally described by Dejerine and Sottas as a hereditary demyelinating sensorimotor polyneuropathy presenting in infancy or early childhood. Affected children are severely weak. Motor NCVs are markedly slowed, typically ≤5–10 m/s. Most cases of CMT3 are caused by point mutations in the genes for PMP-22, MPZ, or ERG-2, which are also the genes responsible for CMT1.

CMT4

CMT4 is extremely rare and is characterized by a severe, childhood-onset sensorimotor polyneuropathy that is usually inherited in an autosomal recessive fashion. Electrophysiologic and histologic evaluations can show demyelinating or axonal features. CMT4 is genetically heterogeneous (Table 438-4).

CMT1X

CMT1X is an X-linked dominant disorder with clinical features similar to CMT1 and CMT2, except that the neuropathy is much more severe in males than in females. CMT1X accounts for ~10–15% of CMT overall. Males usually present in the first two decades of life with atrophy and weakness of the distal arms and legs, areflexia, pes cavus, and hammer toes. Obligate female carriers are frequently asymptomatic, but can develop signs and symptoms of CMT. Onset in females is usually after the second decade of life, and the neuropathy is milder in severity.

NCS reveal features of both demyelination and axonal degeneration. In males, motor NCVs in the arms and legs are moderately slowed (in the low to mid 30-m/s range). About 50% of males with CMT1X have motor NCVs between 15 and 35 m/s with about 80% of these falling between 25 and 35 m/s (intermediate slowing). In contrast, about 80% of females with CMT1X have NCVs in the normal range and 20% have NCVs in the intermediate range. CMT1X is caused by mutations in the

TABLE 438-4 Classification of Charcot-Marie-Tooth Disease and Related Neuropathies

NAME	INHERITANCE	GENE LOCATION	GENE PRODUCT
CMT1			
CMT1A	AD	17p11.2	PMP-22 (usually duplication of gene)
CMT1B	AD	1q21-23	MPZ
CMT1C	AD	16p13.1-p12.3	LITAF
CMT1D	AD	10q21.1-22.1	ERG2
CMT1E (with deafness)	AD	17p11.2	Point mutations in PMP 22 gene
CMT1F	AD	8p13-21	Neurofilament light chain
CMT1G	AD	14q32.33	INF2
CMT1X	X-linked dominant	Xq13	Connexin-32
HNPP	AD	17p11.2	PMP-22
		1q21-23	MPZ
CMT dominant-intermediate (CMTD1)			
CMTD1A	AD	10q24.1-25.1	?
CMTD1B	AD	19.p12-13.2	Dynamin 2
CMTD1C	AD	1p35	YARS
CMTD1D	AD	1q22	MPZ
CMT2			
CMT2A2 (allelic to HMSN VI with optic atrophy)	AD	1p36.2	MFN2
CMT2B	AD	3q13-q22	RAB7
CMT2B1 (allelic to LGMD 1B)	AR	1q21.2	Lamin A/C
CMT2B2	AR and AD	19q13	MED25 for AR; Unknown for AD
CMT2C (with vocal cord and diaphragm paralysis)	AD	12q23-24	TRPV4
CMT2D (allelic to distal SMA5)	AD	7p14	Glycine tRNA synthetase
CMT2E (allelic to CMT 1F)	AD	8p21	Neurofilament light chain
CMT2F	AD	7q11-q21	Heat-shock 27-kDa protein-1
CMT2G	AD	9q31.3-34.2	LRSAM1
CMT2I (allelic to CMT1B)	AD	1q22	MPZ
CMT2J	AD	1q22	MPZ
CMT2H, CMT2K (allelic to CMT4A)	AD	8q13-q21	GDAP1
CMT2L (allelic to distal hereditary motor neuropathy type 2)	AD	12q24	Heat-shock protein 8
CMT2M	AD	16q22	Dynamin-2
CMT2N	AD	16q22.1	AARS
CMT2O	AD	14q32.31	DYNC1H1
CMT2P	AD	9q31.3-34.2	LRSAM1
CMT2P-Okinawa (HSMN2P)	AD	3q13-q14	TFG
CMT2X	X-linked	Xq22-24	PRPS1
CMT3	AD	17p11.2	PMP-22
(Dejerine-Sottas disease, congenital hypomyelinating neuropathy)	AD	1q21-23	MPZ
	AR	10q21.1-22.1	ERG2
	AR	19q13	Periaxon
CMT4			
CMT4A	AR	8q13-21.1	GDAP1
CMT4B1	AR	11q23	MTMR2
CMT4B2	AR	11p15	MTMR13
CMT4C	AR	5q23-33	SH3TC2
CMT4D (HMSN-Lom)	AR	8q24	NDRG1
CMT4E (congenital hypomyelinating neuropathy)	AR	Multiple	Includes PMP22, MPZ, and ERG-2
CMT4F	AR	19q13.1-13.3	Periaxin
CMT4G	AR	10q23.2	HK1
CMT4H	AR	12q12-q13	Frabin
CMT4J	AR	6q21	FIG4
HNA	AD	17q24	SEPT9
HSAN1A	AD	9q22	SPTLC1
HSAN1B	AD	3q21	RAB7
HSAN1C	AD	14q24.3	SPTLC2

(*Continued*)

TABLE 438-4 Classification of Charcot-Marie-Tooth Disease and Related Neuropathies (Continued)

NAME	INHERITANCE	GENE LOCATION	GENE PRODUCT
HSAN1D	AD	14q21.3	ATL1
HSAN1E	AD	19p13.2	DNMT1
HSAN2A	AR	12p13.33	PRKWNK1
HSAN2B	AR	5p15.1	FAM134B
HSAN2C	AR	12q13.13	KIF1A
HSAN2D	AR	2q24.3	SCN9A
HSAN3A	AR	9q21	IKAP
HSAN3B	AR	6p12.1	Dystonin
HSAN4	AR	3q	trkA/NGF receptor
HSAN5	AD or AR	1p11.2-p13.2 2q24.3 3p22.2	NGFb SCN9A SCN11A
HSAN6	AR	6p12.1	Dystonin

Abbreviations: AARS, alanyl-tRNA synthetase; AD, autosomal dominant; AR, autosomal recessive; ATL, atlastin; CMT, Charcot-Marie-Tooth; DNMT1, DNA methyltransferase 1; DYNC1HI, cytoplasmic dynein 1 heavy chain 1; ERG2, early growth response-2 protein; FAM134B, family with sequence similarity 134, member B; FIG4, FDG1-related F actin-binding protein; GDAP1, ganglioside-induced differentiation-associated protein-1; HK1, hexokinase 1; HMSN-P, hereditary motor and sensory neuropathyproximal; HNA, hereditary neuralgic amyotrophy; HNPP, hereditary neuropathy with liability to pressure palsies; HSAN; hereditary sensory and autonomic neuropathy; IFN2, inverted formin-2; IKAP, $_{k}$B kinase complex-associated protein; LGMD, limb girdle muscular dystrophy; LITAF, lipopolysaccharide-induced tumor necrosis factor α factor; LRSAM1, E3 ubiquitin-protein ligase; MED25, mediator 25; MFN2, mitochondrial fusion protein mitofusin 2 gene; MPZ, myelin protein zero protein; MTMR2, myotubularin-related protein-2; NDRG1, N-myc downstream regulated 1; PMP-22, peripheral myelin protein-22; PRKWNK1, protein kinase, lysine deficient 1; PRPS1, phosphoribosylpyrophosphate synthetase 1; RAB7, Ras-related protein 7; SEPT9, Septin 9; SH3TC2, SH3 domain and tetratricopeptide repeats 2; SMA, spinal muscular atrophy; SPTLC, serine palmitoyltransferase long-chain base; TFG, TRK-fused gene; TrkA/NGF, tyrosine kinase A/nerve growth factor; tRNA, transfer ribonucleic acid; TRPV4, transient receptor potential cation channel, subfamily V, member 4; WNK1, WNK lysine deficient; YARS, tyrosyl-tRNA synthetase.

Source: Modified from AA Amato, J Russell: *Neuromuscular Disorders*, 2nd ed. New York, McGraw-Hill, 2016, Table 11-1, pp 265–266.

connexin 32 gene. Connexins are gap junction structural proteins that are important in cell-to-cell communication.

Hereditary Neuropathy with Liability to Pressure Palsies (HNPP) HNPP is an autosomal dominant disorder related to CMT1A. While CMT1A is usually associated with a 1.5-Mb duplication in chromosome 17p11.2 that results in an extra copy of PMP-22 gene, HNPP is caused by inheritance of the chromosome with the corresponding 1.5-Mb deletion of this segment, and thus affected individuals have only one copy of the PMP-22 gene. Patients usually manifest in the second or third decade of life with painless numbness and weakness in the distribution of single peripheral nerves, although multiple mononeuropathies can occur (Pattern 3, Table 438-2). Symptomatic mononeuropathy or multiple mononeuropathies are often precipitated by trivial compression of nerve(s) as can occur with wearing a backpack, leaning on the elbows, or crossing one's legs for even a short period of time. These pressure-related mononeuropathies may take weeks or months to resolve. In addition, some affected individuals manifest with a progressive or relapsing, generalized and symmetric, sensorimotor peripheral neuropathy that resembles CMT.

Hereditary Neuralgic Amyotrophy (HNA) HNA is an autosomal dominant disorder characterized by recurrent attacks of pain, weakness, and sensory loss in the distribution of the brachial plexus often beginning in childhood (Pattern 4, Table 438-2). These attacks are similar to those seen with idiopathic brachial plexitis (see below). Attacks may occur in the postpartum period, following surgery, or at other times of stress. Most patients recover over several weeks or months. Slightly dysmorphic features, including hypotelorism, epicanthal folds, cleft palate, syndactyly, micrognathia, and facial asymmetry, are evident in some individuals. EDx demonstrate an axonal process. HNA is genetically heterogeneous but can be caused by mutations in septin 9 (*SEPT9*). Septins may be important in formation of the neuronal cytoskeleton and have a role in cell division, but it is not known how mutations in SEPT9 lead to HNA.

Hereditary Sensory and Autonomic Neuropathy (HSAN) The HSANs are a very rare group of hereditary neuropathies in which sensory and autonomic dysfunction predominates over muscle weakness, unlike CMT, in which motor findings are most prominent (Pattern 2, Table 438-2; Table 438-4). Nevertheless, affected individuals can develop motor weakness and there can be overlap with CMT. There are no medical therapies available to treat these neuropathies, other than prevention and treatment of mutilating skin and bone lesions.

Of the HSANs, only HSAN1 typically presents in adults. HSAN1 is the most common of the HSANs and is inherited in an autosomal dominant fashion. Affected individuals usually manifest in the second through fourth decades of life. HSAN1 is associated with the degeneration of small myelinated and unmyelinated nerve fibers leading to severe loss of pain and temperature sensation, deep dermal ulcerations, recurrent osteomyelitis, Charcot joints, bone loss, gross foot and hand deformities, and amputated digits. Although most people with HSAN1 do not complain of numbness, they often describe burning, aching, or lancinating pains. Autonomic neuropathy is not a prominent feature, but bladder dysfunction and reduced sweating in the feet may occur. HSAN1A, which is most common, is caused by mutations in the serine palmitoyltransferase long-chain base 1 (*SPTLC1*) gene.

OTHER HEREDITARY NEUROPATHIES (TABLE 438-5)

FABRY'S DISEASE

Fabry's disease (angiokeratoma corporis diffusum) is an X-linked dominant disorder. Although men are more commonly and severely affected, women can also manifest symptoms and signs of the disease. Angiokeratomas are reddish-purple maculopapular lesions that are usually found around the umbilicus, scrotum, inguinal region, and perineum. Burning or lancinating pain in the hands and feet often develops in males in late childhood or early adult life (Pattern 2, Table 438-2). However, the neuropathy is usually overshadowed by complications arising from an associated premature atherosclerosis (e.g., hypertension, renal failure, cardiac disease, and stroke) that often lead to death by the fifth decade of life. Some patients also manifest primarily with a dilated cardiomyopathy.

Fabry's disease is caused by mutations in the a-galactosidase gene that leads to the accumulation of ceramide trihexoside in nerves and blood vessels. A decrease in a-galactosidase activity is evident in leukocytes and cultured fibroblasts. Glycolipid granules may be appreciated in ganglion cells of the peripheral and sympathetic nervous systems and in perineurial cells. Enzyme replacement therapy with a-galactosidase B can improve the neuropathy if patients are treated early, before irreversible nerve fiber loss develops.

TABLE 438-5 Rare Hereditary Neuropathies
Hereditary Disorders of Lipid Metabolism
Metachromatic leukodystrophy
Krabbe's disease (globoid cell leukodystrophy)
Fabry's disease
Adrenoleukodystrophy/adrenomyeloneuropathy
Refsum's disease
Tangier disease
Cerebrotendinous xanthomatosis
Hereditary Ataxias with Neuropathy
Friedreich's ataxia
Vitamin E deficiency
Spinocerebellar ataxia
Abetalipoproteinemia (Bassen-Kornzweig disease)
Disorders of Defective DNA Repair
Ataxia-telangiectasia
Cockayne's syndrome
Giant Axonal Neuropathy
Porphyria
Acute intermittent porphyria (AIP)
Hereditary coproporphyria (HCP)
Variegate porphyria (VP)
Familial Amyloid Polyneuropathy (FAP)
Transthyretin-related
Gelsolin-related
Apolipoprotein A1-related

Source: Modified from AA Amato, J Russell: *Neuromuscular Disorders*, 2nd ed. New York, McGraw-Hill, 2016, Table 12-1, p. 299.

ADRENOLEUKODYSTROPHY/ ADRENOMYELONEUROPATHY

Adrenoleukodystrophy (ALD) and AMN are allelic X-linked dominant disorders caused by mutations in the peroxisomal transmembrane adenosine triphosphate-binding cassette (ABC) transporter gene. Patients with ALD manifest with central nervous system (CNS) abnormalities. However, ~30% of patients with mutations in this gene present with the AMN phenotype that typically manifests in the third to fifth decade of life as mild to moderate peripheral neuropathy combined with progressive spastic paraplegia (Pattern 6, Table 438-2) **(Chap. 434)**. Rare patients present with an adult-onset spinocerebellar ataxia or only with adrenal insufficiency.

EDx is suggestive of a primary axonopathy with secondary demyelination. Nerve biopsies demonstrate a loss of myelinated and unmyelinated nerve fibers with lamellar inclusions in the cytoplasm of Schwann cells. Very long chain fatty acid (VLCFA) levels (C24, C25, and C26) are increased in the urine. Laboratory evidence of adrenal insufficiency is evident in approximately two-thirds of patients. The diagnosis can be confirmed by genetic testing.

Adrenal insufficiency is managed by replacement therapy; however, there is no proven effective therapy for the neurologic manifestations of ALD/AMN. Diets low in VLCFAs and supplemented with Lorenzo's oil (erucic and oleic acids) reduce the levels of VLCFAs and increase the levels of C22 in serum, fibroblasts, and liver; however, several large, open-label trials of Lorenzo's oil failed to demonstrate efficacy.

REFSUM'S DISEASE

Refsum's disease can manifest in infancy to early adulthood with the classic tetrad of (1) peripheral neuropathy, (2) retinitis pigmentosa, (3) cerebellar ataxia, and (4) elevated CSF protein concentration. Most affected individuals develop progressive distal sensory loss and weakness in the legs leading to foot drop by their 20 s (Pattern 2, Table 438-2). Subsequently, the proximal leg and arm muscles may become weak. Patients may also develop sensorineural hearing loss, cardiac conduction abnormalities, ichthyosis, and anosmia.

Serum phytanic acid levels are elevated. Sensory and motor NCS reveal reduced amplitudes, prolonged latencies, and slowed conduction velocities. Nerve biopsy demonstrates a loss of myelinated nerve fibers, with remaining axons often thinly myelinated and associated with onion bulb formation.

Refsum's disease is genetically heterogeneous but autosomal recessive in nature. Classical Refsum's disease with childhood or early adult onset is caused by mutations in the gene that encodes for phytanoyl-CoA α-hydroxylase (*PAHX*). Less commonly, mutations in the gene encoding peroxin 7 receptor protein (PRX7) are responsible. These mutations lead to the accumulation of phytanic acid in the central and peripheral nervous systems. Treatment is removal of phytanic precursors (phytols: fish oils, dairy products, and ruminant fats) from the diet.

TANGIER DISEASE

Tangier disease is a rare autosomal recessive disorder that can present as (1) asymmetric multiple mononeuropathies, (2) a slowly progressive symmetric polyneuropathy predominantly in the legs, or (3) a pseudo-syringomyelia pattern with dissociated sensory loss (i.e., abnormal pain/temperature perception but preserved position/vibration in the arms **[Chap. 434]**). The tonsils may appear swollen and yellowish-orange in color, and there may also be splenomegaly and lymphadenopathy.

Tangier disease is caused by mutations in the ATP-binding cassette transporter 1 (ABC1) gene, which leads to markedly reduced levels of high-density lipoprotein (HDL) cholesterol levels, whereas triacylglycerol levels are increased. Nerve biopsies reveal axonal degeneration with demyelination and remyelination. Electron microscopy demonstrates abnormal accumulation of lipid in Schwann cells, particularly those encompassing umyelinated and small myelinated nerves. There is no specific treatment.

PORPHYRIA

Porphyria is a group of inherited disorders caused by defects in heme biosynthesis **(Chap. 409)**. Three forms of porphyria are associated with peripheral neuropathy: acute intermittent porphyria (AIP), hereditary coproporphyria (HCP), and variegate porphyria (VP). The acute neurologic manifestations are similar in each, with the exception that a photosensitive rash is seen with HCP and VP but not in AIP. Attacks of porphyria can be precipitated by certain drugs (usually those metabolized by the P450 system), hormonal changes (e.g., pregnancy, menstrual cycle), and dietary restrictions.

An acute attack of porphyria may begin with sharp abdominal pain. Subsequently, patients may develop agitation, hallucinations, or seizures. Several days later, back and extremity pain followed by weakness ensues, mimicking GBS (Pattern 1, Table 438-2). Weakness can involve the arms or the legs and can be asymmetric, proximal, or distal in distribution, as well as affecting the face and bulbar musculature. Dysautonomia and signs of sympathetic overactivity are common (e.g., pupillary dilation, tachycardia, and hypertension). Constipation, urinary retention, and incontinence can also be seen.

The CSF protein is typically normal or mildly elevated. Liver function tests and hematologic parameters are usually normal. Some patients are hyponatremic due to inappropriate secretion of antidiuretic hormone **(Chap. 371)**. The urine may appear brownish in color secondary to the high concentration of porphyrin metabolites. Accumulation of intermediary precursors of heme (i.e., d-aminolevulinic acid, porphobilinogen, uroporphobilinogen, coproporphyrinogen, and protoporphyrinogen) is found in urine. Specific enzyme activities can also be measured in erythrocytes and leukocytes. The primary abnormalities on EDx are marked reductions in compound motor action potential (CMAP) amplitudes and signs of active axonal degeneration on needle EMG.

The porphyrias are inherited in an autosomal dominant fashion. AIP is associated with porphobilinogen deaminase deficiency, HCP is caused by defects in coproporphyrin oxidase, and VP is associated with protoporphyrinogen oxidase deficiency. The pathogenesis of the neuropathy is not completely understood. Treatment with glucose and hematin may reduce the accumulation of heme precursors. Intravenous

glucose is started at a rate of 10–20 g/h. If there is no improvement within 24 h, intravenous hematin 2–5 mg/kg per day for 3–14 days should be administered.

FAMILIAL AMYLOID POLYNEUROPATHY

Familial amyloid polyneuropathy (FAP) is phenotypically and genetically heterogeneous and is caused by mutations in the genes for transthyretin (TTR), apolipoprotein A1, or gelsolin (Chap. 108). The majority of patients with FAP have mutations in the TTR gene. Amyloid deposition may be evident in abdominal fat pad, rectal, or nerve biopsies. The clinical features, histopathology, and EDx reveal abnormalities consistent with a generalized or multifocal, predominantly axonal but occasionally demyelinating, polyneuropathy.

Patients with TTR-related FAP usually develop insidious onset of numbness and painful paresthesias in the distal lower limbs in the third to fourth decade of life, although some patients develop the disorder later in life (Pattern 2, Table 438-2). Carpal tunnel syndrome (CTS) is common. Autonomic involvement can be severe, leading to postural hypotension, constipation or persistent diarrhea, erectile dysfunction, and impaired sweating (Pattern 10, Table 438-2). Amyloid deposition also occurs in the heart, kidneys, liver, and corneas. Patients usually die 10–15 years after the onset of symptoms from cardiac failure or complications from malnutrition. Because the liver produces much of the body's TTR, liver transplantation has been used to treat FAP related to TTR mutations. Serum TTR levels decrease after transplantation, and improvement in clinical and EDx features has been reported. Both tafamidis meglumine (20 mg daily) and diflunisal (250 mg twice daily), which prevent misfolding and deposition of mutated TTR, appear to slow the rate of deterioration patients with TTR-related FAP. Recent studies have shown very promising results with gene therapy.

Patients with apolipoprotein A1-related FAP (Van Allen type) usually present in the fourth decade with numbness and painful dysesthesias in the distal limbs. Gradually, the symptoms progress, leading to proximal and distal weakness and atrophy. Although autonomic neuropathy is not severe, some patients develop diarrhea, constipation, or gastroparesis. Most patients die from systemic complications of amyloidosis (e.g., renal failure) 12–15 years after the onset of the neuropathy.

Gelsolin-related amyloidosis (Finnish type) is characterized by the combination of lattice corneal dystrophy and multiple cranial neuropathies that usually begin in the third decade of life. Over time, a mild generalized sensorimotor polyneuropathy develops. Autonomic dysfunction does not occur.

ACQUIRED NEUROPATHIES

PRIMARY OR AL AMYLOIDOSIS (SEE CHAP. 108)

Besides FAP, amyloidosis can also be acquired. In primary or AL amyloidosis, the abnormal protein deposition is composed of immunoglobulin light chains. AL amyloidosis occurs in the setting of multiple myeloma (MM), Waldenström's macroglobulinemia, lymphoma, other plasmacytomas, or lymphoproliferative disorders, or without any other identifiable disease.

Approximately 30% of patients with AL primary amyloidosis present with a polyneuropathy, most typically painful dysesthesias and burning sensations in the feet (Pattern 2, Table 438-2). However, the trunk can be involved, and some patients manifest with a mononeuropathy multiplex pattern. CTS occurs in 25% of patients and may be the initial manifestation. The neuropathy is slowly progressive, and eventually weakness develops along with large-fiber sensory loss. Most patients develop autonomic involvement with postural hypertension, syncope, bowel and bladder incontinence, constipation, impotence, and impaired sweating (Pattern 10, Table 438-2). Patients generally die from their systemic illness (renal failure, cardiac disease).

The monoclonal protein may be composed of IgG, IgA, IgM, or only free light chain. Lambda (λ) is more common than κ light chain (>2:1) in AL amyloidosis. The CSF protein is often increased (with normal cell count), and thus the neuropathy may be mistaken for CIDP (Chap. 439). Nerve biopsies reveal axonal degeneration and amyloid deposition in either a globular or diffuse pattern infiltrating the perineurial, epineurial, and endoneurial connected tissue and in blood vessel walls.

The median survival of patients with primary amyloidosis is <2 years, with death usually from progressive congestive heart failure or renal failure. Chemotherapy with melphalan, prednisone, and colchicine, to reduce the concentration of monoclonal proteins, and autologous stem cell transplantation may prolong survival, but whether the neuropathy improves is controversial.

DIABETIC NEUROPATHY

DM is the most common cause of peripheral neuropathy in developed countries. DM is associated with several types of polyneuropathy: distal symmetric sensory or sensorimotor polyneuropathy, autonomic neuropathy, diabetic neuropathic cachexia, polyradiculoneuropathies, cranial neuropathies, and other mononeuropathies. Risk factors for the development of neuropathy include long-standing, poorly controlled DM and the presence of retinopathy and nephropathy.

Diabetic Distal Symmetric Sensory and Sensorimotor Polyneuropathy (DSPN) DSPN is the most common form of diabetic neuropathy and manifests as sensory loss beginning in the toes that gradually progresses over time up the legs and into the fingers and arms (Pattern 2, Table 438-2). When severe, a patient may develop sensory loss in the trunk (chest and abdomen), initially in the midline anteriorly and later extending laterally. Tingling, burning, deep aching pains may also be apparent. NCS usually show reduced amplitudes and mild to moderate slowing of conduction velocities. Nerve biopsy reveals axonal degeneration, endothelial hyperplasia, and, occasionally, perivascular inflammation. Tight control of glucose can reduce the risk of developing neuropathy or improve the underlying neuropathy. A variety of medications have been used with variable success to treat painful symptoms associated with DSPN, including antiepileptic medications, antidepressants, sodium channel blockers, and other analgesics (Table 438-6).

Diabetic Autonomic Neuropathy Autonomic neuropathy is typically seen in combination with DSPN. The autonomic neuropathy can manifest as abnormal sweating, dysfunctional thermoregulation, dry eyes and mouth, pupillary abnormalities, cardiac arrhythmias, postural hypotension, GI abnormalities (e.g., gastroparesis, postprandial bloating, chronic diarrhea, or constipation), and genitourinary dysfunction (e.g., impotence, retrograde ejaculation, incontinence) (Pattern 10, Table 438-2). Tests of autonomic function are generally abnormal, including sympathetic skin responses and quantitative sudomotor axon reflex testing. Sensory and motor NCS generally demonstrate features described above with DSPN.

Diabetic Radiculoplexus Neuropathy (Diabetic Amyotrophy or Bruns-Garland Syndrome) Diabetic radiculoplexus neuropathy is the presenting manifestation of DM in approximately one-third of patients. Typically, patients present with severe pain in the low back, hip, and thigh in one leg. Rarely, the diabetic polyradiculoneuropathy begins in both legs at the same time (Pattern 4, Table 438-2). Atrophy and weakness of proximal and distal muscles in the affected leg become apparent within a few days or weeks. The neuropathy is often accompanied or heralded by severe weight loss. Weakness usually progresses over several weeks or months, but can continue to progress for 18 months or more. Subsequently, there is slow recovery but many are left with residual weakness, sensory loss, and pain. In contrast to the more typical lumbosacral radiculoplexus neuropathy, some patients develop thoracic radiculopathy or, even less commonly, a cervical polyradiculoneuropathy. CSF protein is usually elevated, while the cell count is normal. ESR is often increased. EDx reveals evidence of active denervation in affected proximal and distal muscles in the affected limbs and in paraspinal muscles. Nerve biopsies may demonstrate axonal degeneration along with perivascular inflammation. Patients with severe pain are sometimes treated in the acute period with glucocorticoids, although a randomized controlled trial has yet to be performed, and the natural history of this neuropathy is gradual improvement.

TABLE 438-6 Treatment of Painful Sensory Neuropathies

THERAPY	ROUTE	DOSE	SIDE EFFECTS
First-Line			
Lidoderm 5% patch	Apply to painful area	Up to 3 patches qd	Skin irritation
Tricyclic antidepressants (e.g., amitriptyline, nortriptyline)	PO	10–100 mg qhs	Cognitive changes, sedation, dry eyes and mouth, urinary retention, constipation
Gabapentin	PO	300–1200 mg tid	Cognitive changes, sedation, peripheral edema
Pregabalin	PO	50–100 mg tid	Cognitive changes, sedation, peripheral edema
Duloxetine	PO	30–60 mg qd	Cognitive changes, sedation, dry eyes, diaphoresis, nausea, diarrhea, constipation
Second-Line			
Carbamazepine	PO	200–400 mg q 6–8 h	Cognitive changes, dizziness, leukopenia, liver dysfunction
Phenytoin	PO	200–400 mg qhs	Cognitive changes, dizziness, liver dysfunction
Venlafaxine	PO	37.5–150 mg/d	Asthenia, sweating, nausea, constipation, anorexia, vomiting, somnolence, dry mouth, dizziness, nervousness, anxiety, tremor, and blurred vision as well as abnormal ejaculation/orgasm and impotence
Tramadol	PO	50 mg qid	Cognitive changes, gastrointestinal upset
Third-Line			
Mexiletine	PO	200–300 mg tid	Arrhythmias
Other Agents			
EMLA cream	Apply cutaneously	qid	Local erythema
2.5% lidocaine			
2.5% prilocaine			
Capsaicin 0.025–0.075% cream	Apply cutaneously	qid	Painful burning skin

Source: Modified from AA Amato, J Russell: *Neuromuscular Disorders*, 2nd ed. New York, McGraw-Hill, 2016, Table 22-3, p 485.

Diabetic Mononeuropathies or Multiple Mononeuropathies

The most common mononeuropathies are median neuropathy at the wrist and ulnar neuropathy at the elbow, but peroneal neuropathy at the fibular head, and sciatic, lateral femoral, cutaneous, or cranial neuropathies also occur (Pattern 3, Table 438-2). In regard to cranial mononeuropathies, seventh nerve palsies are relatively common but may have other, nondiabetic etiologies. In diabetics, a third nerve palsy is most common, followed by sixth nerve, and, less frequently, fourth nerve palsies. Diabetic third nerve palsies are characteristically pupil-sparing **(Chap. 28)**.

HYPOTHYROIDISM

Hypothyroidism is more commonly associated with a proximal myopathy, but some patients develop a neuropathy, most typically CTS. Rarely, a generalized sensory polyneuropathy characterized by painful paresthesias and numbness in both the legs and hands can occur. Treatment is correction of the hypothyroidism.

SJÖGREN'S SYNDROME

Sjögren's syndrome, characterized by the sicca complex of xerophthalmia, xerostomia, and dryness of other mucous membranes, can be complicated by neuropathy **(Chap. 354)**. Most common is a length-dependent axonal sensorimotor neuropathy characterized mainly by sensory loss in the distal extremities (Pattern 2, Table 438-2). A pure small-fiber neuropathy or a cranial neuropathy, particularly involving the trigeminal nerve, can also be seen. Sjögren's syndrome is also associated with sensory neuronopathy/ganglionopathy. Patients with sensory ganglionopathies develop progressive numbness and tingling of the limbs, trunk, and face in a non-length-dependent manner such that symptoms can involve the face or arms more than the legs. The onset can be acute or insidious. Sensory examination demonstrates severe vibratory and proprioceptive loss leading to sensory ataxia.

Patients with neuropathy due to Sjögren's syndrome may have ANAs, SS-A/Ro, and SS-B/La antibodies in the serum, but most do not. NCS demonstrate reduced amplitudes of sensory studies in the affected limbs. Nerve biopsy demonstrates axonal degeneration. Nonspecific perivascular inflammation may be present, but only rarely is there necrotizing vasculitis. There is no specific treatment for neuropathies related to Sjögren's syndrome. When vasculitis is suspected, immunosuppressive agents may be beneficial. Occasionally, the sensory neuronopathy/ganglionopathy stabilizes or improves with immunotherapy, such as IVIg.

RHEUMATOID ARTHRITIS

Peripheral neuropathy occurs in at least 50% of patients with rheumatoid arthritis (RA) and may be vasculitic in nature **(Chap. 351)**. Vasculitic neuropathy can present with a mononeuropathy multiplex (Pattern 3, Table 438-2), a generalized symmetric pattern of involvement (Pattern 2, Table 438-2), or a combination of these patterns **(Chap. 356)**. Neuropathies may also result from drugs used to treat RA (e.g., tumor necrosis blockers, leflunomide). Nerve biopsy often reveals thickening of the epineurial and endoneurial blood vessels as well as perivascular inflammation or vasculitis, with transmural inflammatory cell infiltration and fibrinoid necrosis of vessel walls. The neuropathy is usually responsive to immunomodulating therapies.

SYSTEMIC LUPUS ERYTHEMATOSUS

Between 2 and 27% of individuals with SLE develop a peripheral neuropathy **(Chap. 349)**. Affected patients typically present with a slowly progressive sensory loss beginning in the feet. Some patients develop burning pain and paresthesias with normal reflexes, and NCS suggest a pure small-fiber neuropathy (Pattern 2, Table 438-2). Less common are multiple mononeuropathies presumably secondary to necrotizing vasculitis (Pattern 3, Table 438-2). Rarely, a generalized sensorimotor polyneuropathy meeting clinical, laboratory, electrophysiologic, and histologic criteria for either GBS or CIDP may occur. Immunosuppressive therapy may be beneficial in SLE patients with neuropathy due to vasculitis. Immunosuppressive agents are less likely to be effective in patients with a generalized sensory or sensorimotor polyneuropathy without evidence of vasculitis. Patients with a GBS or CIDP-like neuropathy should be treated accordingly **(Chap. 439)**.

SYSTEMIC SCLEROSIS (SCLERODERMA)

A distal symmetric, mainly sensory, polyneuropathy complicates 5–67% of scleroderma cases (Pattern 2, Table 438-2) **(Chap. 353)**. Cranial mononeuropathies can also develop, most commonly of the trigeminal nerve, producing numbness and dysesthesias in the face. Multiple mononeuropathies also occur (Pattern 3, Table 438-2). The EDx and histologic features of nerve biopsy are those of an axonal sensory greater than motor polyneuropathy.

MIXED CONNECTIVE TISSUE DISEASE (MCTD)

A mild distal axonal sensorimotor polyneuropathy occurs in ~10% of patients with MCTD.

SARCOIDOSIS

The peripheral or CNS is involved in about 5% of patients with sarcoidosis **(Chap. 360)**. The most common cranial nerve involved is the seventh nerve, which can be affected bilaterally. Some patients develop radiculopathy or polyradiculopathy (Pattern 4, Table 438-2). With a generalized root involvement, the clinical presentation can mimic GBS or CIDP. Patients can also present with multiple mononeuropathies (Pattern 3, Table 438-2) or a generalized, slowly progressive, sensory greater than motor polyneuropathy (Pattern 2, Table 438-2). Some have features of a pure small-fiber neuropathy. EDx reveals an axonal neuropathy. Nerve biopsy can reveal noncaseating granulomas infiltrating the endoneurium, perineurium, or epineurium along with lymphocytic necrotizing angiitis. Neurosarcoidosis may respond to treatment with glucocorticoids or other immunosuppressive agents.

HYPEREOSINOPHILIC SYNDROME

Hypereosinophilic syndrome is characterized by eosinophilia associated with various skin, cardiac, hematologic, and neurologic abnormalities. A generalized peripheral neuropathy or a mononeuropathy multiplex occurs in 6–14% of patients (Pattern 2, Table 438-2).

CELIAC DISEASE (GLUTEN-INDUCED ENTEROPATHY OR NONTROPICAL SPRUE)

Neurologic complications, particularly ataxia and peripheral neuropathy, are estimated to occur in 10% of patients with celiac disease **(Chap. 318)**. A generalized sensorimotor polyneuropathy, pure motor neuropathy, multiple mononeuropathies, autonomic neuropathy, small-fiber neuropathy, and neuromyotonia have all been reported in association with celiac disease or antigliadin/antiendomysial antibodies (Patterns 2, 3, and 9; Table 438-2). Nerve biopsy may reveal a loss of large myelinated fibers. The neuropathy may be secondary to malabsorption of vitamins B_{12} and E. However, some patients have no appreciable vitamin deficiencies. The pathogenic basis for the neuropathy in these patients is unclear but may be autoimmune in etiology. The neuropathy does not appear to respond to a gluten-free diet. In patients with vitamin B_{12} or vitamin E deficiency, replacement therapy may improve or stabilize the neuropathy.

INFLAMMATORY BOWEL DISEASE

Ulcerative colitis and Crohn's disease may be complicated by GBS, CIDP, generalized axonal sensory or sensorimotor polyneuropathy, small-fiber neuropathy, or mononeuropathy (Patterns 2 and 3, Table 438-2) **(Chap. 319)**. These neuropathies may be autoimmune, nutritional (e.g., vitamin B_{12} deficiency), treatment related (e.g., metronidazole), or idiopathic in nature. An acute neuropathy with demyelination resembling GBS, CIDP, or multifocal motor neuropathy may occur in patients treated with tumor necrosis factor α blockers.

UREMIC NEUROPATHY

Approximately 60% of patients with renal failure develop a polyneuropathy characterized by length-dependent numbness, tingling, allodynia, and mild distal weakness (Pattern 2, Table 438-2). Rarely, a rapidly progressive weakness and sensory loss very similar to GBS can occur that improves with an increase in the intensity of renal dialysis or with transplantation (Pattern 1, Table 438-2). Mononeuropathies can also occur, the most common of which is CTS. Ischemic monomelic neuropathy (see below) can complicate arteriovenous shunts created in the arm for dialysis (Pattern 3, Table 438-2). EDx in uremic patients reveals features of a length-dependent, primarily axonal, sensorimotor polyneuropathy. Sural nerve biopsies demonstrate a loss of nerve fibers (particularly large myelinated nerve fibers), active axonal degeneration, and segmental and paranodal demyelination. The sensorimotor polyneuropathy can be stabilized by hemodialysis and improved with successful renal transplantation.

CHRONIC LIVER DISEASE

A generalized sensorimotor neuropathy characterized by numbness, tingling, and minor weakness in the distal aspects of primarily the lower limbs commonly occurs in patients with chronic liver failure. EDx studies are consistent with a sensory greater than motor axonopathy. Occasionally patients with severe liver disease develop a combined neuropathy and myopathy. Sural nerve biopsy reveals both segmental demyelination and axonal loss. It is not known if hepatic failure in isolation can cause peripheral neuropathy, as the majority of patients have liver disease secondary to other disorders, such as alcoholism or viral hepatitis, which can also cause neuropathy.

CRITICAL ILLNESS POLYNEUROPATHY

The most common causes of acute generalized weakness leading to admission to a medical intensive care unit (ICU) are GBS and myasthenia gravis (Pattern 1, Table 438-2) **(Chaps. 439 and 440)**. However, weakness developing in critically ill patients while in the ICU is usually caused by critical illness polyneuropathy (CIP) or critical illness myopathy (CIM) or, much less commonly, by prolonged neuromuscular blockade. From a clinical and EDx standpoint, it can be quite difficult to distinguish these disorders. Most specialists believe that CIM is more common. Both CIM and CIP develop as a complication of sepsis and multiple organ failure. They usually present as an inability to wean a patient from a ventilator. A coexisting encephalopathy may limit the neurologic examination, in particular the sensory examination. Muscle stretch reflexes are absent or reduced.

Serum creatine kinase (CK) is usually normal; an elevated serum CK would point to CIM as opposed to CIP. NCS reveal absent or markedly reduced amplitudes of motor and sensory studies in CIP, whereas sensory studies are relatively preserved in CIM. Needle EMG usually reveals profuse positive sharp waves and fibrillation potentials, and it is not unusual in patients with severe weakness to be unable to recruit motor unit action potentials. The pathogenic basis of CIP is not known. Perhaps circulating toxins and metabolic abnormalities associated with sepsis and multiorgan failure impair axonal transport or mitochondrial function, leading to axonal degeneration.

LEPROSY (HANSEN'S DISEASE)

Leprosy, caused by the acid-fast bacteria *Mycobacterium leprae*, is the most common cause of peripheral neuropathy in Southeast Asia, Africa, and South America **(Chap. 174)**. Clinical manifestations range from tuberculoid leprosy at one end of the spectrum to lepromatous leprosy at the other end, with borderline leprosy in between. Neuropathies are most common in patients with borderline leprosy. Superficial cutaneous nerves of the ears and distal limbs are commonly affected. Mononeuropathies, multiple mononeuropathies, or a slowly progressive symmetric sensorimotor polyneuropathy may develop (Patterns 2 and 3, Table 438-2). Sensory NCS are usually absent in the lower limb and are reduced in amplitude in the arms. Motor NCS may demonstrate reduced amplitudes in affected nerves but occasionally can reveal demyelinating features. Leprosy is usually diagnosed by skin lesion biopsy. Nerve biopsy can also be diagnostic, particularly when there are no apparent skin lesions. The tuberculoid form is characterized by granulomas, and bacilli are not seen. In contrast, with lepromatous leprosy, large numbers of infiltrating bacilli, T_H2 lymphocytes, and organism-laden, foamy macrophages with minimal granulomatous infiltration are evident. The bacilli are best appreciated using the Fite stain, where they can be seen as red-staining rods often in clusters free in the endoneurium, within macrophages, or within Schwann cells.

Patients are generally treated with multiple drugs: dapsone, rifampin, and clofazimine. Other medications that are used include thalidomide, pefloxacin, ofloxacin, sparfloxacin, minocycline, and clarithromycin. Patients are generally treated for 2 years. Treatment is sometimes complicated by the so-called reversal reaction, particularly in borderline leprosy. The reversal reaction can occur at any time during treatment and develops because of a shift to the tuberculoid end of the spectrum, with an increase in cellular immunity during treatment. The cellular response is upregulated as evidenced by an increased release of

tumor necrosis factor α, interferon γ, and interleukin 2, with new granuloma formation. This can result in an exacerbation of the rash and the neuropathy as well as in appearance of new lesions. High-dose glucocorticoids blunt this adverse reaction and may be used prophylactically at treatment onset in high-risk patients. Erythema nodosum leprosum (ENL) is also treated with glucocorticoids or thalidomide.

LYME DISEASE

Lyme disease is caused by infection with *Borrelia burgdorferi*, a spirochete usually transmitted by the deer tick *Ixodes dammini* **(Chap. 181)**. Neurologic complications may develop during the second and third stages of infection. Facial neuropathy is most common and is bilateral in about half of cases, which is rare for idiopathic Bell's palsy. Involvement of nerves is frequently asymmetric. Some patients present with a polyradiculoneuropathy or multiple mononeuropathies (Pattern 3 or 4, Table 438-2). EDx is suggestive of a primary axonopathy. Nerve biopsies can reveal axonal degeneration with perivascular inflammation. Treatment is with antibiotics **(Chap. 181)**.

DIPHTHERITIC NEUROPATHY

Diphtheria is caused by the bacteria *Corynebacterium diphtheriae* **(Chap. 145)**. Infected individuals present with flulike symptoms of generalized myalgias, headache, fatigue, low-grade fever, and irritability within a week to 10 days of the exposure. About 20–70% of patients develop a peripheral neuropathy caused by a toxin released by the bacteria. Three to four weeks after infection, patients may note decreased sensation in their throat and begin to develop dysphagia, dysarthria, hoarseness, and blurred vision due to impaired accommodation. A generalized polyneuropathy may manifest 2 or 3 months following the initial infection, characterized by numbness, paresthesias, and weakness of the arms and legs and occasionally ventilatory failure (Pattern 1, Table 438-2). CSF protein can be elevated with or without lymphocytic pleocytosis. EDx suggests a diffuse axonal sensorimotor polyneuropathy. Antitoxin and antibiotics should be given within 48 h of symptom onset. Although early treatment reduces the incidence and severity of some complications (i.e., cardiomyopathy), it does not appear to alter the natural history of the associated peripheral neuropathy. The neuropathy usually resolves after several months.

HUMAN IMMUNODEFICIENCY VIRUS

HIV infection can result in a variety of neurologic complications, including peripheral neuropathies **(Chap. 197)**. Approximately 20% of HIV-infected individuals develop a neuropathy either as a direct result of the virus itself, other associated viral infections (e.g., CMV), or neurotoxicity secondary to antiviral medications (see below). The major presentations of peripheral neuropathy associated with HIV infection include (1) distal symmetric polyneuropathy (DSP), (2) inflammatory demyelinating polyneuropathy (including both GBS and CIDP), (3) multiple mononeuropathies (e.g., vasculitis, CMV-related), (4) polyradiculopathy (usually CMV-related), (5) autonomic neuropathy, and (6) sensory ganglionitis.

HIV-Related Distal Symmetric Polyneuropathy DSP is the most common form of peripheral neuropathy associated with HIV infection and usually is seen in patients with AIDS. It is characterized by numbness and painful paresthesias involving the distal extremities (Pattern 2, Table 438-2). The pathogenic basis for DSP is unknown but is not due to actual infection of the peripheral nerves. The neuropathy may be immune mediated, perhaps caused by the release of cytokines from surrounding inflammatory cells. Vitamin B_{12} deficiency may contribute in some instances but is not a major cause of most cases of DSP. Older antiretroviral agents (e.g., dideoxycytidine, dideoxyinosine, stavudine) are also neurotoxic and can cause a painful sensory neuropathy.

HIV-Related Inflammatory Demyelinating Polyradiculoneuropathy Both acute inflammatory demyelinating polyneuropathy (AIDP) and CIDP can occur as a complication of HIV infection (Pattern 1, Table 438-2). AIDP usually develops at the time of seroconversion, whereas CIDP can occur any time in the course of the infection. Clinical and EDx features are indistinguishable from idiopathic AIDP or CIDP **(Chap. 439)**. In addition to elevated protein levels, lymphocytic pleocytosis is evident in the CSF, a finding that helps distinguish this HIV-associated polyradiculoneuropathy from idiopathic AIDP/CIDP.

HIV-Related Progressive Polyradiculopathy An acute, progressive lumbosacral polyradiculoneuropathy usually secondary to CMV infection can develop in patients with AIDS (Pattern 4, Table 438-2). Patients present with severe radicular pain, numbness, and weakness in the legs, which is usually asymmetric. CSF is abnormal, demonstrating a high protein level, along with a reduced glucose concentration and notably a neutrophilic pleocytosis. EDx studies reveal features of active axonal degeneration. The polyradiculoneuropathy may improve with antiviral therapy.

HIV-Related Multiple Mononeuropathies Multiple mononeuropathies can also develop in patients with HIV infection, usually in the context of AIDS. Weakness, numbness, paresthesias, and pain occur in the distribution of affected nerves (Pattern 3, Table 438-2). Nerve biopsies can reveal axonal degeneration with necrotizing vasculitis or perivascular inflammation. Glucocorticoid treatment is indicated for vasculitis directly due to HIV infection.

HIV-Related Sensory Neuronopathy/Ganglionopathy Dorsal root ganglionitis is a very rare complication of HIV infection, and neuronopathy can be the presenting manifestation. Patients develop sensory ataxia similar to idiopathic sensory neuronopathy/ganglionopathy (Pattern 9, Table 438-2). NCS reveal reduced amplitudes or absence of sensory nerve action potentials (SNAPs).

HERPES VARICELLA-ZOSTER VIRUS

Peripheral neuropathy from herpes varicella-zoster (HVZ) infection results from reactivation of latent virus or from a primary infection **(Chap. 188)**. Two-thirds of infections in adults are characterized by dermal zoster in which severe pain and paresthesias develop in a dermatomal region followed within a week or two by a vesicular rash in the same distribution (Pattern 3, Table 438-2). Weakness in muscles innervated by roots corresponding to the dermatomal distribution of skin lesions occurs in 5–30% of patients. Approximately 25% of affected patients have continued pain (postherpetic neuralgia [PHN]). A large clinical trial demonstrated that vaccination against zoster reduces the incidence of HVZ among vaccine recipients by 51% and reduces the incidence of PHN by 67%. Treatment of PHN is symptomatic (Table 438-6).

CYTOMEGALOVIRUS

CMV can cause an acute lumbosacral polyradiculopathy and multiple mononeuropathies in patients with HIV infection and in other immune deficiency conditions (Pattern 4, Table 438-2) **(Chap. 190)**.

EPSTEIN-BARR VIRUS

EBV infection has been associated with GBS, cranial neuropathies, mononeuropathy multiplex, brachial plexopathy, lumbosacral radiculoplexopathy, and sensory neuronopathies (Patterns 1, 3, 4, and 9, Table 438-2) **(Chap. 189)**.

HEPATITIS VIRUSES

Hepatitis B and C can cause multiple mononeuropathies related to vasculitis, AIDP, or CIDP (Patterns 1 and 3, Table 438-2) **(Chap. 334)**.

NEUROPATHIES ASSOCIATED WITH MALIGNANCY

Patients with malignancy can develop neuropathies due to (1) a direct effect of the cancer by invasion or compression of the nerves, (2) remote or paraneoplastic effect, (3) a toxic effect of treatment, or (4) as a consequence of immune compromise caused by immunosuppressive medications. The most common associated malignancy is lung cancer, but neuropathies also complicate carcinoma of the breast, ovaries, stomach, colon, rectum, and other organs, including the lymphoproliferative system.

PARANEOPLASTIC SENSORY NEURONOPATHY/ GANGLIONOPATHY

Paraneoplastic encephalomyelitis/sensory neuronopathy (PEM/SN) usually complicates small-cell lung carcinoma (Chap. 90). Patients usually present with numbness and paresthesias in the distal extremities that are often asymmetric. The onset can be acute or insidiously progressive. Prominent loss of proprioception leads to sensory ataxia (Pattern 9; Table 438-2). Weakness can be present, usually secondary to an associated myelitis, motor neuronopathy, or concurrent LEMS. Many patients also develop confusion, memory loss, depression, hallucinations or seizures, or cerebellar ataxia. Polyclonal antineuronal antibodies (IgG) directed against a 35- to 40-kDa protein or complex of proteins, the so-called Hu antigen, are found in the sera or CSF in the majority of patients with paraneoplastic PEM/SN. CSF may be normal or may demonstrate mild lymphocytic pleocytosis and elevated protein. PEM/SN is probably the result of antigenic similarity between proteins expressed in the tumor cells and neuronal cells, leading to an immune response directed against both cell types. Treatment of the underlying cancer generally does not affect the course of PEM/SN. However, occasional patients may improve following treatment of the tumor. Unfortunately, plasmapheresis, intravenous immunoglobulin, and immunosuppressive agents have not shown benefit.

NEUROPATHY SECONDARY TO TUMOR INFILTRATION

Malignant cells, in particular leukemia and lymphoma, can infiltrate cranial and peripheral nerves, leading to mononeuropathy, mononeuropathy multiplex, polyradiculopathy, plexopathy, or even a generalized symmetric distal or proximal and distal polyneuropathy (Patterns 1, 2, 3, and 4; Table 438-2). Neuropathy related to tumor infiltration is often painful; it can be the presenting manifestation of the cancer or the heralding symptom of a relapse. The neuropathy may improve with treatment of the underlying leukemia or lymphoma or with glucocorticoids.

NEUROPATHY AS A COMPLICATION OF BONE MARROW TRANSPLANTATION

Neuropathies may develop in patients who undergo bone marrow transplantation (BMT) because of the toxic effects of chemotherapy, radiation, infection, or an autoimmune response directed against the peripheral nerves. Peripheral neuropathy in BMT is often associated with graft-versus-host disease (GVHD). Chronic GVHD shares many features with a variety of autoimmune disorders, and it is possible that an immune-mediated response directed against peripheral nerves is responsible. Patients with chronic GVHD may develop cranial neuropathies, sensorimotor polyneuropathies, multiple mononeuropathies, and severe generalized peripheral neuropathies resembling AIDP or CIDP (Patterns 1, 2, and 3; Table 438-2). The neuropathy may improve by increasing the intensity of immunosuppressive or immunomodulating therapy and resolution of the GVHD.

LYMPHOMA

Lymphomas may cause neuropathy by infiltration or direct compression of nerves or by a paraneoplastic process. The neuropathy can be purely sensory or motor, but most commonly is sensorimotor. The pattern of involvement may be symmetric, asymmetric, or multifocal, and the course may be acute, gradually progressive, or relapsing and remitting (Patterns 1, 2, and 3; Table 438-2). EDx can be compatible with either an axonal or demyelinating process. CSF may reveal lymphocytic pleocytosis and an elevated protein. Nerve biopsy may demonstrate endoneurial inflammatory cells in both the infiltrative and the paraneoplastic etiologies. A monoclonal population of cells favors lymphomatous invasion. The neuropathy may respond to treatment of the underlying lymphoma or immunomodulating therapies.

MULTIPLE MYELOMA

MM usually presents in the fifth to seventh decade of life with fatigue, bone pain, anemia, and hypercalcemia (Chap. 107). Clinical and EDx features of neuropathy occur in as many as 40% of patients. The most common pattern is that of a distal, axonal, sensory, or sensorimotor polyneuropathy (Pattern 2; Table 438-2). Less frequently, a chronic demyelinating polyradiculoneuropathy may develop (Pattern 1; Table 438-2) (see POEMS, Chap. 439). MM can be complicated by amyloid polyneuropathy and should be considered in patients with painful paresthesias, loss of pinprick and temperature discrimination, and autonomic dysfunction (suggestive of a small-fiber neuropathy) and CTS. Expanding plasmacytomas can compress cranial nerves and spinal roots as well. A monoclonal protein, usually composed of γ or μ heavy chains or κ light chains, may be identified in the serum or urine. EDx usually shows reduced amplitudes with normal or only mildly abnormal distal latencies and conduction velocities. A superimposed median neuropathy at the wrist is common. Abdominal fat pad, rectal, or sural nerve biopsy can be performed to look for amyloid deposition. Unfortunately, the treatment of the underlying MM does not usually affect the course of the neuropathy.

NEUROPATHIES ASSOCIATED WITH MONOCLONAL GAMMOPATHY OF UNCERTAIN SIGNIFICANCE (SEE CHAP. 439)

Toxic Neuropathies Secondary to Chemotherapy Many of the commonly used chemotherapy agents can cause a toxic neuropathy (Table 438-7). The mechanisms by which these agents cause toxic neuropathies vary, as does the specific type of neuropathy produced. The risk of developing a toxic neuropathy or more severe neuropathy appears to be greater in patients with a preexisting neuropathy (e.g., CMT disease, diabetic neuropathy) and those who also take other potentially neurotoxic drugs (e.g., nitrofurantoin, isoniazid, disulfiram, pyridoxine). Chemotherapeutic agents usually cause a sensory greater than motor length-dependent axonal neuropathy or neuronopathy/ganglionopathy (Patterns 2 and 9; Table 438-2).

OTHER TOXIC NEUROPATHIES

Neuropathies can develop as complications of toxic effects of various drugs and other environmental exposures (Table 438-8). The more common neuropathies associated with these agents are discussed here.

CHLOROQUINE AND HYDROXYCHLOROQUINE

Chloroquine and hydroxychloroquine can cause a toxic myopathy characterized by slowly progressive, painless, proximal weakness and atrophy, which is worse in the legs than the arms. In addition, neuropathy can also develop with or without the myopathy leading to sensory loss and distal weakness. The "neuromyopathy" usually appears in patients taking 500 mg daily for a year or more but has been reported with doses as low as 200 mg/d. Serum CK levels are usually elevated due to the superimposed myopathy. NCS reveal mild slowing of motor and sensory NCVs with a mild to moderate reduction in the amplitudes, although NCS may be normal in patients with only the myopathy. EMG demonstrates myopathic muscle action potentials (MUAPs), increased insertional activity in the form of positive sharp waves, fibrillation potentials, and occasionally myotonic potentials, particularly in the proximal muscles. Neurogenic MUAPs and reduced recruitment are found in more distal muscles. Nerve biopsy demonstrates autophagic vacuoles within Schwann cells. Vacuoles may also be evident in muscle biopsies. The pathogenic basis of the neuropathy is not known but may be related to the amphiphilic properties of the drug. These agents contain both hydrophobic and hydrophilic regions that allow them to interact with the anionic phospholipids of cell membranes and organelles. The drug-lipid complexes may be resistant to digestion by lysosomal enzymes, leading to the formation of autophagic vacuoles filled with myeloid debris that may in turn cause degeneration of nerves and muscle fibers. The signs and symptoms of the neuropathy and myopathy are usually reversible following discontinuation of medication.

AMIODARONE

Amiodarone can cause a neuromyopathy similar to chloroquine and hydroxychloroquine. The neuromyopathy typically appears after patients have taken the medication for 2–3 years. Nerve biopsy

TABLE 438-7 Toxic Neuropathies Secondary to Chemotherapy

DRUG	MECHANISM OF NEUROTOXICITY	CLINICAL FEATURES	NERVE HISTOPATHOLOGY	EMG/NCS
Vinca alkaloids (vincristine, vinblastine, vindesine, vinorelbine)	Interfere with axonal microtubule assembly; impairs axonal transport	Symmetric, S-M, large-/small-fiber PN; autonomic symptoms common; infrequent cranial neuropathies	Axonal degeneration of myelinated and unmyelinated fibers; regenerating clusters, minimal segmental demyelination	Axonal sensorimotor PN; distal denervation on EMG; abnormal QST, particularly vibratory perception
Cisplatin	Preferential damage to dorsal root ganglia: ? binds to and cross-links DNA ? inhibits protein synthesis ? impairs axonal transport	Predominant large-fiber sensory neuronopathy; sensory ataxia	Loss of large > small myelinated and unmyelinated fibers; axonal degeneration with small clusters of regenerating fibers; secondary segmental demyelination	Low-amplitude or unobtainable SNAPs with normal CMAPs and EMG; abnormal QST, particularly vibratory perception
Taxanes (paclitaxel, docetaxel)	Promotes axonal microtubule assembly; interferes with axonal transport	Symmetric, predominantly sensory PN; large-fiber modalities affected more than small-fiber	Loss of large > small myelinated and unmyelinated fibers; axonal degeneration with small clusters of regenerating fibers; secondary segmental demyelination	Axonal sensorimotor PN; distal denervation on EMG; abnormal QST, particularly vibratory perception
Suramin				
Axonal PN	Unknown; ? inhibition of neurotrophic growth factor binding; ? neuronal lysosomal storage	Symmetric, length-dependent, sensory-predominant PN	None described	Abnormalities consistent with an axonal S-M PN
Demyelinating PN	Unknown; ? immunomodulating effects	Subacute, S-M PN with diffuse proximal and distal weakness; areflexia; increased CSF protein	Loss of large and small myelinated fibers with primary demyelination and secondary axonal degeneration; occasional epi- and endoneurial inflammatory cell infiltrates	Features suggestive of an acquired demyelinating sensorimotor PN (e.g., slow CVs, prolonged distal latencies and F-wave latencies, conduction block, temporal dispersion)
Cytarabine (ARA-C)	Unknown; ? selective Schwann cell toxicity; ? immunomodulating effects	GBS-like syndrome; pure sensory neuropathy; brachial plexopathy	Loss of myelinated nerve fibers; axonal degeneration; segmental demyelination; no inflammation	Axonal, demyelinating, or mixed S-M PN; denervation on EMG
Etoposide (VP-16)	Unknown; ? selective dorsal root ganglia toxicity	Length-dependent, sensory-predominant PN; autonomic neuropathy	None described	Abnormalities consistent with an axonal S-M PN
Bortezomib (Velcade)	Unknown	Length-dependent, sensory, predominantly small-fiber PN	Not reported	Abnormalities consistent with an axonal sensory neuropathy with early small-fiber involvement (abnormal autonomic studies)

Abbreviations: CMAP, compound motor action potential; CSF, cerebrospinal fluid; CVs, conduction velocities; EMG, electromyography; GBS, Guillain-Barré syndrome; NCS, nerve conduction studies; PN, polyneuropathy; QST, quantitative sensory testing; S-M, sensorimotor; SNAP, sensory nerve action potential.

Source: Modified from AA Amato, J Russell: *Neuromuscular Disorders,* 2nd ed. New York, McGraw-Hill, 2016, Table 19-3, p. 439.

demonstrates a combination of segmental demyelination and axonal loss. Electron microscopy reveals lamellar or dense inclusions in Schwann cells, pericytes, and endothelial cells. The inclusions in muscle and nerve biopsies have persisted as long as 2 years following discontinuation of the medication.

COLCHICINE

Colchicine can also cause a neuromyopathy. Patients usually present with proximal weakness and numbness and tingling in the distal extremities. EDx reveals features of an axonal polyneuropathy. Muscle biopsy reveals a vacuolar myopathy, whereas sensory nerves demonstrate axonal degeneration. Colchicine inhibits the polymerization of tubulin into microtubules. The disruption of the microtubules probably leads to defective intracellular movement of important proteins, nutrients, and waste products in muscle and nerves.

THALIDOMIDE

Thalidomide is an immunomodulating agent used to treat MM, GVHD, leprosy, and other autoimmune disorders. Thalidomide is associated with severe teratogenic effects as well as peripheral neuropathy that can be dose-limiting. Patients develop numbness, painful tingling, and burning discomfort in the feet and hands and less commonly muscle weakness and atrophy. Even after stopping the drug for 4–6 years, as many as 50% patients continue to have significant symptoms. NCS demonstrate reduced amplitudes or complete absence of SNAPs, with preserved conduction velocities when obtainable. Motor NCS are usually normal. Nerve biopsy reveals a loss of large-diameter myelinated fibers and axonal degeneration. Degeneration of dorsal root ganglion cells has been reported at autopsy.

PYRIDOXINE (VITAMIN B_6) TOXICITY

Pyridoxine is an essential vitamin that serves as a coenzyme for transamination and decarboxylation. However, at high doses (116 mg/d), patients can develop a severe sensory neuropathy with dysesthesias and sensory ataxia. NCS reveal absent or markedly reduced SNAP amplitudes with relatively preserved CMAPs. Nerve biopsy reveals axonal loss of fiber at all diameters. Loss of dorsal root ganglion cells with subsequent degeneration of both the peripheral and central sensory tracts have been reported in animal models.

ISONIAZID

One of the most common side effects of isoniazid (INH) is peripheral neuropathy. Standard doses of INH (3–5 mg/kg per day) are associated with a 2% incidence of neuropathy, whereas neuropathy develops in at least 17% of patients taking in excess of 6 mg/kg per d. The elderly, malnourished, and "slow acetylators" are at increased risk for developing the neuropathy. INH inhibits pyridoxal phosphokinase, resulting in pyridoxine deficiency and the neuropathy. Prophylactic administration of pyridoxine 100 mg/d can prevent the neuropathy from developing.

ANTIRETROVIRAL AGENTS

The nucleoside analogues zalcitabine (dideoxycytidine or ddC), didanosine (dideoxyinosine or ddI), stavudine (d4T), lamivudine (3TC), and antiretroviral nucleoside reverse transcriptase inhibitor (NRTI) are used to treat HIV infection. One of the major dose-limiting

TABLE 438-8 Toxic Neuropathies

DRUG	MECHANISM OF NEUROTOXICITY	CLINICAL FEATURES	NERVE HISTOPATHOLOGY	EMG/NCS
Misonidazole	Unknown	Painful paresthesias and loss of large- and small-fiber sensory modalities and sometimes distal weakness in length-dependent pattern	Axonal degeneration of large myelinated fibers; axonal swellings; segmental demyelination	Low-amplitude or unobtainable SNAPs with normal or only slightly reduced CMAPs amplitudes
Metronidazole	Unknown	Painful paresthesias and loss of large- and small-fiber sensory modalities and sometimes distal weakness in length-dependent pattern	Axonal degeneration	Low-amplitude or unobtainable SNAPs with normal CMAPs
Chloroquine and hydroxychloroquine	Amphiphilic properties may lead to drug-lipid complexes that are indigestible and result in accumulation of autophagic vacuoles	Loss of large- and small-fiber sensory modalities and distal weakness in length-dependent pattern; superimposed myopathy may lead to proximal weakness	Axonal degeneration with autophagic vacuoles in nerves as well as muscle fibers	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes; distal denervation on EMG; irritability and myopathic-appearing MUAPs proximally in patients with superimposed toxic myopathy
Amiodarone	Amphiphilic properties may lead to drug-lipid complexes that are indigestible and result in accumulation of autophagic vacuoles	Paresthesias and pain with loss of large- and small-fiber sensory modalities and distal weakness in length-dependent pattern; superimposed myopathy may lead to proximal weakness	Axonal degeneration and segmental demyelination with myeloid inclusions in nerves and muscle fibers	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes; can also have prominent slowing of CVs; distal denervation on EMG; irritability and myopathic-appearing MUAPs proximally in patients with superimposed toxic myopathy
Colchicine	Inhibits polymerization of tubulin in microtubules and impairs axoplasmic flow	Numbness and paresthesias with loss of large-fiber modalities in a length-dependent fashion; superimposed myopathy may lead to proximal in addition to distal weakness	Nerve biopsy demonstrates axonal degeneration; muscle biopsy reveals fibers with vacuoles	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes; irritability and myopathic-appearing MUAPs proximally in patients with superimposed toxic myopathy
Podophyllin	Binds to microtubules and impairs axoplasmic flow	Sensory loss, tingling, muscle weakness, and diminished muscle stretch reflexes in length-dependent pattern; autonomic neuropathy	Axonal degeneration	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Thalidomide	Unknown	Numbness, tingling, and burning pain and weakness in a length-dependent pattern	Axonal degeneration; autopsy studies reveal degeneration of dorsal root ganglia	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Disulfiram	Accumulation of neurofilaments and impaired axoplasmic flow	Numbness, tingling, and burning pain in a length-dependent pattern	Axonal degeneration with accumulation of neurofilaments in the axons	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Dapsone	Unknown	Distal weakness that may progress to proximal muscles; sensory loss	Axonal degeneration and segmental demyelination	Low-amplitude or unobtainable CMAPs with normal or reduced SNAP amplitudes
Leflunomide	Unknown	Paresthesias and numbness in a length-dependent pattern	Unknown	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Nitrofurantoin	Unknown	Numbness, painful paresthesias, and severe weakness that may resemble GBS	Axonal degeneration; autopsy studies reveal degeneration of dorsal root ganglia and anterior horn cells	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Pyridoxine (vitamin B_6)	Unknown	Dysesthesias and sensory ataxia; impaired large-fiber sensory modalities on examination	Marked loss of sensory axons and cell bodies in dorsal root ganglia	Reduced amplitudes or absent SNAPs
Isoniazid	Inhibits pyridoxal phosphokinase leading to pyridoxine deficiency	Dysesthesias and sensory ataxia; impaired large-fiber sensory modalities on examination	Marked loss of sensory axons and cell bodies in dorsal root ganglia and degeneration of the dorsal columns	Reduced amplitudes or absent SNAPs and, to a lesser extent, CMAPs
Ethambutol	Unknown	Numbness with loss of large-fiber modalities on examination	Axonal degeneration	Reduced amplitudes or absent SNAPs
Antinucleosides	Unknown	Dysesthesia and sensory ataxia; impaired large-fiber sensory modalities on examination	Axonal degeneration	Reduced amplitudes or absent SNAPs
Phenytoin	Unknown	Numbness with loss of large-fiber modalities on examination	Axonal degeneration and segmental demyelination	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Lithium	Unknown	Numbness with loss of large-fiber modalities on examination	Axonal degeneration	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes

(*Continued*)

TABLE 438-8 Toxic Neuropathies (Continued)

DRUG	MECHANISM OF NEUROTOXICITY	CLINICAL FEATURES	NERVE HISTOPATHOLOGY	EMG/NCS
Acrylamide	Unknown; may be caused by impaired axonal transport	Numbness with loss of large-fiber modalities on examination; sensory ataxia; mild distal weakness	Degeneration of sensory axons in peripheral nerves and posterior columns, spinocerebellar tracts, mammillary bodies, optic tracts, and corticospinal tracts in the CNS	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Carbon disulfide	Unknown	Length-dependent numbness and tingling with mild distal weakness	Axonal swellings with accumulation of neurofilaments	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Ethylene oxide	Unknown; may act as alkylating agent and bind DNA	Length-dependent numbness and tingling; may have mild distal weakness	Axonal degeneration	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Organophosphates	Bind and inhibit neuropathy target esterase	Early features are those of neuromuscular blockade with generalized weakness; later axonal sensorimotor PN ensues	Axonal degeneration along with degeneration of gracile fasciculus and corticospinal tracts	Early: repetitive firing of CMAPs and decrement with repetitive nerve stimulation; late: axonal sensorimotor PN
Hexacarbons	Unknown; may lead to covalent cross-linking between neurofilaments	Acute, severe sensorimotor PN that may resemble GBS	Axonal degeneration and giant axons swollen with neurofilaments	Features of a mixed axonal and/or demyelinating sensorimotor axonal PN—reduced amplitudes, prolonged distal latencies, conduction block, and slowing of CVs
Lead	Unknown; may interfere with mitochondria	Encephalopathy; motor neuropathy (often resembles radial neuropathy with wrist and finger drop); autonomic neuropathy; bluish-black discoloration of gums	Axonal degeneration of motor axons	Reduction of CMAP amplitudes with active denervation on EMG
Mercury	Unknown; may combine with sulfhydryl groups	Abdominal pain and nephrotic syndrome; encephalopathy; ataxia; paresthesias	Axonal degeneration; degeneration of dorsal root ganglia, calcarine, and cerebellar cortex	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Thallium	Unknown	Encephalopathy; painful sensory symptoms; mild loss of vibration; distal or generalized weakness may also develop; autonomic neuropathy; alopecia	Axonal degeneration	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes
Arsenic	Unknown; may combine with sulfhydryl groups	Abdominal discomfort, burning pain, and paresthesias; generalized weakness; autonomic insufficiency; can resemble GBS	Axonal degeneration	Low-amplitude or unobtainable SNAPs with normal or reduced CMAP amplitudes; may have demyelinating features: prolonged distal latencies and slowing of CVs
Gold	Unknown	Distal paresthesias and reduction of all sensory modalities	Axonal degeneration	Low-amplitude or unobtainable SNAPs

Abbreviations: CMAP, compound motor action potential; CVs, conduction velocities; EMG, electromyography; GBS, Guillain-Barré syndrome; MUAP, muscle action potential; NCS, nerve conduction studies; PN, polyneuropathy; S-M, sensorimotor; SNAP, sensory nerve action potential.

Source: Modified from AA Amato, J Russell: *Neuromuscular Disorders,* 2nd ed. New York, McGraw-Hill, 2016, Table 20-1, pp. 449–451.

side effects of these medications is a predominantly sensory, length-dependent, symmetrically painful neuropathy (Pattern 2; Table 438-2). Zalcitabine (ddC) is the most extensively studied of the nucleoside analogues, and at doses >0.18 mg/kg per d, it is associated with a subacute onset of severe burning and lancinating pains in the feet and hands. NCS reveal decreased amplitudes of the SNAPs with normal motor studies. The nucleoside analogues inhibit mitochondrial DNA polymerase, which is the suspected pathogenic basis for the neuropathy. Because of a "coasting effect," patients can continue to worsen even 2–3 weeks after stopping the medication. Following dose reduction, improvement in the neuropathy is seen in most patients after several months (mean time about 10 weeks).

HEXACARBONS (*n*-HEXANE, METHYL *n*-BUTYL KETONE)/GLUE SNIFFER'S NEUROPATHY

n-Hexane and methyl *n*-butyl ketone are water-insoluble industrial organic solvents that are also present in some glues. Exposure through inhalation, accidentally or intentionally (glue sniffing), or through skin absorption can lead to a profound subacute sensory and motor polyneuropathy (Pattern 2; Table 438-2). NCS demonstrate decreased amplitudes of the SNAPs and CMAPs with slightly slow CVs. Nerve biopsy reveals a loss of myelinated fibers and giant axons that are filled with 10-nm neurofilaments. Hexacarbon exposure leads to covalent cross-linking between axonal neurofilaments that result in their aggregation, impaired axonal transport, swelling of the axons, and eventual axonal degeneration.

LEAD

Lead neuropathy is uncommon, but it can be seen in children who accidentally ingest lead-based paints in older buildings and in industrial workers exposed to lead-containing products. The most common presentation of lead poisoning is an encephalopathy; however, symptoms and signs of a primarily motor neuropathy can also occur. The neuropathy is characterized by an insidious and progressive onset of weakness usually beginning in the arms, in particular involving the wrist and finger extensors, resembling a radial neuropathy. Sensation is generally preserved; however, the autonomic nervous system can be affected (Patterns 2, 3, and 10; Table 438-2). Laboratory investigation can reveal a microcytic hypochromic anemia with basophilic stippling of erythrocytes, an elevated serum lead level, and an elevated serum coproporphyrin level. A 24-h urine collection demonstrates elevated levels of lead excretion. The NCS may reveal reduced CMAP amplitudes, while the SNAPs are typically normal. The pathogenic basis may be related to abnormal porphyrin metabolism. The most important

principle of management is to remove the source of the exposure. Chelation therapy with calcium disodium ethylene-diaminetetraacetic acid (EDTA), British anti-Lewisite (BAL), and penicillamine also demonstrates variable efficacy.

MERCURY

Mercury toxicity may occur as a result of exposure to either organic or inorganic mercurials. Mercury poisoning presents with paresthesias in hands and feet that progress proximally and may involve the face and tongue. Motor weakness can also develop. CNS symptoms often overshadow the neuropathy. EDx shows features of a primarily axonal sensorimotor polyneuropathy. The primary site of neuromuscular pathology appears to be the dorsal root ganglia. The mainstay of treatment is removing the source of exposure.

THALLIUM

Thallium can exist in a monovalent or trivalent form and is primarily used as a rodenticide. The toxic neuropathy usually manifests as burning paresthesias of the feet, abdominal pain, and vomiting. Increased thirst, sleep disturbances, and psychotic behavior may be noted. Within the first week, patients develop pigmentation of the hair, an acne-like rash in the malar area of the face, and hyperreflexia. By the second and third weeks, autonomic instability with labile heart rate and blood pressure may be seen. Hyporeflexia and alopecia also occur but may not be evident until the third or fourth week following exposure. With severe intoxication, proximal weakness and involvement of the cranial nerves can occur. Some patients require mechanical ventilation due to respiratory muscle involvement. The lethal dose of thallium is variable, ranging from 8 to 15 mg/kg body weight. Death can result in <48 h following a particularly large dose. NCS demonstrate features of a primarily axonal sensorimotor polyneuropathy. With acute intoxication, potassium ferric ferrocyanide II may be effective in preventing absorption of thallium from the gut. However, there may be no benefit once thallium has been absorbed. Unfortunately, chelating agents are not very efficacious. Adequate diuresis is essential to help eliminate thallium from the body without increasing tissue availability from the serum.

ARSENIC

Arsenic is another heavy metal that can cause a toxic sensorimotor polyneuropathy. The neuropathy manifests 5–10 days after ingestion of arsenic and progresses for several weeks, sometimes mimicking GBS. The presenting symptoms are typically an abrupt onset of abdominal discomfort, nausea, vomiting, pain, and diarrhea followed within several days by burning pain in the feet and hands. Examination of the skin can be helpful in the diagnosis as the loss of the superficial epidermal layer results in patchy regions of increased or decreased pigmentation on the skin several weeks after an acute exposure or with chronic low levels of ingestion. Mee's lines, which are transverse lines at the base of the fingernails and toenails, do not become evident until 1 or 2 months after the exposure. Multiple Mee's lines may be seen in patients with long fingernails who have had chronic exposure to arsenic. Mee's lines are not specific for arsenic toxicity as they can also be seen following thallium poisoning. Because arsenic is cleared from blood rapidly, the serum concentration of arsenic is not diagnostically helpful. However, arsenic levels are increased in the urine, hair, and fingernails of patients exposed to arsenic. Anemia with stippling of erythrocytes is common, and occasionally pancytopenia and aplastic anemia can develop. Increased CSF protein levels without pleocytosis can be seen; this can lead to misdiagnosis as GBS. NCS are usually suggestive of an axonal sensorimotor polyneuropathy; however, demyelinating features can be present. Chelation therapy with BAL has yielded inconsistent results; therefore, it is not generally recommended.

NUTRITIONAL NEUROPATHIES

COBALAMIN (VITAMIN B_{12})

Pernicious anemia is the most common cause of cobalamin deficiency. Other causes include dietary avoidance (vegetarians), gastrectomy, gastric bypass surgery, inflammatory bowel disease, pancreatic insufficiency, bacterial overgrowth, and possibly histamine-2 blockers and proton pump inhibitors. An underappreciated cause of cobalamin deficiency is food-cobalamin malabsorption. This typically occurs in older individuals and results from an inability to adequately absorb cobalamin in food protein. No apparent cause of deficiency is identified in a significant number of patients with cobalamin deficiency. The use of nitrous oxide as an anesthetic agent or as a recreational drug can produce acute cobalamin deficiency neuropathy and subacute combined degeneration.

Complaints of numb hands typically appear before lower extremity paresthesias are noted. A preferential large-fiber sensory loss affecting proprioception and vibration with sparing of small-fiber modalities is present; an unsteady gait reflects sensory ataxia. These features, coupled with diffuse hyperreflexia and absent Achilles reflexes, should always focus attention on the possibility of cobalamin deficiency (Patterns 2 and 6; Table 438-2). Optic atrophy and, in severe cases, behavioral changes ranging from mild irritability and forgetfulness to severe dementia and frank psychosis may appear. The full clinical picture of subacute combined degeneration is uncommon. CNS manifestations, especially pyramidal tract signs, may be missing, and in fact some patients may only exhibit symptoms of peripheral neuropathy.

EDx shows an axonal sensorimotor neuropathy. CNS involvement produces abnormal somatosensory and visual evoked potential latencies. The diagnosis is confirmed by finding reduced serum cobalamin levels. In up to 40% of patients, anemia and macrocytosis are lacking. Serum methylmalonic acid and homocysteine, the metabolites that accumulate when cobalamin-dependent reactions are blocked, are elevated. Antibodies to intrinsic factor are present in ~60%, and antiparietal cell antibodies in about 90%, of individuals with pernicious anemia.

Cobalamin deficiency can be treated with various regimens of cobalamin. One typical regimen consists of 1000 µg cyanocobalamin IM weekly for 1 month and monthly thereafter. Patients with food cobalamin malabsorption can absorb free cobalamin and therefore can be treated with oral cobalamin supplementation. An oral cobalamin dose of 1000 µg per day should be sufficient. Treatment for cobalamin deficiency usually does not completely reverse the clinical manifestations, and at least 50% of patients exhibit some permanent neurologic deficit.

THIAMINE DEFICIENCY

Thiamine (vitamin B_1) deficiency is an uncommon cause of peripheral neuropathy in developed countries. It is now most often seen as a consequence of chronic alcohol abuse, recurrent vomiting, total parenteral nutrition, and bariatric surgery. Thiamine deficiency polyneuropathy can occur in normal, healthy young adults who do not abuse alcohol but who engage in inappropriately restrictive diets. Thiamine is water-soluble. It is present in most animal and plant tissues, but the greatest sources are unrefined cereal grains, wheat germ, yeast, soybean flour, and pork. Beriberi means "I can't, I can't" in Singhalese, the language of natives of what was once part of the Dutch East Indies (now Sri Lanka). *Dry beriberi* refers to neuropathic symptoms. The term *wet beriberi* is used when cardiac manifestations predominate (in reference to edema). Beriberi was relatively uncommon until the late 1800s when it became widespread among people for whom rice was a dietary mainstay. This epidemic was due to a new technique of processing rice that removed the germ from the rice shaft, rendering the so-called polished rice deficient in thiamine and other essential nutrients.

Symptoms of neuropathy follow prolonged deficiency. These begin with mild sensory loss and/or burning dysesthesias in the toes and feet and aching and cramping in the lower legs. Pain may be the predominant symptom. With progression, patients develop features of a nonspecific generalized polyneuropathy, with distal sensory loss in the feet and hands.

Blood and urine assays for thiamine are not reliable for diagnosis of deficiency. Erythrocyte transketolase activity and the percentage increase in activity (in vitro) following the addition of thiamine pyrophosphate (TPP) may be more accurate and reliable. EDx shows nonspecific findings of an axonal sensorimotor polyneuropathy. When a diagnosis of thiamine deficiency is made or suspected, thiamine

replacement should be provided until proper nutrition is restored. Thiamine is usually given intravenously or intramuscularly at a dose of 100 mg/d. Although cardiac manifestations show a striking response to thiamine replacement, neurologic improvement is usually more variable and less dramatic.

VITAMIN E DEFICIENCY

The term *vitamin E* is usually used for a-tocopherol, the most active of the four main types of vitamin E. Because vitamin E is present in animal fat, vegetable oils, and various grains, deficiency is usually due to factors other than insufficient intake. Vitamin E deficiency usually occurs secondary to lipid malabsorption or in uncommon disorders of vitamin E transport. One hereditary disorder is abetalipoproteinemia, a rare autosomal dominant disorder characterized by steatorrhea, pigmentary retinopathy, acanthocytosis, and progressive ataxia. Patients with cystic fibrosis may also have vitamin E deficiency secondary to steatorrhea. There are genetic forms of isolated vitamin E deficiency not associated with lipid malabsorption. Vitamin E deficiency may also occur as a consequence of various cholestatic and hepatobiliary disorders as well as short-bowel syndromes resulting from the surgical treatment of intestinal disorders.

Clinical features may not appear until many years after the onset of deficiency. The onset of symptoms tends to be insidious, and progression is slow. The main clinical features are spinocerebellar ataxia and polyneuropathy, thus resembling Friedreich's ataxia or other spinocerebellar ataxias. Patients manifest progressive ataxia and signs of posterior column dysfunction, such as impaired joint position and vibratory sensation. Because of the polyneuropathy, there is hyporeflexia, but plantar responses may be extensor as a result of the spinal cord involvement (Patterns 2 and 6; Table 438-2). Other neurologic manifestations may include ophthalmoplegia, pigmented retinopathy, night blindness, dysarthria, pseudoathetosis, dystonia, and tremor. Vitamin E deficiency may present as an isolated polyneuropathy, but this is very rare. The yield of checking serum vitamin E levels in patients with isolated polyneuropathy is extremely low, and this test should not be part of routine practice.

Diagnosis is made by measuring a-tocopherol levels in the serum. EDx shows features of an axonal neuropathy. Treatment is replacement with oral vitamin E, but high doses are not needed. For patients with isolated vitamin E deficiency, treatment consists of 1500–6000 IU/d in divided doses.

VITAMIN B_6 DEFICIENCY

Vitamin B_6, or pyridoxine, can produce neuropathic manifestations from both deficiency and toxicity. Vitamin B_6 toxicity was discussed above. Vitamin B_6 deficiency is most commonly seen in patients treated with isoniazid or hydralazine. The polyneuropathy of vitamin B_6 is nonspecific, manifesting as a generalized axonal sensorimotor polyneuropathy. Vitamin B_6 deficiency can be detected by direct assay. Vitamin B_6 supplementation with 50–100 mg/d is suggested for patients being treated with isoniazid or hydralazine. This same dose is appropriate for replacement in cases of nutritional deficiency.

PELLAGRA (NIACIN DEFICIENCY)

Pellagra is produced by deficiency of niacin. Although pellagra may be seen in alcoholics, this disorder has essentially been eradicated in most Western countries by means of enriching bread with niacin. Nevertheless, pellagra continues to be a problem in a number of underdeveloped regions, particularly in Asia and Africa, where corn is the main source of carbohydrate. Neurologic manifestations are variable; abnormalities can develop in the brain and spinal cord as well as peripheral nerves. When peripheral nerves are involved, the neuropathy is usually mild and resembles beriberi. Treatment is with niacin 40–250 mg/d.

COPPER DEFICIENCY

A syndrome that has only recently been described is myeloneuropathy secondary to copper deficiency. Most patients present with lower limb paresthesias, weakness, spasticity, and gait difficulties (Pattern 6; Table 438-2). Large-fiber sensory function is impaired, reflexes are brisk, and plantar responses are extensor. In some cases, light touch and pinprick sensation are affected, and NCS indicate sensorimotor axonal polyneuropathy in addition to myelopathy.

Hematologic abnormalities are a known complication of copper deficiency; these can include microcytic anemia, neutropenia, and occasionally pancytopenia. Because copper is absorbed in the stomach and proximal jejunum, many cases of copper deficiency occur in the setting of prior gastric surgery. Excess zinc is an established cause of copper deficiency. Zinc upregulates enterocyte production of metallothionine, which results in decreased absorption of copper. Excessive dietary zinc supplements or denture cream containing zinc can produce this clinical picture. Other potential causes of copper deficiency include malnutrition, prematurity, total parenteral nutrition, and ingestion of copper-chelating agents.

Following oral or IV copper replacement, some patients show neurologic improvement, but this may take many months or not occur at all. Replacement consists of oral copper sulfate or gluconate 2 mg one to three times a day. If oral copper replacement is not effective, elemental copper in the copper sulfate or copper chloride forms can be given as 2 mg IV daily for 3–5 days, then weekly for 1–2 months until copper levels normalize. Thereafter, oral daily copper therapy can be resumed. In contrast to the neurologic manifestations, most of the hematologic indices normalize in response to copper replacement therapy.

NEUROPATHY ASSOCIATED WITH GASTRIC SURGERY

Polyneuropathy may occur following gastric surgery for ulcer, cancer, or weight reduction. This usually occurs in the context of rapid, significant weight loss and recurrent, protracted vomiting. The clinical picture is one of acute or subacute sensory loss and weakness. Neuropathy following weight loss surgery usually occurs in the first several months after surgery. Weight reduction surgical procedures include gastrojejunostomy, gastric stapling, vertical banded gastroplasty, and gastrectomy with Roux-en-Y anastomosis. The initial manifestations are usually numbness and paresthesias in the feet (Pattern 2; Table 438-2). In many cases, no specific nutritional deficiency factor is identified.

Management consists of parenteral vitamin supplementation, especially including thiamine. Improvement has been observed following supplementation, parenteral nutritional support, and reversal of the surgical bypass. The duration and severity of deficits before identification and treatment of neuropathy are important predictors of final outcome.

CRYPTOGENIC (IDIOPATHIC) SENSORY AND SENSORIMOTOR POLYNEUROPATHY

Cryptogenic (idiopathic) sensory and sensorimotor polyneuropathy (CSPN) is a diagnosis of exclusion, established after a careful medical, family, and social history; neurologic examination; and directed laboratory testing. Despite extensive evaluation, the cause of polyneuropathy in as many as 50% of all patients is idiopathic. CSPN should be considered a distinct diagnostic subset of peripheral neuropathy. The onset of CSPN is predominantly in the sixth and seventh decades. Patients complain of distal numbness, tingling, and often burning pain that invariably begins in the feet and may eventually involve the fingers and hands. Patients exhibit a distal sensory loss to pinprick, touch, and vibration in the toes and feet, and occasionally in the fingers (Pattern 2; Table 438-2). It is uncommon to see significant proprioception deficits, even though patients may complain of gait unsteadiness. However, tandem gait may be abnormal in a minority of cases. Neither subjective nor objective evidence of weakness is a prominent feature. Most patients have evidence of both large- and small-fiber loss on neurologic examination and EDx. Approximately 10% of patients have only evidence of small-fiber involvement. The ankle muscle stretch reflex is frequently absent, but in cases with predominantly small-fiber loss, this may be preserved. The EDx findings range from isolated SNAP abnormalities (usually with loss of amplitude), to evidence for an axonal sensorimotor neuropathy, to a completely normal study (if primarily small fibers are involved). Therapy primarily involves the control of neuropathic pain (Table 438-6) if present. These drugs should not be used if the patient has only numbness and tingling but no pain.

Although no treatment is available that can reverse an idiopathic distal peripheral neuropathy, the prognosis is good. Progression often does not occur or is minimal, with sensory symptoms and signs progressing proximally up to the knees and elbows. The disorder does not lead to significant motor disability over time. The relatively benign course of this disorder should be explained to patients.

MONONEUROPATHIES/PLEXOPATHIES/RADICULOPATHIES (PATTERN 3; TABLE 438-2)

MEDIAN NEUROPATHY

CTS is a compression of the median nerve in the carpal tunnel at the wrist. The median nerve enters the hand through the carpal tunnel by coursing under the transverse carpal ligament. The symptoms of CTS consist of numbness and paresthesias variably in the thumb, index, middle, and half of the ring finger. At times, the paresthesias can include the entire hand and extend into the forearm or upper arm or can be isolated to one or two fingers. Pain is another common symptom and can be located in the hand and forearm and, at times, in the proximal arm. CTS is common and often misdiagnosed as thoracic outlet syndrome. The signs of CTS are decreased sensation in the median nerve distribution; reproduction of the sensation of tingling when a percussion hammer is tapped over the wrist (Tinel's sign) or the wrist is flexed for 30–60 s (Phalen's sign); and weakness of thumb opposition and abduction. EDx is extremely sensitive and shows slowing of sensory and, to a lesser extent, motor median potentials across the wrist. Ultrasound can show focal swelling of the median nerve at the wrist. Treatment options consist of avoidance of precipitating activities; control of underlying systemic-associated conditions if present; nonsteroidal anti-inflammatory medications; neutral (volar) position wrist splints, especially for night use; glucocorticoid/anesthetic injection into the carpal tunnel; and surgical decompression by dividing the transverse carpal ligament. The surgical option should be considered if there is a poor response to nonsurgical treatments; if there is thenar muscle atrophy and/or weakness; and if there are significant denervation potentials on EMG.

Other proximal median neuropathies are very uncommon and include the pronator teres syndrome and anterior interosseous neuropathy. These often occur as a partial form of brachial plexitis.

ULNAR NEUROPATHY AT THE ELBOW—"CUBITAL TUNNEL SYNDROME"

The ulnar nerve passes through the condylar groove between the medial epicondyle and the olecranon. Symptoms consist of paresthesias, tingling, and numbness in the medial hand and half of the fourth and the entire fifth fingers, pain at the elbow or forearm, and weakness. Signs consist of decreased sensation in an ulnar distribution, Tinel's sign at the elbow, and weakness and atrophy of ulnar-innervated hand muscles. The Froment sign indicates thumb adductor weakness and consists of flexion of the thumb at the interphalangeal joint when attempting to oppose the thumb against the lateral border of the second digit. EDx may show slowing of ulnar motor NCV across the elbow with prolonged ulnar sensory latencies. Ultrasound can show swelling of the ulnar nerve around the elbow as well. Treatment consists of avoiding aggravating factors, using elbow pads, and surgery to decompress the nerve in the cubital tunnel. Ulnar neuropathies can also rarely occur at the wrist in the ulnar (Guyon) canal or in the hand, usually after trauma.

RADIAL NEUROPATHY

The radial nerve winds around the proximal humerus in the spiral groove and proceeds down the lateral arm and enters the forearm, dividing into the posterior interosseous nerve and superficial nerve. The symptoms and signs consist of wrist drop; finger extension weakness; thumb abduction weakness; and sensory loss in the dorsal web between the thumb and index finger. Triceps and brachioradialis strength is often normal, and triceps reflex is often intact. Most cases of radial neuropathy are transient compressive (neuropraxic) injuries that recover spontaneously in 6–8 weeks. If there has been prolonged compression and severe axonal damage, it may take several months to recover. Treatment consists of cock-up wrist and finger splints, avoiding further compression, and physical therapy to avoid flexion contracture. If there is no improvement in 2–3 weeks, an EDx study is recommended to confirm the clinical diagnosis and determine the degree of severity.

LATERAL FEMORAL CUTANEOUS NEUROPATHY (MERALGIA PARESTHETICA)

The lateral femoral cutaneous nerve arises from the upper lumbar plexus (spinal levels L2/3), crosses through the inguinal ligament near its attachment to the iliac bone, and supplies sensation to the anterior lateral thigh. The neuropathy affecting this nerve is also known as meralgia paresthetica. Symptoms and signs consist of paresthesias, numbness, and occasionally pain in the lateral thigh. Symptoms are increased by standing or walking and are relieved by sitting. There is normal strength, and knee reflexes are intact. The diagnosis is clinical, and further tests usually are not performed. EDx is only needed to rule out lumbar plexopathy, radiculopathy, or femoral neuropathy. If the symptoms and signs are classic, EMG is not necessary. Symptoms often resolve spontaneously over weeks or months, but the patient may be left with permanent numbness. Treatment consists of weight loss and avoiding tight belts. Analgesics in the form of a lidocaine patch, nonsteroidal agents, and occasionally medications for neuropathic pain can be used (Table 438-6). Rarely, locally injecting the nerve with an anesthetic can be tried. There is no role for surgery.

FEMORAL NEUROPATHY

Femoral neuropathies can arise as complications of retroperitoneal hematoma, lithotomy positioning, hip arthroplasty or dislocation, iliac artery occlusion, femoral arterial procedures, infiltration by hematogenous malignancy, penetrating groin trauma, pelvic surgery including hysterectomy and renal transplantation, and diabetes (a partial form of lumbosacral diabetic plexopathy); some cases are idiopathic. Patients with femoral neuropathy have difficulty extending their knee and flexing the hip. Sensory symptoms occurring either on the anterior thigh and/or medial leg occur in only half of reported cases. A prominent painful component is the exception rather than the rule, may be delayed, and is often self-limited in nature. The quadriceps (patellar) reflex is diminished.

SCIATIC NEUROPATHY

Sciatic neuropathies commonly complicate hip arthroplasty, pelvic procedures in which patients are placed in a prolonged lithotomy position, trauma, hematomas, tumor infiltration, and vasculitis. In addition, many sciatic neuropathies are idiopathic. Weakness may involve all motions of the ankles and toes as well as flexion of the leg at the knee; abduction and extension of the thigh at the hip are spared. Sensory loss occurs in the entire foot and the distal lateral leg. The ankle jerk and on occasion the internal hamstring reflex are diminished or more typically absent on the affected side. The peroneal subdivision of the sciatic nerve is typically involved disproportionately to the tibial counterpart. Thus, patients may have only ankle dorsiflexion and eversion weakness with sparing of knee flexion, ankle inversion, and plantar flexion; these features can lead to misdiagnosis of a common peroneal neuropathy.

PERONEAL NEUROPATHY

The sciatic nerve divides at the distal femur into the tibial and peroneal nerve. The common peroneal nerve passes posterior and laterally around the fibular head, under the fibular tunnel. It then divides into the superficial peroneal nerve, which supplies the ankle evertor muscles and sensation over the anterolateral distal leg and dorsum of the foot, and the deep peroneal nerve, which supplies ankle dorsiflexors and toe extensor muscles and a small area of sensation dorsally in the area of the first and second toes.

Symptoms and signs consist of foot drop (ankle dorsiflexion, toe extension, and ankle eversion weakness) and variable sensory loss, which may involve the superficial and deep peroneal pattern. There is usually no pain. Onset may be on awakening in the morning. Peroneal neuropathy needs to be distinguished from L5 radiculopathy. In L5

TABLE 438-9 Causes of Radiculopathy

- Herniated nucleus pulposus
- Degenerative joint disease
- Rheumatoid arthritis
- Trauma
- Vertebral body compression fracture
- Pott's disease
- Compression by extradural mass (e.g., meningioma, metastatic tumor, hematoma, abscess)
- Primary nerve tumor (e.g., neurofibroma, schwannoma, neurinoma)
- Carcinomatous meningitis
- Perineurial spread of tumor (e.g., prostate cancer)
- Acute inflammatory demyelinating polyradiculopathy
- Chronic inflammatory demyelinating polyradiculopathy
- Sarcoidosis
- Amyloidoma
- Diabetic radiculopathy
- Infection (Lyme disease, herpes zoster, HIV, cytomegalovirus, syphilis, schistosomiasis, strongyloides)
- Arachnoiditis (e.g., postsurgical)
- Radiation

radiculopathy, ankle invertors and evertors are weak and needle EMG reveals denervation. EDx can help localize the lesion. Peroneal motor conduction velocity shows slowing and amplitude drop across the fibular head. Management consists of rapid weight loss and avoiding leg crossing. Footdrop is treated with an ankle brace. A knee pad can be worn over the lateral knee to avoid further compression. Most cases spontaneously resolve over weeks or months.

RADICULOPATHIES

Radiculopathies are most often due to compression from degenerative joint disease and herniated disks, but there are a number of unusual etiologies (Table 438-9). Degenerative spine disease affects a number of different structures, which narrow the diameter of the neural foramen or canal of the spinal column and compromise nerve root integrity; these are discussed in detail in Chap. 14.

PLEXOPATHIES (PATTERN 4; TABLE 438-2)

BRACHIAL PLEXUS

The brachial plexus is composed of three trunks (upper, middle, and lower), with two divisions (anterior and posterior) per trunk (Fig. 438-2). Subsequently, the trunks divide into three cords (medial, lateral, and posterior), and from these arise the multiple terminal nerves innervating the arm. The anterior primary rami of C5 and C6 fuse to form the upper trunk; the anterior primary ramus of C7 continues as the middle trunk, while the anterior rami of C8 and T1 join to form the lower trunk. There are several disorders commonly associated with brachial plexopathy.

Immune-Mediated Brachial Plexus Neuropathy Immune-mediated brachial plexus neuropathy (IBPN) goes by various terms, including *acute brachial plexitis, neuralgic amyotrophy*, and *Parsonage-Turner syndrome*. IBPN usually presents with an acute onset of severe pain in the shoulder region. The intense pain usually lasts several days to a few weeks, but a dull ache can persist. Individuals who are affected may not appreciate weakness of the arm early in the course because the pain limits movement. However, as the pain dissipates, weakness and often sensory loss are appreciated. Attacks can occasionally recur.

Clinical findings are dependent on the distribution of involvement (e.g., specific trunk, divisions, cords, or terminal nerves). The most common pattern of IBPN involves the upper trunk or a single or multiple mononeuropathies primarily involving the suprascapular, long thoracic, or axillary nerves. Additionally, the phrenic and anterior interosseous nerves may be concomitantly affected. Any of these nerves may also be affected in isolation. EDx is useful to confirm and localize the site(s) of involvement. Empirical treatment of severe pain with glucocorticoids is often used in the acute period.

Brachial Plexopathies Associated with Neoplasms Neoplasms involving the brachial plexus may be primary nerve tumors, local cancers expanding into the plexus (e.g., Pancoast lung tumor or lymphoma), and metastatic tumors. Primary brachial plexus tumors are less common than the secondary tumors and include schwannomas, neurinomas, and neurofibromas. Secondary tumors affecting the brachial plexus are more common and are always malignant. These may arise from local tumors, expanding into the plexus. For example, a Pancoast tumor of the upper lobe of the lung may invade or compress the lower trunk, whereas a primary lymphoma arising from the cervical or axillary lymph nodes may also infiltrate the plexus. Pancoast tumors typically present as an insidious onset of pain in the upper arm, sensory disturbance in the medial aspect of the forearm and hand, and weakness and atrophy of the intrinsic hand muscles along with an ipsilateral Horner's syndrome. Chest computed tomography (CT) scans or magnetic resonance imaging (MRI) can demonstrate extension of the tumor into the plexus. Metastatic involvement of the brachial plexus may occur with spread of breast cancer into the axillary lymph nodes and local spread into the nearby nerves.

Perioperative Plexopathies (Median Sternotomy) The most common surgical procedures associated with brachial plexopathy as a complication are those that involve median sternotomies (e.g., open-heart surgeries and thoracotomies). Brachial plexopathies occur in as many as 5% of patients following a median sternotomy and typically affect the lower trunk. Thus, individuals manifest with sensory disturbance affecting the medial aspect of forearm and hand along with weakness of the intrinsic hand muscles. The mechanism is related to the stretch of the lower trunk, so most individuals who are affected recover within a few months.

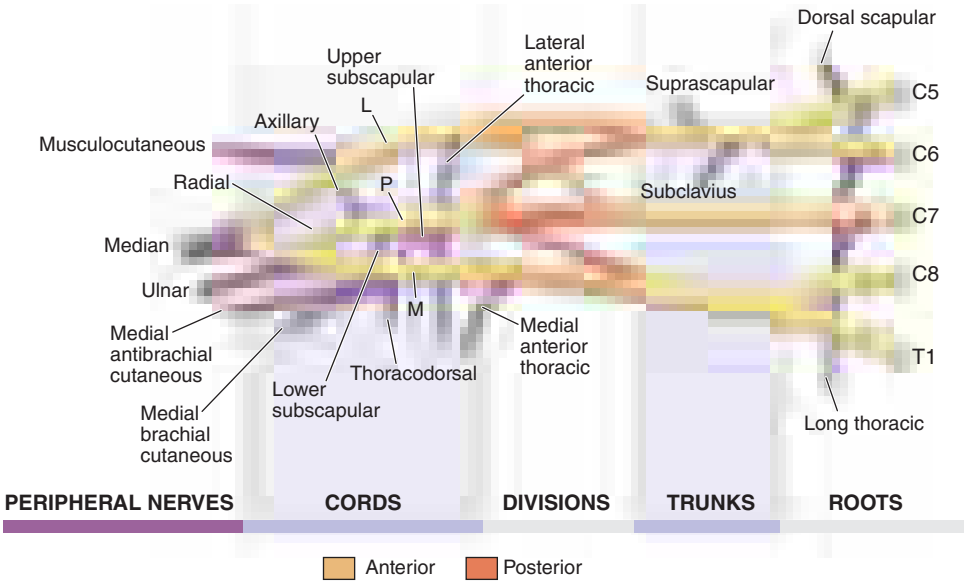


FIGURE 438-2 Brachial plexus anatomy. L, lateral; M, medial; P, posterior. *(From J Goodgold: Anatomical Correlates of Clinical Electromyography. Baltimore, Williams and Wilkins, 1974, p. 56, with permission.)*

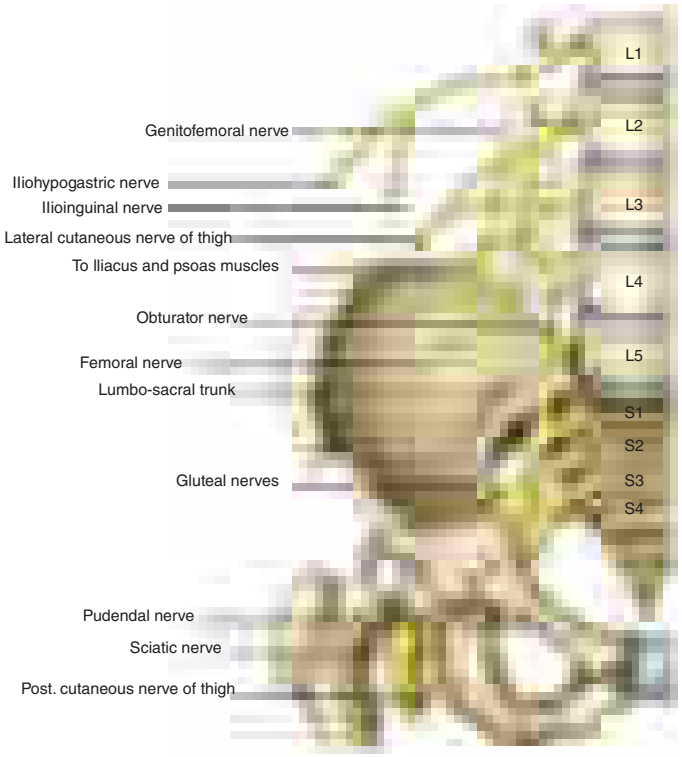


FIGURE 438-3 Lumbosacral plexus. *(From AA Amato, JA Russell (eds): Neuromuscular Disorders, 2nd ed. McGraw-Hill Education, 2016, Table 24-3, p. 541, with permission.)*

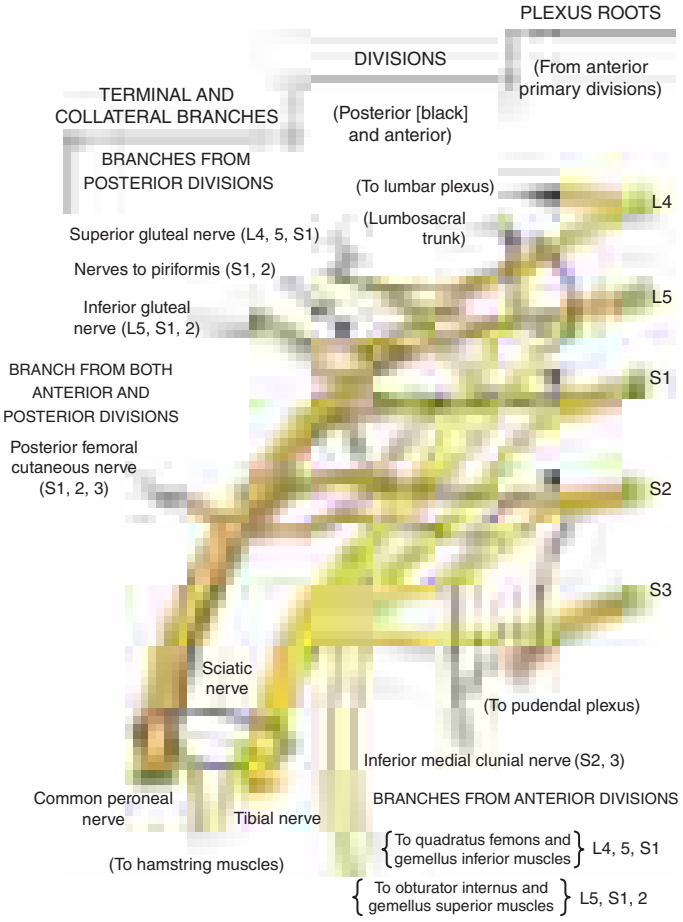


FIGURE 438-4 Lumbosacral trunk sacral plexus, and sciatic nerve. *(From AA Amato, JA Russell (eds): Neuromuscular Disorders, 2nd ed. McGraw-Hill Education, 2016, Table 24-4, p. 542, with permission.)*

Lumbosacral Plexus The lumbar plexus arises from the ventral primary rami of the first to the fourth lumbar spinal nerves (Fig. 438-3). These nerves pass downward and laterally from the vertebral column within the psoas major muscle. The femoral nerve derives from the dorsal branches of the second to the fourth lumbar ventral rami. The obturator nerve arises from the ventral branches of the same lumbar rami. The lumbar plexus communicates with the sacral plexus by the lumbosacral trunk, which contains some fibers from the fourth and all of the fibers from the fifth lumbar ventral rami (Fig. 438-4).

The sacral plexus is the part of the lumbosacral plexus that is formed by the union of the lumbosacral trunk with the ventral rami of the first to fourth sacral nerves. The plexus lies on the posterior and posterolateral wall of the pelvis with its components converging toward the sciatic notch. The lateral trunk of the sciatic nerve (which forms the common peroneal nerve) arises from the union of the dorsal branches of the lumbosacral trunk (L4, L5) and the dorsal branches of the S1 and S2 spinal nerve ventral rami. The medial trunk of the sciatic nerve (which forms the tibial nerve) derives from the ventral branches of the same ventral rami (L4-S2).

LUMBOSACRAL PLEXOPATHIES

Plexopathies are typically recognized when motor, sensory, and if applicable, reflex deficits occur in multiple nerve and segmental distributions confined to one extremity. If localization within the lumbosacral plexus can be accomplished, designation as a lumbar plexopathy, a sacral plexopathy, a lumbosacral trunk lesion, or a pan-plexopathy is the best localization that can be expected. Although lumbar plexopathies may be bilateral, usually occurring in a stepwise and chronologically dissociated manner, sacral plexopathies are more likely to behave in this manner due to their closer anatomic proximity. The differential diagnosis of plexopathy includes disorders of the conus medullaris and cauda equina (polyradiculopathy). If there is a paucity of pain and sensory involvement, motor neuron disease should be considered as well.

The causes of lumbosacral plexopathies are listed in Table 438-10. Diabetic radiculopathy (discussed above) is a fairly common cause of painful leg weakness. Lumbosacral plexopathies are a well-recognized complication of retroperitoneal hemorrhage. Various primary and metastatic malignancies can affect the lumbosacral plexus as well; these include carcinoma of the cervix, endometrium, and ovary; osteosarcoma; testicular cancer; MM; lymphoma; acute myelogenous leukemia; colon cancer; squamous cell carcinoma of the rectum; adenocarcinoma of unknown origin; and intraneural spread of prostate cancer.

RECURRENT NEOPLASTIC DISEASE OR RADIATION-INDUCED PLEXOPATHY

The treatment for various malignancies is often radiation therapy, the field of which may include parts of the brachial plexus. It can be difficult in such situations to determine if a new brachial or lumbosacral plexopathy is related to tumor within the plexus or from radiation-induced nerve damage. Radiation can be associated with microvascular abnormalities and fibrosis of surrounding tissues, which can damage

TABLE 438-10 Lumbosacral Plexopathies: Etiologies

- Retroperitoneal hematoma
- Psoas abscess
- Malignant neoplasm
- Benign neoplasm
- Radiation
- Amyloid
- Diabetic radiculoplexus neuropathy
- Idiopathic radiculoplexus neuropathy
- Sarcoidosis
- Aortic occlusion/surgery
- Lithotomy positioning
- Hip arthroplasty
- Pelvic fracture
- Obstetric injury

the axons and the Schwann cells. Radiation-induced plexopathy can develop months or years following therapy and is dose dependent.

Tumor invasion is usually painful and more commonly affects the lower trunk, whereas radiation injury is often painless and affects the upper trunk. Imaging studies such as MRI and CT scans are useful but can be misleading, especially when there is small microscopic invasion of the plexus. EMG can be informative if myokymic discharges are appreciated, as this finding strongly suggests radiation-induced damage.

EVALUATION AND TREATMENT OF PLEXOPATHIES

Most patients with plexopathies will undergo both imaging with MRI and EDx evaluations. Severe pain from acute idiopathic lumbosacral plexopathy may respond to a short course of glucocorticoids.

FURTHER READING

Amato AA, Russell J: *Neuromuscular Disorders,* 2nd ed. New York, McGraw-Hill, 2016.

Arnold WD, Elsheikh B: Entrapment neuropathies. Neurol Clin 31:405, 2013.

Barohn RJ, Amato AA. Pattern-recognition approach to neuropathy and neuronopathy. Neurol Clin 31:343, 2013.

England JD et al: Evaluation of distal symmetric polyneuropathy: The role of autonomic testing, nerve biopsy, and skin biopsy (an evidence-based review). Muscle Nerve 39:106, 2009.

England JD et al: Evaluation of distal symmetric polyneuropathy: The role of laboratory and genetic testing (an evidence-based review). Muscle Nerve 39:116, 2009.

Klein CJ et al: Inherited neuropathies: Clinical overview and update. Muscle Nerve 48:604, 2013.

Pasnoor M et al: Diabetic neuropathy part 1: Overview and symmetric phenotypes. Neurol Clin 31:425, 2013.

Pasnoor M et al: Diabetic neuropathy part 2: Proximal and asymmetric phenotypes. Neurol Clin 31:447, 2013.

Saporta AS et al: Charcot-Marie-Tooth disease subtypes and genetic testing strategies. Ann Neurol 69:22, 2011.

Waldfogel JM et al: Pharmacotherapy for diabetic peripheral neuropathy pain and quality of life: A systematic review. Neurology 87:978, 2016.

439 Guillain-Barré Syndrome and Other Immune-Mediated Neuropathies

Stephen L. Hauser, Anthony A. Amato

GUILLAIN-BARRÉ SYNDROME

Guillain-Barré syndrome (GBS) is an acute, frequently severe, and fulminant polyradiculoneuropathy that is autoimmune in nature. It occurs year-round at a rate of between 1 and 4 cases per 100,000 annually; in the United States, ~5000–6000 cases occur per year. Males are at slightly higher risk for GBS than females, and in Western countries, adults are more frequently affected than children.

Clinical Manifestations GBS manifests as a rapidly evolving areflexic motor paralysis with or without sensory disturbance. The usual pattern is an ascending paralysis that may be first noticed as rubbery legs. Weakness typically evolves over hours to a few days and is frequently accompanied by tingling dysesthesias in the extremities. The legs are usually more affected than the arms, and facial diparesis is present in 50% of affected individuals. The lower cranial nerves are also frequently involved, causing bulbar weakness with difficulty handling secretions and maintaining an airway; the diagnosis in these patients may initially be mistaken for brainstem ischemia. Pain in the neck, shoulder, back, or diffusely over the spine is also common in the early stages of GBS, occurring in ~50% of patients. Most patients require hospitalization, and in different series, up to 30% require ventilatory assistance at some time during the illness. The need for mechanical ventilation is associated with more severe weakness on admission, a rapid tempo of progression, and the presence of facial and/or bulbar weakness during the first week of symptoms. Fever and constitutional symptoms are absent at the onset and, if present, cast doubt on the diagnosis. Deep tendon reflexes attenuate or disappear within the first few days of onset. Cutaneous sensory deficits (e.g., loss of pain and temperature sensation) are usually relatively mild, but functions subserved by large sensory fibers, such as deep tendon reflexes and proprioception, are more severely affected. Bladder dysfunction may occur in severe cases but is usually transient. If bladder dysfunction is a prominent feature and comes early in the course or there is a sensory level on examination, diagnostic possibilities other than GBS should be considered, particularly spinal cord disease **(Chap. 434)**. Once clinical worsening stops and the patient reaches a plateau (almost always within 4 weeks of onset), further progression is unlikely.

Autonomic involvement is common and may occur even in patients whose GBS is otherwise mild. The usual manifestations are loss of vasomotor control with wide fluctuations in blood pressure, postural hypotension, and cardiac dysrhythmias. These features require close monitoring, and management and can be fatal. Pain is another common feature of GBS; in addition to the acute pain described above, a deep aching pain may be present in weakened muscles that patients liken to having overexercised the previous day. Other pains in GBS include dysesthetic pain in the extremities as a manifestation of sensory nerve fiber involvement. These pains are self-limited and often respond to standard analgesics **(Chap. 10)**.

Several subtypes of GBS are recognized, as determined primarily by electrodiagnostic (Edx) and pathologic distinctions **(Table 439-1)**. The most common variant is acute inflammatory demyelinating polyneuropathy (AIDP). Additionally, there are two axonal variants, which are often clinically severe—the acute motor axonal neuropathy (AMAN) and acute motor sensory axonal neuropathy (AMSAN) subtypes. In addition, a range of limited or regional GBS syndromes are also encountered. Notable among these is the Miller Fisher syndrome (MFS), which presents as rapidly evolving ataxia and areflexia of limbs without weakness, and ophthalmoplegia, often with pupillary paralysis. The MFS variant accounts for ~5% of all cases and is strongly associated with antibodies to the ganglioside GQ1b (see "Immunopathogenesis," below). Other regional variants of GBS include (1) pure sensory forms; (2) ophthalmoplegia with anti-GQ1b antibodies as part of severe motor-sensory GBS; (3) GBS with severe bulbar and facial paralysis, sometimes associated with antecedent cytomegalovirus (CMV) infection and anti-GM2 antibodies; and (4) acute pandysautonomia **(Chap. 432)**.

Antecedent Events Approximately 70% of cases of GBS occur 1–3 weeks after an acute infectious process, usually respiratory or gastrointestinal. Culture and seroepidemiologic techniques show that 20–30% of all cases occurring in North America, Europe, and Australia are preceded by infection or reinfection with *Campylobacter jejuni*. A similar proportion is preceded by a human herpes virus infection, often CMV or Epstein-Barr virus. Other viruses (e.g., HIV, hepatitis E, Zika) and also *Mycoplasma pneumoniae* have been identified as agents involved in antecedent infections, as have recent immunizations. The swine influenza vaccine, administered widely in the United States in 1976, is the most notable example. Influenza vaccines in use from 1992 to 1994, however, resulted in only one additional case of GBS per million persons vaccinated, and the more recent seasonal influenza vaccines appear to confer a GBS risk of <1 per million. Epidemiologic studies looking at H1N1 vaccination demonstrated at most only a slight increased risk of GBS. Meningococcal vaccinations (Menactra) does not appear to carry an increased risk. Older-type rabies vaccine, prepared in nervous system tissue, is implicated as a trigger of GBS in developing countries where it is still used; the mechanism is presumably immunization against neural antigens. GBS also occurs more

TABLE 439-1 Subtypes of Guillain-Barré Syndrome (GBS)

SUBTYPE	FEATURES	ELECTRODIAGNOSIS	PATHOLOGY
Acute inflammatory demyelinating polyneuropathy (AIDP)	Adults affected more than children; 90% of cases in Western world; recovery rapid; anti-GM1 antibodies (<50%)	Demyelinating	First attack on Schwann cell surface; widespread myelin damage, macrophage activation, and lymphocytic infiltration; variable secondary axonal damage
Acute motor axonal neuropathy (AMAN)	Children and young adults; prevalent in China and Mexico; may be seasonal; recovery rapid; anti-GD1a antibodies	Axonal	First attack at motor nodes of Ranvier; macrophage activation, few lymphocytes, frequent periaxonal macrophages; extent of axonal damage highly variable
Acute motor sensory axonal neuropathy (AMSAN)	Mostly adults; uncommon; recovery slow, often incomplete; closely related to AMAN	Axonal	Same as AMAN, but also affects sensory nerves and roots; axonal damage usually severe
Miller Fisher syndrome (MFS)	Adults and children; ophthalmoplegia, ataxia, and areflexia; anti-GQ1b antibodies (90%)	Axonal or demyelinating	Few cases examined; resembles AIDP

frequently than can be attributed to chance alone in patients with lymphoma (including Hodgkin's disease), in HIV-seropositive individuals, and in patients with systemic lupus erythematosus (SLE).

C. jejuni has also been implicated in summer outbreaks of AMAN among children and young adults exposed to chickens in rural China. Infection by Zika virus recently has been implicated in the increased incidence of GBS in Brazil and other endemic regions.

Immunopathogenesis Several lines of evidence support an autoimmune basis for acute inflammatory demyelinating polyneuropathy (AIDP), the most common and best-studied type of GBS; the concept extends to all of the subtypes of GBS (Table 439-1).

It is likely that both cellular and humoral immune mechanisms contribute to tissue damage in AIDP. T cell activation is suggested by the finding that elevated levels of cytokines and cytokine receptors are present in serum (interleukin [IL] 2, soluble IL-2 receptor) and in cerebrospinal fluid (CSF) (IL-6, tumor necrosis factor α, interferon γ). AIDP is also closely analogous to an experimental T cell–mediated immunopathy designated *experimental allergic neuritis* (EAN). EAN is induced in laboratory animals by immune sensitization against protein fragments derived from peripheral nerve proteins, and in particular against the P2 protein. Based on analogy to EAN, it was initially thought that AIDP was likely to be primarily a T cell–mediated disorder; however, abundant data now suggest that autoantibodies directed against T-cell independent nonprotein determinants may be central to many cases.

Circumstantial evidence suggests that all GBS results from immune responses to nonself antigens (infectious agents, vaccines) that misdirect to host nerve tissue through a resemblance-of-epitope (molecular mimicry) mechanism **(Fig. 439-1)**. The neural targets are likely to be

Subtypes and variants	IgG autoantibodies to
Guillain-Barré syndrome	
Acute inflammatory demyelinating polyneuropathy	None
Facial variant: Facial diplegia and paresthesia	None
Acute motor axonal neuropathy	GM1, GD1a
More and less extensive forms	
Acute motor-sensory axonal neuropathy	GM1, GD1a
Acute motor-conduction-block neuropathy	GM1, GD1a
Pharyngeal-cervical-brachial weakness	GT1a>GQ1b>>GD1a
Miller Fisher syndrome	GQ1b, GT1a
Incomplete forms	
Acute ophthalmoparesis (without ataxia)	GQ1b, GT1a
Acute ataxic neyropathy (without ophthalmoplegia)	GQ1b, GT1a
CNS variant: Bickerstaff's brainstem encephalitis	GQ1b, GT1a

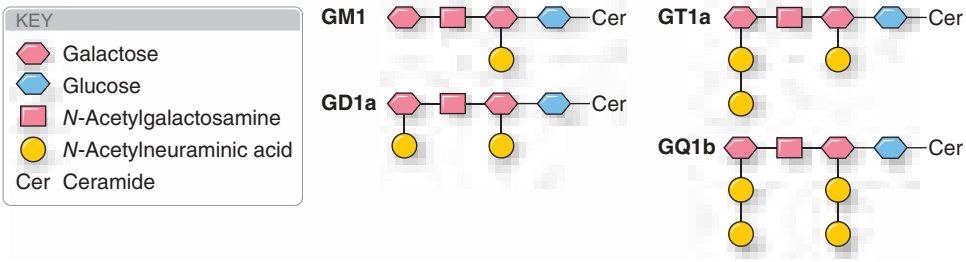


FIGURE 439-1 Spectrum of disorders in Guillain–Barré Syndrome and associated antiganglioside antibodies. IgG autoantibodies against GM1 or GD1a are strongly associated with AMAN, as well as the more extensive AMSAN, and the less extensive acute motor-conduction-block neuropathy. IgG anti-GQ1b antibodies, which cross-react with GT1a, are strongly associated with Miller Fisher syndrome, its incomplete forms (acute ophthalmoparesis [without ataxia] and acute ataxic neuropathy [without ophthalmoplegia]), and its more extensive form, Bickerstaff's brain-stem encephalitis. Pharyngeal–cervical–brachial weakness is categorized as a localized form of acute motor axonal neuropathy or an extensive form of Miller Fisher syndrome. Half of patients with pharyngeal–cervical–brachial weakness have IgG anti-GT1a antibodies, which often cross-react with GQ1b. IgG anti-GD1a antibodies have also been detected in a small percentage of patients. The anti-GQ1b antibody syndrome includes Miller Fisher syndrome, acute ophthalmoparesis, acute ataxic neuropathy, Bickerstaff's brain-stem encephalitis, and pharyngeal–cervical–brachial weakness. The presence of clinical overlap also indicates that Miller Fisher syndrome is part of a continuous spectrum with these conditions. Patients who have had Guillain–Barré syndrome overlapped with Miller Fisher syndrome or with its related conditions have IgG antibodies against GM1 or GD1a as well as against GQ1b or GT1a, supporting a link between AMAN and anti-GQ1b syndrome. AMAN, acute motor axonal neuropathy; AMSAN, acute motor–sensory axonal neuropathy. *(Adapted from N Yuki, H-P Hartung: Guillain-Barré syndrome. N Engl J Med 366:2294, 2012, Figure 1.)*

TABLE 439-2 Principal Antiglycolipid Antibodies Implicated in Immune Neuropathies

CLINICAL PRESENTATION	ANTIBODY TARGET	USUAL ISOTYPE
Acute Immune Neuropathies (Guillain-Barré Syndrome)		
Acute inflammatory demyelinating polyneuropathy (AIDP)	No clear patterns GM1 most common	IgG (polyclonal)
Acute motor axonal neuropathy (AMAN)	GD1a, GM1, GM1b, GalNAc–GD1a (<50% for any)	IgG (polyclonal)
Miller Fisher syndrome (MFS)	GQ1b (>90%)	IgG (polyclonal)
Acute pharyngeal cervicobrachial neuropathy (APCBN)	GT1a (? most)	IgG (polyclonal)
Chronic Immune Neuropathies		
Chronic inflammatory demyelinating polyneuropathy (CIDP) (75%)	<10% to CNTN1 or NF155 and even more rare to P0, myelin P2 protein, orPMP22,	IgG4 with CNTN1 and NF155
CIDP-M (MGUS associated) (25%)	Neural binding sites	IgG, IgA (monoclonal)
Chronic sensory > motor neuropathy	SPGP, SGLPG (on MAG) (50%)	IgM (monoclonal)
	Uncertain (50%)	IgM (monoclonal)
Multifocal motor neuropathy (MMN)	GM1, GalNAc–GD1a, others (25–50%)	IgM (polyclonal, monoclonal)
Chronic sensory ataxic neuropathy	GD1b, GQ1b, and other b-series gangliosides	IgM (monoclonal)

Abbreviations: CIDP-M, CIDP with a monoclonal gammopathy; CNTN1, contactin-1; MAG, myelin-associated glycoprotein; MGUS, monoclonal gammopathy of undetermined significance; NF155, neurofascin-155.

Source: Modified from HJ Willison, N Yuki: Brain 125:2591, 2002.

glycoconjugates, specifically gangliosides **(Table 439-2; Fig. 439-2)**. Gangliosides are complex glycosphingolipids that contain one or more sialic acid residues; various gangliosides participate in cell-cell interactions (including those between axons and glia), modulation of receptors, and regulation of growth. They are typically exposed on the plasma membrane of cells, rendering them susceptible to an antibody-mediated attack. Gangliosides and other glycoconjugates are present in large quantity in human nervous tissues and in key sites, such as nodes of Ranvier. Antiganglioside antibodies, most frequently to GM1, are common in GBS (20–50% of cases), particularly in AMAN and AMSAN and in those cases preceded by *C. jejuni* infection. Some AIDP autoantibodies may recognize glycolipid heterocomplexes, rather than single species, present on cell membranes. Furthermore, isolates of *C. jejuni* from stool cultures of patients with GBS have surface glycolipid structures that antigenically cross react with gangliosides, including GM1, concentrated in human nerves. Sialic acid residues from pathogenic *C. jejuni* strains can also trigger activation of dendritic cells via signaling through toll-like receptor 4 (TLR4), promoting B cell differentiation and further amplifying humoral autoimmunity. Another line of evidence implicating humoral autoimmunity is derived from cases of GBS that followed intravenous administration of bovine brain gangliosides for treatment of various neuropathies; 5–15 days after injection, some recipients developed AMAN with high titers of anti-GM1 antibodies that recognized epitopes at nodes of Ranvier and motor endplates. Experimentally, anti-GM1 antibodies can trigger complement-mediated injury at paranodal axon-glial junctions, disrupting the clustering of sodium channels and likely contributing to conduction block (see "Pathophysiology," below).

Anti-GQ1b IgG antibodies are found in >90% of patients with MFS (Table 439-2; Fig. 439-2), and titers of IgG are highest early in the course. Anti-GQ1b antibodies are not found in other forms of GBS unless there is extraocular motor nerve involvement. A possible explanation for this association is that extraocular motor nerves are enriched in GQ1b gangliosides in comparison to limb nerves. In addition, a monoclonal anti-GQ1b antibody raised against *C. jejuni* isolated from a patient with MFS blocked neuromuscular transmission experimentally.

Taken together, these observations provide strong but still inconclusive evidence that autoantibodies play an important pathogenic role in GBS. Although antiganglioside antibodies have been studied most intensively, other antigenic targets may also be important. Proof that these antibodies are pathogenic requires that they be capable of mediating disease following direct passive transfer to naïve hosts; this has not yet been demonstrated, although one case of possible maternal-fetal transplacental transfer of GBS has been described.

In AIDP, an early step in the induction of tissue damage appears to be complement deposition along the outer surface of the Schwann cell. Activation of complement initiates a characteristic vesicular disintegration of the myelin sheath and also leads to recruitment of activated macrophages, which participate in damage to myelin and axons. In AMAN, the pattern is different in that complement is deposited along with IgG at the nodes of Ranvier along large motor axons. Interestingly, in cases of AMAN, antibodies against GD1a appear to have a fine specificity that favors binding to motor rather than sensory nerve roots, even though this ganglioside is expressed on both fiber types.

Pathophysiology In the demyelinating forms of GBS, the basis for flaccid paralysis and sensory disturbance is conduction block. This finding, demonstrable electrophysiologically, implies that the axonal connections remain intact. Hence, recovery can take place rapidly as remyelination occurs. In severe cases of demyelinating GBS, secondary axonal degeneration usually occurs; its extent can be estimated electrophysiologically. More secondary axonal degeneration correlates with a slower rate of recovery and a greater degree of residual disability. When a severe primary axonal pattern is encountered electrophysiologically, the implication is that axons have degenerated and become disconnected from their targets, specifically the neuromuscular junctions, and must therefore regenerate for recovery to take place. In motor axonal cases in which recovery is rapid, the lesion is thought to be localized to preterminal motor branches, allowing regeneration and reinnervation to take place quickly. Alternatively, in mild cases, collateral sprouting and reinnervation from surviving motor axons near the neuromuscular junction may begin to reestablish physiologic continuity with muscle cells over a period of several months.

Laboratory Features CSF findings are distinctive, consisting of an elevated CSF protein level (1–10 g/L [100–1000 mg/dL]) without accompanying pleocytosis. The CSF is often normal when symptoms have been present for ≤48 h; by the end of the first week, the level of protein is usually elevated. A transient increase in the CSF white cell count (10–100/μL) occurs on occasion in otherwise typical GBS; however, a sustained CSF pleocytosis suggests an alternative diagnosis (viral myelitis) or a concurrent diagnosis such as unrecognized HIV infection, leukemia or lymphoma with infiltration of nerves, or neurosarcoidosis. Edx features are mild or absent in the early stages of GBS and lag behind the clinical evolution. In AIDP, the earliest features are prolonged F-wave latencies, prolonged distal latencies, and reduced amplitudes of compound muscle action potentials (CMAPs), probably owing to the predilection for involvement of nerve roots and distal motor nerve terminals early in the course. Later, slowing of conduction velocity, conduction block, and temporal dispersion may be appreciated (Table 439-1). Occasionally, sensory nerve action potentials (SNAPs) may be normal in the feet (e.g., sural nerve) when abnormal in the arms. This is also a sign that the patient does not have one of the more typical "length-dependent" polyneuropathies. In cases with primary axonal pathology, the principal Edx finding is reduced amplitude of CMAPs (and also SNAPS with AMSAN) without conduction slowing or prolongation of distal latencies.

Diagnosis GBS is a descriptive entity. The diagnosis of AIDP is made by recognizing the pattern of rapidly evolving paralysis with areflexia, absence of fever or other systemic symptoms, and characteristic antecedent events. In 2011, the Brighton Collaboration developed

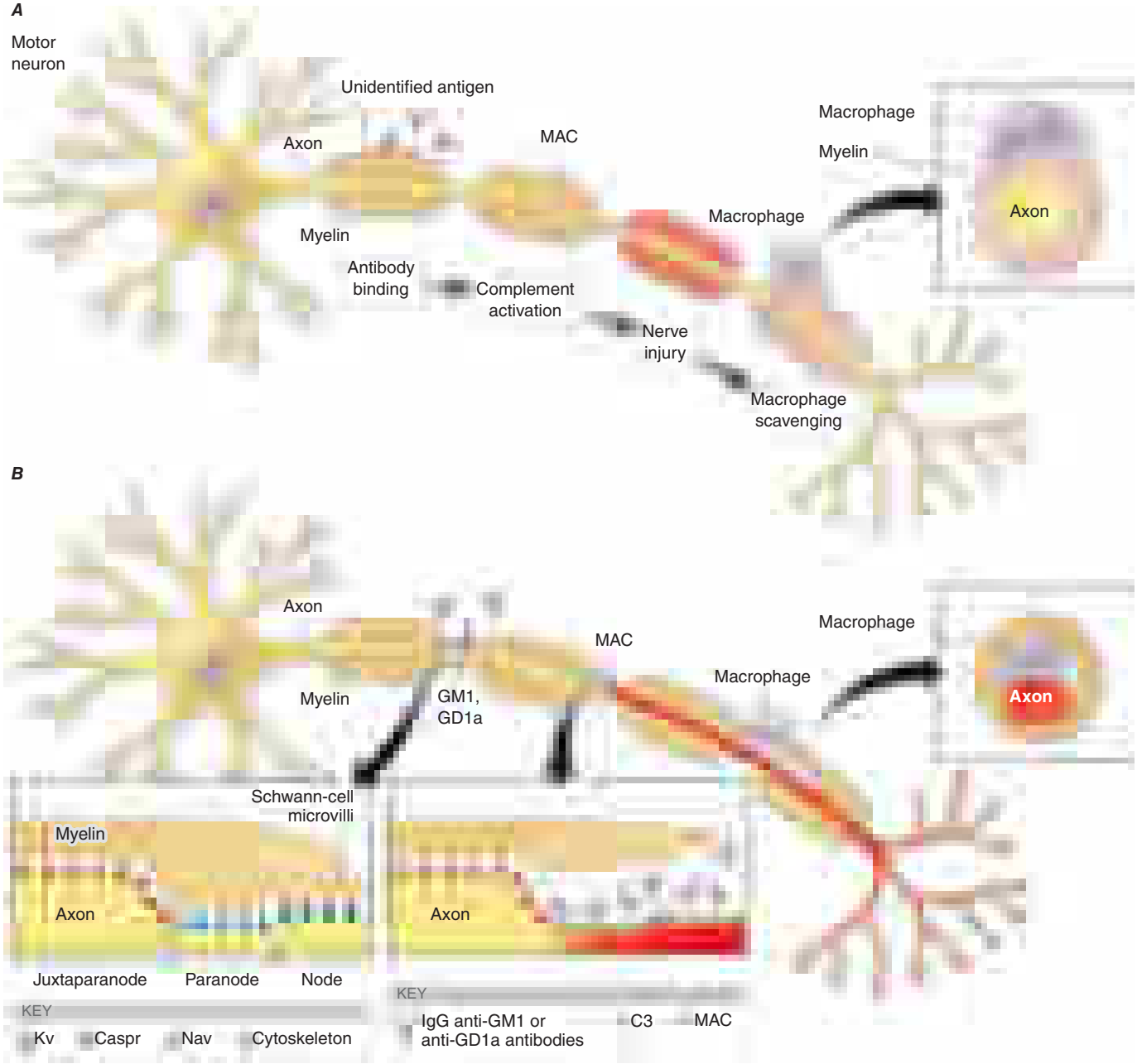


FIGURE 439-2 Possible immune mechanisms in GBS. Panel ***A*** shows the immunopathogenesis of AIDP. Although autoantigens have yet to be unequivocally identified, autoantibodies may bind to myelin antigens and activate complement. This is followed by the formation of membrane-attack complex (MAC) on the outer surface of Schwann cells and the initiation of vesicular degeneration. Macrophages subsequently invade myelin and act as scavengers to remove myelin debris. Panel ***B*** shows the immunopathogenesis of acute axonal forms of GBS (AMAN and AMSAN). Myelinated axons are divided into four functional regions: the nodes of Ranvier, paranodes, juxtaparanodes, and internodes. Gangliosides GM1 and GD1a are strongly expressed at the nodes of Ranvier, where the voltage-gated sodium (Nav) channels are localized. Contactin-associated protein (Caspr) and voltage-gated potassium (Kv) channels are respectively present at the paranodes and juxtaparanodes. IgG anti-GM1 or anti-GD1a autoantibodies bind to the nodal axolemma, leading to MAC formation. This results in the disappearance of Nav clusters and the detachment of paranodal myelin, which can lead to nerve-conduction failure and muscle weakness. Axonal degeneration may follow at a later stage. Macrophages subsequently invade from the nodes into the periaxonal space, scavaenging the injured axons. *(Adapted from N Yuki, H-P Hartung: Guillain-Barré syndrome. N Engl J Med 366:2294, 2012, Figure 2.)*

a new set of case definitions for GBS in response to needs of epidemiologic studies of vaccination and assessing risks of GBS (Table 439-3). These criteria have subsequently been validated. Other disorders that may enter into the differential diagnosis include acute myelopathies (especially with prolonged back pain and sphincter disturbances); diphtheria (early oropharyngeal disturbances); Lyme polyradiculitis and other tick-borne paralyses; porphyria (abdominal pain, seizures, psychosis); vasculitic neuropathy (check erythrocyte sedimentation rate, described below); poliomyelitis (fever and meningismus common); West Nile virus; CMV polyradiculitis (in immunocompromised patients); critical illness neuropathy or myopathy; neuromuscular junction disorders such as myasthenia gravis and botulism (pupillary reactivity lost early); poisonings with organophosphates, thallium, or arsenic; paralytic shellfish poisoning; or severe hypophosphatemia (rare). Laboratory tests are helpful primarily to exclude mimics of GBS. Edx features may be minimal, and the CSF protein level may not rise until the end of the first week. If the diagnosis is strongly suspected, treatment should be initiated without waiting for evolution of the characteristic Edx and CSF findings to occur. GBS patients with risk factors for HIV or with CSF pleocytosis should have a serologic test for HIV.

TREATMENT

Guillain-Barré Syndrome

In the vast majority of patients with GBS, treatment should be initiated as soon after diagnosis as possible. Each day counts; ~2 weeks after the first motor symptoms, it is not known whether

TABLE 439-3 Brighton Criteria for Diagnosis of Guillain-Barré Syndrome (GBS) and Miller Fisher Syndrome

Clinical case definitions for diagnosis of GBS

Level 1 of diagnostic certainty

Bilateral AND flaccid weakness of the limbs

AND

Decreased or absent deep tendon reflexes in weak limbs

AND

Monophasic illness pattern and interval between onset and nadir of weakness between 12 h and 28 days and subsequent clinical plateau

AND

Electrophysiologic findings consistent with GBS

AND

Cytoalbuminologic dissociation (i.e., elevation of CSF protein level above laboratory normal value AND CSF total white cell count <50 cells/μL)

AND

Absence of an identified alternative diagnosis for weakness

Level 2 of diagnostic certainty

Bilateral AND flaccid weakness of the limbs

AND

Decreased or absent deep tendon reflexes in weak limbs

AND

Monophasic illness pattern and interval between onset and nadir of weakness between 12 h and 28 days and subsequent clinical plateau

AND

CSF total white cell count <50 cells/μL (with or without CSF protein elevation above laboratory normal value)

OR

If CSF not collected or results not available, electrophysiologic studies consistent with GBS

AND

Absence of identified alternative diagnosis for weakness

Level 3 of diagnostic certainty

Bilateral and flaccid weakness of the limbs

AND

Decreased or absent deep tendon reflexes in weak limbs

AND

Monophasic illness pattern and interval between onset and nadir of weakness between 12 h and 28 days and subsequent clinical plateau

AND

Absence of identified alternative diagnosis for weakness

Clinical case definitions for diagnosis of Miller Fisher syndrome

Level 1 of diagnostic certainty

Bilateral ophthalmoparesis and bilateral reduced or absent tendon reflexes, and ataxia

AND

Absence of limb weakness

AND

Monophasic illness pattern and interval between onset and nadir of weakness between 12 h and 28 days and subsequent clinical plateau

AND

Cytoalbuminologic dissociation (i.e., elevation of cerebrospinal protein above the laboratory normal and total CSF white cell count <50 cells/μL)

AND

Nerve conduction studies are normal, OR indicate involvement of sensory nerves only

AND

No alterations in consciousness or corticospinal tract signs

AND

Absence of identified alternative diagnosis

Level 2 of diagnostic certainty

Bilateral ophthalmoparesis and bilateral reduced or absent tendon reflexes and ataxia

AND

Absence of limb weakness

AND

Monophasic illness pattern and interval between onset and nadir of weakness between 12 h and 28 days and subsequent clinical plateau

AND

CSF with a total white cell count <50 cells/μL) (with or without CSF protein elevation above laboratory normal value)

OR

Nerve conduction studies are normal, OR indicate involvement of sensory nerves only

AND

No alterations in consciousness or corticospinal tract signs

AND

Absence of identified alternative diagnosis

Level 3 of diagnostic certainty

Bilateral ophthalmoparesis and bilateral reduced or absent tendon reflexes and ataxia

AND

Absence of limb weakness

AND

Monophasic illness pattern and interval between onset and nadir of weakness between 12 h and 28 days and subsequent clinical plateau

AND

No alterations in consciousness or corticospinal tract signs

AND

Absence of identified alternative diagnosis

Abbreviation: CSF, cerebrospinal fluid.

Source: From JJ Sejvar et al: Guillain-Barré syndrome and Fisher syndrome: Case definitions and guidelines for collection, analysis, and presentation of immunization safety data. Vaccine 29:599, 2011. Validation study published by C Fokke et al: Diagnosis of Guillain-Barré syndrome and validation of Brighton criteria. Brain 137:33, 2014.

immunotherapy is still effective. If the patient has already reached the plateau stage, then treatment probably is no longer indicated, unless the patient has severe motor weakness and one cannot exclude the possibility that an immunologic attack is still ongoing. Either high-dose intravenous immune globulin (IVIg) or plasmapheresis can be initiated, as they are equally effective for typical GBS. A combination of the two therapies is not significantly better than either alone. IVIg is often the initial therapy chosen because of its ease of administration and good safety record. Anecdotal data have also suggested that IVIg may be preferable to plasma exchange (PE) for the AMAN and MFS variants of GBS. IVIg is administered as five daily infusions for a total dose of 2 g/kg body weight. There is some evidence that GBS autoantibodies are neutralized by anti-idiotypic antibodies present in IVIg preparations, perhaps accounting for the therapeutic effect. A course of plasmapheresis usually consists of ~40–50 mL/kg PE 4–5 times over 7–10 days. Meta-analysis of randomized clinical trials indicates that treatment reduces the need for mechanical ventilation by nearly half (from 27 to 14% with PE) and increases the likelihood of full recovery at 1 year (from 55 to 68%). Functionally significant improvement may occur toward the end of the first week of treatment or may be delayed for several weeks. The lack of noticeable improvement following a course of IVIg or PE is not an indication to treat with the alternate treatment. However, there are occasional patients who are treated early in the course of GBS and improve, who then relapse within a month. Brief retreatment with the original therapy is usually effective in such cases. Glucocorticoids have not been found to be effective in GBS. Occasional patients with very mild forms of GBS, especially those who appear to have already reached a plateau when initially seen, may be managed conservatively without IVIg or PE.

In the worsening phase of GBS, most patients require monitoring in a critical care setting, with particular attention to vital capacity, heart rhythm, blood pressure, nutrition, deep-vein thrombosis prophylaxis, cardiovascular status, early consideration (after 2 weeks of intubation) of tracheotomy, and chest physiotherapy. As noted, ~30% of patients with GBS require ventilatory assistance, sometimes for prolonged periods of time (several weeks or longer). Frequent turning and assiduous skin care are important, as are daily range-of-motion exercises to avoid joint contractures and daily reassurance as to the generally good outlook for recovery.

Prognosis and Recovery Approximately 85% of patients with GBS achieve a full functional recovery within several months to a year, although minor findings on examination (such as areflexia) may persist and patients often complain of continued symptoms, including fatigue. The mortality rate is <5% in optimal settings; death usually results from secondary pulmonary complications. The outlook is worst in patients with severe proximal motor and sensory axonal damage. Such axonal damage may be either primary or secondary in nature (see "Pathophysiology," above), but in either case successful regeneration cannot occur. Other factors that worsen the outlook for recovery are advanced age, a fulminant or severe attack, and a delay in the onset of treatment. Between 5 and 10% of patients with typical GBS have one or more late relapses; many of these cases are then classified as chronic inflammatory demyelinating polyneuropathy (CIDP).

CHRONIC INFLAMMATORY DEMYELINATING POLYNEUROPATHY

CIDP is distinguished from GBS by its chronic course. In other respects, this neuropathy shares many features with the common demyelinating form of GBS, including elevated CSF protein levels and the Edx findings of acquired demyelination. Most cases occur in adults, and males are affected slightly more often than females. The incidence of CIDP is lower than that of GBS, but due to the protracted course, the prevalence is greater.

Clinical Manifestations Onset is usually gradual over a few months or longer, but in a few cases, the initial attack is indistinguishable from that of GBS. An acute-onset form of CIDP may mimic GBS but should be considered if it deteriorates >9 weeks after onset or relapses at least three times. Symptoms are both motor and sensory in most cases. Weakness of the limbs is usually symmetric but can be strikingly asymmetric in multifocal acquired demyelinating sensory and motor (MADSAM) neuropathy variant (Lewis-Sumner syndrome) in which discrete peripheral nerves are involved. There is considerable variability from case to case. Some patients experience a chronic progressive course, whereas others, usually younger patients, have a relapsing and remitting course. Some have only motor findings, and a small proportion present with a relatively pure syndrome of sensory ataxia. Tremor occurs in ~10% and may become more prominent during periods of subacute worsening or improvement. A small proportion have cranial nerve findings, including external ophthalmoplegia. CIDP tends to ameliorate over time with treatment; the result is that many years after onset, nearly 75% of patients have reasonable functional status. Death from CIDP is uncommon.

Diagnosis The diagnosis rests on characteristic clinical, CSF, and electrophysiologic findings. The CSF is usually acellular with an elevated protein level, sometimes several times normal. As with GBS, a CSF pleocytosis should lead to the consideration of HIV infection, leukemia or lymphoma, and neurosarcoidosis. Edx findings reveal variable degrees of conduction slowing, prolonged distal latencies, distal and temporal dispersion of CMAPs, and conduction block as the principal features. In particular, the presence of conduction block is a certain sign of an acquired demyelinating process. Evidence of axonal loss, presumably secondary to demyelination, is present in >50% of patients. Serum protein electrophoresis with immunofixation is indicated to search for monoclonal gammopathy and associated conditions (see "Monoclonal Gammopathy of Undetermined Significance," below). In all patients with presumptive CIDP, it is also reasonable to exclude vasculitis, collagen vascular disease (especially SLE), chronic hepatitis, HIV infection, amyloidosis, and diabetes mellitus. Other associated conditions include inflammatory bowel disease and lymphoma.

Pathogenesis Biopsy typically reveals little inflammation and onion-bulb changes (imbricated layers of attenuated Schwann cell processes surrounding an axon) that result from recurrent demyelination and remyelination (Fig. 439-1). The response to therapy suggests that CIDP is immune-mediated; CIDP responds to glucocorticoids, whereas GBS does not. Passive transfer of demyelination into experimental animals has been accomplished using IgG purified from the serum of some patients with CIDP, lending support for a humoral autoimmune pathogenesis. A minority of patients have serum antibodies against P0, myelin P2 protein, or PMP22 (proteins whose genes are mutated in certain forms of hereditary Charcot-Marie-Tooth neuropathy). More recently antibodies of IgG4 isotype directed against contactin-1 (CNTN1) or neurofascin-155 (NF155) have been associated with early axonal damage, severe distal motor involvement or sensory ataxia with tremor, and a poor response to IVIg. CNTN1 and its partner contactin-associated protein-1 (CASPR1) interact with NF155 at paranodal axoglial junctions. Passive transfer of IgG4 CNTN1 antibodies produces paranodal damage and ataxia in rodents. It is also of interest that a CIDP-like illness developed spontaneously in the nonobese diabetic (NOD) mouse when the immune co-stimulatory molecule B7-2 (CD86) was genetically deleted; this suggests that CIDP can result from altered triggering of T cells by antigen-presenting cells.

As many as 25% of patients with clinical features of CIDP also have a monoclonal gammopathy of undetermined significance (MGUS). Cases associated with monoclonal IgA or IgG kappa usually respond to treatment as favorably as cases without a monoclonal gammopathy. Patients with IgM-kappa monoclonal gammopathy and antibodies directed against myelin-associated glycoprotein (MAG) have a distinct demyelinating polyneuropathy with more sensory findings, usually only distal weakness, and a poor response to immunotherapy.

TREATMENT

Chronic Inflammatory Demyelinating Polyneuropathy

Most authorities initiate treatment for CIDP when progression is rapid or walking is compromised. If the disorder is mild, management can be expectant, awaiting spontaneous remission. Controlled studies have shown that high-dose IVIg, PE, and glucocorticoids are all more effective than placebo. Initial therapy is usually with IVIg, administered as 2.0 g/kg body weight given in divided doses over 2–5 days; three monthly courses are generally recommended before concluding a patient is a treatment failure. If the patient responds, the infusion intervals can be gradually increased or the dosage decreased (e.g., starting at 1 g/kg every 3–4 weeks). PE, which appears to be as effective as IVIg, is initiated at 2-3 treatments per week for 6 weeks; periodic re-treatment may also be required. Treatment with glucocorticoids is another option (60–80 mg prednisone PO daily for 1–2 months, followed by a gradual dose reduction of 10 mg per month as tolerated), but long-term adverse effects including bone demineralization, gastrointestinal bleeding, and cushingoid changes are problematic. As many as one-third of patients with CIDP fail to respond adequately to the initial therapy chosen; a different treatment should then be tried. Patients who fail therapy with IVIg, PE, and glucocorticoids may benefit from treatment with immunosuppressive agents such as azathioprine, methotrexate, cyclosporine, and cyclophosphamide, either alone or as adjunctive therapy. CIDP associated with anti-CNTN1 and NF155 antibodies is typically refractory to IVIg, but a few studies suggest a response to rituximab. Use of these therapies requires periodic reassessment of their risks and benefits. In patients with a CIDP-like neuropathy who fail to respond to treatment, it is important to evaluate for POEMS syndrome (*p*olyneuropathy, *o*rganomegaly, *e*ndocrinopathy, *m*onoclonal gammopathy, *s*kin changes; see below).

MULTIFOCAL MOTOR NEUROPATHY

Multifocal motor neuropathy (MMN) is a distinctive but uncommon neuropathy that presents as slowly progressive motor weakness and atrophy evolving over years in the distribution of selected nerve trunks, associated with sites of persistent focal motor conduction block in the same nerve trunks. Sensory fibers are relatively spared. The arms are affected more frequently than the legs, and >75% of all patients are male. Some cases have been confused with lower motor neuron forms of amyotrophic lateral sclerosis **(Chap. 429)**. Less than 50% of patients present with high titers of polyclonal IgM antibody to the ganglioside GM1. It is uncertain how this finding relates to the discrete foci of persistent motor conduction block, but high concentrations of GM1 gangliosides are normal constituents of nodes of Ranvier in peripheral nerve fibers. Pathology reveals demyelination and mild inflammatory changes at the sites of conduction block.

Most patients with MMN respond to high-dose IVIg (dosages as for CIDP, above); periodic re-treatment is required (usually at least monthly) to maintain the benefit. Some refractory patients have responded to rituximab or cyclophosphamide. Glucocorticoids and PE are not effective.

NEUROPATHIES WITH MONOCLONAL GAMMOPATHY

MULTIPLE MYELOMA

Clinically overt polyneuropathy occurs in ~5% of patients with the commonly encountered type of multiple myeloma, which exhibits either lytic or diffuse osteoporotic bone lesions. These neuropathies are sensorimotor, are usually mild and slowly progressive but may be severe, and generally do not reverse with successful suppression of the myeloma. In most cases, Edx and pathologic features are consistent with a process of axonal degeneration.

In contrast, myeloma with osteosclerotic features, although representing only 3% of all myelomas, is associated with polyneuropathy in one-half of cases. These neuropathies, which may also occur with solitary plasmacytoma, are distinct because they (1) are usually demyelinating in nature and resemble CIDP; (2) often respond to radiation therapy or removal of the primary lesion; (3) are associated with different monoclonal proteins and light chains (almost always lambda as opposed to primarily kappa in the lytic type of multiple myeloma); (4) are typically refractory to standard treatments of CIDP; and (5) may occur in association with other systemic findings including thickening of the skin, hyperpigmentation, hypertrichosis, organomegaly, endocrinopathy, anasarca, and clubbing of fingers. These are features of POEMS syndrome. Levels of vascular endothelial growth factor (VEGF) are increased in the serum, and this factor is thought to somehow play a pathogenic role in this syndrome. Treatment of the neuropathy is best directed at the osteosclerotic myeloma using surgery, radiotherapy, chemotherapy, or autologous peripheral blood stem cell transplantation.

Neuropathies are also encountered in other systemic conditions with gammopathy, including Waldenström's macroglobulinemia, primary systemic amyloidosis, and cryoglobulinemic states (mixed essential cryoglobulinemia, some cases of hepatitis C).

MONOCLONAL GAMMOPATHY OF UNDETERMINED SIGNIFICANCE

Chronic polyneuropathies occurring in association with MGUS are usually associated with the immunoglobulin isotypes IgG, IgA, and IgM. Most patients present with isolated sensory symptoms in their distal extremities and have Edx features of an axonal sensory or sensorimotor polyneuropathy. These patients otherwise resemble idiopathic sensory polyneuropathy, and the MGUS might just be coincidental. They usually do not respond to immunotherapies designed to reduce the concentration of the monoclonal protein. Some patients, however, present with generalized weakness and sensory loss and Edx studies indistinguishable from CIDP without monoclonal gammopathy (see "Chronic Inflammatory Demyelinating Polyneuropathy," above), and their response to immunosuppressive agents is also similar. An exception is the syndrome of IgM kappa monoclonal gammopathy associated with an indolent, long-standing, sometimes static sensory neuropathy, frequently with tremor and sensory ataxia. Most patients are male and aged >50 years. In the majority, the monoclonal IgM immunoglobulin binds to a normal peripheral nerve constituent, MAG, found in the paranodal regions of Schwann cells. Binding appears to be specific for a polysaccharide epitope that is also found in other normal peripheral nerve myelin glycoproteins, P0 and PMP22, and also in other normal nerve-related glycosphingolipids (Fig. 439-1). In the MAG-positive cases, IgM paraprotein is incorporated into the myelin sheaths of affected patients and widens the spacing of the myelin lamellae, thus producing a distinctive ultrastructural pattern. Demyelination and remyelination are the hallmarks of the lesions, but axonal loss develops over time. These anti-MAG polyneuropathies are typical refractory to immunotherapy. In a small proportion of patients (30% at 10 years), MGUS will in time evolve into frankly malignant conditions such as multiple myeloma or lymphoma.

VASCULITIC NEUROPATHY

Peripheral nerve involvement is common in polyarteritis nodosa (PAN), appearing in half of all cases clinically and in 100% of cases at postmortem studies **(Chap. 356)**. The most common pattern is multifocal (asymmetric) motor-sensory neuropathy (mononeuropathy multiplex) due to ischemic lesions of nerve trunks and roots; however, some cases of vasculitic neuropathy present as a distal, symmetric sensorimotor polyneuropathy. Symptoms of neuropathy are a common presenting complaint in patients with PAN. The Edx findings are those of an axonal process. Small- to medium-sized arteries of the vasa nervorum, particularly the epineural vessels, are affected in PAN, resulting in a widespread ischemic neuropathy. A high frequency of neuropathy occurs in eosinophilic granulomatosis with polyangiitis (Churg-Strauss syndrome [CSS]).

Systemic vasculitis should always be considered when a subacute or chronically evolving mononeuropathy multiplex occurs in conjunction with constitutional symptoms (fever, anorexia, weight loss, loss of energy, malaise, and nonspecific pains). Diagnosis of suspected vasculitic neuropathy is made by a combined nerve and muscle biopsy, with serial section or skip-serial techniques.

Approximately one-third of biopsy-proven cases of vasculitic neuropathy are "nonsystemic" in that the vasculitis appears to affect only peripheral nerves. Constitutional symptoms are absent, and the course is more indolent than that of PAN. The erythrocyte sedimentation rate may be elevated, but other tests for systemic disease are negative. Nevertheless, clinically silent involvement of other organs is likely, and vasculitis is frequently found in muscle biopsied at the same time as nerve.

Vasculitic neuropathy may also be seen as part of the vasculitis syndrome occurring in the course of other connective tissue disorders **(Chap. 356)**. The most frequent is rheumatoid arthritis, but ischemic neuropathy due to involvement of vasa nervorum may also occur in mixed cryoglobulinemia, Sjögren's syndrome, granulomatosis with polyangiitis (formerly known as Wegener's), hypersensitivity angiitis, SLE, and progressive systemic sclerosis.

Some vasculitides are associated with antineutrophil cytoplasmic antibodies (ANCAs), which in turn, are subclassified as cytoplasmic (cANCA) or perinuclear (pANCA). cANCAs are directed against proteinase 3 (PR3), whereas pANCAs target myeloperoxidase (MPO). PR3/cANCAs are associated with eosinophilic granulomatosis with polyangiitis, whereas MPO/pANCAs are typically associated with microscopic polyangiitis, CSS, and less commonly PAN. Of note, MPO/pANCA has also been seen in minocycline-induced vasculitis.

Management of these neuropathies, including the "nonsystemic" vasculitic neuropathy, consists of treatment of the underlying condition as well as the aggressive use of glucocorticoids and cyclophosphamide. Use of these immunosuppressive agents has resulted in dramatic improvements in outcome, with 5-year survival rates now >80%. Clinical trials found that the combination of rituximab and glucocorticoids is not inferior to cyclophosphamide and glucocorticoids. Thus, combination therapy with glucocorticoids and rituximab is recommended as the

standard initial treatment, particularly for ANCA-associated vasculitis. A recent study demonstrated that mepolizumab, an anti-interleukin-5 monoclonal antibody, when added to standard care is effective for treatment of eosinophilic granulomatosis with polyangiitis.

ANTI-Hu PARANEOPLASTIC NEUROPATHY (CHAP. 90)

This uncommon immune-mediated disorder manifests as a sensory neuronopathy (i.e., selective damage to sensory nerve bodies in dorsal root ganglia). The onset is often asymmetric with dysesthesias and sensory loss in the limbs that soon progress to affect all limbs, the torso, and the face. Marked sensory ataxia, pseudoathetosis, and inability to walk, stand, or even sit unsupported are frequent features and are secondary to the extensive deafferentation. Subacute sensory neuronopathy may be idiopathic, but more than half of cases are paraneoplastic, primarily related to lung cancer, and most of those are small-cell lung cancer (SCLC). Diagnosis of the underlying SCLC requires awareness of the association, testing for the paraneoplastic antibody, and often positron emission tomography (PET) scanning for the tumor. The target antigens are a family of RNA-binding proteins (HuD, HuC, and Hel-N1) that in normal tissues are only expressed by neurons. The same proteins are usually expressed by SCLC, triggering in some patients an immune response characterized by antibodies and cytotoxic T cells that cross-react with the Hu proteins of the dorsal root ganglion neurons, resulting in immune-mediated neuronal destruction. An encephalomyelitis may accompany the sensory neuronopathy and presumably has the same pathogenesis. Neurologic symptoms usually precede, by ≤6 months, the identification of SCLC. The sensory neuronopathy runs its course in a few weeks or months and stabilizes, leaving the patient disabled. Most cases are unresponsive to treatment with glucocorticoids, IVIg, PE, or immunosuppressant drugs.

FURTHER READING

Amato AA, Russell JA (eds): *Neuromuscular Disorders*, 2nd ed. New York, McGraw-Hill, 2016, pp 320–383.

Collins MP: The vasculitic neuropathies: an update. Curr Opin Neurol 25:573, 2012.

Devaux JJ et al: Neurofascin-155 IgG4 in chronic inflammatory demyelinating polyneuropathy. Neurology 86:800, 2016.

Fokke C et al: Diagnosis of Guillain-Barré syndrome and validation of Brighton criteria. Brain 137:33, 2014.

Manso C et al: Contactin-1 IgG4 antibodies cause paranode dismantling and conduction defects. Brain 139:1700, 2016.

Querol L, Illa I: Paranodal and other autoantibodies in chronic inflammatory neuropathies. Curr Opin Neurol 28:474, 2015.

Willison HJ et al: Guillain-Barré syndrome. Lancet 388: 717, 2016.

Yuki N, Hartung H-P: Guillain-Barré syndrome. N Engl J Med 366:2294, 2012.

440 Myasthenia Gravis and Other Diseases of the Neuromuscular Junction

Anthony A. Amato

Myasthenia gravis (MG) is a neuromuscular junction (NMJ) disorder characterized by weakness and fatigability of skeletal muscles. The underlying defect is a decrease in the number of available acetylcholine receptors (AChRs) at NMJs due to an antibody-mediated autoimmune attack. Treatment now available for MG is highly effective, although a specific cure has remained elusive.

PATHOPHYSIOLOGY

At the NMJ **(Fig. 440-1, Video 440-1)**, acetylcholine (ACh) is synthesized in the motor nerve terminal and stored in vesicles (quanta). When an action potential travels down a motor nerve and reaches the nerve terminal, ACh from 150 to 200 vesicles is released and combines with AChRs that are densely packed at the peaks of postsynaptic folds. The AChR consists of five subunits (2α, 1β, 1δ, 1γ, or ε) arranged around a central pore. When ACh combines with the binding sites on the α subunits of the AChR, the channel in the AChR opens, permitting the rapid entry of cations, chiefly sodium, which produces depolarization at the end-plate region of the muscle fiber. If the depolarization is sufficiently large, it initiates an action potential that is propagated along the muscle fiber, triggering muscle contraction. This process is rapidly terminated by hydrolysis of ACh by acetylcholinesterase (AChE), which is present within the synaptic folds, and by diffusion of ACh away from the receptor.

In MG, the fundamental defect is a decrease in the number of available AChRs at the postsynaptic muscle membrane. In addition, the postsynaptic folds are flattened, or "simplified." These changes result in decreased efficiency of neuromuscular transmission. Therefore, although ACh is released normally, it produces small end-plate potentials that may fail to trigger muscle action potentials. Failure of transmission results in weakness of muscle contraction.

The amount of ACh released per impulse normally declines on repeated activity (termed *presynaptic rundown*). In the myasthenic patient, the decreased efficiency of neuromuscular transmission combined with the normal rundown results in the activation of fewer and fewer muscle fibers by successive nerve impulses and hence increasing weakness, or *myasthenic fatigue*. This mechanism also accounts for the decremental response to repetitive nerve stimulation seen on electrodiagnostic testing.

MG is an autoimmune disorder most commonly caused by anti-AChR antibodies. The anti-AChR antibodies reduce the number of available AChRs at NMJs by three distinct mechanisms: (1) accelerated turnover of AChRs by a mechanism involving cross-linking and rapid endocytosis of the receptors; (2) damage to the postsynaptic muscle membrane by the antibody in collaboration with complement; and (3) blockade of the active site of the AChR (i.e., the site that normally binds Ach). An immune response to muscle-specific kinase (MuSK), a protein involved in AChR clustering at the NMJ, can also result in MG, with reduction of AChRs demonstrated experimentally. Anti-MuSK antibody occurs in about 10% of patients (about 40% of AChR antibody negative patients), while 1–3% have antibodies to another protein at the NMJ—low-density lipoprotein receptor-related protein 4 (LRP4)—that is also important for clustering of AChRs. The pathogenic antibodies are IgG and are T cell dependent. Thus, immunotherapeutic strategies directed against either the antibody-producing B cells or helper T cells are effective in this antibody-mediated disease.

How the autoimmune response is initiated and maintained in MG is not completely understood, but the thymus appears to play a role in this process. The thymus is abnormal in ~75% of patients with AChR antibody–positive MG; in ~65% the thymus is "hyperplastic," with the presence of active germinal centers detected histologically, although the hyperplastic thymus is not necessarily enlarged. An additional 10% of patients have thymic tumors (thymomas). Muscle-like cells within the thymus (myoid cells), which express AChRs on their surface, may serve as a source of autoantigen and trigger the autoimmune reaction within the thymus gland.

CLINICAL FEATURES

MG has a prevalence as high as 200 in 100,000. It affects individuals in all age groups, but peak incidences occur in women in their twenties and thirties and in men in their fifties and sixties. Overall, women are affected more frequently than men, in a ratio of ~3:2. The cardinal features are *weakness* and *fatigability* of muscles. The weakness increases during repeated use (fatigue) or late in the day and may improve following rest or sleep. The course of MG is often variable. Exacerbations and remissions may occur, particularly during the first few years after the onset of the disease. Unrelated infections or systemic disorders can lead to increased myasthenic weakness and may precipitate "crisis" (see below).

The distribution of muscle weakness often has a characteristic pattern. The cranial muscles, particularly the lids and extraocular muscles (EOMs), are typically involved early in the course of MG; diplopia and ptosis are common initial complaints. Facial weakness produces a "snarling" expression when the patient attempts to smile. Weakness in chewing is most noticeable after prolonged effort, as in chewing meat. Speech may have a nasal timbre caused by weakness of the palate or a dysarthric "mushy" quality due to tongue weakness. Difficulty in swallowing may occur as a result of weakness of the palate, tongue, or pharynx, giving rise to nasal regurgitation or aspiration of liquids or food. Bulbar weakness is especially prominent in MuSK antibody–positive MG. In ~85% of patients, the weakness becomes generalized, affecting the limb muscles as well. If weakness remains restricted to the EOMs for 3 years, it is likely that it will not become generalized, and these patients are said to have *ocular MG*. The limb weakness in MG is often proximal and may be asymmetric. Despite the muscle weakness, deep tendon reflexes are preserved. If ventilatory weakness becomes requires respiratory assistance, the patient is said to be in *crisis*.

DIAGNOSIS AND EVALUATION (TABLE 440-1)

The diagnosis is suspected on the basis of weakness and fatigability in the typical distribution described above, without loss of reflexes or impairment of sensation or other neurologic function. The suspected diagnosis should always be confirmed definitively before treatment is undertaken; this is essential because (1) other treatable conditions may closely resemble MG and (2) the treatment of MG may involve surgery and the prolonged use of drugs with potentially adverse side effects.

Ice-pack Test If a patient has ptosis, application of a pack of ice over a ptotic eye often results in improvement if the ptosis is due to an NMJ defect. This is hypothesized to be due to less depletion of quanta of AChR in the cold and reduced activity of AChE at the NMJ. It is a quick and easy test to do in the clinic or at the bedside of a hospitalized patient.

Autoantibodies Associated with MG As previously mentioned, anti-AChR antibodies are detectable in the serum of ~85% of all myasthenic patients but in only about 50% of patients with weakness confined to the ocular muscles. The presence of anti-AChR antibodies is virtually diagnostic of MG, but a negative test does not exclude the disease. The measured level of anti-AChR antibody does not correspond well with the severity of MG in different patients. Antibodies to MuSK are present in ~40% of AChR antibody–negative patients with generalized MG. MuSK antibodies are rarely present in AChR

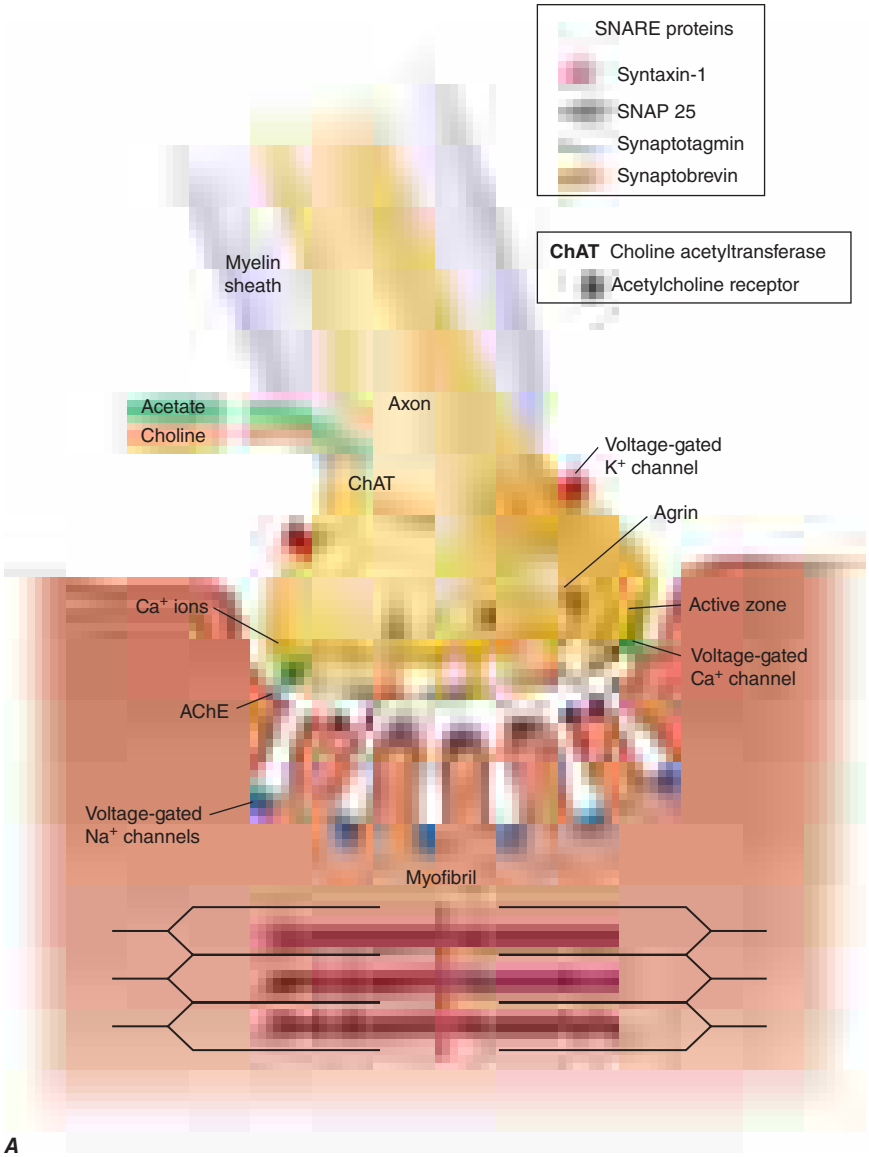
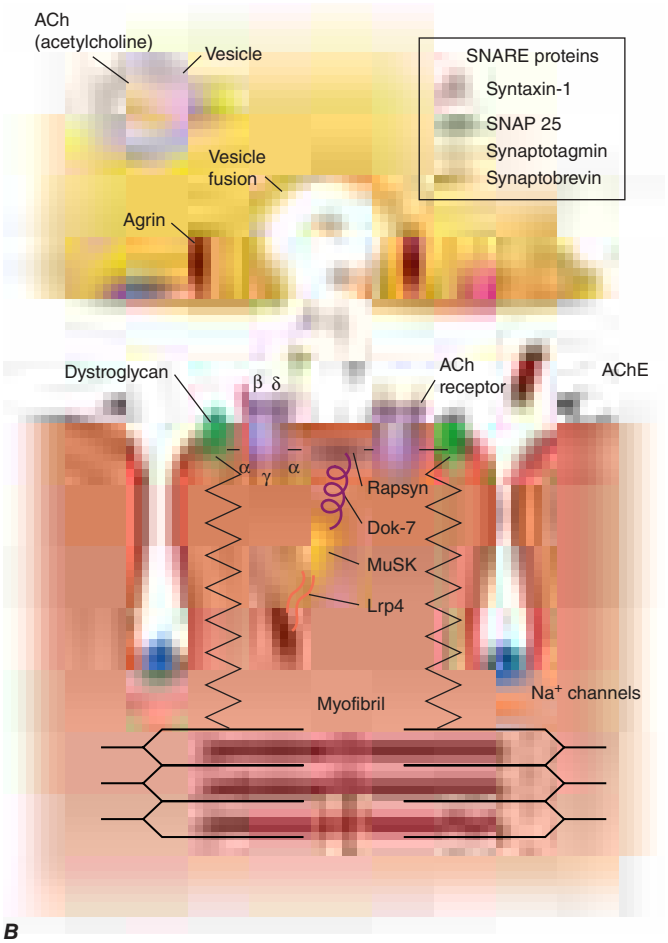
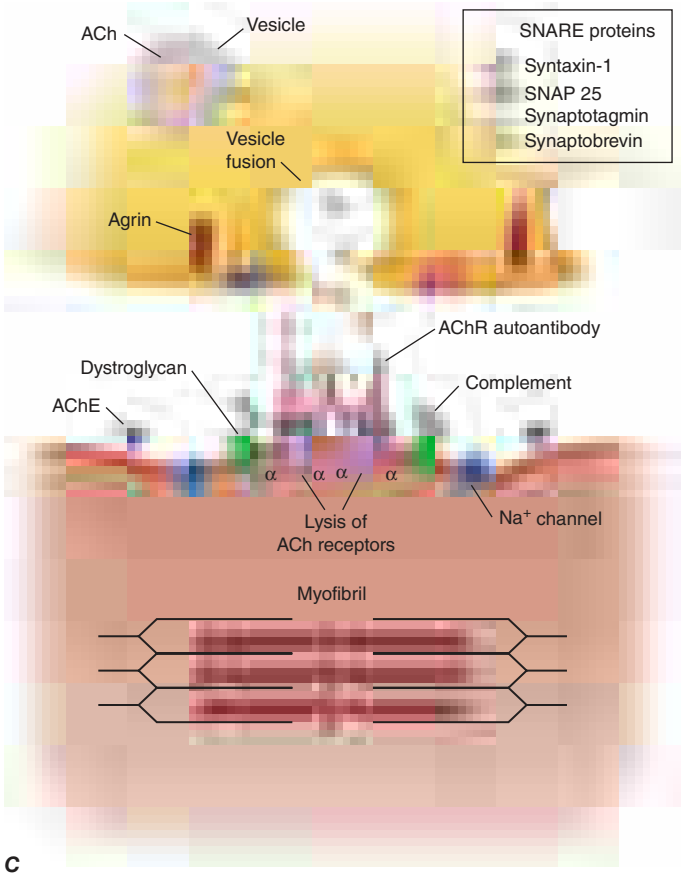


FIGURE 440-1 Illustrations of (*A*) a normal presynaptic neuromuscular junction, (*B*) a normal postsynapatic terminal, and (*C*) a myasthenic neuromuscular junction. AChE, acetylcholinesterase. See text for description of normal neuromuscular transmission. The myasthenia gravis (MG) junction demonstrates a reduced number of acetylcholine receptors (AChRs); flattened, simplified postsynaptic folds; and a widened synaptic space. See Video 440-1 also. *(From AA Amato, J Russell: Neuromuscular Disorders, 2nd ed. New York, McGraw-Hill, 2016, Figures 25-3 [p 588], 25-4 [p 589], and 25-5 [p 590]; with permission.)*



B



C

FIGURE 440-1 (*Continued*)

TABLE 440-1 Diagnosis of Myasthenia Gravis (MG)

History
Diplopia, ptosis, dysarthria, dysphagia, dyspnea
Weakness in characteristic distribution: proximal limbs, neck extensors, generalized
Fluctuation and fatigue: worse with repeated activity, improved by rest
Effects of previous treatments
Physical examination
Evaluation for ptosis at rest and following one minute of exercise, extraocular muscles and subjective diplopia, orbicularis oculi and oris strength, jaw opening and closure
Assessment of muscle strength in neck and extremities
Weakness following repeated shoulder abduction
Vital capacity measurement
Absence of other neurologic signs
Laboratory testing
Anti-AChR radioimmunoassay: ~85% positive in generalized MG; 50% in ocular MG; definite diagnosis if positive; negative result does not exclude MG; ~40% of AChR antibody–negative patients with generalized MG have anti-MuSK antibodies
Repetitive nerve stimulation: decrement of >10% at 3 Hz: highly probable
Single-fiber electromyography: blocking and jitter, with normal fiber density; confirmatory, but not specific
Edrophonium chloride (Enlon®) 2 mg + 8 mg IV; highly probable diagnosis if unequivocally positive
For ocular or cranial MG: exclude intracranial lesions by CT or MRI

Abbreviations: AChR, acetylcholine receptor; CT, computed tomography; MRI, magnetic resonance imaging; MuSK, muscle-specific tyrosine kinase.

antibody–positive patients or in patients with MG limited to ocular muscles. These antibodies may interfere with clustering of AChRs at NMJs. A small proportion of MG patients without antibodies to AChR or MuSK have antibodies to LRP4. Interestingly, antibodies against agrin have recently been found in some patients with MG. Agrin is a protein derived from motor nerves that normally binds to LRP4 and important for normal clustering of AChRs at NMJ. Additionally, anti-striated muscle antibodies directed against titin and other skeletal muscle components are found in ~30% of myasthenic without thymoma, 24% of thymoma patients without myasthenia, and 70–80% of patients with both myasthenia and thymoma. Furthermore, antibodies directed against Netrin-1 receptors and Caspr2 (contactin-associated protein-like 2) often coexist and are associated in patients with thymoma who have MG and neuromyotonia or Morvan syndrome.

Electrodiagnostic Testing Repetitive nerve stimulation may provide helpful diagnostic evidence of MG. Anti-AChE medication should be stopped 6–12 h before testing. It is best to test weak muscles or proximal muscle groups. Electrical stimulation is delivered at a rate of two or three per second to the appropriate nerves, and action potentials are recorded from the muscles. In normal individuals, the amplitude of the evoked muscle action potentials does not change by >10% at these rates of stimulation. However, in myasthenic patients, there is a rapid reduction of >10% in the amplitude of the evoked responses.

Anticholinesterase Test Drugs that inhibit the enzyme AChE allow ACh to interact repeatedly with the limited number of AChRs in MG, producing improvement in muscle strength. Edrophonium is used most commonly for diagnostic testing because of the rapid onset (30 s) and short duration (~5 min) of its effect. An objective end point must be selected to evaluate the effect of edrophonium, such as weakness of

EOMs, impairment of speech, or the length of time that the patient can maintain the arms in forward. An initial IV dose of 2 mg of edrophonium is given. If definite improvement occurs, the test is considered positive and is terminated. If there is no change, the patient is given an additional 8 mg IV. The dose is administered in two parts because some patients react to edrophonium with side effects such as nausea, diarrhea, salivation, fasciculations, and rarely with severe symptoms of syncope or bradycardia. Atropine (0.6 mg) should be drawn up in a syringe and ready for IV administration if these symptoms become troublesome. The edrophonium test is now reserved for patients with clinical findings that are suggestive of MG but who have negative antibody, electrodiagnostic testing, or ice-pack test. False-positive tests occur in occasional patients with other neurologic disorders, such as amyotrophic lateral sclerosis **(Chap. 429)**, and in placebo-reactors. False-negative or equivocal tests may also occur.

Pulmonary Function Tests (Chap. 278) Measurements of ventilatory function are valuable because of the frequency and seriousness of respiratory impairment in myasthenic patients.

Differential Diagnosis Other conditions that cause weakness of the cranial and/or somatic musculature include the nonautoimmune congenital myasthenia, drug-induced myasthenia, Lambert-Eaton myasthenic syndrome (LEMS), neurasthenia, hyperthyroidism (Graves' disease), botulism, intracranial mass lesions, oculopharyngeal dystrophy, and mitochondrial myopathy (Kearns-Sayre syndrome, progressive external ophthalmoplegia). Treatment with penicillamine (used for scleroderma or rheumatoid arthritis) and check point inhibitors for cancer may also result in autoimmune MG. Aminoglycoside antibiotics or procainamide can cause exacerbation of weakness in myasthenic patients; very large doses can cause neuromuscular weakness in normal individuals.

The *congenital myasthenic syndromes (CMS) comprise* a rare heterogeneous group of disorders of the NMJ that are not autoimmune but rather are due to genetic mutations in which virtually any component of the NMJ may be affected. Alterations in function of the presynaptic nerve terminal, in the various subunits of the AChR, AChE, or the other molecules involved in end-plate development or maintenance, have been identified in the different forms of CMS. These disorders share many of the clinical features of autoimmune MG, including weakness and fatigability of proximal or distal extremity muscles, and often involving EOMs and the eyelids similar to the distribution in autoimmune MG. CMS should be suspected when symptoms of myasthenia have begun in infancy or childhood, but they can present in early adulthood. As in acquired autoimmune MG, repetitive nerve stimulation is associated with a decremental response. Some forms (e.g., AChE deficiency, prolonged open channel syndrome) have a feature of after-discharges which are not seen in MG. An additional clue is the absence of AChR and MuSK antibodies though these are absent in ~10% of generalized MG patients (so-called double seronegative MG) antibodies.

The prevalence of CMS is estimated at ~3.8 per 100,000. The most common genetic defects occur in the ε subunit of the AChR, accounting for ~50% of CMS cases, with mutations in the genes encoding for rapsin, COLQ, DOK7, agrin, and GFPT together accounting for ~40%. In most of the recessively inherited forms of CMS, the mutations are heteroallelic; that is, different mutations affecting each of the two alleles are present. Features of the most common forms of CMS are summarized in **Table 440-2**. Molecular analysis is required for precise elucidation of the defect; this may lead to helpful treatment as well as genetic counseling. Some forms of CMS improve with AChE inhibitors, while others (e.g., slow channel syndrome, AChE deficiency, DOK-7 related CMS) actually worsen. Fluoxetine and quinidine can be useful for slow channel syndrome, and albuterol for mutations affecting AChE, DOK-7, rapsyn, and agrin. Additionally, ephedrine and 3,4 diaminopyridine (3,4 DAP) may be of benefit in some forms of CMS.

LEMS is a presynaptic disorder of the NMJ that can cause weakness similar to that of MG. The proximal muscles of the lower limbs are most commonly affected, but other muscles may be involved as well. Cranial nerve findings, including ptosis of the eyelids and diplopia, occur in up to 70% of patients and resemble features of MG. However, the two conditions are usually readily distinguished because patients with LEMS have depressed or absent reflexes and experience autonomic changes such as dry mouth and impotence. Nerve stimulation produces an initial low-amplitude compound muscle action potential and, at low rates of repetitive stimulation (2–3 Hz), a decremental responses as seen in MG; however, at high rates (20–50 Hz), or following brief exercise, incremental responses occur. LEMS is caused by autoantibodies directed against P/Q-type calcium channels at the motor nerve terminals detected in ~85% of LEMS patients. These autoantibodies impair the release of ACh from nerve terminals. In young adults, particularly women, LEMS is not associated with an underlying cancer. However, in older adults, most LEMS is associated with malignancy, most commonly small-cell lung cancer (SCLC) The tumor cells may express calcium channels that stimulate the autoimmune response. Treatment of LEMS involves plasmapheresis and immunotherapy, as for MG. 3,4-Diaminopyridine (3,4-DAP) and pyridostigmine can also help with symptoms. 3,4-DAP acts by blocking potassium channels, which results in prolonged depolarization of the motor nerve terminals and thus enhances ACh release. Pyridostigmine prolongs the action of ACh, allowing repeated interactions with AChRs.

Botulism **(Chap. 148)** is due to potent bacterial toxins produced by any of eight different strains of *Clostridium botulinum*. The toxins enzymatically cleave specific proteins essential for the release of ACh from the motor nerve terminal, thereby interfering with neuromuscular transmission. Most commonly, botulism is caused by ingestion of improperly prepared food containing toxin. Rarely, the nearly ubiquitous spores of *C. botulinum* may germinate in wounds. In infants, the spores may germinate in the gastrointestinal (GI) tract and release toxin, causing muscle weakness. Patients present with myasthenia-like bulbar weakness (e.g., diplopia, dysarthria, dysphagia) and lack sensory symptoms and signs. Weakness may generalize to the limbs and may result in respiratory failure. Reflexes are present early, but they may be diminished as the disease progresses. Mentation is normal. Autonomic findings include paralytic ileus, constipation, urinary retention, dilated or poorly reactive pupils, and dry mouth. The demonstration of toxin in serum by bioassay is definitive, but the results usually take a relatively long time to be completed and may be negative. Nerve stimulation studies reveal reduced compound muscle action potential (CMAP) amplitudes that increase following high-frequency repetitive stimulation. Treatment includes ventilatory support and aggressive inpatient supportive care (e.g., nutrition, deep vein thrombosis prophylaxis) as needed. Antitoxin should be given as early as possible to be effective and can be obtained through the Centers for Disease Control and Prevention. A preventive vaccine is available for laboratory workers or other highly exposed individuals.

Neurasthenia is the historic term for a myasthenia-like fatigue syndrome without an organic basis. These patients may present with subjective symptoms of weakness and fatigue, but muscle testing usually reveals the "give-away weakness" characteristic of nonorganic disorders; the complaint of fatigue in these patients means tiredness or apathy rather than decreasing muscle power on repeated effort. *Hyperthyroidism* is readily diagnosed or excluded by tests of thyroid function, which should be carried out routinely in patients with suspected MG. Abnormalities of thyroid function (hyper- or hypothyroidism) may increase myasthenic weakness. Diplopia resembling that in MG may occasionally be due to an intracranial mass lesion that compresses nerves to the EOMs (e.g., sphenoid ridge meningioma), but magnetic resonance imaging (MRI) of the head and orbits usually reveals the lesion.

Progressive external ophthalmoplegia is a rare condition resulting in weakness of the EOMs, which may be accompanied by weakness of the proximal muscles of the limbs and other systemic features. Most patients with this condition have mitochondrial disorders that can be detected on muscle biopsy **(Chap. 441)**.

TABLE 440-2 Congenital Myasthenic Syndromes

CMS SUBTYPE	GENE	CLINICAL FEATURES	ELECTROPHYSIOLOGICAL FEATURES	RESPONSE TO ACHE INHIBITORS	TREATMENT
Presynaptic Disorders					
CMS with paucity of ACh release	CHAT CHT	AR; early onset, respiratory failure at birth, episodic apnea, improvement with age	Decremental response to RNS	Improve	AChE inhibitors; 3,4 DAP
Synaptic Disorders					
AChE deficiency	COLQ	AR; early onset; variable severity; axial weakness with scoliosis; apnea; +/- EOM involvement, slow or absent pupillary responses	After discharges on nerve stimulation and decrement on RNS	Worsen	Albuterol; ephedrine; 3,4 DAP; avoid AChE inhibitors
Postsynaptic Disorders Involving AChR Deficiency or Kinetics					
Primary AChR deficiency	AChR subunit genes	AR; early onset; variable severity; fatigue; typical MG features	Decremental response to RNS	Improve	AChE inhibitors; 3,4 DAP
AChR kinetic disorder: slow channel syndrome	AChR subunit genes	AD; onset childhood to early adult; weak forearm extensors and neck; respiratory weakness; variable severity	After discharges on nerve stimulation and decrement on RNS	Worsen	Fluoxetine and quinidine; avoid AChE inhibitors
AChR kinetic disorder: fast channel syndrome	AChR subunit genes	AR; early onset; mild to severe; ptosis, EOM involvement; weakness and fatigue	Decremental response to RNS	Improve	AChE inhibitors; caution with 3,4 DAP
Postsynaptic Disorders Involving Abnormal Clustering/Function of AChR					
	DOK 7	AR; limb girdle weakness with ptosis but no EOM involvement	Decremental response to RNS	Variable	Albuterol; ephedrine; may worsen with AChE inhibitors
	Rapsyn	AR; early onset with hypotonia, respiratory failure, and arthrogryposis at birth to early adult onset resembling MG	Decremental response to RNS	Variable	Albuterol
	Agrin	AR; limb girdle or distal weakness, apnea	Decremental response to RNS	Variable	Albuterol; may worsen with AChE inhibitors
	MuSK	AR; congenital or childhood onset of ptosis, EOM and progressive limb girdle weakness	Decremental response to RNS	Variable	Variable response to AChE inhibitors and 3,4, DAP Positive response to albuterol
	LPR4	AR; congenital onset with hypotonia; ventilatory failure, mild ptosis, and EOM weakness	Decremental response to RNS	Worsen	Worsen with AChE inhibitors
Other Postsynaptic Disorders					
Limb-girdle CMS with tubular aggregates	GFPT1; DPAGT1; ALG2; ALG14; DPAGT1	AR; limb girdle weakness usually without ptosis or EOM weakness; onset in infancy or early adult	Decremental response to RNS	Variable	Albuterol; ephedrine; variable response to AChE inhibitors and 3,4, DAP; albuterol
Congenital muscular dystrophy with myasthenia	Plectin	AR; infantile or childhood onset of generalized weakness including ptosis and EOM; epidermolysis bullosa simplex; elevated CK	Decremental response to RNS	Variable	No response to AChE and 3,4 DAP

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; CHAT, choline acetyl transferase; CHT, sodium-dependent high-affinity choline transport 1; CMA, congenital myasthenic syndrome; COLQ, collaganic tail of endplate acetylcholinesterase; 3,4 DAP, 3,4-diaminopyridine; Dok7, downstream of tyrosine kinase 7; DPAGT1, UDP-N-acetylglucosamine-dolichyl-phosphate N-acetylglucosamine phosphotransferase; GFPT1, glutamine-fructose-6-phosphate aminotransferase 1; LRP4, lipoprotein receptor-related protein 4; MuSK, muscle specific kinase; RNS, repetitive nerve stimulation

Source: From AA Amato, J Russell: *Neuromuscular Disorders*, 2nd ed. New York: McGraw-Hill, 2016, p 627, Figure 26-2; with permission.

Search for Associated Conditions (Table 440-3) Myasthenic patients have an increased incidence of several associated disorders. Thymic abnormalities occur in ~75% of AChR antibody–positive patients, as noted above. Neoplastic change (thymoma) may produce enlargement of the thymus, which is detected by chest computed tomography (CT). A thymic shadow on CT scan may normally be present through young adulthood, but enlargement of the thymus in a patient age >40 years is highly suspicious of thymoma. Hyperthyroidism occurs in 3–8% of patients and may aggravate the myasthenic weakness. Thyroid function tests should be obtained in all patients with suspected MG. Other autoimmune disorders, most commonly systemic lupus erythematosus and rheumatoid arthritis, can coexist with MG; associations also occur with neuromyelitis optica, neuromyotonia, Morvan's syndrome (encephalitis, insomnia, confusion, hallucinations, autonomic dysfunction, and neuromyotonia), rippling muscle disease, granulomatous myositis/myocarditis, and chronic inflammatory demyelinating polyneuropathy.

An infection of any kind can exacerbate typical MG, and should be sought carefully in patients with relapses. Because of the side effects of glucocorticoids and other immunotherapies used in the treatment of MG, a thorough medical investigation should be undertaken, searching specifically for evidence of chronic or latent infection (such as tuberculosis or hepatitis), hypertension, diabetes, renal disease, and glaucoma.

TREATMENT

Myasthenia Gravis

The prognosis has improved strikingly as a result of advances in treatment. Nearly all myasthenic patients can be returned to full

TABLE 440-3 Disorders Associated with Myasthenia Gravis and Recommended Laboratory Tests

Associated disorders
Disorders of the thymus: thymoma, hyperplasia
Other autoimmune neurological disorders: chronic inflammatory demyelinating polyneuropathy, neuromyelitis optica
Other autoimmune disorders: Hashimoto's thyroiditis, Graves' disease, rheumatoid arthritis, systemic lupus erythematosus, skin disorders, family history of autoimmune disorder
Disorders or circumstances that may exacerbate myasthenia gravis: hyperthyroidism or hypothyroidism, occult infection, medical treatment for other conditions (see Table 440-4)
Disorders that may interfere with therapy: tuberculosis, diabetes, peptic ulcer, gastrointestinal bleeding, renal disease, hypertension, asthma, osteoporosis, obesity
Recommended laboratory tests or procedures
CT or MRI of chest
Tests for antinuclear antibodies, rheumatoid factor
Thyroid function tests
Testing for tuberculosis
Fasting blood glucose, hemoglobin A1c
Pulmonary function tests
Bone densitometry

Abbreviations: CT, computed tomography; MRI, magnetic resonance imaging.

productive lives with proper therapy. The most useful treatments for MG include anticholinesterase medications, immunosuppressive agents, thymectomy, plasmapheresis, and intravenous immunoglobulin (IVIg) **(Fig. 440-2)**.

ANTICHOLINESTERASE MEDICATIONS

Anticholinesterase medication produces at least partial improvement in most myasthenic patients, although improvement is complete in only a few. Patients with anti-MuSK MG generally obtain less benefit from anticholinesterase agents than those with AChR antibodies and may actually worsen. Pyridostigmine is the most widely used anticholinesterase drug and is initiated at a dosage of 30–60 mg three to four times daily. The beneficial action of oral pyridostigmine begins within 15–30 min and lasts for 3–4 h, but individual responses vary. The frequency and amount of the dose should be tailored to the patient's individual requirements throughout the day. For example, patients with weakness in chewing and swallowing may benefit by taking the medication before meals so that peak strength coincides with mealtimes. Long-acting pyridostigmine may occasionally be useful to get the patient through the night but should not be used for daytime medication because of variable absorption. The maximum useful dose of pyridostigmine rarely exceeds 300 mg daily. Overdosage with anticholinesterase medication may cause increased weakness and other side effects. In some patients, muscarinic side effects of the anticholinesterase medication (diarrhea, abdominal cramps, salivation, nausea) may limit the dose tolerated. Atropine/diphenoxylate or loperamide is useful for the treatment of GI symptoms.

THYMECTOMY

Two separate issues should be distinguished: (1) surgical removal of thymoma, and (2) thymectomy as a treatment for MG. Surgical removal of a thymoma is necessary because of the possibility of local tumor spread, although most thymomas are histologically benign. Until recently there was a debate regarding the role of thymectomy in non-thymotous MG, but a recent large international trial of extended transternal thymectomy in non-thymomatous AChR antibody positive, generalized MG demonstrated that participants who underwent thymectomy had improved strength and function, required less prednisone and additions of second line agents (e.g., azathioprine), and fewer hospitalizations for exacerbations. Whether or not less invasive thymectomy may be beneficial is unknown. Also, patients with ocular myasthenia, MuSK-positive, and seronegative MG were excluded from the study; retrospective

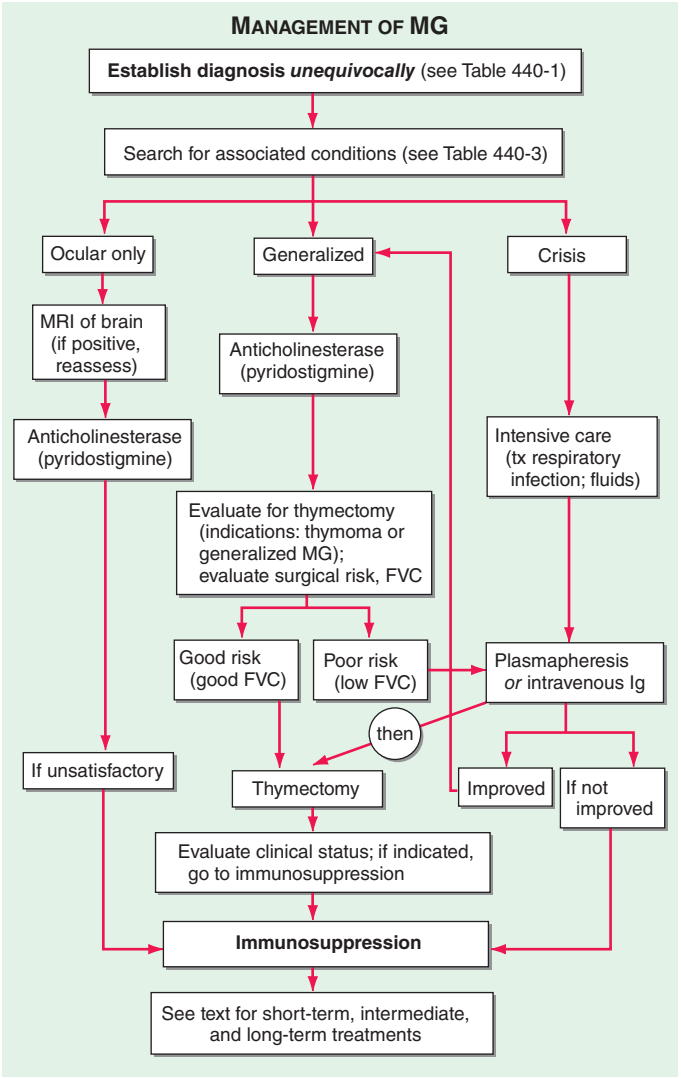


FIGURE 440-2 Algorithm for the management of myasthenia gravis. FVC, forced vital capacity; MRI, magnetic resonance imaging.

and anecdotal evidence suggest that these patients may not benefit from thymectomy. Thymectomy should never be carried out as an emergency procedure, but only when the patient is adequately prepared. If necessary, treatment with IVIg or plasmapheresis may be used before surgery to maximize strength in weak patients.

IMMUNOSUPPRESSION

Immunosuppression using one or more of the available agents is effective in nearly all patients with MG. The choice of drugs or other immunomodulatory treatments should be guided by the relative benefits and risks for the individual patient and the urgency of treatment. It is helpful to develop a treatment plan based on short-term, intermediate-term, and long-term objectives. For example, if immediate improvement is essential either because of the severity of weakness or because of the patient's need to return to activity as soon as possible, IVIg should be administered or plasmapheresis should be undertaken. For the intermediate term, glucocorticoids and cyclosporine or tacrolimus generally produce clinical improvement within a period of 1–3 months. The beneficial effects of azathioprine and mycophenolate mofetil usually begin after many months (as long as a year), but these drugs have advantages for the long-term treatment of patients with MG. There is a growing body of evidence that rituximab is effective in many MG patients, especially those with MuSK antibody.

Glucocorticoid Therapy Glucocorticoids, when used properly, produce improvement in myasthenic weakness in the great majority of patients. To minimize adverse side effects, prednisone should

be given in a single dose rather than in divided doses throughout the day. In patients with only mild or moderate weakness, the initial dose should be relatively low (15–25 mg/d) to avoid the early weakening that occurs in perhaps 10–15% of patients treated initially with a high-dose regimen. The dose is increased stepwise, as tolerated by the patient (usually by 5 mg/d at 2–3 day intervals), until there is marked clinical improvement or a dose of 50–60 mg/d is reached. In patients with more severe weakness and those already in the hospital, starting at a high dose is reasonable. Patients are maintained on the dose that seems to control their symptoms for about a month, and then the dosage is slowly tapered (no faster than 10 mg a month until on 20 mg daily and then by 2.5–5 mg a month) to determine the minimum effective dose, and close monitoring is required. Some patients are able to be managed without the addition of other immunotherapies. Patients on long-term glucocorticoid therapy must be followed carefully to prevent or treat adverse side effects. The most common errors in glucocorticoid treatment of myasthenic patients include (1) insufficient persistence—improvement may be delayed and gradual; (2) tapering the dosage too early, too rapidly, or excessively; and (3) lack of attention to prevention and treatment of side effects.

The management of patients treated with glucocorticoids is discussed in Chap. 379.

Other Immunotherapies Mycophenolate mofetil, azathioprine, cyclosporine, tacrolimus, rituximab, and occasionally cyclophosphamide are effective in many patients, either alone or in various combinations.

Mycophenolate mofetil is widely used because of its presumed effectiveness and relative lack of side effects. A dose of 1–1.5 g bid is recommended. Its mechanism of action involves inhibition of purine synthesis by the de novo pathway. Since lymphocytes have only the de novo pathway, but lack the alternative salvage pathway that is present in all other cells, mycophenolate inhibits proliferation of lymphocytes but not proliferation of other cells. It does not kill or eliminate preexisting autoreactive lymphocytes, and therefore clinical improvement may be delayed for many months to a year, until the preexisting autoreactive lymphocytes die spontaneously. The advantage of mycophenolate lies in its relative lack of adverse side effects, with only occasional production of GI symptoms, rare development of leukopenia, and very small risks of malignancy or progressive multifocal leukoencephalopathy inherent in nearly all immunosuppressive treatments. Although two published studies did not show positive outcomes, most experts attribute the negative results to flaws in the trial designs, and mycophenolate is widely used for long-term treatment of myasthenic patients.

Azathioprine has long been used for MG and a randomized, clinical trial demonstrated that it was effective in reducing the dosage of prednisone necessary to control symptoms. However, the beneficial effect of azathioprine can take a year or more to become evident. Approximately 10–15% of patients are unable to tolerate azathioprine because of idiosyncratic reactions consisting of flulike symptoms of fever and malaise, bone marrow suppression, or abnormalities of liver function. An initial dose of 50 mg/d is given for about a week to test for these side effects. If this dose is tolerated, it is increased gradually to about 2–3 mg/kg of total body weight, or until the white blood count falls to 3000–4000/μL. Allopurinol should never be used in combination with azathioprine because the two drugs share a common degradation pathway; the result may be severe bone marrow suppression due to increased effects of the azathioprine.

The calcineurin inhibitors cyclosporine and tacrolimus seem to be effective in MG and appear to work more rapidly than azathioprine and mycophenolate. However, they are associated with more frequent severe side effects including hypertension and nephrotoxicity. The usual dose of cyclosporine is 4–5 mg/kg per d, and the average dose of tacrolimus is 0.07–0.1 mg/kg per d, given in two equally divided doses. "Trough" blood levels are measured 12 h after the evening dose. The therapeutic range for the trough level of cyclosporine is 150–200 ng/L, and for tacrolimus, it is 5–15 ng/L.

Rituximab (Rituxan) is a monoclonal antibody that binds to the CD20 molecule on B lymphocytes. It has been widely used for the treatment of B cell lymphomas and has also proven successful in the treatment of several autoimmune diseases including rheumatoid arthritis, pemphigus, and some IgM-related neuropathies. There is an increasing literature on the benefit of rituximab in MG. It appears particularly effective in MuSK antibody–positive MG, although some patients with AChR antibody MG also respond. A large NIH sponsored trial is underway in AChR-positive MG. The usual dose is 1 g IV on two occasions 2 weeks apart. Periodically, a repeat course needs to be administered; some MuSK patients go 2–3 years between infusions.

Eculizumab is a monoclonal antibody that binds membrane attack complex and was beneficial in a small pilot study of MG patients. The results of a large phase 3 clinical trial were recently published and largely positive leading to recent FDA approval. The drug is administered intravenously every 2 weeks.

For the rare refractory MG patient, a course of high-dose cyclophosphamide may induce long-lasting benefit by "rebooting" the immune system. At high doses, cyclophosphamide eliminates mature lymphocytes but spares hematopoietic precursors (stem cells), because they express the enzyme aldehyde dehydrogenase, which hydrolyzes cyclophosphamide. This procedure is reserved for refractory patients and should be administered only in a facility fully familiar with this approach. Maintenance immunotherapy after rebooting is usually required to sustain the beneficial effect.

PLASMAPHERESIS AND INTRAVENOUS IMMUNOGLOBULIN

Plasmapheresis has been used therapeutically in MG. Plasma, which contains the pathogenic antibodies, is mechanically separated from the blood cells, which are returned to the patient. A course of five exchanges (3–4 L per exchange) is generally administered over a 10- to 14-day period. Plasmapheresis produces a short-term reduction in anti-AChR antibodies, with clinical improvement in many patients. It is useful as a temporary expedient in seriously affected patients or to improve the patient's condition prior to surgery (e.g., thymectomy).

The indications for the use of IVIg are the same as those for plasma exchange: to produce rapid improvement to help the patient through a difficult period of myasthenic weakness or prior to surgery. This treatment has the advantages of not requiring special equipment or large-bore venous access. The usual dose is 2 g/kg, which is typically administered >2–5 days. Improvement occurs in ~70% of patients, beginning during treatment or within a week, and continuing for weeks to months. The mechanism of action of IVIg is not known; the treatment has no consistent effect on the measurable amount of circulating AChR antibody. Adverse reactions are generally not serious but may include headache, fluid overload, and rarely aseptic meningitis or renal failure. IVIg or plasma exchange is occasionally used in combination with other immunosuppressive therapy for maintenance treatment of difficult MG.

MANAGEMENT OF MYASTHENIC CRISIS

Myasthenic crisis is defined as an exacerbation of weakness sufficient to endanger life; it usually includes ventilatory failure caused by diaphragmatic and intercostal muscle weakness. Treatment should be carried out in intensive care units staffed with teams experienced in the management of MG. The possibility that deterioration could be due to excessive anticholinesterase medication ("cholinergic crisis") is best excluded by temporarily stopping anticholinesterase drugs. The most common cause of crisis is intercurrent infection. This should be treated immediately because the mechanical and immunologic defenses of the patient can be assumed to be compromised. The myasthenic patient with fever and early infection should be treated like other immunocompromised patients. Early and effective antibiotic therapy, ventilatory assistance, and pulmonary physiotherapy are essentials of the treatment program. As discussed above, plasmapheresis or IVIg is frequently helpful in hastening recovery.

TABLE 440-4 Drugs with Interactions in Myasthenia Gravis (MG)

Drugs That May Exacerbate MG
Antibiotics
Aminoglycosides: e.g., streptomycin, tobramycin, kanamycin
Quinolones: e.g., ciprofloxacin, levofloxacin, ofloxacin, gatifloxacin
Macrolides: e.g., erythromycin, azithromycin
Nondepolarizing muscle relaxants for surgery
D-Tubocurarine (curare), pancuronium, vecuronium, atracurium
Beta-blocking agents
Propranolol, atenolol, metoprolol
Local anesthetics and related agents
Procaine, Xylocaine in large amounts
Procainamide (for arrhythmias)
Botulinum toxin
Botox exacerbates weakness
Quinine derivatives
Quinine, quinidine, chloroquine, mefloquine (Lariam)
Magnesium
Decreases acetylcholine release
Penicillamine
May cause MG
Check point inhibitors
May cause MG and other autoimmune neuromuscular disorders (e.g., myositis, inflammatory neuropathy)
Drugs with Important Interactions in MG
Cyclosporine and Tacrolimus
Broad range of drug interactions, which may raise or lower levels.
Azathioprine
Avoid allopurinol—combination may result in myelosuppression.

DRUGS TO AVOID IN MYASTHENIC PATIENTS

Many drugs can potentially exacerbate weakness in patients with MG (Table 440-4). As a rule, the listed drugs should be avoided whenever possible.

PATIENT ASSESSMENT

To evaluate the effectiveness of treatment as well as drug-induced side effects, it is important to assess the patient's clinical status systematically at baseline and on repeated interval examinations. Spirometry with determination of forced vital capacity and mean inspiratory and expiratory pressures are important to follow.

PROGNOSIS

Approximately 20% of patients with MG can be tapered off all immunotherapies and achieve a sustained remission. There does not appear to be a correlation with maximal disease severity and chance for remission. Thymectomy may increase the likelihood of achieving remission in anti-AChR MG, but the large randomized trial was too short in duration to examine this endpoint; rather, the results revealed only that thymectomy was efficacious and led to less use of glucocorticoids and second line agents. Mortality from MG diminished greatly during the twentieth century, changing from a "grave" illness with mortality of nearly 70% a century ago, to 2–30% by the 1950s, with contemporary estimates in the 1–2% range. Anti-MuSK patients typically do not experience myasthenic crises, but are generally more difficult to treat than anti-AChR MG. However, as noted above, recent series suggest that rituximab is effective in this subgroup. Non-paraneoplastic LEMS is usually responsive to immunotherapy and symptomatic treatment with pyridostigmine and 3,4 DAP. In older adults, LEMS is most often paraneoplastic, and screening for an underlying tumor is indicated. Recent studies suggest that survival in patients with LEMS has improved, for uncertain reasons and likely not due to earlier diagnosis and treatment of the tumor. There is wide variability in age of onset, severity, and prognosis of the many types of CMS.

GLOBAL ISSUES

The incidence of MG and its subtpes vary in different populations, for example occurring in ~2–10/10^6 individuals in the United States and the Netherlands, and up to 20/10^6 in Spain. Estimates of prevalence in different parts of the world range widely from 2–200/10^6. The age of onset may also be influenced by geographical and/or ethnic differences. Juvenile onset MG is uncommon in Western populations but may represent more than half of cases in Asians. MuSK MG appearts to be more common in the Mediterranean area of Europe than in northern Europe and is also more common in the northern regions of East Asia than in the southern regions.

FURTHER READING

Amato AA, Russell JA: *Neuromuscular Disorders*, 2nd ed. New York: McGraw-Hill; 2016, pp. 581–655.

Beeson D: Congenital myasthenic syndromes: Recent advances. Curr Opin Neurol 29:565, 2016.

Garg N et al: Late presentations of congenital myasthenic syndromes: How many do we miss? Muscle Nerve 54:721, 2016.

Gilhus NE: Myasthenia Gravis. N Engl J Med 375:2570, 2016.

Howard JF Jr et al: Safety and efficacy of eculizumab in anti-acetylcholine receptor antibody-positive refractory generalised myasthenia gravis (REGAIN): A phase 3, randomised, double-blind, placebo-controlled, multicentre study. Lancet Neurol 16:976, 2017.

Nicolle MW: Myasthenia gravis and Lambert-Eaton myasthenic syndrome. Continuum (Minneap Minn) 22(6, Muscle and neuromuscular junction disorders):1978, 2016.

Tandan R et al: Rituximab treatment of myasthenia gravis: A systematic review. Muscle Nerve 56:185, 2017.

Titulaer MJ et al: Lambert-Eaton myasthenic syndrome: From clinical characteristics to therapeutic strategies. Lancet Neurol 10:1098, 2011.

Wolfe GI et al: Randomized trial of thymectomy in myasthenia gravis. N Engl J Med 375:511, 2016.

VIDEO 440-1 Myasthenia gravis and other diseases of the neuromuscular junction.

441 Muscular Dystrophies and Other Muscle Diseases

Anthony A. Amato, Robert H. Brown, Jr.

Myopathies are disorders with structural changes or functional impairment of muscle and can be differentiated from other diseases of the motor unit (e.g., lower motor neuron or neuromuscular junction pathologies) by characteristic clinical and laboratory findings. **Myasthenia gravis and related disorders are discussed in Chap. 440; inflammatory myopathies are discussed in Chap. 358.**

CLINICAL FEATURES

The most important aspect of assessing individuals with neuromuscular disorders is taking a thorough history of the patient's symptoms, disease progression, past medical and family history as well as performing a detailed neurologic examination. Based on this and additional laboratory workup (e.g., serum creatine kinase [CK], electromyography [EMG]) one can usually localize the site of the lesion to muscle (as opposed to motor neurons, peripheral nerves, or neuromuscular junction) and the pattern of muscle involvement. It is this pattern of muscle involvement which is most useful in narrowing the differential diagnosis (Table 441-1). Most myopathies present with proximal, symmetric limb weakness with preserved reflexes and sensation. However, asymmetric and predominantly distal weakness can be seen in some myopathies. An associated sensory loss suggests a peripheral

TABLE 441-1 Myopathies by Pattern of Weakness/Muscle Involvement

Proximal (Limb-Girdle) Weakness
Most dystrophies (e.g., dystrophinophies, limb-girdle, myotonic dystrophy type 2, rare FSHD)
Congenital myopathies (e.g., central core, multiminicore, centronuclear, nemaline rod)
Metabolic myopathies (e.g., glycogen and lipid storage diseases)
Mitochondrial myopathies
Inflammatory myopathies (DM, PM, IMNM)
Toxic myopathies (see Table 441-6)
Endocrine myopathies
Neuromuscular junction disorders (myasthenia gravis, LEMS, congenital myasthenia, botulism, **see Chap. 440**)
Distal Weakness
Distal muscular dystrophies/myofibrillar myopathy (see Table 441-5)
Congenital myopathies (e.g., late onset centronuclear and nemaline rod myopathies)
Metabolic
Glygogen storage disease (e.g., brancher and debrancher deficiency, rarely McArdle disease)
Lipid storage disease (e.g., neutral lipid storage myopathy, multiacyldehydrogenase deficiency)
NMJ disorders (e.g., rare myasthenia gravis and congenital myasthenia)
Proximal Arm/Distal Leg Weakness (Scapuloperoneal or Humeroperonal) Weakness
Facioscapulohumeral muscular dystrophy (FSHD)
Scapuloperoneal myopathy and neuropathy
Myofibrillar myopathies
Emery-Dreifuss muscular dystrophy (EDMD)
Bethlem myopathy
Distal Arm/Proximal Leg Weakness
Inclusion body myositis (usually wrist and finger flexors in arms, hip flexors and knee extensors in legs, and asymmetric)
Myotonic dystrophy (uncommon presentation)
Axial Muscle Weakness
Inflammatory (cervico-brachial myositis)
sIBM and hIBM
Myotonic dystrophy 2
Isolated neck extensor myopathy/bent spine syndrome
FSHD
Late onset central core (RYR1 mutations)
Sporadic lat onset nemaline rod with or without a monoclonal gammopathy
Metabolic (late onset Pompe, McArdle disease, lipid storage, mitochondrial)
Hyperparathyroidism/ostemalacia/vitamin D deficiency
Myasthenia gravis
Eye Muscle Weakness (Ptosis/Ophthalmoparesis)
Ptosis without ophthalmoparesis
Myotonic dystrophy
Congenital myopathies
Neuromuscular junction disorders
Ptosis with ophthalmoparesis
Oculopharyngeal dystrophy
Mitochondrial myopathy
hIBM type 3
Neuromuscular junction disorders
Episodic Weakness or Myoglobinuria
Related to exercise
Glycogenoses (e.g., McArdle disease, etc.)
Lipid disorders (e.g., CPT2 deficiency)
Mitochondrial myopathies (e.g., cytrochrome B deficiency)
Not Related to Exercise
Malignant hyperthermia
Drugs/toxins (e.g., statins)
Prolonged/intensive eccentric exercise
Inflammatory (e.g., PM/DM—rare, viral/bacterial infections)
Delayed or unrelated to exercise
Periodic paralysis (e.g., hereditary hyper- or hyokalemic, thyrotoxic, associated renal tubular acidosis, acquired electrolyte imbalance)
NMJ disorders
Muscle Stiffness/Decreased Ability To Relax
Myotonic dystrophy 1 and 2
Myotonia congenita
Paramyotonia congenita
Hyperkalemic periodic paralysis with myotonia
Potassium aggravated myotonia
Schwartz-Jampel syndrome
Other: rippling muscle disease (acquired and hereditary), acquired neuromyotonia (Isaacs syndrome), stiff-person syndrome, Brody disease

neuropathy or a central nervous system (CNS) abnormality (e.g., myelopathy) rather than a myopathy. On occasion, disorders affecting the motor nerve cell bodies in the spinal cord (anterior horn cell disease), the neuromuscular junction, or peripheral nerves can mimic findings of myopathy.

Muscle Weakness Symptoms of muscle weakness can be either intermittent or persistent. Disorders causing *intermittent weakness* (Table 441-1 and **Fig. 441-1**) include myasthenia gravis, periodic paralyses (hypokalemic or hyperkalemic), and metabolic energy deficiencies of glycolysis (especially myophosphorylase deficiency), fatty acid utilization (carnitine palmitoyltransferase [CPT] deficiency), and some mitochondrial myopathies. The states of energy deficiency cause activity-related muscle breakdown accompanied by myoglobinuria.

Most muscle disorders cause *persistent weakness* (Table 441-1 and **Fig. 441-2**). In the majority of these, including most types of muscular dystrophy, polymyositis, and dermatomyositis, the proximal muscles are weaker than the distal and are symmetrically affected, and the facial muscles are spared, a pattern referred to as *limb-girdle weakness*. The differential diagnosis is more restricted for other patterns of weakness. Facial weakness (difficulty with eye closure and impaired smile) and scapular winging **(Fig. 441-3)** are characteristic of facioscapulohumeral dystrophy (FSHD). Facial and distal limb weakness associated with hand grip myotonia is virtually diagnostic of myotonic dystrophy type 1. When other cranial nerve muscles are weak, causing ptosis or extraocular muscle weakness, the most important disorders to consider include neuromuscular junction disorders, oculopharyngeal muscular dystrophy, mitochondrial myopathies, or some of the congenital myopathies (Table 441-1). A pathognomonic pattern characteristic of inclusion body myositis is atrophy and weakness of the flexor forearm (e.g., wrist and finger flexors) and quadriceps muscles that is often asymmetric. Less frequently, but important diagnostically, is the presence of a dropped head syndrome indicative of selective neck extensor muscle weakness. The most important neuromuscular diseases associated with this pattern of weakness include myasthenia gravis, amyotrophic lateral sclerosis, late-onset core or nemaline rod myopathy, hyperparathyroidism, focal myositis, and some forms of inclusion body myopathy. A final pattern, recognized because of preferential distal extremity weakness, is seen in the distal myopathies.

It is important to examine functional capabilities to help disclose certain patterns of weakness (Table 441-1 and **Table 441-2**). The Gower sign **(Fig. 441-4)** is particularly useful. Observing the gait of an individual may disclose a lordotic posture caused by combined trunk and hip weakness, frequently exaggerated by toe walking **(Fig. 441-5)**. A waddling gait is caused by the inability of weak hip muscles to prevent hip drop or hip dip. Hyperextension of the knee (genu recurvatum or back-kneeing) is characteristic of quadriceps muscle weakness; and a steppage gait, due to footdrop, accompanies distal weakness.

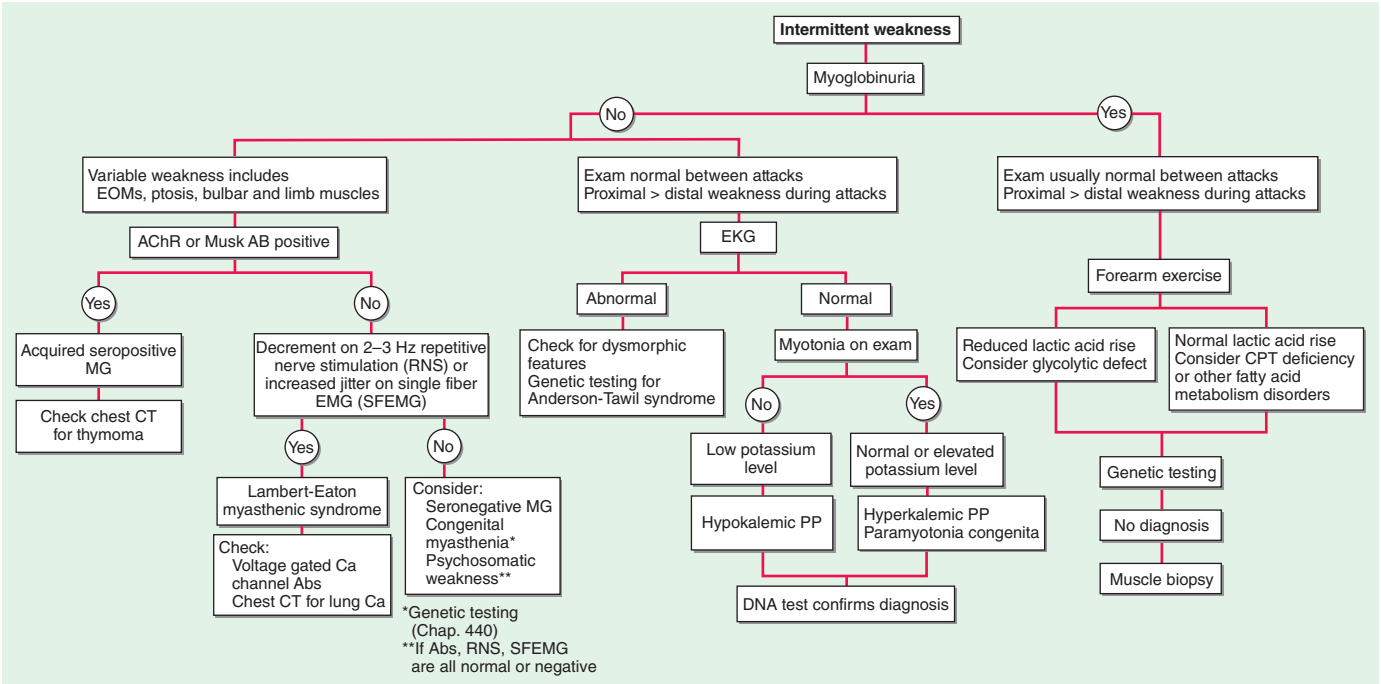


FIGURE 441-1 Diagnostic evaluation of intermittent weakness. AChR AB, acetylcholine receptor antibody; CPT, carnitine palmitoyltransferase; EOMs, extraocular muscles; MG, myasthenia gravis; PP, periodic paralysis.

Any disorder causing muscle weakness may be accompanied by *fatigue*, referring to an inability to maintain or sustain a force (pathologic fatigability). This condition must be differentiated from asthenia, a type of fatigue caused by excess tiredness or lack of energy. Associated symptoms may help differentiate asthenia and pathologic fatigability. Asthenia is often accompanied by a tendency to avoid physical activities, complaints of daytime sleepiness, necessity for frequent naps, and difficulty concentrating on activities such as reading. There may be feelings of overwhelming stress and depression. In contrast, pathologic fatigability occurs in disorders of neuromuscular transmission and in disorders altering energy production, including defects in glycolysis, lipid metabolism, or mitochondrial energy production. Pathologic fatigability also occurs in chronic myopathies because of difficulty accomplishing a task with less muscle. Pathologic fatigability is accompanied by abnormal clinical or laboratory findings. Fatigue without those supportive features almost never indicates a primary muscle disease.

Muscle Pain (Myalgias), Cramps, and Stiffness Some myopathies can be associated with muscle pain, cramps, contractures, stiff or rigid muscles, or inability to relax the muscles (e.g., myotonia) (Table 441-1). *Muscle cramps* are abrupt in onset, short in duration, triggered by voluntary muscle contraction, and may cause abnormal posturing of the joint. Muscle cramps often occur in neurogenic disorders, especially motor neuron disease **(Chap. 429)**, radiculopathies and polyneuropathies **(Chap. 438)**, but are not a feature of most primary muscle diseases.

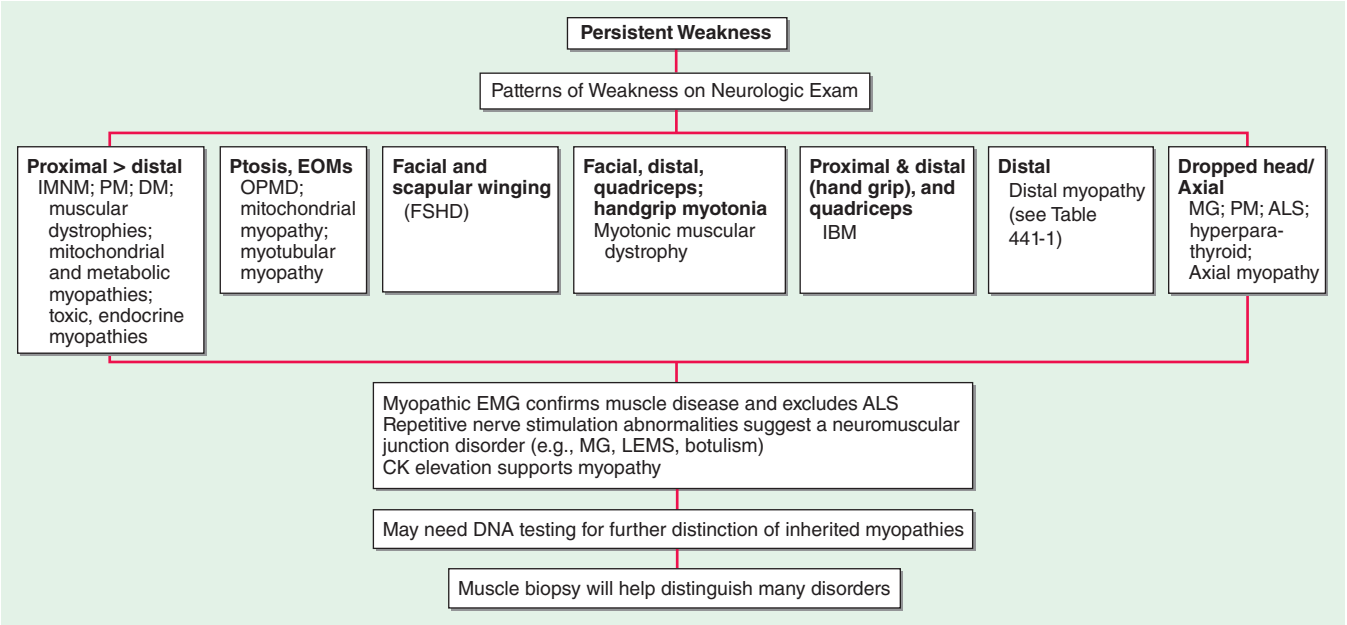


FIGURE 441-2 Diagnostic evaluation of persistent weakness. Examination reveals one of seven patterns of weakness. The pattern of weakness in combination with the laboratory evaluation leads to a diagnosis. ALS, amyotrophic lateral sclerosis; CK, creatine kinase; DM, dermatomyositis; EMG, electromyography; EOMs, extraocular muscles; FSHD, facioscapulohumeral dystrophy; IBM, inclusion body myositis; IMNM, immune-mediated necrotizing myopathy; MG, myasthenia gravis; OPMD, oculopharyngeal muscular dystrophy; PM, polymyositis.

FIGURE 441-3 Facioscapulohumeral dystrophy with prominent scapular winging.

A *muscle contracture* is different from a muscle cramp. In both conditions, the muscle becomes hard, but a contracture is associated with energy failure in glycolytic disorders. The muscle is unable to relax after an active muscle contraction. The EMG shows electrical silence. Confusion is created because contracture also refers to a muscle that cannot be passively stretched to its proper length (fixed contracture) because of fibrosis. In some muscle disorders, especially in Emery-Dreifuss muscular dystrophy and Bethlem myopathy, fixed contractures occur early and represent distinctive features of the disease.

Myotonia is a condition of prolonged muscle contraction followed by slow muscle relaxation. It always follows muscle activation (action myotonia), usually voluntary, but may be elicited by mechanical stimulation (percussion myotonia) of the muscle. Myotonia typically causes difficulty in releasing objects after a firm grasp. In myotonic muscular dystrophy type 1 (DM1), distal weakness usually accompanies myotonia, whereas in DM2, proximal muscles are more affected. Myotonia also occurs with *myotonia congenita* (a chloride channel disorder), but in this condition muscle weakness is not prominent. Myotonia may also be seen in individuals with sodium channel mutations (*hyperkalemic periodic paralysis* or *potassium-sensitive myotonia*). Another sodium channelopathy, *paramyotonia congenita* (PC), also is associated with muscle stiffness. In contrast to other disorders associated with myotonia in which the myotonia is eased by repetitive activity, PC is named for a paradoxical phenomenon whereby the myotonia worsens with repetitive activity. Potassium-aggravated myotonia is an allelic disorder in which myotonia is brought on by consumption of too much potassium-containing foods.

TABLE 441-2 Observations on Examination That Disclose Muscle Weakness

FUNCTIONAL IMPAIRMENT	MUSCLE WEAKNESS
Inability to forcibly close eyes	Upper facial muscles
Impaired pucker	Lower facial muscles
Inability to raise head from prone position	Neck extensor muscles
Inability to raise head from supine position	Neck flexor muscles
Inability to raise arms above head	Proximal arm muscles (may be only scapular stabilizing muscles)
Inability to walk without hyperextending knee (back-kneeing or genu recurvatum)	Knee extensor muscles
Inability to walk with heels touching the floor (toe walking)	Shortening of the Achilles tendon
Inability to lift foot while walking (steppage gait or footdrop)	Anterior compartment of leg
Inability to walk without a waddling gait	Hip muscles
Inability to get up from the floor without climbing up the extremities (Gowers' sign)	Hip, thigh, and trunk muscles
Inability to get up from a chair without using arms	Hip muscles

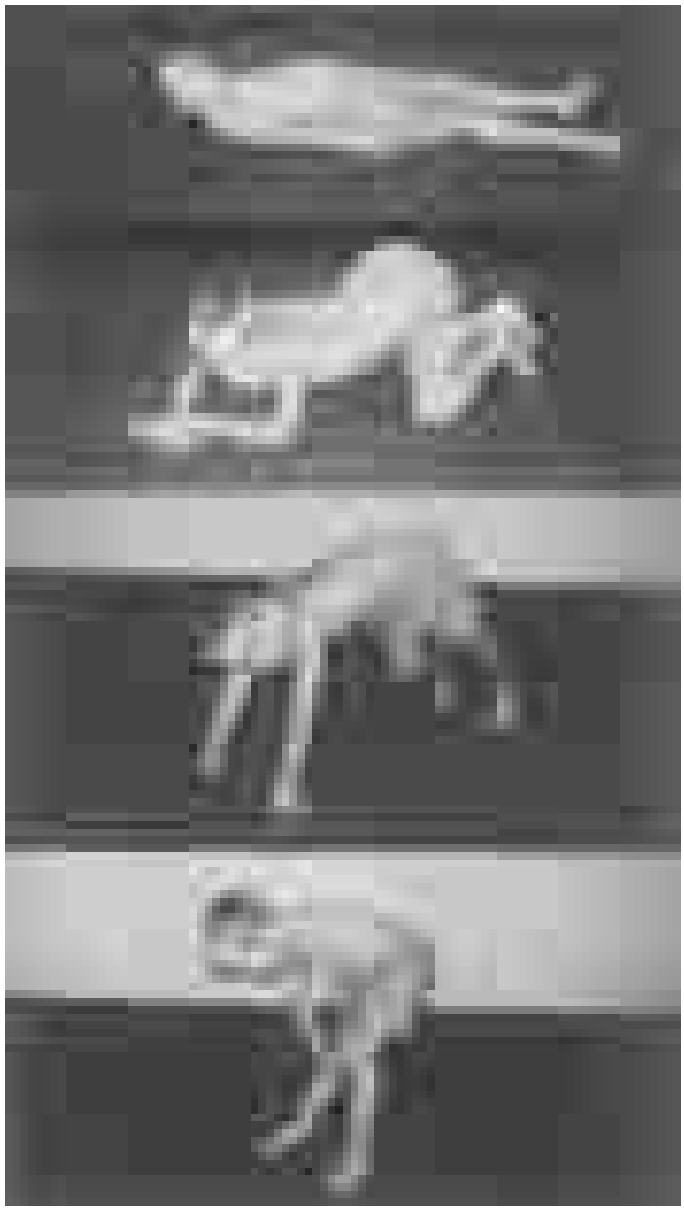


FIGURE 441-4 Gowers sign showing a patient using his arms to climb up the legs in attempting to get up from the floor.

Muscle stiffness can refer to different phenomena. Some patients with inflammation of joints and periarticular surfaces feel stiff. This condition is different from the disorders of hyperexcitable motor nerves causing stiff or rigid muscles. In *stiff-person syndrome*, spontaneous discharges of the motor neurons of the spinal cord cause involuntary muscle contractions mainly involving the axial (trunk) and proximal lower extremity muscles. The gait becomes stiff and labored, with hyperlordosis of the lumbar spine. Superimposed episodic muscle spasms are precipitated by sudden movements, unexpected noises, and emotional upset. The muscles relax during sleep. Serum antibodies against glutamic acid decarboxylase are present in approximately two-thirds of cases. In *acquired neuromyotonia* (*Isaacs syndrome*), there is hyperexcitability of the peripheral nerves manifesting as continuous muscle fiber activity in the form of widespread fasciculations and myokymia with impaired muscle relaxation. Muscles of the leg are stiff, and the constant contractions of the muscle cause increased sweating of the extremities. This peripheral nerve hyperexcitability is mediated by antibodies that target voltage-gated potassium channels.

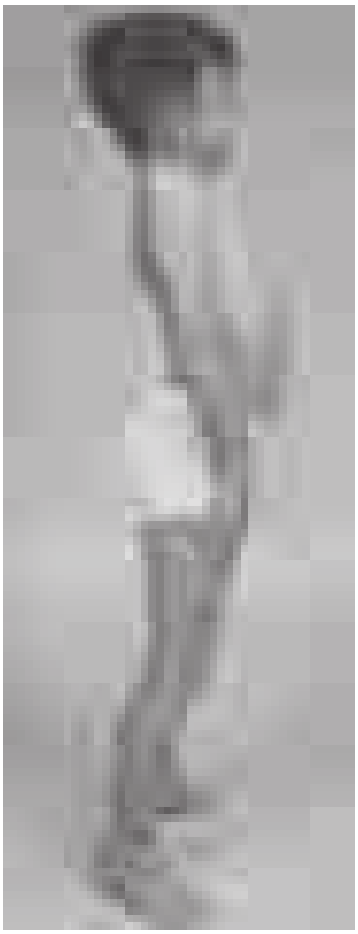


FIGURE 441-5 Lordotic posture, exaggerated by standing on toes, associated with trunk and hip weakness.

There are two painful muscle conditions of particular importance, neither of which is associated with muscle weakness. *Fibromyalgia* is a common, yet poorly understood myofascial pain syndrome in which patients complain of severe muscle pain and tenderness, severe fatigue, and often poor sleep. Serum CK, erythrocyte sedimentation rate (ESR), EMG, and muscle biopsy are normal **(Chap. 366)**. *Polymyalgia rheumatica* occurs mainly in patients aged >50 years and is characterized by stiffness and pain in the shoulders, lower back, hips, and thighs **(Chap. 356)**. The ESR and CRP are elevated, while serum CK, EMG, and muscle biopsy are normal.

Muscle Enlargement and Atrophy In most myopathies muscle tissue is replaced by fat and connective tissue, but the size of the muscle is usually not affected. However, in many limb-girdle muscular dystrophies, enlarged calf muscles are typical. The enlargement represents true muscle hypertrophy; thus the term *pseudohypertrophy* should be avoided when referring to these patients. The calf muscles remain very strong even late in the course of these disorders. Muscle enlargement can also result from infiltration by sarcoid granulomas, amyloid deposits, bacterial and parasitic infections, and focal myositis. In contrast, muscle atrophy is characteristic of other myopathies. In dysferlinopathies (LGMD2B) and anoctaminopathies (LGMD2L), there is a predilection for early atrophy of the gastrocnemius muscles, particularly the medial aspect. Atrophy of the humeral muscles is characteristic of FSHD and EDMD.

LABORATORY EVALUATION

Various tests can be used to evaluate a suspected myopathy, including CK levels, endocrine studies (e.g, thyroid function tests, parathyroid hormone and vitamin D levels), autoantibodies (associated with myositis and systemic disorders), forearm exercise test, muscle biopsy, and genetic testing. Electrodiagnostic studies can be useful to differentiate muscle disorders from other motor unit diseases but in most instances do not help distinguish the specific type of myopathy.

Serum Enzymes CK is the most sensitive measure of muscle damage. The MM isoenzyme predominates in skeletal muscle, whereas creatine kinase-myocardial bound (CK-MB) is the marker for cardiac muscle. Serum CK can be elevated in normal individuals without provocation, presumably on a genetic basis or after strenuous activity, trauma, a prolonged muscle cramp, or a generalized seizure. Aspartate aminotransferase (AST), alanine aminotransferase (ALT), aldolase, and lactic dehydrogenase (LDH) are enzymes sharing an origin in both muscle and liver. Problems arise when the levels of these enzymes are found to be elevated in a routine screening battery, leading to the erroneous assumption that liver disease is present when in fact muscle could be the cause. An elevated γ-glutamyl transferase (GGT) helps to establish a liver origin because this enzyme is not found in muscle.

Electrodiagnostic Studies EMG, repetitive nerve stimulation, and nerve conduction studies (NCS) **(Chap. 438)** are helpful in differentiating myopathies from neuropathies and neuromuscular junction diseases. Routine NCS are typically normal in myopathies but reduced amplitudes of compound muscle action potentials may be seen in atrophied muscles. The needle EMG may reveal irritability on needle insertion and spontaneously that is suggestive of a necrotizing myopathy (inflammatory myopathies, dystrophies, toxic myopathies, myotonic myopathies), whereas a lack of irritability is characteristic of long-standing myopathic disorders (muscular dystrophies with severe fibro-fatty replacement, endocrine myopathies, disuse atrophy, and many of the metabolic myopathies). In addition, the EMG may demonstrate myotonic discharges that will narrow the differential diagnosis (Table 441-1). Another important EMG finding is the presence of short-duration, small-amplitude, polyphasic motor unit action potentials (MUAPs). In myopathies, the MUAPs fire early but at a normal rate to compensate for the loss of individual muscle fibers, whereas in neurogenic disorders the MUAPs fire faster. An EMG is usually normal in steroid or disuse myopathy, both of which are associated with type 2 fiber atrophy; this is because the EMG preferentially assesses the physiologic function of type 1 fibers. The EMG can supplement the clinical examination in choosing an appropriately affected muscle to biopsy.

Genetic Testing This is increasingly available and is the gold standard for diagnosing patients with hereditary myopathies.

Forearm Exercise Test With exercise-induced muscle pain and myoglobinuria, there may be a defect in glycolysis. For safety, the test should not be performed under ischemic conditions to avoid an unnecessary insult to the muscle, causing rhabdomyolysis. The test is performed by placing a small indwelling catheter into an antecubital vein. A baseline blood sample is obtained for lactic acid and ammonia. The forearm muscles are exercised by asking the patient to vigorously open and close the hand for 1 min. Blood is then obtained at intervals of 1, 2, 4, 6, and 10 min for comparison with the baseline sample. A three- to fourfold rise of lactic acid is typical. The simultaneous measurement of ammonia serves as a control because it should also rise with exercise. In patients with myophosphorylase deficiency and certain other glycolytic defects, the lactic acid rise will be absent or below normal, while the rise in ammonia will reach control values. If there is lack of effort, neither lactic acid nor ammonia will rise. Patients with selective failure to increase ammonia may have myoadenylate deaminase deficiency. This condition has been reported to be a cause of myoglobinuria, but deficiency of this enzyme in asymptomatic individuals makes interpretation controversial.

Muscle Biopsy Muscle biopsy is extremely helpful in evaluation of acquired myopathies but is performed less frequently in suspected hereditary myopathies as genetic testing has become more widely available. Almost any superficial muscle can be biopsied, but it is important to biopsy one that is affected clinically but not too severely (for example grade 4 out of 5 strength or movement against moderate resistance by manual muscle testing **[Chap. 415]**). A specific diagnosis can be established in many disorders.

HEREDITARY MYOPATHIES

Muscular dystrophy refers to a group of hereditary progressive diseases each with unique phenotypic and genetic features (Tables 441-3, 441-4, 441-5, and Fig. 441-6). The prognosis of dystrophies is slow progressive weakness, though the severity and course is variable between and even within subtypes. Some are associated with cardiac and ventilatory muscle involvement, which are the leading cause of mortality. Unfortunately, there are no specific medical therapies for most of the muscular dystrophies and treatment is aimed at maintaining function with physical and occupational therapy. Non-invasive ventilation and tracheostomy may be warranted. Those with cardiomyopathy may require afterload reduction, antiarrhythmic agents, pacemakers or intracardiac defibrillators, and occasionally cardiac transplantation. We will focus primarily on those that manifest in adulthood.

DUCHENNE AND BECKER MUSCULAR DYSTROPHY (DMD AND BMD)

DMD and BMD are X-linked recessive muscular dystrophies caused by mutations in the *dystrophin* gene. Affecting 1/3,000 male births, DMD is the most common mutational disease affecting boys. The incidence of BMD is ~5 per 100,000.

Clinical Features Proximal muscles, especially of the lower extremities, are prominently involved in both disorders. This becomes evident in DMD very early; boys with DMD have difficulty climbing stairs and never run well. As the disease progresses, weakness becomes more generalized. Hypertrophy of muscles, particularly in the calves, is an early and prominent finding. Most patients with BMD first experience difficulties between ages 5 and 15 years, although onset in the third or fourth decade or even later can occur. Life expectancy for DMD and BMD is reduced, but most BMD cases survive into the fourth or fifth decade. Mental retardation may occur in both disorders, but is less common in BMD. Cardiac involvement is common in both DMD and BMD and may result in heart failure; some BMD patients manifest with only heart failure. Other less common presentations of dystrophinopathy are asymptomatic hyper-CK-emia, myalgias without weakness, and myoglobinuria.

Laboratory Features Serum CK levels are usually elevated. Muscle biopsies demonstrate dystrophic features. Western blot analysis of muscle biopsy samples demonstrate absent dystrophin in DMD, or reduction in levels or size of dystrophin in BMD. In both disorders, mutations can be established using DNA from peripheral blood leukocytes. In most cases, muscle biopsies are no longer performed when DMD or BMD is suspected, as genetic testing is less invasive, less costly, and routinely available. Deletions within or duplications of the *dystrophin* gene are common in both DMD and BMD; in ~95% of cases, the mutation does not alter the translational reading frame of messenger RNA. These "in-frame" mutations allow for production of some dystrophin, which accounts for the presence of altered rather than absent dystrophin on Western blot analysis and a milder clinical phenotype.

TREATMENT

Duchenne and Becker Muscular Dystrophy

Glucocorticoids slow progression in DMD but their use has not been adequately studied in Becker dystrophy. Physical and occupational therapy are important in helping maintain function. As patients

TABLE 441-3 Autosomal Dominant Limb-Girdle Muscular Dystrophies (LGMDs)

DISEASE	CLINICAL FEATURES	LABORATORY FEATURES	ABNORMAL PROTEIN
LGMD1A	Onset second to eighth decade Muscle weakness affects proximal and distal limb muscles, vocal cords, and pharyngeal muscles	Serum CK 2× normal EMG myopathic and may have pseudomotonic discharges Muscle biopsy: features of MFM	Myotilin
LGMD1B	Onset first or second decade Proximal lower limb weakness and cardiomyopathy with conduction defects Some cases indistinguishable from Emery-Dreifuss muscular dystrophy with joint contractures	Serum CK 3–5× normal EMG myopathic	Lamin A/C
LGMD1C	Onset in early childhood Proximal weakness Gower sign, calf hypertrophy, rippling muscles Exercise-related muscle cramps	Serum CK 4–25× normal EMG myopathic	Caveolin-3
LGMD1D	Onset second to sixth decade Proximal and distal muscle weakness	Serum CK 2–3× normal EMG myopathic Muscle biopsy: features of MFM	DNAJB6
LGMD1E	Onset first to sixth decade Proximal or distal muscle weakness Cardiomyopathy and arrhythmias	Serum CK 2–4× normal EMG myopathic and may have pseudomotonic discharges Muscle biopsy: features of MFM	Desmin
LGMD1F	Onset infancy to sixth decade Proximal or distal weakness May have early contractures resembling Emery-Dreifuss syndrome	Serum CK normal to 20× normal EMG myopathic Muscle biopsy may show enlarged nuclei with central pallor, rimmed vacuoles, and filamentous inclusions	TNPO3
LGMD1G	Onset teens to sixth decade Proximal weakness contractures of fingers and toes	Muscle biopsies show rimmed vaculoes CK normal to 9× normal	HNRNPDL
LGMD1H	Onset teens to sixth decade. Proximal weakness	Muscle biopsy with many COX negative/SDH positive fibers	
LGMD1I	Onset in teens to eighties, proximal weakness, scapular winging Allelic to LGMD2A but milder phenotype	CK normal to 50× normal EMG myopathic	Calpain-3

Abbreviations: CK, creatine kinase; EMG, electromyography; MFM, myofibrillar myopathy; NCS, nerve conduction studies; HNRNPDL, heterogeneous nuclear ribonucleoprotein D-like protein.

TABLE 441-4 Autosomal Recessive Limb-Girdle Muscular Dystrophies (LGMDs)

DISEASE	CLINICAL FEATURES	LABORATORY FEATURES	ABNORMAL PROTEIN
LGMD2A	Onset first or second decade Scapular winging; no calf hypertrophy; no cardiac or respiratory muscle weakness Proximal and distal weakness; may have contractures at elbows, wrists, and fingers	Serum CK 3–15× normal EMG myopathic Muscle biopsy may show lobulated muscle fibers	Calpain-3
LGMD2B	Onset second or third decade Proximal muscle weakness at onset, later distal (calf) muscles affected Miyoshi's myopathy is variant of LGMD2B with calf muscles affected at onset	Serum CK 3–100× normal EMG myopathic Inflammation on muscle biopsy may simulate polymyositis	Dysferlin
LGMD2C–F	Onset in childhood to teenage years Clinical condition similar to Duchenne and Becker muscular dystrophies Cognitive function normal	Serum CK 5–100× normal EMG myopathic	γ, α, β, δ sarcoglycans
LGMD2G	Onset age 10–15 Proximal and distal muscle weakness	Serum CK 3–17× normal EMG myopathic Muscle biopsy may show rimmed vacuoles	Telethonin
LGMD2H	Onset first to third decade Allelic to sarcotubular congenital myopathy Proximal muscle weakness	Serum CK 2–25× normal EMG myopathic Muscle biopsy reveals dilated T-tubules	*TRIM32* gene
LGMD2I	Onset first to third decade Clinical condition similar to Duchenne or Becker dystrophies Cardiomyopathy and respiratory failure may occur early before significant weakness Cognitive function normal	Serum CK 10–30× normal EMG myopathic	Fukutin-related protein
LGMD2J[a]	Onset first to third decade Proximal lower limb weakness Mild distal weakness Progressive weakness causes loss of ambulation	Serum CK 1.5–2× normal EMG myopathic Muscle biopsy reveals rimmed vacuoles	Titin
LGMD2K	Usually presents in infancy as Walker-Warburg syndrome but can present in early adult life with proximal weakness and only minor CNS abnormalities	CK 10–20× normal EMG myopathic	POMT1
LGMD2L	Presents in childhood or adult life May manifest with quadriceps atrophy and myalgia Some present with early involvement of the calves in the second decade of life, resembling Miyoshi myopathy type 1 (dysferlinopathy)	CK 8–20× normal EMG myopathic	Anoctamin 5
LGMD2M	Usually presents in infancy as Fukuyama congenital muscular dystrophy but can present in early adult life with proximal weakness and only minor CNS abnormalities	CK 10–50× normal EMG myopathic	Fukutin
LGMD2N	Usually presents in infancy as muscle-eye-brain disease but can present in early adult life with proximal weakness and only minor CNS abnormalities	CK 5–20× normal EMG myopathic	POMGnT1
LGMD2O	Usually presents in infancy as Walker-Warburg syndrome but can present in early adult life with proximal weakness and only minor CNS abnormalities	CK 5–20× normal EMG myopathic	POMT2
LGMD2P	One case reported presenting in early childhood	CK >10× normal	α-Dystroglycan
LGMD2Q	Onset in infancy to fourth decade; proximal weakness; may have ptosis and extraocular weakness; epidermolysis bullosa (also considered a congenital myasthenic syndrome)	CK variable, but usually only mildly elevated EMG myopathic Repetitive nerve stimulation may show decrement	Plectin 1
LGMD2R	See LGMD1E (Table 441-6)	See LGMD1E	Desmin
LGMD2S	Onset in infancy to sixth decade Proximal weakness Eye abnormalities common; truncal ataxia and chorea Mild to moderate intellectual disability Hutterite descent	CK 1.5–20× normal	TRAPC11
LGMD2T	Onset in early childhood to fourth decade Proximal weakness CNS abnormalities, cataracts, cardiomyopathy, and neuromuscular junction dysfunction	CK3-> 10× normal EMG myopathic	GMPPB

[a]Udd type distal myopathy is a form of titin deficiency with only distal muscle weakness (see Table 441-9).

Abbreviations: CK, creatine kinase; EMG, electromyography; GMPPB; guanosine diphosphate (GDP)-mannose pyrophosphorylase B; NCS, nerve conduction studies; POMGNT1, O-linked mannose beta 1,2-N-acetylglucosaminyltransferase; POMT1, protein-O-mannosyltransferase 1; POMT2, protein-O-mannosyltransferase 2; TNPO3, transportin 3; TRAPC11, transport (trafficking) protein particle complex, subunit 11.

TABLE 441-5 Distal Myopathies

DISEASE	CLINICAL FEATURES	LABORATORY FEATURES	ABNORMAL PROTEIN
Welander distal myopathy	Onset in fifth decade Weakness begins in hands Slow progression with spread to distal lower extremities Lifespan normal	Serum CK 2–3× normal EMG myopathic NCS normal Muscle biopsy shows dystrophic features and rimmed vacuoles	AD TIA1
Tibial muscular dystrophy (Udd)	Onset fourth to eighth decade Distal lower extremity weakness (tibial distribution) Upper extremities usually normal Lifespan normal	Serum CK 2–4× normal EMG myopathic NCS normal Muscle biopsy shows dystrophic features and rimmed vacuoles Titin absent in M-line of muscle	AD Titin AR (associated with more proximal weakness—LGMD2J)
Markesbery-Griggs distal myopathy	Onset fourth to eighth decade Distal lower extremity weakness (tibial distribution) with progression to distal arms and proximal muscles	Serum CK is usually mildly elevated EMG reveals irritative myopathy Muscle biopsies demonstrate rimmed vacuoles and features of MFM	AD Z-band alternatively spliced PDX motif-containing protein (ZASP)
Laing distal myopathy	Onset childhood to third decade Distal lower extremity weakness (anterior tibial distribution) and neck flexors affected early May have cardiomyopathy	Serum CK is normal or slightly elevated Muscle biopsies do not typically show rimmed vacuoles, but may show hyaline bodies with accumulation of myosin Large deposits of myosin heavy chain are seen in type 1 muscle fibers	AD Myosin heavy chain 7
GNE myopathy (Nonaka distal myopathy and autosomal recessive hereditary inclusion body myopathy)	Onset: second to third decade Distal lower extremity weakness (anterior tibial distribution) Mild distal upper limb weakness may be present early Progression to other muscles sparing quadriceps Ambulation may be lost in 10–15 y	Serum CK 3–10× normal EMG myopathic NCS normal Dystrophic features on muscle biopsy plus rimmed vacuoles and 15- to 19-nm filaments within vacuoles	AR *GNE* gene: UDP-*N*-acetylglucosamine 2-epimerase/*N*-acetylmannosamine kinase Allelic to hereditary inclusion body myopathy
Miyoshi myopathy[a]	Onset: second to third decade Lower extremity weakness in posterior compartment muscles Progression leads to weakness in other muscle groups Ambulation lost after 10–15 y in about one-third of cases	Serum CK 20–100× normal EMG myopathic NCS normal Muscle biopsy shows nonspecific dystrophic features often with prominent inflammatory cell infiltration; no rimmed vacuoles	AR Dysferlin (allelic to LGMD2B) ANO-5 (allelic to LGMD2L)
Williams myopathy	Distal lower extremity weakness (anterior tibial distribution)	Muscle biopsy may show rimmed vacuoles and features of MFM	X-linked Filamin-C
Myofibrillar myopathies	Onset from early childhood to late adult life Weakness may be distal, proximal, or generalized Cardiomyopathy and respiratory involvement is not uncommon	Serum CKs can be normal or moderately elevated EMG is myopathic and often associated with myotonic discharges Muscle biopsy demonstrates abnormal accumulation of desmin and other proteins, rimmed vacuoles, and myofibrillar degeneration	Genetically heterogeneous AD Myotilin (also known as LGMD 1A) ZASP (see Markesbery-Griggs distal myopathy) Filamin-C Desmin Alpha B crystallin Bag3 Titin DNAJB6 TNPO3 AR Desmin X-linked FHL1

[a]Miyoshi myopathy phenotype may also be seen with mutations in ANO-5 that encodes for anoctamin 5 (allelic to LGMD2L).

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; CK, creatine kinase; EMG, electromyography; MFM, myofibrillar myopathy; NCS, nerve conduction studies.

often die from the associated cardiomyopathy, it is important to follow patients with a cardiologist and treat appropriately. Recent studies suggest that there is clinical benefit in selected cases of DMD from short oligonucleotides that permit skipping of mutant exons, leading to expression of a short but nonetheless functional dystrophin protein. In parallel, other studies suggest that small molecules may permit read-through of protein-truncating mutations in some Duchenne cases, again with clinical benefit.

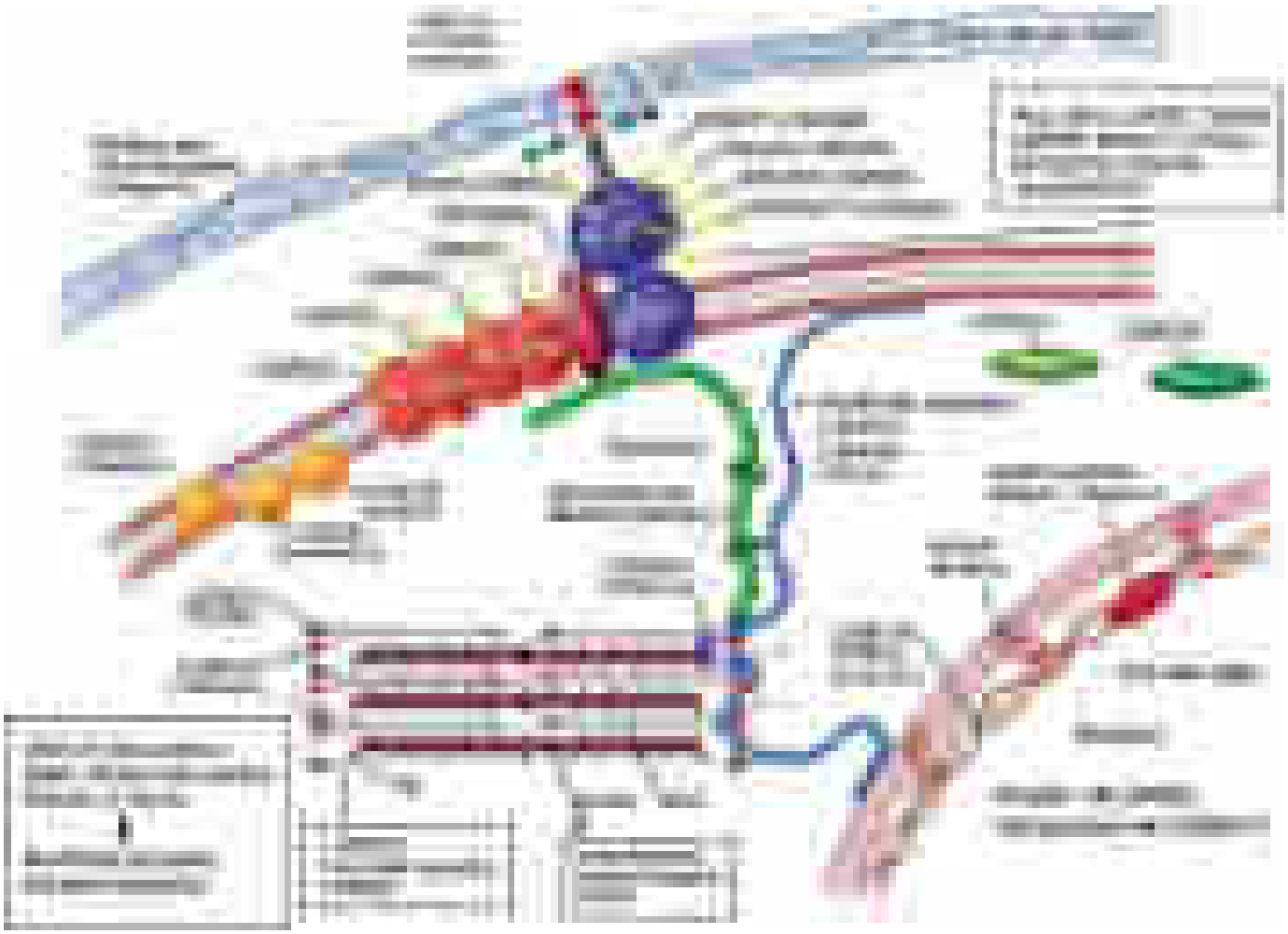


FIGURE 441-6 Proteins involved in the muscular dystrophies. This schematic shows the location of various sarcolemmal, sarcomeric, nuclear, and enzymatic proteins associated with muscular dystrophies. The diseases associated with mutations in the genes responsible for encoding these proteins are shown in boxes. Dystrophin, via its interaction with the dystroglycan complex, connects the actin cytoskeleton to the extracellular matrix. Extracellularly, the sarcoglycan complex interacts with biglycan, which connects this complex to the dystroglycan complex and the extracellular matrix collagen. Various enzymes are important in the glycosylation of the α-dystroglycan and mediate its binding to the extracellular matrix and usually cause a congenital muscular dystrophy with severe brain and eye abnormalities but may cause milder LGMD phenotype. Mutations in genes that encode for sarcomeric and Z-disk proteins cause forms of LGMD and distal myopathies (including myofibrillar myopathy, forms of hereditary inclusion body myopathy) as well as nemaline rod myopathy and other "congenital" myopathies. Mutations affecting nuclear membrane proteins are responsible for most forms of EDMD. Mutations in other nuclear genes cause other forms of dystrophy. *(Used with permission from AA Amato, J Russell (eds): Neuromuscular Disorders, 2nd ed. New York, McGraw-Hill, 2016, Figure 27-1, p 657.)*

LIMB-GIRDLE MUSCULAR DYSTROPHY

The LGMDs are a genetically heterogenous group of dystrophies in which males and females are affected equally, with onset ranging from late in the first decade to the fourth decade. The LGMDs typically manifest with progressive weakness of pelvic and shoulder girdle musculature and are often clinically indistinguishable from DMD and BMD. Respiratory insufficiency from weakness of the diaphragm may occur, as may cardiomyopathy. Serum CKs are elevated and the EMG is myopathic. Muscle biopsy reveal dystrophic features, but the findings are not specific to differentiate subtypes from one another unless immunohistochemistry is employed (e.g., immunostaining for various sarcoglycans, dysferlin, alpha-dystroglycan, merosin) or there are features to suggest one of the myofibrillar myopathies. Nonetheless, definitive diagnosis requires genetic testing.

A systematic classification of LGMD is based on autosomal dominant (LGMD1) and autosomal recessive (LGMD2) inheritance. Superimposed on the backbone of LGMD1 and LGMD2, the classification uses a sequential alphabetical lettering system (LGMD1A, LGMD2A, etc.) based on genotype. This results in an ever-expanding list of conditions summarized in Tables 441-3 and 441-4. The prevalence of LGMD ranges from 80 to 700 per 100,000 while estimated prevalence of individual specific subtypes LGMDs vary. The most common types of adult-onset LGMD are calpainopathy (LGMD2A), Fukutin-related protein (FKRP) deficiency (LGMD2I), and anoctaminopathy (LGMD2L). LGMD2A, the most common cause of LGMD in those with ancestry from Spain, France, Italy, and Great Britain, is associated with marked scapular winging, lack of calf muscle hypertrophy, and lack of cardiac and lung involvement. LGMD2I is more common in those from northern European ancestry, is associated with calf muscle hypertrophy, and can have cardiac and lung involvement out of proportion to extremity weakness. LGMD2L is becoming increasingly recognized as a common form of LGMD; as seen in dysferlinopathies (LGMD2B and Miyoshi myopathy type 1), anoctaminopathy has an early predilection for medial calf atrophy and weakness.

EMERY-DREIFUSS MUSCULAR DYSTROPHY (EDMD)

There are at least five genetically distinct forms of EDMD. Emerin mutations are the most common cause of X-linked EDMD, although mutations in *FHL1* may also be associated with a similar phenotype, which is X-linked as well. Mutations involving the gene for lamin A/C are the most common cause of autosomal dominant EDMD (also known as LGMD1B) and are also a common cause of hereditary cardiomyopathy. Less commonly, autosomal dominant EDMD has been reported with mutations in *nesprin-1*, *nesprin-2*, and *TMEM43* genes.

Clinical Features Prominent contractures can be recognized in early childhood and teenage years, often preceding muscle weakness. The contractures persist throughout the course of the disease and

are present at the elbows, ankles, and neck. Muscle weakness affects humeral and peroneal muscles at first and later spreads to a limb-girdle distribution (Table 441-1). The cardiomyopathy is potentially life threatening and may result in sudden death. A spectrum of atrial rhythm and conduction defects includes atrial fibrillation and paralysis and atrioventricular heart block. Some patients have a dilated cardiomyopathy. Female carriers of the X-linked variant may manifest with a cardiomyopathy.

Laboratory Features Serum CK is usually slightly elevated and the EMG is myopathic. Muscle biopsy usually shows nonspecific dystrophic features, although cases associated with *FHL1* mutations have features of myofibrillar myopathy. Immunohistochemistry reveals absent emerin staining of myonuclei in X-linked EDMD due to *emerin* mutations. ECGs demonstrate atrial and atrioventricular rhythm disturbances.

X-linked EDMD usually arises from defects in the *emerin* gene encoding a nuclear envelope protein. *FHL1* mutations are also a cause of X-linked scapuloperoneal dystrophy, but can also present with an X-linked form of EDMD. The autosomal dominant disease can be caused by mutations in the *LMNA* gene encoding lamin A and C; in the synaptic nuclear envelope protein 1 (*SYNE1*) or 2 (*SYNE2*) encoding nesprin-1 and nesprin-2, respectively; and in *TMEM43* encoding LUMA. These proteins are essential components of the filamentous network underlying the inner nuclear membrane. Loss of structural integrity of the nuclear envelope from defects in emerin, lamin A/C, nesprin-1, nesprin-2, and LUMA accounts for overlapping phenotypes.

TREATMENT

Emery-Dreifuss Muscular Dystrophy

Supportive care should be offered for neuromuscular disability, including ambulatory aids, if necessary. Stretching of contractures is difficult. Management of cardiomyopathy and arrhythmias (e.g., early use of a defibrillator or cardiac pacemaker) may be lifesaving.

MYOTONIC DYSTROPHY

There are two distinct forms of myotonic dystrophy (*dystrophia myotonica* or DM), namely myotonic dystrophy type 1 (DM1), and myotonic dystrophy type 2 (DM2) also called *proximal myotonic myopathy* (PROMM).

Clinical Features The clinical expression of DM1 varies widely and involves many systems other than muscle. Affected patients may have a "hatchet-faced" appearance due to temporalis, masseter, and facial muscle atrophy and weakness. Frontal baldness is frequent. Weakness of wrist and fingers occurs early as does footdrop. Proximal muscles are less affected. Palatal, pharyngeal, and tongue involvement can lead to dysarthria and dysphagia. Some patients have diaphragm and intercostal muscle weakness, resulting in ventilatory insufficiency. Myotonia is usually apparent by the age of 5 years and is best demonstrable by percussion of the thenar eminence or asking patients to close their fingers very tightly and then relax.

ECG abnormalities include first-degree heart block and more extensive conduction system involvement. Complete heart block and sudden death can occur. Congestive heart failure occurs infrequently but may result from cor pulmonale secondary to respiratory failure. Other associated features include intellectual impairment, hypersomnia, posterior subcapsular cataracts, gonadal atrophy, insulin resistance, and decreased esophageal and colonic motility.

Congenital myotonic dystrophy is a more severe form of DM1 and occurs in ~25% of infants of affected mothers. It is characterized by severe facial and bulbar weakness, transient neonatal respiratory insufficiency, and mental retardation.

DM2 or PROMM involves mainly proximal muscles. Other features of the disease overlap with DM1, including cataracts, testicular atrophy, insulin resistance, constipation, hypersomnia, and cognitive defects. Cardiac conduction defects occur but are less common. The hatchet face and frontal baldness are also less consistent features. A very striking difference is the failure to clearly identify a congenital form of DM2.

Laboratory Features The diagnosis of myotonic dystrophy can usually be made on the basis of clinical findings. Serum CK levels may be normal or mildly elevated. EMG evidence of myotonia is present in most cases of DM1 but is more patchy in DM2. Muscle biopsy is not typically performed for diagnosis but is sometimes done when the clinical features and electrophysiological features are not recognized. The major histopathological feature in both DM1 and DM2 are numerous internalized nuclei can be seen in individual muscle fibers combined with many atrophic fibers with pyknotic nuclear clumps.

DM1 and DM2 are autosomal dominant disorders. DM1 is transmitted by an intronic mutation consisting of an unstable expansion of a CTG trinucleotide repeat in a serine-threonine protein kinase gene (named *DMPK*). An increase in the severity of the disease phenotype in successive generations (genetic anticipation) is accompanied by an increase in the number of trinucleotide repeats. The unstable triplet repeat in myotonic dystrophy can be used for prenatal diagnosis. Congenital disease occurs almost exclusively in infants born to affected mothers.

DM2 is caused by a DNA expansion mutation consisting of a CCTG repeat in intron 1 of the *CNBP* gene encoding the CCHC-type zinc finger nucleic acid binding protein. The DNA expansions in DM1 and DM2 impair muscle function by a toxic gain of function of the mutant mRNA. In both DM1 and DM2, the mutant RNA appears to form intranuclear inclusions composed of aberrant RNA. These RNA inclusions sequester RNA-binding proteins essential for proper splicing of a variety of other mRNAs. This leads to abnormal transcription of multiple proteins in a variety of tissues/organ systems, in turn causing the systemic manifestations of DM1 and DM2.

TREATMENT

Myotonic Dystrophy

The myotonia in DM1 and DM2 is usually not so bothersome to warrant treatment, but when it is mexiletine may be helpful. A cardiac pacemaker or implantable cardioverter defibrillator should be considered for patients with significant arrhythmia. Molded ankle-foot orthoses help stabilize gait in patients with foot drop. Excessive daytime somnolence with or without sleep apnea is not uncommon. Sleep studies, noninvasive respiratory support (biphasic positive airway pressure [BiPAP]), and treatment with modafinil may be beneficial.

FACIOSCAPULOHUMERAL (FSH) MUSCULAR DYSTROPHY

There are two forms of FSHD that have similar pathogenesis. Most patients have FSHD type 1 (95%), whereas ~5% have FSHD2. Both forms are clinically and histopathologically identical. The prevalence FSHD is ~5 per 100,000 individuals.

Clinical Features FSHD typically presents in childhood or young adulthood. In most cases, facial weakness is the initial manifestation, appearing as an inability to smile, whistle, or fully close the eyes. Loss of scapular stabilizer muscles makes arm elevation difficult. Scapular winging (Fig. 441-3) becomes apparent with attempts at abduction and forward movement of the arms. Biceps and triceps muscles may be severely affected, with relative sparing of the deltoid muscles. Weakness is invariably worse for wrist extension than for wrist flexion, and weakness of the anterior compartment muscles of the legs may lead to footdrop. In 20% of patients, weakness progresses to involve the pelvic muscles, and severe functional impairment and possible wheelchair dependency result. The heart is not involved but there can be ventilatory muscle weakness in 5% of affected individuals. There is an increased incidence of nerve deafness. *Coats' disease*, a disorder consisting of telangiectasia, exudation, and retinal detachment, also occurs.

Laboratory Features The serum CK level may be normal or mildly elevated. EMG and muscle biopsy show nonspecific abnormalities but on occasion can reveal a prominent inflammatory infiltrate leading to an incorrect diagnosis of myositis (Chap. 358).

FSHD1 is associated with deletions of tandem 3.3-kb repeats at 4q35. The deletion reduces the number of repeats to a fragment of <35 kb in most patients. Within these repeats lies the *DUX4* gene, which usually is not expressed after early muscle development. In patients with FSHD1 these deletions in the setting of a specific polymorphism leads to hypomethylation of the region and toxic expression of the *DUX4* gene. In patients with FSHD2, there is no deletion, but a mutation in *SMCHD1*, leading to hypomethylation of the DUX4 region and the permissive expression of the *DUX4* gene. In both FSHD1 and FSHD2, there is overexpression of the *DUX4* transcript.

TREATMENT

Facioscapulohumeral Muscular Dystrophy

No specific treatment is available; ankle-foot orthoses are helpful for footdrop. Scapular stabilization procedures improve scapular winging but may not improve function.

OCULOPHARYNGEAL DYSTROPHY (OPMD)

OPMD represents one of several disorders characterized by progressive external ophthalmoplegia, which consists of slowly progressive ptosis and limitation of eye movements with sparing of pupillary reactions for light and accommodation. Patients usually do not complain of diplopia, in contrast to patients having conditions with a more acute onset of ocular muscle weakness (e.g., myasthenia gravis).

Clinical Features OPMD has a late onset; it usually presents in the fourth to sixth decade with ptosis or dysphagia. The extraocular muscle impairment is less prominent in the early phase but may become severe over time. The swallowing problem may lead to aspiration. Weakness of the neck and proximal extremities can develop but is usually mild in degree.

Laboratory Features The serum CK level may be two to three times normal. EMG can identify myopathic changes in weak muscles. Muscle biopsies are no longer necessary for diagnosis in most cases, but when performed demonstrate muscle fibers with rimmed vacuoles. On electron microscopy, a distinctive feature of OPMD is the presence of 8.5 nm tubular filaments in some muscle cell nuclei.

OPMD is an autosomal dominant disorder that has a high incidence in certain populations (e.g., French-Canadians, individuals of Spanish ancestry, and Ashkenazi Jews). The molecular defect in OPMD is an expansion of a polyalanine repeat tract in a poly-RNA-binding protein (*PABP2*) gene.

TREATMENT

Oculopharyngeal Dystrophy

Dysphagia can lead to significant undernourishment and aspiration. Cricopharyngeal myotomy may improve swallowing. Eyelid crutches can improve vision when ptosis obstructs vision; candidates for ptosis surgery must be carefully selected—those with severe facial weakness are not suitable.

DISTAL MYOPATHIES / DYSTROPHIES

The distal myopathies are notable for their preferential distal distribution of muscle weakness in contrast to most muscle conditions associated with proximal weakness. The major distal myopathies are summarized in Tables 441-1 and 441-5.

Clinical Features *Welander, Udd*, and *Markesbery-Griggs type distal myopathies* are all late-onset, dominantly inherited disorders of distal limb muscles, usually beginning after age 40 years. Welander distal myopathy preferentially involves the wrist and finger extensors, whereas the others are associated with anterior tibial weakness leading to progressive footdrop. *Laing distal myopathy* is also a dominantly inherited disorder heralded by tibial weakness; however, it is distinguished by onset in childhood or early adult life. GNE myopathy (also known as *Nonaka distal myopathy* and autosomal recessive hereditary inclusion body myopathy) and *Miyoshi myopathy* are distinguished by autosomal recessive inheritance and onset in the late teens or twenties. GNE and Williams myopathy produce prominent anterior tibial weakness, whereas Miyoshi myopathy is unique in that gastrocnemius muscles are preferentially affected at onset. Finally, the *myofibrillar myopathies* (MFMs) are a clinically and genetically heterogeneous group of muscular dystrophies that can be associated with prominent distal weakness; they can be inherited in an autosomal dominant or recessive pattern. Of note, Markesbery-Griggs myopathy (caused by mutations in ZASP) and LGMD1B (caused by mutations in myotilin) are subtypes of myofibrillar myopathy (MFM).

Laboratory Features Serum CK levels are markedly elevated in Miyoshi myopathy, but in the other conditions, serum CK is only slightly increased. EMGs are myopathic and can be irritable with myotonic discharges in MFM. Muscle biopsy shows nonspecific dystrophic features and, with the exception of Laing and Miyoshi myopathies, often shows rimmed vacuoles. MFM is associated with the accumulation of dense inclusions and amorphous material best seen on Gomori trichrome staining along with myofibrillar disruption on electron microscopy. Immune staining sometimes demonstrates accumulation of desmin and other proteins in MFM, large deposits of myosin heavy chain in the subsarcolemmal region of type 1 muscle fibers in Laing myopathy, and reduced or absent dysferlin in Miyoshi myopathy type 1.

TREATMENT

Distal Myopathies

Occupational therapy is offered for loss of hand function; ankle-foot orthoses can support distal lower limb muscles. The MFMs can be associated with cardiomyopathy (congestive heart failure or arrhythmias) and respiratory failure that may require medical management. Laing-type distal myopathy can also be associated with a cardiomyopathy.

DISORDERS OF MUSCLE ENERGY METABOLISM

There are two principal sources of energy for skeletal muscle—fatty acids and glucose. Abnormalities in either glucose or lipid utilization can be associated with distinct clinical presentations that can range from an acute, painful syndrome with rhabdomyolysis and myoglobinuria to a chronic, progressive muscle weakness simulating muscular dystrophy (Table 441-1). As with the muscular dystrophies there are no specific medical treatments available.

GLYCOGEN STORAGE AND GLYCOLYTIC DEFECTS

Disorders of Glycogen Storage Causing Progressive Weakness • α-GLUCOSIDASE, OR ACID MALTASE, DEFICIENCY (POMPE DISEASE) Three clinical forms of α-glucosidase, or acid maltase, deficiency (*type II glycogenosis*) can be distinguished. The infantile form is the most common, with onset of symptoms in the first 3 months of life. Infants develop severe muscle weakness, cardiomegaly, hepatomegaly, and respiratory insufficiency. Glycogen accumulation in motor neurons of the spinal cord and brainstem contributes to muscle weakness. Death usually occurs by 1.5 years of age. In the childhood form, the picture resembles Duchenne muscular dystrophy with delayed motor milestones resulting from proximal limb muscle weakness and involvement of respiratory muscles. The heart may be involved, but the liver and brain are unaffected. The adult form usually begins in the third or fourth decade but can present as late as the seventh decade. Ventilatory weakness can be the initial and only manifestation in 20–30% of late-onset cases.

The serum CK level is 2–10 times normal in infantile or childhood-onset Pompe's disease but can be normal in adult-onset cases. EMG can demonstrate muscle membrane irritability, particularly in the paraspinal muscles. The muscle biopsy in infants typically reveals vacuoles containing glycogen and the lysosomal enzyme acid phosphatase. Electron microscopy reveals membrane-bound and free tissue glycogen. However, muscle biopsies in late-onset Pompe disease may demonstrate only nonspecific abnormalities. Enzyme analysis of dried blood spots is a sensitive technique to screen for Pompe's disease. A definitive diagnosis is established by genetic testing.

Pompe disease is inherited as an autosomal recessive disorder caused by mutations of the α-*glucosidase* gene. Enzyme replacement therapy (ERT) with IV recombinant human α-glucosidase is beneficial in infantile-onset Pompe disease. In late-onset cases, ERT has a more modest benefit.

OTHER GLYCOGEN STORAGE DISEASES WITH PROGRESSIVE WEAKNESS In *de-branching enzyme deficiency* (*type III glycogenosis*), a slowly progressive form of muscle weakness can develop after puberty. Rarely, myoglobinuria may be seen. Patients are usually diagnosed in infancy, however, because of hypotonia and delayed motor milestones; hepatomegaly, growth retardation, and hypoglycemia are other manifestations. *Branching enzyme deficiency* (*type IV glycogenosis*) is a rare and fatal glycogen storage disease characterized by failure to thrive and hepatomegaly. Hypotonia and muscle wasting may be present, but the skeletal muscle manifestations are minor compared to liver failure.

Disorders of Glycolysis Causing Exercise Intolerance

Several glycolytic defects are associated with recurrent myoglobinuria. The most common is *McArdle disease* caused by mutations in the *PYGM* gene leading to *myophosphorylase deficiency*. Symptoms of muscle pain and stiffness usually begin in adolescence. With severe episodes myoglobinuria can occur.

Certain features help distinguish some enzyme defects. In McArdle disease, exercise tolerance can be enhanced by a slow induction phase (warm-up) or brief periods of rest, allowing for the start of the "second-wind" phenomenon (switching to utilization of fatty acids). Varying degrees of hemolytic anemia accompany deficiencies of both phosphofructokinase (mild) and phosphoglycerate kinase (severe). In phosphoglycerate kinase deficiency, the usual clinical presentation is a seizure disorder associated with mental retardation; exercise intolerance is an infrequent manifestation.

In all of these conditions, the serum CK levels fluctuate widely and may be elevated even during symptom-free periods. CK levels >100 times normal are expected accompanying myoglobinuria. A forearm exercise test reveals a blunted rise in venous lactate with a normal rise in ammonia. A definitive diagnosis of glycolytic disease can be made by muscle biopsy with appropriate staining and enzyme assays, but genetic testing is now done in lieu of biopsy in most cases.

Training may enhance exercise tolerance, perhaps by increasing perfusion to muscle. Dietary intake of free glucose or fructose prior to activity may improve function but care must be taken to avoid obesity from ingesting too many calories.

LIPID AS AN ENERGY SOURCE AND ASSOCIATED DEFECTS

Lipid is an important muscle energy source during rest and during prolonged, submaximal exercise. Oxidation of fatty acids occurs in the mitochondria. To enter the mitochondria, a fatty acid must first be converted to an "activated fatty acid," acyl-CoA. The acyl-CoA must be linked with carnitine by the enzyme CPT for transport into the mitochondria.

Carnitine Palmitoyltransferase 2 Deficiency CPT2 deficiency is the most common recognizable cause of recurrent myoglobinuria. Onset is usually in the teenage years or early twenties. Muscle pain and myoglobinuria typically occur after prolonged exercise but can also be precipitated by fasting or infections; up to 20% of patients do not exhibit myoglobinuria, however. Strength is normal between attacks. In contrast to disorders caused by defects in glycolysis, in which muscle cramps follow short, intense bursts of exercise, the muscle pain in CPT2 deficiency does not occur until the limits of utilization have been exceeded and muscle breakdown has already begun.

Serum CK levels and EMG findings are both usually normal between episodes. A normal rise of venous lactate during forearm exercise distinguishes this condition from glycolytic defects. Muscle biopsy does not show lipid accumulation and is usually normal between attacks. The diagnosis requires direct measurement of muscle CPT or genetic testing. Attempts to improve exercise tolerance with frequent meals and a low-fat, high-carbohydrate diet, or by substituting medium-chain triglycerides in the diet, have not proven to be beneficial.

MITOCHONDRIAL MYOPATHIES

Mitochondria play a key role in energy production. Oxidation of the major nutrients derived from carbohydrate, fat, and protein leads to the generation of reducing equivalents. The latter are transported through the respiratory chain in the process known as *oxidative phosphorylation*. The energy generated by the oxidation-reduction reactions of the respiratory chain is stored in an electrochemical gradient coupled to ATP synthesis.

A novel feature of mitochondria is their genetic composition. Each mitochondrion possesses a DNA genome that is distinct from that of the nuclear DNA. Human mitochondrial DNA (mtDNA) consists of a double-strand, circular molecule comprising 16,569 base pairs. It codes for 22 transfer RNAs, 2 ribosomal RNAs, and 13 polypeptides of the respiratory chain enzymes. The genetics of mitochondrial diseases differ from the genetics of chromosomal disorders. The DNA of mitochondria is directly inherited from the cytoplasm of the gametes, mainly from the oocyte. The sperm contributes very little of its mitochondria to the offspring at the time of fertilization. Thus, mitochondrial genes are derived almost exclusively from the mother, accounting for maternal inheritance of some mitochondrial disorders.

Patients with mitochondrial myopathies have clinical manifestations that usually fall into three groups: chronic progressive external ophthalmoplegia (CPEO), skeletal muscle–CNS syndromes, and pure myopathy simulating muscular dystrophy or metabolic myopathy. Unfortunately, no specific medical therapies are clearly beneficial, although Co-enzyme Q10 supplements are often prescribed.

Kearn-Sayre Syndrome (KSS) This is a widespread multi-organ system disorder with a defined triad of clinical findings: onset before age 20, CPEO, and pigmentary retinopathy, plus one or more of the following features: complete heart block, cerebrospinal fluid (CSF) protein >1 g/L (100 mg/dL), or cerebellar ataxia. The cardiac disease includes syncopal attacks and cardiac arrest related to the abnormalities in the cardiac conduction system: prolonged intraventricular conduction time, bundle branch block, and complete atrioventricular block. Death attributed to heart block occurs in ~20% of the patients. Varying degrees of progressive limb muscle weakness and easy fatigability affect activities of daily living. Many affected individuals have intellectual disabilities. Endocrine abnormalities are also common, including gonadal dysfunction in both sexes with delayed puberty, short stature, and infertility. Diabetes mellitus occurs in ~13% of KSS patients. Other less common endocrine disorders include thyroid disease, hyperaldosteronism, Addison disease, and hypoparathyroidism.

Serum CK and lactate levels are normal or slightly elevated. EMG is myopathic. NCS may be abnormal related to an associated neuropathy. Muscle biopsies reveal ragged red fibers and cytochrome oxidase (COX) negative fibers. By electron microscopy, there are increased numbers of mitochondria that often appear enlarged and contain paracrystalline inclusions.

KSS is a sporadic disorder caused by single mtDNA deletions that are presumed to arise spontaneously in the ovum or zygote. The most common deletion, occurring in about one-third of patients, removes 4977 bp of contiguous mtDNA. Monitoring for cardiac conduction defects is critical. Prophylactic pacemaker implantation is indicated when ECGs demonstrate a bifascicular block.

Progressive External Ophthalmoplegia (PEO) PEO can be caused by nuclear DNA mutations affecting mtDNA and thus inherited in a Mendelian fashion or by mutations in mtDNA. Onset is usually

after puberty. Fatigue, exercise intolerance, dysphagia, and complaints of muscle weakness are typical. The neurologic examination confirms the ptosis and ophthalmoplegia, usually asymmetric in distribution. Patients do not complain of diplopia. Mild facial, neck flexor, and proximal weakness are typical. Rarely, respiratory muscles may be progressively affected and may be the direct cause of death.

Serum CK and lactate can be normal or mildly elevated. The EMG can myopathic. Ragged red and COX-negative fibers are prominently displayed in the muscle biopsy.

This autosomal dominant form of CPEO is most commonly caused by mutations in the genes encoding adenine nucleotide translocator 1 (*ANT1*), twinkle gene (*C10orf2*), and mtDNA polymerase 1(*POLG1*). Autosomal recessive PEO can also be caused by mutations in *POLG1*. Point mutations have been identified within various mitochondrial tRNA (Leu, Ile, Asn, Trp) genes in families with maternal inheritance of PEO.

There is no specific medical treatment available; exercise may improve function but will depend on the patient's ability to participate.

Myoclonic Epilepsy with Ragged Red Fibers (MERRF) The onset of MERRF is variable, ranging from late childhood to middle adult life. Characteristic features include myoclonic epilepsy, cerebellar ataxia, and progressive proximal muscle weakness. The seizure disorder is an integral part of the disease and may be the initial symptom. Cerebellar ataxia precedes or accompanies epilepsy. Other more variable features include dementia, peripheral neuropathy, optic atrophy, hearing loss, and diabetes mellitus.

Serum CK levels and lactate may be normal or elevated. EMG is myopathic, and in some patients NCS show a neuropathy. The electroencephalogram is abnormal, corroborating clinical findings of epilepsy. Typical ragged red fibers are seen on muscle biopsy. MERRF is caused by maternally inherited point mutations of mitochondrial tRNA genes. The most common mutation found in 80% of MERRF patients is an A to G substitution at nucleotide 8344 of tRNA lysine (A8344G $tRNA^{lys}$). Only supportive treatment is possible, with special attention to epilepsy.

Mitochondrial Myopathy, Encephalopathy, Lactic Acidosis, and Stroke-like Episodes (MELAS) MELAS is the most common mitochondrial encephalomyopathy. The term *stroke-like* is appropriate because the cerebral lesions do not conform to a strictly vascular distribution. The onset in the majority of patients is before age 20. Seizures, usually partial motor or generalized, are common and may represent the first clearly recognizable sign of disease. The cerebral insults that resemble strokes cause hemiparesis, hemianopia, and cortical blindness. A presumptive stroke occurring before age 40 should place this mitochondrial encephalomyopathy high in the differential diagnosis. Associated conditions include hearing loss, diabetes mellitus, hypothalamic pituitary dysfunction causing growth hormone deficiency, hypothyroidism, and absence of secondary sexual characteristics. In its full expression, MELAS leads to dementia, a bedridden state, and a fatal outcome. Serum lactic acid is typically elevated.

The CSF protein is also increased but is usually ≤1 g/L (100 mg/dL). Muscle biopsies show ragged red fibers. Neuroimaging demonstrates basal ganglia calcification in a high percentage of cases. Focal lesions that mimic infarction are present predominantly in the occipital and parietal lobes. Strict vascular territories are not respected, and cerebral angiography fails to demonstrate lesions of the major cerebral blood vessels.

MELAS is usually caused by maternally inherited point mutations of mitochondrial tRNA genes. The A3243G point mutation in $tRNA^{Leu(UUR)}$ is the most common, occurring in ~80% of MELAS cases. No specific treatment is available. Supportive treatment is essential for the strokelike episodes, seizures, and endocrinopathies.

Mitochondrial DNA Depletion Syndromes Mitochondrial DNA depletion syndrome (MDS) is a heterogeneous group of disorders that are inherited in an autosomal recessive fashion and can present in infancy or adults. MDS can be caused by mutations in several genes (*TK2*, *DGUOK*, *RRM2B*, *TYMP*, *SUCLA1*, and *SUCLA2*) that lead to depletion of mitochondrial deoxyribonucleotides (dNTP) necessary for mtDNA replication. The other major cause of MDS is a set of mutations in genes essential for mtDNA replication (e.g., *POLG1* and *C10orf2*). The clinical phenotypes associated with MDS vary. Patients may develop a severe encephalopathy (e.g., Leigh's syndrome), PEO, an isolated myopathy, myo-neuro-gastrointestinal-encephalopathy (MNGIE), and a sensory neuropathy with ataxia.

DISORDERS OF MUSCLE MEMBRANE EXCITABILITY

Muscle membrane excitability is affected in a group of disorders referred to as *channelopathies*. These disorders usually present with episodic muscle weakness (periodic paralysis) and sometimes myotonia or paramyotonia (Table 441-1).

CALCIUM CHANNEL DISORDERS OF MUSCLE

Hypokalemic Periodic Paralysis (HypoKPP) This is an autosomal dominant disorder with onset in adolescence. Males are more often affected because of decreased penetrance in females. Episodic weakness with onset after age 25 is almost never due to periodic paralyses, with the exception of thyrotoxic periodic paralysis. Attacks are often provoked by meals high in carbohydrates or sodium and may accompany rest following prolonged exercise. Weakness usually affects proximal limb muscles more than distal. Ocular and bulbar muscles are less likely to be affected. Respiratory muscles are usually spared, but when they are involved, the condition may prove fatal. Weakness may take as long as 24 h to resolve. Life-threatening cardiac arrhythmias related to hypokalemia may occur during attacks. As a late complication, patients commonly develop severe, disabling proximal lower extremity weakness.

Attacks of thyrotoxic periodic paralysis resemble those of primary HypoKPP. Despite a higher incidence of thyrotoxicosis in women, men, particularly those of Asian descent, are more likely to manifest this complication. Attacks abate with treatment of the underlying thyroid condition.

A low serum potassium level during an attack, excluding secondary causes, establishes the diagnosis. In the midst of an attack of weakness, motor conduction studies may demonstrate reduced amplitudes, whereas EMG may show electrical silence in severely weak muscles. In between attacks, the EMG and routine NCS are normal. However, a long exercise NCS test may demonstrate decrementing amplitudes.

HypoKPP type 1 is the most common form, and is caused by mutations in the voltage-sensitive, skeletal muscle calcium channel gene, *CALCL1A3*. Approximately 10% of cases are HypoKPP type 2, arising from mutations in the voltage-sensitive sodium channel gene (*SCN4A*). In both forms the mutations lead to an abnormal gating pore current that predisposes the muscle cell to depolarize when potassium levels are low.

TREATMENT

Hypokalemic Periodic Paralysis

Mild attacks usually do not require medical treatment. However, severe attacks of weakness can be improved by the administration of potassium. Oral KCl (0.2–0.4 mmol/kg) can be given every 30 min. Only rarely is IV therapy necessary (e.g., when swallowing problems or vomiting is present). The long-term goal of therapy is to avoid attacks. Patients should be made aware of the importance of a low-carbohydrate, low-sodium diet, and consequences of intense exercise. Prophylactic administration of acetazolamide or dichlorphenamide can reduce attacks of periodic weakness. However, in patients with HypoKPP type 2, attacks of weakness can be exacerbated with these medications.

SODIUM CHANNEL DISORDERS OF MUSCLE

Hyperkalemic Periodic Paralysis (HyperKPP) The term *hyperkalemic* is misleading because patients are often normokalemic

during attacks. That attacks are precipitated by potassium administration best defines the disease. The onset is usually in the first decade; males and females are affected equally. Attacks are brief and mild, usually lasting 30 min to several hours. Weakness affects proximal muscles, sparing bulbar muscles. Attacks are precipitated by rest following exercise and fasting.

Potassium may be slightly elevated or normal during an attack. As in HypoKPP, NCS in HyperKPP muscle may demonstrate reduced motor amplitudes and the EMG may be silent in very weak muscles. A long exercise NCS test can reveal diminished amplitudes as well. The EMG may demonstrate myotonic discharges. HyperKPP is caused by mutations of the voltage-gated sodium channel *SCN4A* gene. Acetazolamide or dichlorphenamide can reduce the frequency and severity of attacks. Mexiletine to be helpful in patients with significant clinical myotonia.

Paramyotonia Congenita In PC, the attacks of weakness are cold-induced or occur spontaneously and are mild. Myotonia is a prominent feature but worsens with muscle activity (paradoxical myotonia). This is in contrast to classic myotonia in which exercise alleviates the condition. Attacks of weakness are seldom severe enough to require emergency room treatment. Over time patients develop interattack weakness as they do in other forms of periodic paralysis.

Serum CK is usually mildly elevated. Routine NCS are normal. Short exercise NCS test may be abnormal however, and cooling of the muscle often dramatically reduces the amplitude of the compound muscle action potentials. EMG reveals diffuse myotonic potentials in PC. Upon local cooling of the muscle, the myotonic discharges disappear as the patient becomes unable to activate MUAPs.

PC is inherited as an autosomal dominant condition; voltage-gated sodium channel mutations are responsible, and thus this disorder is allelic with HyperKPP. Mexiletine is reported to be helpful in reducing the myotonia.

POTASSIUM CHANNEL DISORDERS

Andersen-Tawil Syndrome This rare disease is characterized by episodic weakness, cardiac arrhythmias, and dysmorphic features (short stature, scoliosis, clinodactyly, hypertelorism, small or prominent low-set ears, micrognathia, and broad forehead). The cardiac arrhythmias are potentially serious and life threatening. They include long QT, ventricular ectopy, bidirectional ventricular arrhythmias, and tachycardia. The disease is most commonly caused by mutations of the inwardly rectifying potassium channel (*Kir 2.1*) gene that heighten muscle cell excitability. The episodes of weakness may differ between patients because of potassium variability. Acetazolamide may decrease the attack frequency and severity.

CHLORIDE CHANNEL DISORDERS

Two forms of this disorder, autosomal dominant (*Thomsen disease*) and autosomal recessive (*Becker disease*) are both caused by mutations in the chloride channel 1 gene (*CLCN1*). Symptoms are noted in infancy and early childhood. The severity lessens in the third to fourth decade. Myotonia is worsened by cold and improved by activity. The gait may appear slow and labored at first but improves with walking. In Thomsen disease, muscle strength is normal, but in Becker disease, which is usually more severe, there may be muscle weakness. Muscle hypertrophy is usually present. Myotonic discharges are prominently displayed by EMG recordings. Serum CK is normal or mildly elevated. Mexiletine is helpful in relieving the myotonia.

ENDOCRINE AND METABOLIC MYOPATHIES

Endocrinopathies can cause weakness, but fatigue is more common than true weakness. The serum CK level is often normal (except in hypothyroidism) and the muscle histology is characterized by atrophy rather than destruction of muscle fibers. Nearly all endocrine myopathies respond to treatment.

THYROID DISORDERS

Hypothyroidism (Chap. 376) Patients with hypothyroidism have frequent muscle complaints, and about one-third have proximal muscle weakness. Muscle cramps, pain, and stiffness are common. Some patients have enlarged muscles. Features of slow muscle contraction and relaxation occur in 25% of patients; the relaxation phase of muscle stretch reflexes is characteristically prolonged and best observed at the ankle or biceps brachii reflexes. The serum CK level is often elevated (up to 10 times normal). EMG is typically normal. Muscle biopsy shows no distinctive morphologic abnormalities.

Hyperthyroidism (Chap. 377) Patients who are thyrotoxic commonly have proximal muscle weakness, but they rarely complain of myopathic symptoms. Activity of deep tendon reflexes may be enhanced. Fasciculations may be apparent and, when coupled with increased muscle stretch reflexes, may lead to an erroneous diagnosis of amyotrophic lateral sclerosis. A form of hypokalemic periodic paralysis can occur in patients who are thyrotoxic. Mutations in the *KCNJ18* gene that encodes for the inwardly rectifying potassium channel, Kir 2.6, have been discovered in up to a third of cases.

PARATHYROID DISORDERS (SEE ALSO CHAP. 403)

Hyperparathyroidism Proximal muscle weakness, muscle wasting, and brisk muscle stretch reflexes are the main features of this endocrinopathy. Some patients develop neck extensor weakness (part of the dropped head syndrome). Serum CK levels are usually normal or slightly elevated. Serum parathyroid hormone levels are elevated, while vitamin D and calcium levels are usually reduced. Muscle biopsies show only mild type 2 fiber atrophy.

Hypoparathyroidism An overt myopathy due to hypocalcemia rarely occurs. Neuromuscular symptoms are usually related to localized or generalized tetany. Serum CK levels may be increased secondary to muscle damage from sustained tetany. Hyporeflexia or areflexia is usually present and contrasts with the hyperreflexia in hyperparathyroidism.

ADRENAL DISORDERS (SEE ALSO CHAP. 379)

Conditions associated with glucocorticoid excess cause a myopathy; steroid myopathy is the most commonly diagnosed endocrine muscle disease. Proximal muscle weakness combined with a cushingoid appearance are the key clinical features. Serum CK and EMG are normal. Muscle biopsy, not typically done for diagnostic purposes, reveals type 2b muscle fiber atrophy. In primary hyperaldosteronism (*Conn's syndrome*), neuromuscular complications are due to potassium depletion. The clinical picture is one of persistent muscle weakness. Longstanding hyperaldosteronism may lead to proximal limb weakness and wasting. Serum CK levels may be elevated, and a muscle biopsy may demonstrate necrotic fibers. These changes relate to hypokalemia and are not a direct effect of aldosterone on skeletal muscle.

PITUITARY DISORDERS (SEE ALSO CHAP. 373)

Patients with acromegaly usually have mild proximal weakness. Muscles often appear enlarged but exhibit decreased force generation. The duration of acromegaly, rather than the serum growth hormone levels, correlates with the degree of myopathy.

DIABETES MELLITUS (SEE ALSO CHAP. 398)

Neuromuscular complications of diabetes mellitus are most often related to neuropathy. The only notable myopathy is ischemic infarction of leg muscles, usually involving one of the thigh muscles but on occasion affecting the distal leg. This condition occurs in patients with poorly controlled diabetes and presents with the abrupt onset of pain, tenderness, and edema of a thigh or calf. The area of muscle infarction is hard and indurated. The muscles most often affected include the vastus lateralis, thigh adductors, and biceps femoris. Computed tomography (CT) or MRI can demonstrate focal abnormalities in the affected muscle. Diagnosis by imaging is preferable to muscle biopsy, if possible, as hemorrhage into the biopsy site can occur.

MYOPATHIES OF SYSTEMIC ILLNESS

Systemic illnesses such as chronic respiratory, cardiac, or hepatic failure are frequently associated with severe muscle wasting and complaints

of weakness. Fatigue is usually a more significant problem than weakness, which is typically mild.

DRUG-INDUCED OR TOXIC MYOPATHIES

The most common toxic myopathies are caused by the cholesterol-lowering agents and glucocorticoids. Others impact practice to a lesser degree but are important to consider in specific situations. Table 441-6 provides a comprehensive list of drug-induced myopathies with their distinguishing features.

MYOPATHY FROM LIPID-LOWERING AGENTS

All classes of lipid-lowering agents have been implicated in muscle toxicity, including HMG-CoA reductase inhibitors (statins), and to a much lesser extent, fibrates, niacin, and ezetimibe. Myalgia and elevated CKs are the most common manifestations. Rarely patients exhibit proximal weakness or myoglobinuria. Concomitant use of statins with fibrates and cyclosporine increase the risk of severe myotoxicity. EMG demonstrates irritability and myopathic units and muscle biopsies reveal necrotic muscle fibers in weak muscles. Severe myalgia, weakness, marked elevations in serum CK (>3–5 times baseline), and myoglobinuria are indications for stopping the drug. Patients usually improve with drug cessation, although this may take several weeks. Rare cases continue to progress after the offending agent is discontinued. It is possible that in such cases the statin may have triggered an immune-mediated necrotizing myopathy, as these individuals require aggressive immunotherapy (e.g., prednisone and sometimes other agents) to improve and often relapse when these therapies are discontinued (Chap. 358). Interestingly, antibodies directed against HMG-CoA reductase have been identified in many of these cases.

TABLE 441-6 Drug-Induced Myopathies

DRUGS	MAJOR TOXIC REACTION
Lipid-lowering agents HMG-CoA reductase inhibitors Fibric acid derivatives Niacin (nicotinic acid)	Drugs belonging to all three of the major classes of lipid-lowering agents can produce a spectrum of toxicity: asymptomatic serum creatine kinase elevation, myalgias, exercise-induced pain, rhabdomyolysis, and myoglobinuria.
Glucocorticoids	Acute, high-dose glucocorticoid treatment can cause acute quadriplegic myopathy. These high doses of steroids are often combined with nondepolarizing neuromuscular blocking agents but the weakness can occur without their use. Chronic steroid administration produces predominantly proximal weakness.
Nondepolarizing neuromuscular blocking agents	Acute quadriplegic myopathy can occur with or without concomitant glucocorticoids.
Zidovudine	Mitochondrial myopathy with ragged red fibers
Drugs of abuse Alcohol Amphetamines Cocaine Heroin Phencyclidine Meperidine	All drugs in this group can lead to widespread muscle breakdown, rhabdomyolysis, and myoglobinuria. Local injections cause muscle necrosis, skin induration, and limb contractures.
Autoimmune myopathy Statins Check point inhibitors D-Penicillamine	Use of statins may cause an immune mediated necrotizing myopathy associated with HMG-CoA reductase antibodies. Check point inhibitors can be complicated by myositis, myasthenia gravis, and immune-mediated neuropathies. Myasthenia gravis has also been reported with penicillamine.
Amphophilic cationic drugs Amiodarone Chloroquine Hydroxychloroquine	All amphophilic drugs have the potential to produce painless, proximal weakness associated with necrosis and autophagic vacuoles in the muscle biopsy.
Antimicrotubular drugs Colchicine	This drug produces painless, proximal weakness especially in the setting of renal failure. Muscle biopsy shows necrosis and fibers with autophagic vacuoles.

GLUCOCORTICOID-RELATED MYOPATHIES

Glucocorticoid myopathy occurs with chronic treatment or as "acute quadriplegic" myopathy secondary to high-dose IV glucocorticoid use. Chronic administration produces proximal weakness accompanied by cushingoid manifestations, which can be quite debilitating; the chronic use of prednisone at a daily dose of ≥30 mg/d is most often associated with toxicity. Patients taking fluorinated glucocorticoids (triamcinolone, betamethasone, dexamethasone) appear to be at especially high risk for myopathy. In chronic steroid myopathy, the serum CK is usually normal. Serum potassium may be low. The muscle biopsy in chronic cases shows preferential type 2 muscle fiber atrophy; this is not reflected in the EMG, which is usually normal.

Patients receiving high-dose IV glucocorticoids for status asthmaticus, chronic obstructive pulmonary disease, organ transplantation, or other indications may develop severe generalized weakness (critical illness myopathy). This myopathy, also known as acute quadriplegic myopathy, can also occur in the setting of sepsis. Involvement of the diaphragm and intercostal muscles causes ventilatory muscle weakness and is usually appreciated when patients are unable to be weaned off a ventilatory in the ICU. NCS demonstrate reduced compound muscle action potentials in the setting of relatively preserved sensory potentials. EMG can demonstrate abnormal insertional and spontaneous activity and early recruitment of myopathic appearing units in those muscles that can be activated. Muscle biopsy can show a distinctive loss of thick filaments (myosin) by electron microscopy. Treatment is withdrawal of glucocorticoids and physical therapy but the recovery is slow. Patients require supportive care and rehabilitation.

OTHER DRUG-INDUCED MYOPATHIES

Certain drugs produce painless, largely proximal, muscle weakness. These drugs include the amphophilic cationic drugs (amiodarone, chloroquine, hydroxychloroquine) and antimicrotubular drugs (colchicine) (Table 441-6). Muscle biopsy can be useful in the identification of toxicity because autophagic vacuoles are prominent pathologic features of these toxins.

GLOBAL ISSUES

As previously discussed, certain dystrophies have an increased prevalence in different parts of the world. LGMD2A is the most common LGMD in individuals from Spain, France, Italy, and Great Britain; LGMD2I is more common in those with northern European ancestry. GNE myopathy is the most common form of distal myopathy in Japan but is also prevalent in the Ashkenazi population. OPMD is most comment in those with ancestry from Spain and French-Canada as well as among Ashkenazi. Epidemiological studies are lacking regarding other forms of myopathy and their prevalence in different areas of the world.

FURTHER READING

Amato AA, Russell JA (eds): *Neuromuscular Disorders*, 2nd ed. New York, McGraw-Hill Education, 2016.

Finsterer J et al: Secondary myopathy due to systemic diseases. Acta Neurol Scand 134:388, 2016.

Iyadurai SJ, Kissel JT: The limb-girdle muscular dystrophies and the dystrophinopathies. Continuum (Minneap Minn) 22:1954, 2016.

Katzberg HD, Kassardjian CD: Toxic and endocrine myopathies. Continuum (Minneap Minn) 22:1815, 2016.

Mah JK et al: A systematic review and meta-analysis on the epidemiology of the muscular dystrophies. Can J Neurol Sci 43:163, 2016.

Narayanaswami P et al: Summary of evidence-based guideline: Diagnosis and treatment limb-girdle and distal muscular dystrophies. Neurology 83:1453, 2014.

Rosow LK, Amato AA: The role of electrodiagnostic testing, imaging, and muscle biopsy in the investigation of muscle disease. Continuum (Minneap Minn) 22:1787, 2016.

Sansone VA: The dystrophic and nondystrophic myotonias. Continuum (Minneap Minn) 22:1889, 2016.

Sansone VA et al: Randomized, placebo-controlled trials of dichlorphenamide in periodic paralysis. Neurology 86:1408, 2016.
Statland JM, Tawil R: Facioscapulohumeral muscular dystrophy. Continuum (Minneap Minn) 22:1916, 2016.
Tawil R et al: Evidence-based guideline summary: Evaluation, diagnosis, and management of facioscapulohumeral muscular dystrophy: Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology and the Practice Issues Review Panel of the American Association of Neuromuscular & Electrodiagnostic Medicine. Neurology 85:357, 2015.

Section 4 Chronic Fatigue Syndrome

442 Chronic Fatigue Syndrome

Gijs Bleijenberg, Jos W. M. van der Meer

DEFINITION

Chronic fatigue syndrome (CFS) is a disorder characterized by persistent and unexplained fatigue resulting in severe impairment in daily functioning. Besides intense fatigue, most patients with CFS report concomitant symptoms such as pain, cognitive dysfunction, and unrefreshing sleep. Additional symptoms can include headache, sore throat, tender lymph nodes, muscle aches, joint aches, post-exertional malaise, feverishness, difficulty sleeping, psychiatric problems, allergies, and abdominal cramps. Criteria for the diagnosis of CFS have been developed by the U.S. Centers for Disease Control and Prevention **(Table 442-1)**. The Institute of Medicine (IOM; now the National Academy of Medicine) has recently changed the diagnostic criteria and proposed the name systemic exercise intolerance disease (SEID). To date, however, no studies have been reported to demonstrate the usefulness of this change.

EPIDEMIOLOGY

CFS is seen worldwide, with adult prevalence rates varying between 0.2 and 0.4%. In the United States, the prevalence is higher among women (~75% of cases), members of minority groups (African and Native Americans), and individuals with lower levels of education and occupational status. CFS has been reported to be associated with an increase in mortality from suicide. The mean age of onset is between 29 and 35 years. Many patients probably go undiagnosed and/or do not seek help.

TABLE 442-1 Diagnostic Criteria for Chronic Fatigue Syndrome

Characteristic Persistent or Relapsing Unexplained Chronic Fatigue
Fatigue lasts for at least 6 months.
Fatigue is of new or definite onset.
Fatigue is not the result of an organic disease or of continuing exertion.
Fatigue is not alleviated by rest.
Fatigue results in a substantial reduction in previous occupational, educational, social, and personal activities.
Four or more of the following symptoms are concurrently present for 6 months: impaired memory or concentration, sore throat, tender cervical or axillary lymph nodes, muscle pain, pain in several joints, new headaches, unrefreshing sleep, or malaise after exertion.
Exclusion Criteria
Medical condition explaining fatigue
Major depressive disorder (psychotic features) or bipolar disorder
Schizophrenia, dementia, or delusional disorder
Anorexia nervosa, bulimia nervosa
Alcohol or substance abuse
Severe obesity (body mass index >40)

TABLE 442-2 Predisposing, Precipitating, and Perpetuating Factors in Chronic Fatigue Syndrome

Predisposing Factors
Childhood trauma (sexual, physical, emotional abuse; emotional and physical neglect)
Physical inactivity during childhood
Premorbid psychiatric illness or psychopathology
Premorbid hyperactivity
↓
Precipitating Factors
Somatic events: infection (e.g., mononucleosis, Q fever, Lyme disease), surgery, pregnancy
Psychosocial stress, life events
↓
Perpetuating Factors
Non-acknowledgment by physician
Negative self-efficacy
Strong physical attributions
Strong focus on bodily symptoms
Fear of fatigue
Lack of social support
Low physical activity pattern

ETIOLOGY

There are numerous hypotheses about the etiology of CFS; there is no definitively identified cause. Distinguishing between predisposing, precipitating, and perpetuating factors in CFS helps to provide a framework for understanding this complex condition **(Table 442-2)**.

Predisposing Factors Physical inactivity and trauma in childhood tend to increase the risk of CFS in adults. Neuroendocrine dysfunction may be associated with childhood trauma, reflecting a biological correlate of vulnerability. Psychiatric illness and physical hyperactivity in adulthood raise the risk of CFS in later life. Twin studies suggest a familial predisposition to CFS, but research into causative genes has yielded variable results. In a recent systematic review, a number of single nucleotide polymorphisms encoding for TNFalpha, IL-1beta, IL-6 and IL-4 were found to be associated with increased fatigue in CFS, cancer-related fatigue, and other disease-related fatigue states.

Precipitating Factors Physical or psychological stress may elicit the onset of CFS. A substantial number of patients report an infection (often a flulike illness or infectious mononucleosis) as the trigger of their fatigue. Relatively high percentages of CFS cases follow Q fever and Lyme disease. However, no differences in Epstein-Barr virus load and immunologic reactivity were found between individuals who developed CFS and those who did not. While antecedent infections are associated with CFS, a direct microbial causality is unproven and unlikely. Patients also often report other precipitating somatic events such as serious injury, surgery, pregnancy, or childbirth. Serious life events, such as the loss of a loved one or a job, military combat, and other stressful situations, may also precipitate CFS. One-third of all patients cannot recall a trigger.

Perpetuating Factors. Once CFS has developed, numerous factors may impede recovery. Physicians may contribute to chronicity by ordering unnecessary diagnostic procedures, by persistently suggesting psychological causes, and by not acknowledging CFS as a diagnosis.

A patient's focus on symptoms and avoidance of activities may perpetuate symptoms. A firm belief in a physical cause, a strong focus on bodily sensations, and a poor sense of control over symptoms may also

prolong or exacerbate the fatigue and functional impairment. In most patients, inactivity is caused by negative illness perceptions rather than by poor physical fitness. Solicitous behavior of others may reinforce a patient's illness-related perceptions and behavior. A lack of social support is another known perpetuating factor.

PATHOPHYSIOLOGY

The pathophysiology of CFS is unclear. Neuroimaging studies have suggested that CFS is associated with reduced gray matter volume, but recent studies seem to indicate that pain rather than fatigue is associated with these changes. Functional MRI data have suggested that abnormal patterns of activation correlate with self-reported problems with information processing. Neurophysiologic studies have shown altered CNS activation patterns during muscle contraction.

Evidence for immunologic dysfunction is inconsistent. Modest elevations in titers of antinuclear antibodies, reductions in immunoglobulin subclasses, deficiencies in mitogen-driven lymphocyte proliferation, reductions in natural killer cell activity, disturbances in cytokine production, and shifts in lymphocyte subsets have been described. None of these immune findings has been firmly established and none of these changes appear in most patients, nor does any correlate with the severity of CFS. In theory, symptoms of CFS could result from excessive production of a cytokine, such as interleukin 1, that induces asthenia and other flulike symptoms; however, compelling data in support of this hypothesis are lacking. There is some evidence that CFS patients have mild hypocortisolism, the degree of which is associated with a poorer response to cognitive-behavioral therapy (CBT). Discrepancies in perceived and actual cognitive performance are consistent findings in patients with CFS.

DIAGNOSIS

In addition to a thorough history, a systematic physical examination is warranted to exclude disorders causing fatigue (e.g., endocrine disorders, neoplasms, heart failure). The heart rate of CFS patients is often slightly above normal, but postural hypotension, which has been included in the diagnostic criteria put forward by the IOM (see above), is not more common in CFS patients than in controls. Laboratory tests serve primarily to exclude other diagnoses; no test can diagnose CFS. The following laboratory screen usually suffices: complete blood count; erythrocyte sedimentation rate; C-reactive protein; serum creatinine, electrolytes, calcium, and iron; blood glucose; creatine kinase; liver function tests; thyroid-stimulating hormone; anti-gliadin antibodies; and urinalysis. Serology for viral or bacterial infections usually is not helpful. No diagnostic abnormalities have been identified on MRI or CT scans. Extensive, unfocused, and expensive testing in a search for the "hidden" cause of the fatigue is not productive. CFS is a constellation of symptoms with no pathognomonic features and remains a diagnosis of exclusion.

Bipolar disorders, schizophrenia, and substance abuse exclude a diagnosis of CFS, as do eating disorders, unless these health problems have been resolved ≥5 years before symptom onset. In addition, CFS is excluded if the chronic fatigue developed immediately after a depressive episode. Depression developing in the course of the fatigue, however, does not preclude CFS. Concurrent psychiatric disorders, especially anxiety and mood disorders, are present in 30–60% of cases.

INITIAL MANAGEMENT

In cases of suspected CFS, the clinician should acknowledge the impact of the patient's symptoms on daily functioning. Disbelief or denial can provoke an exacerbation of genuine symptoms, which in turn strengthens the clinician's disbelief, leading to an unfortunate cycle of miscommunication. The possibility of CFS should be considered if a patient fulfills all criteria (Table 442-1) and if other diagnoses have been excluded.

The patient should be asked to describe the symptoms (fatigue and accompanying symptoms) and their duration as well as their consequences (reduction in daily activities). To assess symptom severity and the extent of daily-life impairment, the patient should describe a typical day, from waking to retiring, and, for comparison, an average day prior to symptom onset. Next, potential fatigue-precipitating factors are sought. The severity of fatigue is commonly difficult to assess quantitatively; a brief questionnaire is often helpful **(Fig. 442-1)**.

The patient should be informed of the current understanding of precipitating and perpetuating factors and effective treatments and provided general advice about disease management. If CBT for CFS is not available as an initial option (see below) and depression and anxiety are present, these symptoms should be treated. For patients with headache, diffuse pain, and feverishness, nonsteroidal anti-inflammatory drugs may be helpful. Even modest improvements in symptoms can

How have you felt during the last two weeks?

Please rate all four statements and per statement check the box that reflects your situation best.

		Yes, that is true							No, that is not true
1.	I feel tired								
2.	I tire easily								
3.	I feel fit								
4.	Physically I feel exhausted								

Scoring:
1, 2 and 4:

Yes, that is true							No, that is not true
7	6	5	4	3	2	1	

3: Reversed

Sum scores >18 indicate severe fatigue

FIGURE 442-1 Shortened fatigue questionnaire.

make an important difference in the patient's degree of self-sufficiency and ability to appreciate life's pleasures.

Controlled therapeutic trials have established that acyclovir, fludrocortisone, galantamine, modafinil, and IV immunoglobulin, among other agents, offer no significant benefit in CFS. Countless anecdotes circulate regarding other traditional and nontraditional therapies. It is important to guide patients away from those therapeutic modalities that are toxic, expensive, or unreasonable.

The patient should be encouraged to maintain regular sleep patterns, to remain as active as possible, and to gradually return to previous levels of exercise and other activity (work).

TREATMENT

Chronic Fatigue Syndrome

CBT and graded exercise therapy (GET) have been found to be the only beneficial interventions in CFS. Some patient groups argue against these approaches because of the implication that CFS is a purely mental disorder. CBT is a psychotherapeutic approach directed at changing unhealthy disease-perpetuating patterns of thoughts and behaviors. It includes educating the patient about the etiologic model, setting goals, restoring fixed bedtimes and wake-up times, challenging and changing fatigue- and activity-related concerns, reducing a focus on symptoms, spreading activities evenly throughout the day, gradually increasing physical activity, planning a return to work, and resuming other activities. The intervention, typically consisting of 12–14 sessions over 6 months performed by an experienced cognitive behavior therapist, helps CFS patients gain control over their symptoms.

GET targets deconditioning and exercise intolerance and usually involves a home exercise program that continues for 3–5 months. Walking or cycling is systematically increased, with set goals for maximal heart rates. Evidence that deconditioning is the basis for symptoms in CFS is lacking, however. CBT and GET appear to improve fatigue primarily by changing the patient's perception of the fatigue and also by reducing the focus on symptoms.

CBT and GET seem equally effective, however not all patients benefit from these interventions. Predictors of poor outcome are insufficient motivation for the treatment, medical (including psychiatric) comorbidities, current disability claims, and severe pain. CBT offered in an early stage of the illness reduces the burden of CFS for the patient as well as for society in terms of decreased medical and disability-related costs.

PROGNOSIS

Full recovery from untreated CFS is rare: the median annual recovery rate is 5% (range, 0–31%), and the median improvement rate is 39% (range, 8–63%). Patients with an underlying psychiatric disorder and those who continue to attribute their symptoms to an undiagnosed medical condition have poorer outcomes.

FURTHER READING

Blundell S et al: Chronic fatigue syndrome and circulating cytokines: A systematic review. Brain Behav Immun 50:186, 2015.

Castell BD et al: Cognitive behavioral therapy and graded exercise for chronic fatigue syndrome: A meta-analysis. Clin Psychol Sci Prac 18:311, 2011.

Committee on the Diagnostic Criteria for Myalgic Encephalomyelitis/Chronic Fatigue Syndrome; Board on the Health of Select Populations; Institute of Medicine. Beyond myalgic encephalomyelitis/chronic fatigue syndrome: Redefining an illness. National Academies Press, Washington, DC, 2015.

Larun L et al: Exercise therapy for chronic fatigue syndrome. Cochrane Database Syst Rev 12:CD003200, 2016.

443 Biology of Psychiatric Disorders

Robert O. Messing, Eric J. Nestler, Matthew W. State

Psychiatric disorders are central nervous system diseases characterized by disturbances in emotion, cognition, motivation, and socialization. They are highly heritable, with genetic risk comprising 20–90% of disease vulnerability. As a result of their prevalence, early onset, and persistence, they contribute substantially to the burden of illness worldwide. All psychiatric disorders are broad heterogeneous syndromes that currently lack well-defined neuropathology and bona fide biologic markers. Therefore, diagnoses continue to be made solely from clinical observations using criteria in the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)*, of the American Psychiatric Association.

There is increasing agreement that the classification of psychiatric illnesses in DSM does not accurately reflect the underlying biology of these disorders. Uncertainties in diagnosis complicate efforts to study the genetic basis and attendant neurobiological mechanisms underlying mental illness, though recent advances in genomic and neuroscience technologies along with the consolidation of very large patient cohorts have, for multiple disorders, led to major progress in these realms. In addition, there have been recent efforts to address the limitations of a categorical nosology directly through the development of an alternative diagnostic scheme, termed Research Domain Criteria (RDoC). This system classifies mental illness on the basis of core behavioral abnormalities shared across several syndromes—such as psychosis (loss of reality) or anhedonia (decreased ability to experience pleasure)—and the associated brain circuitry that controls these behavioral domains. It is anticipated that that such classifications will assist in defining the biologic basis of key symptoms. Other factors that have impeded progress in understanding mental illness include the lack of access to pathologic brain tissue except upon death and the inherent limitations of animal models for disorders defined largely by behavioral abnormalities (e.g., hallucinations, delusions, guilt, suicidality) that are inaccessible in animals.

Despite these limitations, the past decade has been marked by real progress. Neuroimaging methods are beginning to provide evidence of brain pathology; genome-wide association studies and high-throughput sequencing are reliably identifying genes and genomic loci that confer risk for severe forms of mental illness; and investigations of better validated animal models, leveraging a host of new methods to study molecular, cellular and circuit level processes, are offering new insight into disease pathogenesis. There is also excitement in the utility of neurons and brain organoids induced in vitro from patient-derived pluripotent stem cells, providing novel ways to study disease pathophysiology and screen for new treatments. There is consequently justified optimism that the field of psychiatry will better integrate behaviorally defined syndromes with an understanding of biological substrates in a way that will drive the development of improved treatments and eventually cures and preventive measures. This chapter describes several examples of recent discoveries in basic neuroscience and genetics that have informed our current understanding of disease mechanisms in psychiatry.

NEUROGENETICS

Because the human brain can only be examined indirectly during life, genome analyses have been extremely important for obtaining molecular clues about the pathogenesis of psychiatric disorders. Moreover,

the identification of germ-line risk alleles and mutations provides potential traction on the question of cause versus effect. In other types of cross-sectional studies, it may be impossible to determine whether a phenotype or biomarker observed in affected humans or model systems reflects an etiological factor or a compensatory response. In contrast, germ-line genetic risk is present before the brain develops—at least theoretically allowing for experiments to address temporal sequencing.

A wealth of new information has been made possible by recent technological developments that have permitted affordable, large-scale genome-wide association studies and high-throughput sequencing. As an example of the latter, significant progress has been made in the genetics of autism spectrum disorders (ASDs), which are a heterogeneous group of neurodevelopmental diseases that share clinical features of impaired social communication and restricted, repetitive patterns of behavior. ASDs are highly heritable; concordance rates in monozygotic twins (~60–90%) are five- to tenfold higher than in dizygotic twins and siblings, and first-degree relatives show approximately tenfold increased risk compared with the general population. ASDs are also genetically heterogeneous. At present, ~70 individual risk genes, along with dozens of submicroscopic deletions and duplications often containing multiple genes, have been identified, almost exclusively through the study of rare, large-effect new (de novo) mutations **(Fig. 443-1)**. All told, genes and genomic regions vulnerable to these types of mutations account for about 20–30% of formerly idiopathic cases that present in the clinic, although none individually accounts for >1%. In addition, ~10% of individuals with ASD have well-described intellectual disability syndromes including *fragile X, Rett syndrome* and *tuberous sclerosis* **(Chap. 86)**. However, it appears that most of the risk for ASD in the population involves true polygenic inheritance. There is considerable evidence, for example, that >50% of the genetic liability is carried in common alleles of very small individual effect. To date, however, studies of many thousands of cases have yet to identify a reproducible association of a specific nucleotide polymorphism (SNP) using gold standard genome-wide association methods—although with continually increasing cohort sizes, and thus power, this is certain to change in the near future.

Amidst the genetic heterogeneity that has so far been identified, common themes have emerged that inform pathogenesis of ASDs. For instance, many identified rare mutations are in genes that encode proteins involved in synaptic function and early transcriptional regulation (Fig. 443-1) and have a clear relationship to activity-dependent neural responses that can affect the development of neural systems underlying cognition and social behaviors. One particularly intriguing hypothesis is that these genes may lead to ASD risk by changing the balance of excitatory versus inhibitory synaptic signaling in local and extended circuits and by altering mechanisms that control brain growth. Some mutations affect genes (e.g., *PTEN, TSC1*, and *TSC2*) that negatively regulate signaling from several types of extracellular stimuli, including those transduced by receptor tyrosine kinases. Their dysregulation can alter neuronal growth as well as synaptic development and function. Finally, several recent studies have focused on the question of when and where multiple functionally diverse risk genes converge with respect to human brain development. Interestingly, these studies have thus far tended to overlap with expression patterns of glutamatergic neurons in mid-fetal cortex (Fig. 443-1). Given the pleiotropic biological effects of the ASD genes identified to date, an understanding of the developmental or "spatiotemporal" dimensions of risk is likely to serve as a useful complement to studies of the function of individual genes. In short, it may turn out that when and where genetic variation has its impact in the developing brain may be as important as the key processes that are identified.

With further understanding of pathogenesis and the definition of specific ASD subtypes, there is reason to believe that effective therapies will be identified. Work in mouse models has already demonstrated that some autism-like behaviors can be reversed, even in fully developed adult animals, by modifying the underlying pathology; these results encourage hope for many affected individuals. Treatments that target excitation-inhibition imbalance or altered mRNA translation appear to offer early promise. For example, the genes *TSC1, TSC2*, and *PTEN* are negative regulators of signaling through the target of rapamycin complex 1 (TORC1), which regulates protein synthesis. Rapamycin, a selective inhibitor of TORC1, can reverse several behavioral and synaptic defects in mice carrying null mutations in these genes. Another example is fragile X syndrome, which is the leading cause of inherited autism and mental disability and is due to mutations in *FMR1* that result in loss of the encoded fragile X mental retardation protein (FMRP). FMRP is a polyribosome-associated mRNA-binding protein that represses the translation of a subset (~5%) of all mRNAs, several of which encode proteins that comprise the postsynaptic density, including the metabotropic glutamate receptor 5 (mGluR5). Treatment of *Fmr1* knockout mice with mGluR5 antagonists reduces several behavioral and morphologic abnormalities in these mice; these promising preclinical results have led to ongoing trials of mGluR5 antagonists in humans with fragile X and other ASD syndromes. While early clinical data have been disappointing, this work nonetheless illustrates a potential path forward in therapeutics development.

The ability to catalog common genetic variants and assay them on array-based platforms and, more recently, to carry out whole-exome sequencing has allowed investigators to leverage very large patient cohorts to detect genetic risk loci for schizophrenia and bipolar disorder with genome-wide significance. In contrast to ASD, where the lion's share of early success in gene identification has resulted from the study of rare large effect de novo mutations, much of gene discovery to date for these syndromes has resulted from genome-wide association studies of common inherited polymorphisms. It is noteworthy that there is also striking overlap among the submicroscopic deletions and duplications, called copy number variants (CNVs), that have been found to carry large risks for ASD, schizophrenia, and bipolar disorders, as well as epilepsy and intellectual disability.

To date more than a hundred distinct genomic regions, marked by associated SNPs, have been identified in schizophrenia, some of which show risk as well for bipolar disorder. Several of the identified genes are parts of molecular complexes, such as voltage-gated calcium channels (in particular, *CACNA1C* and *CACNB2*) and the postsynaptic density of excitatory synapses. Genes that promote risk for addiction and depression have also begun to emerge from large studies. The best-established susceptibility locus for addiction is the *CHRNA5-A3-B4* nicotinic acetylcholine receptor gene cluster on chromosome 15 associated with nicotine and alcohol addiction. Recent genome-wide association studies of depression have required hundreds of thousands of cases and controls to identify the first statistically significant loci using state-of-the-art approaches. These findings collectively point to the tremendous heterogeneity of depressive disorders as well as the very small biological effects conferred by any individual common allele.

A recurrent theme that has emerged from genetic studies of psychiatric disorders is phenotypic pleiotropy, namely, that many genes are associated with multiple psychiatric syndromes. For example, mutations in *MECP2, FMR1*, and *TSC1* and *TSC2* (see **Table 443-1** for abbreviations) can cause mental retardation without ASD, others in *MECP2* can cause obsessive-compulsive and attention-deficit hyperactivity disorders, some alleles of *NRXN1* are associated with symptoms of both ASD and schizophrenia, and common polymorphisms in *CACNA1C* are strongly associated with both schizophrenia and bipolar disorder. Likewise, duplication of chromosome 16p is associated with both schizophrenia and autism, whereas deletions in the DiGeorge's (velocardiofacial) syndrome region are associated with schizophrenia, autism, and bipolar disorder. The association of genes and genomic regions with multiple syndromes attests to the complexity of psychiatric disorders, the very large gap between molecular mechanisms and the current categorical diagnostic schemes, and the influence of additional factors that combine to specify the ultimate phenotype. The latter might include polygenic "background," variations in regulatory regions of the genome that determine cell-type specificity and timing of gene expression, protective variants, stochastic events, and epigenetic effects.

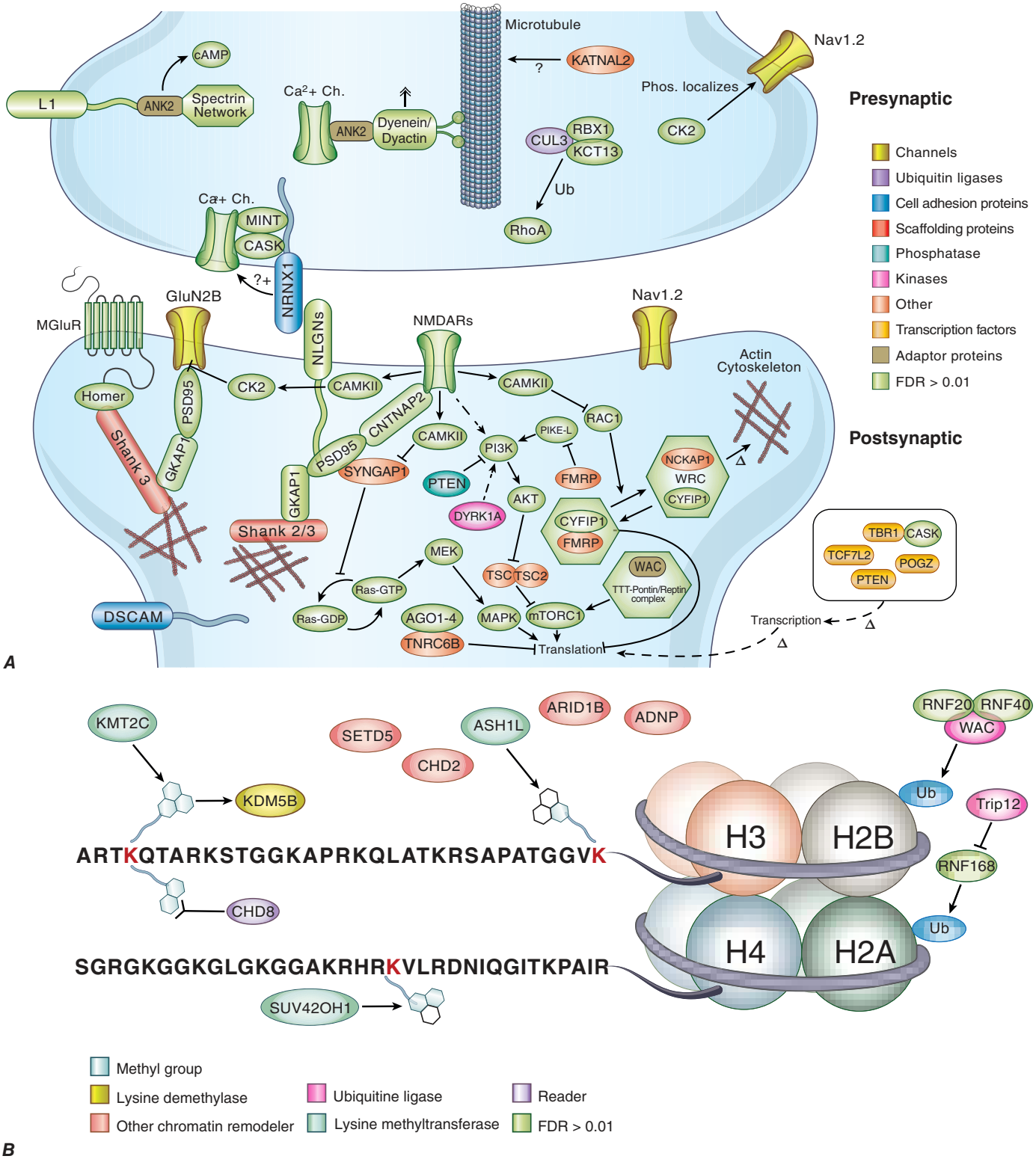


FIGURE 443-1 Functional characteristics and developmental convergence of autism spectrum disorders (ASDs) associated genes: genes associated with risk for ASD based on recurrent de novo mutations are shown in ***A*** and ***B*** (Sanders et al: *Neuron* 2015). Those genes encoding proteins meeting criteria for the highest confidence statistical association (false discovery rate [FDR] < 0.01) are highlighted with respect to their putative functions. Additional interacting and functionally related molecules that do not meet this threshold are shown in green. As a group, genes with FDR <0.01 carry large effects, conveying approximately a twentyfold increase in risk. Multiple gene ontology analyses of ASD genes have highlighted both pre- and postsynaptic molecules (***A***) and chromatin modifiers (***B***) as points of enrichment. In ***C***, an alternative strategy for grouping ASD risk genes is highlighted, based on their spatiotemporal expression patterns as opposed to putative functions. One analytic strategy, illustrated in ***C***, leveraged only high confidence ASD genes and examined their developmental expression patterns using the BrainSpan dataset. Convergence for ASD risk was identified in deep layer (V and VI) excitatory neurons in mid-fetal human cortex. Multiple analyses have similarly found glutamatergic neurons in mid-fetal prefrontal cortex as one point of convergence, with somewhat less agreement on layer-specificity and potential additional spatiotemporal points of convergence. *(Figure drawn by Montana Morris and Sarah Pyle.)*

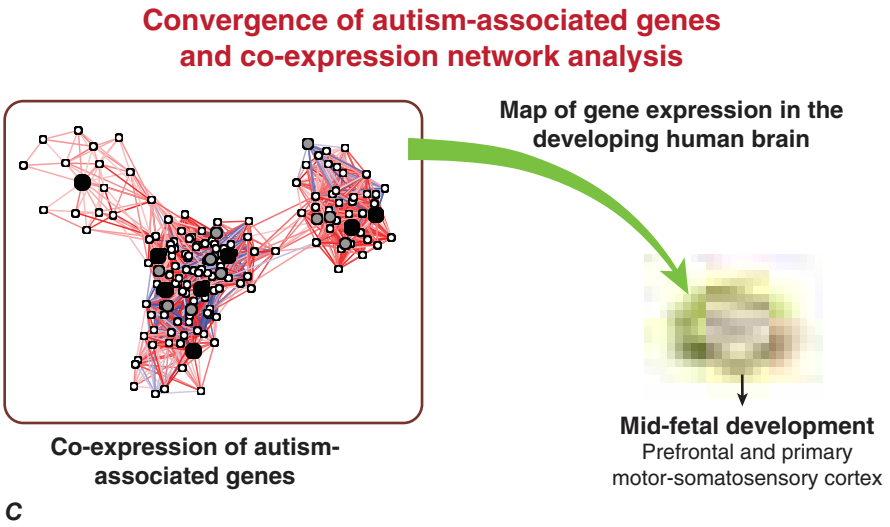


FIGURE 443-1 (*Continued*)

SIGNAL TRANSDUCTION

Studies of signal transduction have revealed numerous intracellular signaling pathways that are perturbed in psychiatric disorders, and such research has provided insight into development of new therapeutic agents. For example, lithium is a highly effective drug for bipolar disorder and competes with magnesium to inhibit numerous magnesium-dependent enzymes, including the enzyme GSK3β and several enzymes involved in phosphoinositide signaling that lead to activation of protein kinase C. These findings have led to discovery programs focused on developing GSK3β or protein kinase C inhibitors as potential novel treatments for mood disorders, although none have demonstrated clinical efficacy to date.

The observations that tricyclic antidepressants (e.g., imipramine) inhibit serotonin and/or norepinephrine reuptake and that monoamine oxidase inhibitors (e.g., tranylcypromine) are effective antidepressants initially led to the view that depression is caused by a deficiency of these monoamines. However, this hypothesis has not been substantiated. A cardinal feature of these drugs is that long-term (weeks to months) administration is needed for their antidepressant effects. This means that their short-term actions, namely promotion of serotonin or norepinephrine function, are not per se antidepressant but rather induce a cascade of adaptations in the brain that underlie their slowly developing clinical effects. The nature of these therapeutic drug-induced adaptations has not been identified with certainty. One theory holds that, in a subset of depressed patients who display upregulation of the hypothalamic-pituitary-adrenal (HPA) axis characterized by increased secretion of corticotropin-releasing factor (CRF) and glucocorticoids, excessive glucocorticoids cause atrophy of hippocampal neurons, which is associated with reduced hippocampal volumes seen clinically. Chronic antidepressant administration might reverse this atrophy by increasing brain-derived neurotrophic factor (BDNF) or a host of other neurotrophic factors in hippocampus. A role for stress-induced decreases in the generation of newly born hippocampal granule cell neurons, and its reversal by antidepressants through BDNF or other growth factors, has also been suggested.

A major advance in recent years has been the identification of several rapidly acting antidepressants with non-monoamine-based mechanisms of action. The best established is ketamine, a noncompetitive antagonist of *N*-methyl-D-aspartate (NMDA) glutamate receptors, which exerts rapid (hours) and robust antidepressant effects in severely depressed patients who have not responded to other treatments. Ketamine, which at higher doses is psychotomimetic and anesthetic, exerts these antidepressant effects at lower doses with minimal side effects. However, the response to ketamine is transient, which has led to several approaches to maintain treatment response, such as repeated ketamine delivery. The mechanism underlying ketamine's antidepressant action is not known, and its action as an NMDA receptor antagonist has recently been called into question. Nevertheless, ketamine's striking clinical efficacy has stimulated animal research on the role of glutamate neurotransmission and synaptic plasticity in key limbic regions. Recent evidence supports a role for TORC1 or BDNF activation, as blockade of either blocks the antidepressant-like effects of ketamine in animal models. Mechanisms by which ketamine activates these signaling cascades are currently an active area of investigation.

A major goal in the field of drug abuse has been to identify neuroadaptive mechanisms that lead from recreational use to addiction. Such research has determined that repeated intake of abused drugs induces specific changes in cellular signal transduction, leading to changes in synaptic strength (long-term potentiation or depression) and neuronal structure (altered dendritic branching or cell soma size) within the brain's reward circuitry. These drug-induced modifications are mediated in part by changes in gene expression, achieved by regulation of transcription factors (e.g., CREB [cAMP response element-binding protein] and ΔFosB [a Fos family protein]) and their target genes. Such alterations in gene expression are associated with lasting alterations in epigenetic modifications, including histone acetylation and

TABLE 443-1 Initial Actions of Drugs of Abuse

DRUG	NEUROTRANSMITTER AFFECTED	DRUG TARGET (ACTION)
Opiates	Endorphins, enkephalins	μ- and δ-opioid receptors (agonist)
Psychostimulants (cocaine, amphetamine, methamphetamine)	Dopamine	Dopamine transporter (antagonist—cocaine; reverse transport—amphetamine, methamphetamine)
Nicotine	Acetylcholine	Nicotinic cholinergic receptors (agonist)
Ethanol	GABA	$GABA_A$ receptors (positive allosteric modulator)
	Glutamate	NMDA glutamate receptors (antagonist)
	Acetylcholine	Nicotinic cholinergic receptors (allosteric modulator)
	Serotonin	5HT-3 receptor (positive allosteric modulator)
	Others	Calcium-activated K^+ channel (activator)
Marijuana	Endocannabinoids (anandamide, 2-arachidonoylglycerol)	CB_1 receptor (agonist)
Phencyclidine	Glutamate	NMDA glutamate receptor (antagonist)

Abbreviations: GABA, γ-aminobutyric acid; NMDA, *N*-methyl-D-aspartate

methylation and DNA methylation. These adaptations provide opportunities for developing treatments targeted to drug-addicted individuals. The fact that the spectrum of these adaptations differs in part depending on the particular addictive substance used raises hope that treatments could be developed that are specific for different classes of addictive drugs and less likely to disturb basic mechanisms that govern normal motivation and reward.

Increasingly, causal relationships are being established between individual molecular and cellular adaptations and specific behavioral abnormalities that characterize the addicted state. For example, acute activation of μ-opioid receptors by morphine or other opiates activates $G_{i/o}$ proteins, leading to inhibition of adenylyl cyclase (AC), resulting in reduced cyclic AMP (cAMP) production, protein kinase A (PKA) activation, and activation of the transcription factor CREB. Repeated administration of these drugs **(Fig. 443-2)** evokes a homeostatic response involving upregulation of ACs and PKA and increased activation of CREB. Such upregulation of cAMP-CREB signaling has been identified in the locus coeruleus (LC), periaqueductal gray, ventral tegmental area (VTA), nucleus accumbens (NAc), and several other CNS regions, and contributes to opiate craving and signs of opiate withdrawal. The fact that endogenous opioid peptides do not produce tolerance and dependence, while morphine and heroin do, may relate to the observation that, unlike endogenous opioids, morphine and heroin are weak inducers of μ-opioid receptor desensitization and endocytosis. Therefore, these drugs cause prolonged receptor activation and inhibition of ACs, which provides a powerful stimulus for the upregulation of cAMP-CREB signaling that characterizes the opiate-dependent state.

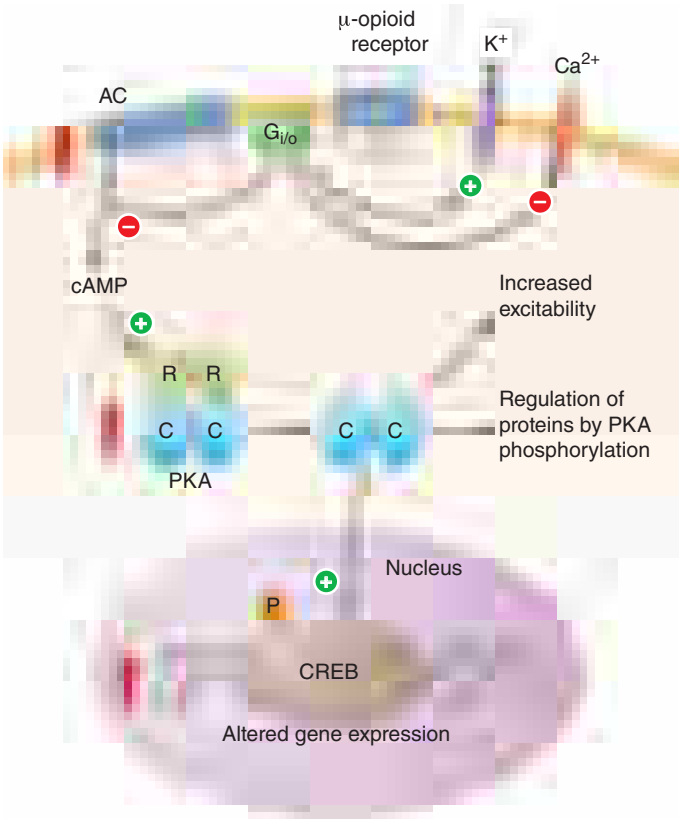


FIGURE 443-2 Opiate action in the locus coeruleus (LC). Binding of opiate agonists to μ-opioid receptors catalyzes nucleotide exchange on G_i and G_o proteins, leading to inhibition of adenylyl cyclase (AC), neuronal hyperpolarization via activation of K^+ channels, and inhibition of neurotransmitter release via inhibition of Ca^{2+} channels. Inhibition of AC reduces protein kinase A (PKA) activity and phosphorylation of several PKA substrate proteins, thereby altering their function. For example, opiates reduce phosphorylation of the cAMP response element-binding protein (CREB), which initiates longer term changes in neuronal function. Chronic administration of opiates increases levels of AC isoforms, PKA catalytic (C) and regulatory (R) subunits, and the phosphorylation of several proteins, including CREB (indicated by *red arrows*). These changes contribute to the altered phenotype of the drug-addicted state. For example, the excitability of LC neurons is increased by enhanced cAMP signaling. Activation of CREB causes upregulation of AC isoforms and tyrosine hydroxylase, the rate-limiting enzyme in catecholamine biosynthesis.

SYSTEMS NEUROSCIENCE

The study of interconnected brain circuits that drive behavior has been greatly advanced through newer methods in brain imaging that have documented abnormalities in neural function and connectivity in psychiatric disorders. Electroceutical devices, which use electrical or magnetic stimulation to control neuronal activity, have had some success in depression, obsessive compulsive disorder, pain, and addiction. The past decade has also witnessed the development of revolutionary new techniques—optogenetics, designer receptors and ligands—that provide unprecedented temporal and spatial control of neural circuits. The development of genetically encoded calcium detectors and electrode arrays has allowed in vivo monitoring of thousands of neurons in multiple brain regions simultaneously. Advances in histology and microscopy now permit three-dimensional imaging of specific proteins in the intact brain, while advances in endoscopic microscopy allow imaging of hundreds of neurons within deep brain structures in awake, freely moving animals. These new methods promise to revolutionize our ability to understand the circuit basis of brain function.

Positron emission tomography (PET), diffusion tensor imaging (DTI), and functional magnetic resonance imaging (fMRI) have identified neural circuits that contribute to psychiatric disorders, for example, defining the neural circuitry of mood within the brain's limbic system **(Fig. 443-3)**. Integral to this system are the NAc (important also for brain reward—see below), amygdala, hippocampus, and regions of prefrontal cortex. Recent optogenetic research in animals, where the activity of specific types of neurons in defined circuits can be controlled with light, has confirmed the importance of this limbic circuitry in controlling depression-related behavioral abnormalities. Given that many symptoms of depression (so-called neurovegetative symptoms) involve physiologic functions, a key role for the hypothalamus is presumed as well. A subset of depressed individuals shows a small reduction in hippocampal size, as noted above. In addition, brain imaging investigations have revealed increased activation of the amygdala by negative stimuli and reduced activation of the NAc by rewarding stimuli. There is also evidence for altered activity in prefrontal cortex, such as hyperactivity of subgenual area 25 in anterior cingulate cortex. Such findings have led to trials of deep brain stimulation (DBS) of either the NAc or subgenual area 25, which appears to be therapeutic in some severely depressed individuals.

In schizophrenia, structural and functional imaging studies have confirmed earlier pathologic studies that show enlargement of the ventricular system and reduction of cortical and subcortical gray matter in frontal and temporal lobes and in the limbic system. Functional imaging studies show reduced metabolic (presumably neural) activity in the dorsolateral prefrontal cortex at rest and when performing tests of executive function, including working memory. There is also evidence for impaired structural and task-related functional connectivity, mainly in frontal and temporal lobes. The reduction in cortical thickness seen in schizophrenia is associated with increased cell packing density and reduced neuropil (defined as axons, dendrites, and glial cell processes) without an apparent change in neuronal cell number. Specific classes of interneurons in prefrontal cortex consistently show reduced expression of the gene encoding the enzyme glutamic acid decarboxylase 1 (*GAD1*), which synthesizes γ-aminobutyric acid (GABA), the principal inhibitory neurotransmitter in the brain. Recently, results from well-powered genome-wide association studies point to synaptic pruning as a potential contributing mechanism. In the region of the genome most strongly statistically associated with schizophrenia risk, variations in the relative expression of two isotypes of complement component 4, C4A and C4B, have been found to account for a significant proportion of this genetic signal. Studies of loss of C4 in mice show deficient synaptic pruning, leading to the hypothesis that increased expression of C4A in humans may result in excessive synaptic pruning. Such results point to the potential for a gene-driven understanding of pathophysiology; however, the findings also leave some important

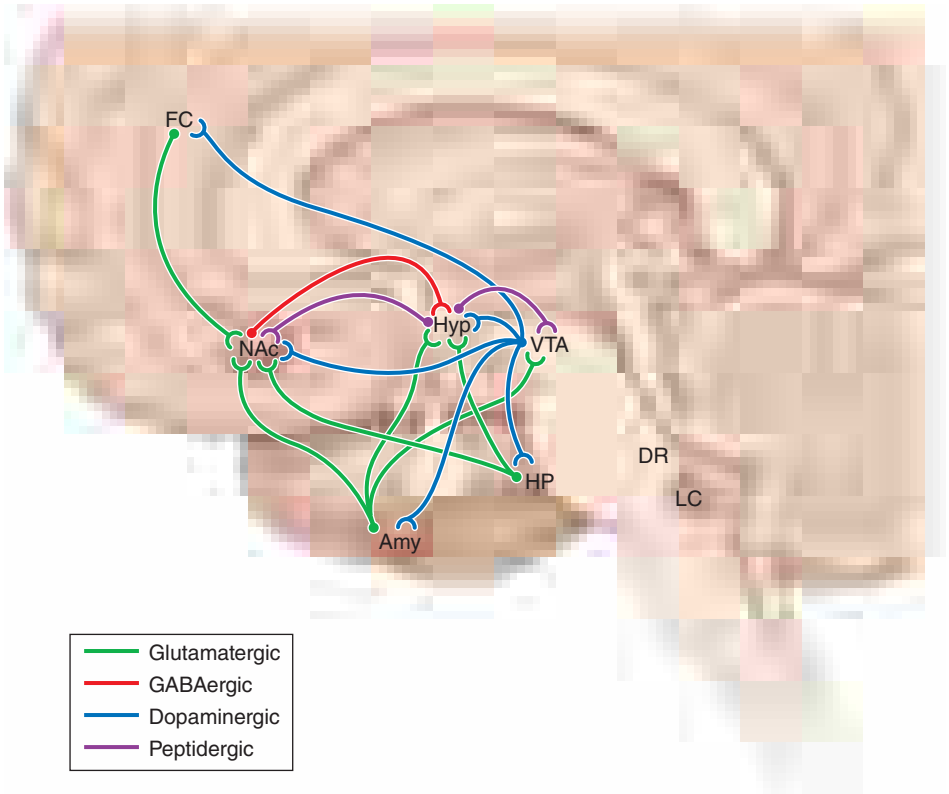


FIGURE 443-3 Neural circuitry of depression and addiction. The figure shows a simplified summary of a series of limbic circuits in brain that regulate mood and motivation and are implicated in depression and addiction. Shown in the figure are the hippocampus (HP) and amygdala (Amy) in the temporal lobe, regions of prefrontal cortex, NAc, and hypothalamus (Hyp). Only a subset of the known interconnections among these brain regions is shown. Also shown is the innervation of several of these brain regions by monoaminergic neurons. The ventral tegmental area (VTA) provides dopaminergic input to each of the limbic structures. Norepinephrine (from the locus coeruleus [LC]) and serotonin (from the dorsal raphe [DR] and other raphe nuclei) innervate all of the regions shown. In addition, there are strong connections between the hypothalamus and the VTA–NAc pathway. Important peptidergic projections from the hypothalamus include those from the arcuate nucleus that release β-endorphin and melanocortin and from the lateral hypothalamus that release orexin.

questions unanswered. The strongest effect haplotype in humans still only accounts for a very small increase in risk, with an odds ratio of less than 1.3. In contrast, having a sibling with schizophrenia increases risk approximately tenfold. In short, whether this allele reflects a driving pathophysiological mechanism remains to be determined. Moreover, humans have diverged at the C4 locus compared with rodents such that only a single C4 isotype is present in the mouse, preventing any analysis of the putative effects of changing the ratio of C4A to C4B—the phenomenon associated with disease risk in humans. Nonetheless, all the aforementioned findings support the notion that schizophrenia is a developmental neurodegenerative disorder with some evidence pointing to loss of cortical interneurons in frontal and temporal lobes.

Work in rodent and nonhuman primate models of addiction has established the brain's reward regions as key neural substrates for the acute actions of drugs of abuse and for addiction induced in vulnerable individuals by repeated drug administration (Fig. 443-3). Midbrain dopamine neurons in the VTA function normally as rheostats of reward: they are activated by natural rewards (food, sex, social interaction) or even by the expectation of such rewards, and many are suppressed by the absence of an expected reward or by aversive stimuli. These neurons thereby transmit crucial survival signals to the rest of the limbic brain to promote reward-related behavior, including motor responses to seek and obtain the rewards (NAc), memories of reward-related cues (amygdala, hippocampus), and executive control of obtaining rewards (prefrontal cortex).

Drugs of abuse alter neurotransmission through initial actions at different classes of ion channels, neurotransmitter receptors, or neurotransmitter transporters (Table 443-1). Studies in animal models have demonstrated that although the initial targets differ, the actions of these drugs converge on the brain's reward circuitry by promoting dopamine neurotransmission in the NAc and other limbic targets of the VTA. In addition, some drugs promote activation of opioid and cannabinoid receptors, which modulate this reward circuitry. By these mechanisms, drugs of abuse produce powerful rewarding signals, which, after repeated drug administration, corrupt a vulnerable brain's reward circuitry in ways that promote addiction. Three major pathologic adaptations have been described. First, drugs produce tolerance and dependence in reward circuits, which promote escalating drug intake and a negative emotional state during drug withdrawal that promotes relapse. Second, sensitization to the rewarding effects of the drugs and associated cues is seen during prolonged abstinence and also triggers relapse. Third, executive function is impaired in such a way as to increase impulsivity and compulsivity, both of which promote relapse.

Imaging studies in humans confirm that addictive drugs, as well as craving for them, activate the brain's reward circuitry. In addition, patients who abuse alcohol or psychostimulants show reduced gray matter in the prefrontal cortex as well as reduced activity in anterior cingulate and orbitofrontal cortex during tasks of attention and inhibitory control. It is thought that damage to these cortical areas contributes to addiction by impairing decision-making and increasing impulsivity.

NEUROINFLAMMATION

There is increasing evidence for the involvement of inflammatory mechanisms in a wide range of psychiatric syndromes. For example, a subset of depressed patients display elevated blood levels of interleukin 6 (IL-6), tumor necrosis factor α (TNF-α), and other cytokines. Moreover, rodents exposed to chronic stress exhibit similar increases in peripheral cytokines, and peripheral or central delivery of those cytokines to normal rodents increases their susceptibility to chronic stress. These findings have led to the novel idea of using peripheral cytokines as biomarkers of a subtype of depression and the potential utility of developing new antidepressants that oppose cytokine action.

Recent evidence has also linked proinflammatory signaling in the brain to addiction, particularly to alcohol. Human alcoholism is associated with impaired innate immunity, increases in circulating proinflammatory cytokines, and increases in brain expression of several immune-related genes. Many of these genes are expressed by astrocytes and microglia, and by neurons under certain pathologic conditions, where they play important roles in modifying neuronal function and plasticity. For example, cytokine monocyte chemotactic protein-1 (MCP-1) modulates the release of certain neurotransmitters and, when administered into the VTA, increases neuronal excitability, promotes dopamine release, and increases locomotor activity. Recent gene expression studies of alcohol drinking in mice have identified a network of regulated neuroimmune proteins in brain, and a role in regulation of alcohol consumption has been recently validated for several, including chemokines MCP-1 and chemokine (C-C motif) ligand 3 (CCL3), beta-2 microglobulin, CD14, IL-1 receptor antagonist, and cathepsins S and F. This work has led to discovery of anti-inflammatory medications that reduce alcohol intake in animals, such as agonists of peroxisome proliferator-activated receptors (PPARs), which are transcription factors that repress key inflammatory signaling molecules such as nuclear factor κB (NFκB) and nuclear factor of activated T cells (NFAT). A major focus of current

research is to define the sites and mechanisms by which proinflammatory cytokines impair brain function to elicit a depressive episode or promote drug abuse.

CONCLUSIONS

This brief narrative illustrates the substantial progress that is being made in understanding the genetic and neurobiological basis of mental illness. It is anticipated that biologic measures will be used increasingly to diagnose and subtype psychiatric disorders and that targeted therapeutics will become available for these complex conditions.

FURTHER READING

Gandal MJ et al: The road to precision psychiatry: Translating genetics into disease mechanisms. Nat Neurosci 19:1397, 2016.

Koob GF, Volkow ND: Neurobiology of addiction: A neurocircuitry analysis. Lancet Psychiatry 3:760, 2016.

Rajasethupathy P et al: Targeting neural circuits. Cell 165:524, 2016.

Ron D, Barak S: Molecular mechanisms underlying alcohol-drinking behaviours. Nat Rev Neurosci 17:576, 2016.

Sanders SJ et al: Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. Neuron 87:1215, 2015.

Schizophrenia Working Group of the Psychiatric Genomics Consortium: Biological insights from 108 schizophrenia-associated genetic loci. Nature 511:421, 2014.

Wohleb ES et al: Integrating neuroimmune systems in the neurobiology of depression. Nat Rev Neurosci 17:497, 2016.

444 Psychiatric Disorders

Victor I. Reus

Psychiatric disorders are common in medical practice and may present either as a primary disorder or as a comorbid condition. The prevalence of mental or substance use disorders in the United States is ~30%, but only one-third of affected individuals are currently receiving treatment. Global burden of disease statistics indicates that 4 of the 10 most important causes of morbidity and attendant health care costs worldwide are psychiatric in origin.

Changes in health care delivery underscore the need for primary care physicians to assume responsibility for the initial diagnosis and treatment of the most common mental disorders. Prompt diagnosis is essential to ensure that patients have access to appropriate medical services and to maximize the clinical outcome. Validated patient-based questionnaires have been developed that systematically probe for signs and symptoms associated with the most prevalent psychiatric diagnoses and guide the clinician into targeted assessment. The Primary Care Evaluation of Mental Disorders (PRIME-MD; and a self-report form, the Patient Health Questionnaire) and the Symptom-Driven Diagnostic System for Primary Care (SDDS-PC) are inventories that require only 10 min to complete and link patient responses to the formal diagnostic criteria of anxiety, mood, somatoform, and eating disorders and to alcohol abuse or dependence.

A physician who refers patients to a psychiatrist should know not only when doing so is appropriate but also how to refer because societal misconceptions and the stigma of mental illness impede the process. Primary care physicians should base referrals to a psychiatrist on the presence of signs and symptoms of a mental disorder and not simply on the absence of a physical explanation for a patient's complaint. The physician should discuss with the patient the reasons for requesting the referral or consultation and provide reassurance that he or she will continue to provide medical care and work collaboratively with the mental health professional. Consultation with a psychiatrist or transfer of care is appropriate when physicians encounter evidence of psychotic symptoms, mania, severe depression, or anxiety; symptoms of posttraumatic stress disorder (PTSD); suicidal or homicidal preoccupation; or a failure to respond to first-order treatment. This chapter reviews the clinical assessment and treatment of some of the most common mental disorders presenting in primary care and is based on the *Diagnostic and Statistical Manual of Mental Disorders*, Fifth Edition (DSM-5), the framework for categorizing psychiatric illness used in the United States. **Eating disorders are discussed later in this chapter, and the biology of psychiatric and addictive disorders is discussed in Chap. 443.**

GLOBAL CONSIDERATIONS

The DSM-5 and the tenth revision of the International Classification of Diseases (ICD-10), which is used more commonly worldwide, have taken somewhat differing approaches to the diagnosis of mental illness, but considerable effort has been expended to provide an operational translation between the two nosologies. Both systems are in essence purely descriptive and emphasize clinical pragmatism, in distinction to the Research Domain Criteria (RDOC) proposed by National Institute of Mental Health, which aspires to provide a causal framework for classification of behavioral disturbance. None of these diagnostic systems has as yet achieved adequate validation. The Global Burden of Disease Study 2010, using available epidemiologic data, nevertheless has reinforced the conclusion that, regardless of nosologic differences, mental and substance abuse disorders are the major cause of life-years lost to disability among all medical illnesses. There is general agreement that high-income countries will need to build capacity in professional training in low- and middle-income countries in order to provide an adequate balanced care model for the delivery of evidence-based therapies for mental disorders. Recent surveys that indicate a dramatic increase in mental disorder prevalence in rapidly developing countries, such as China, may reflect both an increased recognition of the issue, but also the consequence of social turmoil, stigma, and historically inadequate resources. The need for improved prevention strategies and for more definitive and effective interventional treatments remains a global concern.

ANXIETY DISORDERS

Anxiety disorders, the most prevalent psychiatric illnesses in the general community, are present in 15–20% of medical clinic patients. Anxiety, defined as a subjective sense of unease, dread, or foreboding, can indicate a primary psychiatric condition or can be a component of, or reaction to, a primary medical disease. The primary anxiety disorders are classified according to their duration and course and the existence and nature of precipitants.

When evaluating the anxious patient, the clinician must first determine whether the anxiety antedates or postdates a medical illness or is due to a medication side effect. Approximately one-third of patients presenting with anxiety have a medical etiology for their psychiatric symptoms, but an anxiety disorder can also present with somatic symptoms in the absence of a diagnosable medical condition.

PANIC DISORDER

Clinical Manifestations Panic disorder is defined by the presence of recurrent and unpredictable panic attacks, which are distinct episodes of intense fear and discomfort associated with a variety of physical symptoms, including palpitations, sweating, trembling, shortness of breath, chest pain, dizziness, and a fear of impending doom or death. Paresthesias, gastrointestinal distress, and feelings of unreality are also common. Diagnostic criteria require at least 1 month of concern or worry about the attacks or a change in behavior related to them. The lifetime prevalence of panic disorder is 2–3%. Panic attacks have a sudden onset, developing within 10 min and usually resolving over the course of an hour, and they occur in an unexpected fashion. Some may occur when waking from sleep. The frequency and severity of panic attacks vary, ranging from once a week to clusters of attacks separated by months of well-being. The first attack is usually outside the home, and onset is typically in late adolescence to early adulthood. In some individuals, anticipatory anxiety develops over time and results in a

generalized fear and a progressive avoidance of places or situations in which a panic attack might recur. *Agoraphobia*, which occurs commonly in patients with panic disorder, is an acquired irrational fear of being in places where one might feel trapped or unable to escape. It may, however, be diagnosed even if panic disorder is not present. Typically, it leads the patient into a progressive restriction in lifestyle and, in a literal sense, in geography. Frequently, patients are embarrassed that they are housebound and dependent on the company of others to go out into the world and do not volunteer this information; thus, physicians will fail to recognize the syndrome if direct questioning is not pursued.

Differential Diagnosis A diagnosis of panic disorder is made after a medical etiology for the panic attacks has been ruled out. A variety of cardiovascular, respiratory, endocrine, and neurologic conditions can present with anxiety as the chief complaint. Patients with true panic disorder will often focus on one specific feature to the exclusion of others. For example, 20% of patients who present with syncope as a primary medical complaint have a primary diagnosis of a mood, anxiety, or substance abuse disorder, the most common being panic disorder. The differential diagnosis of panic disorder is complicated by a high rate of comorbidity with other psychiatric conditions, especially alcohol and benzodiazepine abuse, which patients initially use in an attempt at self-medication. Some 75% of panic disorder patients will also satisfy criteria for major depression at some point in their illness.

When the history is nonspecific, physical examination and focused laboratory testing must be used to rule out anxiety states resulting from medical disorders such as pheochromocytoma, thyrotoxicosis, or hypoglycemia. Electrocardiogram (ECG) and echocardiogram may detect some cardiovascular conditions associated with panic such as paroxysmal atrial tachycardia and mitral valve prolapse. In two studies, panic disorder was the primary diagnosis in 43% of patients with chest pain who had normal coronary angiograms and was present in 9% of all outpatients referred for cardiac evaluation. Panic disorder has also been diagnosed in many patients referred for pulmonary function testing or with symptoms of irritable bowel syndrome.

Etiology and Pathophysiology The etiology of panic disorder is unknown but appears to involve a genetic predisposition, altered autonomic responsivity, and social learning. Panic disorder shows familial aggregation; the disorder is concordant in 30–45% of monozygotic twins, and genome-wide screens have identified suggestive risk loci. Acute panic attacks appear to be associated with increased noradrenergic discharges in the locus coeruleus. Intravenous infusion of sodium lactate evokes an attack in two-thirds of panic disorder patients, as do the α_2-adrenergic antagonist yohimbine, cholecystokinin tetrapeptide (CCK-4), and carbon dioxide inhalation. It is hypothesized that each of these stimuli activates a pathway involving noradrenergic neurons in the locus coeruleus and serotonergic neurons in the dorsal raphe. Resting state fMRI has identified abnormalities in the default mode network involving the medial temporal lobe, with greater activation in the sensorimotor cortex in panic disorder and in amygdala-frontal connectivity in social anxiety disorder. Agents that block serotonin reuptake can prevent attacks. Patients with panic disorder have a heightened sensitivity to somatic symptoms, which triggers increasing arousal, setting off the panic attack; accordingly, therapeutic intervention involves altering the patient's cognitive interpretation of anxiety-producing experiences as well as preventing the attack itself.

TREATMENT

Panic Disorder

Achievable goals of treatment are to decrease the frequency of panic attacks and to reduce their intensity. The cornerstone of drug therapy is antidepressant medication **(Tables 444-1 through 444-3)**. Selective serotonin reuptake inhibitors (SSRIs) benefit the majority of panic disorder patients and do not have the adverse effects of tricyclic antidepressants (TCAs). Fluoxetine, paroxetine, sertraline, and the selective serotonin-norepinephrine reuptake inhibitor (SNRI) venlafaxine have received approval from the U.S. Food and Drug Administration (FDA) for this indication. These drugs should be started at one-third to one-half of their usual antidepressant dose (e.g., 5–10 mg fluoxetine, 25–50 mg sertraline, 10 mg paroxetine, venlafaxine 37.5 mg). Monoamine oxidase inhibitors (MAOIs) are also effective and may specifically benefit patients who have comorbid features of atypical depression (i.e., hypersomnia and weight gain). Insomnia, orthostatic hypotension, and the need to maintain a low-tyramine diet (avoidance of cheese and wine) have limited their use, however. Antidepressants typically take 2–6 weeks to become effective, and doses may need to be adjusted based on the clinical response.

Because of anticipatory anxiety and the need for immediate relief of panic symptoms, benzodiazepines are useful early in the course of treatment and sporadically thereafter **(Table 444-4)**. FDA-approved agents include alprazolam and clonazepam. A recent Cochrane review found no difference between antidepressants and benzodiazepines in response rate, although benzodiazepines were somewhat better tolerated by patients. In treatment resistant cases, short-term augmentation with aripiprazole, divalproex sodium, or pindolol has some evidence for efficacy. There also is no clear difference in short-term efficacy between psychological therapies and antidepressant or benzodiazepine treatment, alone or in combination,

Early psychotherapeutic intervention and education aimed at symptom control enhance the effectiveness of drug treatment. Patients can be taught breathing techniques, be educated about physiologic changes that occur with panic, and learn to expose themselves voluntarily to precipitating events in a treatment program spanning 12–15 sessions. Homework assignments and monitored compliance are important components of successful treatment. Once patients have achieved a satisfactory response, drug treatment should be maintained for 1–2 years to prevent relapse. Controlled trials indicate a success rate of 75–85%, although the likelihood of complete remission is somewhat lower.

GENERALIZED ANXIETY DISORDER

Clinical Manifestations Patients with generalized anxiety disorder (GAD) have persistent, excessive, and/or unrealistic worry associated with muscle tension, impaired concentration, autonomic arousal, feeling "on edge" or restless, and insomnia **(Table 444-5)**. Onset is usually before age 20 years, and a history of childhood fears and social inhibition may be present. The lifetime prevalence of GAD is 5–6%; the risk is higher in first-degree relatives of patients with the diagnosis. Interestingly, family studies indicate that GAD and panic disorder segregate independently. More than 80% of patients with GAD also suffer from major depression, dysthymia, or social phobia. Comorbid substance abuse is common in these patients, particularly alcohol and/or sedative/hypnotic abuse. Patients with GAD worry excessively over minor matters, with life-disrupting effects; unlike in panic disorder, complaints of shortness of breath, palpitations, and tachycardia are relatively rare.

Etiology and Pathophysiology All anxiogenic agents act on the γ-aminobutyric acid $(GABA)_A$ receptor/chloride ion channel complex, implicating this neurotransmitter system in the pathogenesis of anxiety and panic attacks. Benzodiazepines are thought to bind two separate $GABA_A$ receptor sites: type I, which has a broad neuroanatomic distribution, and type II, which is concentrated in the hippocampus, striatum, and neocortex. The antianxiety effects of the various benzodiazepines are influenced by their relative binding to alpha 2 and 3 subunits of the $GABA_A$ receptor, and sedation and memory impairment to the alpha 1 subunit. Serotonin (5-hydroxytryptamine [5-HT]), and 3α-reduced neuroactive steroids (allosteric modulators of $GABA_A$) also appear to have a role in anxiety, and buspirone, a partial $5\text{-}HT_{1A}$ receptor agonist, and certain $5\text{-}HT_{2A}$ and $5\text{-}HT_{2C}$ receptor antagonists (e.g., nefazodone) may have beneficial effects.

TABLE 444-1 Antidepressants

NAME	USUAL DAILY DOSE, mg	SIDE EFFECTS	COMMENTS
SSRIs			
Fluoxetine (Prozac)	10–80	Headache; nausea and other GI effects; jitteriness; insomnia; sexual dysfunction; can affect plasma levels of other medicines (except sertraline); akathisia rare	Once-daily dosing, usually in the morning; fluoxetine has very long half-life; must not be combined with MAOIs
Sertraline (Zoloft)	50–200		
Paroxetine (Paxil)	20–60		
Fluvoxamine (Luvox)	100–300		
Citalopram (Celexa)	20–60		
Escitalopram (Lexapro)	10–30		
TCAs and Tetracyclics			
Amitriptyline (Elavil)	150–300	Anticholinergic (dry mouth, tachycardia, constipation, urinary retention, blurred vision); sweating; tremor; postural hypotension; cardiac conduction delay; sedation; weight gain	Once-daily dosing, usually qhs; blood levels of most TCAs available; can be lethal in overdose (lethal dose = 2 g); nortriptyline best tolerated, especially by elderly
Nortriptyline (Pamelor)	50–200		
Imipramine (Tofranil)	150–300		
Desipramine (Norpramin)	150–300		
Doxepin (Sinequan)	150–300		FDA approved for OCD
Clomipramine (Anafranil)	150–300		
Maprotiline (Ludiomil)	25–150		
Mixed Norepinephrine/Serotonin Reuptake Inhibitors (SNRI) and Receptor Blockers			
Venlafaxine (Effexor)	75–375	Nausea; dizziness; dry mouth; headaches; increased blood pressure; anxiety and insomnia	Bid-tid dosing (extended release available); lower potential for drug interactions than SSRIs; contraindicated with MAOIs
Desvenlafaxine (Pristiq)	50–400	Nausea, dizziness, insomnia	Primary metabolite of venlafaxine; no increased efficacy with higher dosing
Duloxetine (Cymbalta)	40–60	Nausea, dizziness, headache, insomnia, constipation	May have utility in treatment of neuropathic pain and stress incontinence
Mirtazapine (Remeron)	15–45	Somnolence, weight gain; neutropenia rare	Once-a-day dosing
Vilazodone (Viibryd)	40	Nausea, diarrhea, headache; dosage adjustment if given with CYP3A4 inhibitor/stimulator	Also 5-HT_{1a} receptor partial agonist
Vortioxetine (Brintellix)	5–20	Nausea, diarrhea, sweating, headache; low incidence of sedation or weight gain	No specific p450 effects; 5-HT_{3a} and 5-HT_{7} receptor antagonist, 5-HT_{1b} partial agonist, and 5-HT_{1a} agonist
Levomilnacipran (Fetzima)	40–120	Nausea, constipation, sweating; rare increase in blood pressure/pulse	Most noradrenergic of SNRIs
Mixed-Action Drugs			
Bupropion (Wellbutrin)	250–450	Jitteriness; flushing; seizures in at-risk patients; anorexia; tachycardia; psychosis	Tid dosing, but sustained release also available; fewer sexual side effects than SSRIs or TCAs; may be useful for adult ADD
Trazodone (Desyrel)	200–600	Sedation; dry mouth; ventricular irritability; postural hypotension; priapism rare	Useful in low doses for sleep because of sedating effects with no anticholinergic side effects
Trazodone extended release (Oleptro)	150–375	Daytime somnolence, dizziness, nausea	
Amoxapine (Asendin)	200–600	Sexual dysfunction	Lethality in overdose; EPS possible
MAOIs			
Phenelzine (Nardil)	45–90	Insomnia; hypotension; edema; anorgasmia; weight gain; neuropathy; hypertensive crisis; toxic reactions with SSRIs; narcotics	May be more effective in patients with atypical features or treatment-refractory depression
Tranylcypromine (Parnate)	20–50		
Isocarboxazid (Marplan)	20–60		Less weight gain and hypotension than phenelzine
Transdermal selegiline (Emsam)	6–12	Local skin reaction hypertension	No dietary restrictions with 6-mg dose

Abbreviations: ADD, attention deficit disorder; EPS, extrapyramidal symptoms; FDA, U.S. Food and Drug Administration; GI, gastrointestinal; MAOIs, monoamine oxidase inhibitors; OCD, obsessive-compulsive disorder; SSRIs, selective serotonin reuptake inhibitors; TCAs, tricyclic antidepressants.

TREATMENT

Generalized Anxiety Disorder

A combination of pharmacologic and psychotherapeutic interventions is most effective in GAD, but complete symptomatic relief is rare. A short course of a benzodiazepine is usually indicated, preferably lorazepam, oxazepam, or alprazolam. (The first two of these agents are metabolized via conjugation rather than oxidation and thus do not accumulate if hepatic function is impaired; the latter also has limited active metabolites.) Treatment should be initiated at the lowest dose possible and prescribed on an as-needed basis as symptoms warrant. Benzodiazepines differ in their milligram per kilogram potency, half-life, lipid solubility, metabolic pathways, and presence of active metabolites. Agents that are absorbed rapidly and are lipid soluble, such as diazepam, have a rapid onset of action and a higher abuse potential. Benzodiazepines should generally not be prescribed for >4–6 weeks because of the development of tolerance and the risk of abuse and dependence. Withdrawal must be closely monitored as relapses can occur. It is important to warn patients that concomitant use of alcohol or other sedating drugs may exacerbate side effects and impair their ability to function. An optimistic approach that encourages the patient to clarify environmental precipitants, anticipate his or her reactions, and plan effective response strategies is an essential element of therapy.

Adverse effects of benzodiazepines generally parallel their relative half-lives. Longer-acting agents, such as diazepam, chlordiazepoxide, flurazepam, and clonazepam, tend to accumulate active metabolites, with resultant sedation, impairment of cognition,

TABLE 444-2 Management of Antidepressant Side Effects

SYMPTOMS	COMMENTS AND MANAGEMENT STRATEGIES
Gastrointestinal	
Nausea, loss of appetite	Usually short-lived and dose-related; consider temporary dose reduction or administration with food and antacids
Diarrhea	Famotidine, 20–40 mg/d
Constipation	Wait for tolerance; try diet change, stool softener, exercise; avoid laxatives
Sexual dysfunction	Consider dose reduction; drug holiday
Anorgasmia/ impotence; impaired ejaculation	Bethanechol, 10–20 mg, 2 h before activity, or cyproheptadine, 4–8 mg 2 h before activity, or bupropion, 100 mg bid, or amantadine, 100 mg bid/tid
Orthostasis	Tolerance unlikely; increase fluid intake, use calf exercises/support hose; fludrocortisone, 0.025 mg/d
Anticholinergic	Wait for tolerance
Dry mouth, eyes	Maintain good oral hygiene; use artificial tears, sugar-free gum
Tremor/jitteriness	Antiparkinsonian drugs not effective; use dose reduction/slow increase; lorazepam, 0.5 mg bid, or propranolol, 10–20 mg bid
Insomnia	Schedule all doses for the morning; trazodone, 50–100 mg qhs
Sedation	Caffeine; schedule all dosing for bedtime; bupropion, 75–100 mg in afternoon
Headache	Evaluate diet, stress, other drugs; try dose reduction; amitriptyline, 50 mg/d
Weight gain	Decrease carbohydrates; exercise; consider fluoxetine
Loss of therapeutic benefit over time	Related to tolerance? Increase dose or drug holiday; add amantadine, 100 mg bid, buspirone, 10 mg tid, or pindolol, 2.5 mg bid

and poor psychomotor performance. Shorter-acting compounds, such as alprazolam, lorazepam, and oxazepam, can produce daytime anxiety, early morning insomnia, and, with discontinuation, rebound anxiety and insomnia. Although patients develop tolerance to the sedative effects of benzodiazepines, they are less likely to habituate to the adverse psychomotor effects. Withdrawal from the longer half-life benzodiazepines can be accomplished through gradual, stepwise dose reduction (by 10% every 1–2 weeks) over 6–12 weeks. It is usually more difficult to taper patients off shorter-acting benzodiazepines. Physicians may need to switch the patient to a benzodiazepine with a longer half-life or use an adjunctive medication such as a beta blocker or carbamazepine, before attempting to discontinue the benzodiazepine. Withdrawal reactions vary in severity and duration; they can include depression, anxiety, lethargy, diaphoresis, autonomic arousal, and, rarely, seizures.

TABLE 444-3 Possible Drug Interactions with Selective Serotonin Reuptake Inhibitors

AGENT	EFFECT
Monoamine oxidase inhibitors	Serotonin syndrome—absolute contraindication
Serotonergic agonists, e.g., tryptophan, fenfluramine, tryptans	Potential serotonin syndrome
Drugs that are metabolized by P450 isoenzymes: tricyclics, other SSRIs, antipsychotics, beta blockers, codeine, triazolobenzodiazepines, calcium channel blockers	Delayed metabolism resulting in increased blood levels and potential toxicity
Drugs that are bound tightly to plasma proteins, e.g., warfarin	Increased bleeding secondary to displacement
Drugs that inhibit the metabolism of SSRIs by P450 isoenzymes, e.g., quinidine	Increased SSRI side effects

Abbreviation: SSRIs, selective serotonin reuptake inhibitors.

Buspirone is a nonbenzodiazepine anxiolytic agent. It is nonsedating, does not produce tolerance or dependence, does not interact with benzodiazepine receptors or alcohol, and has no abuse or disinhibition potential. However, it requires several weeks to take effect and requires thrice-daily dosing. Patients who were previously responsive to a benzodiazepine are unlikely to rate buspirone as equally effective, but patients with head injury or dementia who have symptoms of anxiety and/or agitation may do well with this agent. Escitalopram, paroxetine, duloxetine, and venlafaxine are FDA approved for the treatment of GAD, usually at doses that are comparable to their efficacy in major depression, and may be preferable to usage of benzodiazepines in the treatment of chronic anxiety. Benzodiazepines are contraindicated during pregnancy and breast-feeding.

Anticonvulsants with GABAergic properties may also be effective against anxiety. Gabapentin, oxcarbazepine, tiagabine, pregabalin, and divalproex have all shown some degree of benefit in a variety of anxiety-related syndromes in off-label usage.

PHOBIC DISORDERS

Clinical Manifestations The cardinal feature of phobic disorders is a marked and persistent fear of objects or situations, exposure to which results in an immediate anxiety reaction. The patient avoids the phobic stimulus, and this avoidance usually impairs occupational or social functioning. Panic attacks may be triggered by the phobic stimulus or may occur spontaneously. Unlike patients with other anxiety disorders, individuals with phobias usually experience anxiety only in specific situations. Common phobias include fear of closed spaces (claustrophobia), fear of blood, and fear of flying. Social phobia is distinguished by a specific fear of social or performance situations in which the individual is exposed to unfamiliar individuals or to possible examination and evaluation by others. Examples include having to converse at a party, use public restrooms, and meet strangers. In each case, the affected individual is aware that the experienced fear is excessive and unreasonable given the circumstance. The specific content of a phobia may vary across gender, ethnic, and cultural boundaries.

Phobic disorders are common, affecting ~7–9% of the population. Twice as many females are affected than males. Full criteria for diagnosis are usually satisfied first in early adulthood, but behavioral avoidance of unfamiliar people, situations, or objects dating from early childhood is common.

In one study of female twins, concordance rates for agoraphobia, social phobia, and animal phobia were found to be 23% for monozygotic twins and 15% for dizygotic twins. A twin study of fear conditioning, a model for the acquisition of phobias, demonstrated a heritability of 35–45%. Animal studies of fear conditioning have indicated that processing of the fear stimulus occurs through the lateral nucleus of the amygdala, extending through the central nucleus and projecting to the periaqueductal gray region, lateral hypothalamus, and paraventricular hypothalamus.

TREATMENT

Phobic Disorders

Beta blockers (e.g., propranolol, 20–40 mg orally 2 h before the event) are particularly effective in the treatment of "performance anxiety" (but not general social phobia) and appear to work by blocking the peripheral manifestations of anxiety such as perspiration, tachycardia, palpitations, and tremor. MAOIs alleviate social phobia independently of their antidepressant activity, and paroxetine, sertraline, and venlafaxine have received FDA approval for treatment of social anxiety. Benzodiazepines can be helpful in reducing fearful avoidance, but the chronic nature of phobic disorders limits their usefulness.

Behaviorally focused psychotherapy is an important component of treatment because relapse rates are high when medication is used as the sole treatment. Cognitive-behavioral strategies are based

TABLE 444-4 Anxiolytics

NAME	EQUIVALENT PO DOSE, mg	ONSET OF ACTION	HALF-LIFE, h	COMMENTS
Benzodiazepines				
Diazepam (Valium)	5	Fast	20–70	Active metabolites; quite sedating
Flurazepam (Dalmane)	15	Fast	30–100	Flurazepam is a prodrug; metabolites are active; quite sedating
Triazolam (Halcion)	0.25	Intermediate	1.5–5	No active metabolites; can induce confusion and delirium, especially in elderly
Lorazepam (Ativan)	1	Intermediate	10–20	No active metabolites; direct hepatic glucuronide conjugation; quite sedating; FDA approved for anxiety with depression
Alprazolam (Xanax)	0.5	Intermediate	12–15	Active metabolites; not too sedating; FDA approved for panic disorder and anxiety with depression; tolerance and dependence develop easily; difficult to withdraw
Chlordiazepoxide (Librium)	10	Intermediate	5–30	Active metabolites; moderately sedating
Oxazepam (Serax)	15	Slow	5–15	No active metabolites; direct glucuronide conjugation; not too sedating
Temazepam (Restoril)	15	Slow	9–12	No active metabolites; moderately sedating
Clonazepam (Klonopin)	0.5	Slow	18–50	No active metabolites; moderately sedating; FDA approved for panic disorder
Clorazepate (Tranxene)	15	Fast	40–200	Low sedation; unreliable absorption
Nonbenzodiazepines				
Buspirone (BuSpar)	7.5	2 weeks	2–3	Active metabolites; tid dosing—usual daily dose 10–20 mg tid; nonsedating; no additive effects with alcohol; useful for controlling agitation in demented or brain-injured patients

Abbreviation: FDA, U.S. Food and Drug Administration.

on the finding that distorted perceptions and interpretations of fear-producing stimuli play a major role in perpetuation of phobias. Individual and group therapy sessions teach the patient to identify specific negative thoughts associated with the anxiety-producing situation and help to reduce the patient's fear of loss of control. In desensitization therapy, hierarchies of feared situations are constructed, and the patient is encouraged to pursue and master gradual exposure to the anxiety-producing stimuli.

Patients with social phobia, in particular, have a high rate of comorbid alcohol abuse, as well as of other psychiatric conditions (e.g., eating disorders), necessitating the need for parallel management of each disorder if anxiety reduction is to be achieved.

TABLE 444-5 Diagnostic Criteria for Generalized Anxiety Disorder

A. Excessive anxiety and worry (apprehensive expectation), occurring more days than not for at least 6 months, about a number of events or activities (such as work or school performance).

B. The individual finds it difficult to control the worry.

C. The anxiety and worry are associated with three (or more) of the following six symptoms (with at least some symptoms present for more days than not for the past 6 months): (1) restlessness or feeling keyed up or on edge; (2) being easily fatigued; (3) difficulty concentrating or mind going blank; (4) irritability; (5) muscle tension; (6) sleep disturbance (difficulty falling or staying asleep, or restless, unsatisfying sleep).

D. The anxiety, worry, or physical symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.

E. The disturbance is not attributable to the physiologic effects of a substance (e.g., a drug of abuse, a medication) or another medical condition (e.g., hyperthyroidism).

F. The disturbance is not better explained by another mental disorder (e.g., anxiety or worry about having panic attacks in panic disorder, negative evaluation in social anxiety disorder [social phobia], contamination or other obsessions in obsessive-compulsive disorder, separation from attachment figures in separation anxiety disorder, reminders of traumatic events in posttraumatic stress disorder, gaining weight in anorexia nervosa, physical complaints in somatic symptom disorder, perceived appearance flaws in body dysmorphic disorder, having a serious illness in illness anxiety disorder, or the content of delusional beliefs in schizophrenia or delusional disorder).

Source: Diagnostic and Statistical Manual of Mental Disorders, 5th ed. Washington, DC, American Psychiatric Association, 2013.

STRESS DISORDERS

Clinical Manifestations Patients may develop anxiety after exposure to extreme traumatic events such as the threat of personal death or injury or the death of a loved one. The reaction may occur shortly after the trauma (*acute stress disorder*) or be delayed and subject to recurrence (PTSD) **(Table 444-6)**. In both syndromes, individuals experience associated symptoms of detachment and loss of emotional responsivity. The patient may feel depersonalized and unable to recall specific aspects of the trauma, although typically it is reexperienced through intrusions in thought, dreams, or flashbacks, particularly when cues of the original event are present. Patients often actively avoid stimuli that precipitate recollections of the trauma and demonstrate a resulting increase in vigilance, arousal, and startle response. Patients with stress disorders are at risk for the development of other disorders related to anxiety, mood, and substance abuse (especially alcohol). Between 5 and 10% of Americans will at some time in their life satisfy criteria for PTSD, with women more likely to be affected than men. A validated 4-item screen for PTSD (PC-PTSD) is available.

Risk factors for the development of PTSD include a past psychiatric history and personality characteristics of high neuroticism and extroversion. Twin studies show a substantial genetic influence on all symptoms associated with PTSD, with less evidence for an environmental effect.

Etiology and Pathophysiology It is hypothesized that in PTSD there is excessive release of norepinephrine from the locus coeruleus in response to stress and increased noradrenergic activity at projection sites in the hippocampus and amygdala. These changes theoretically facilitate the encoding of fear-based memories. Greater sympathetic responses to cues associated with the traumatic event occur in PTSD, although pituitary adrenal responses are blunted. In addition to fear learning, changes in threat detection (insula overactivity), executive function, emotional regulation and contextual learning have been documented.

TREATMENT

Stress Disorders

Acute stress reactions are usually self-limited, and treatment typically involves the short-term use of benzodiazepines and supportive/expressive psychotherapy. The chronic and recurrent nature of PTSD,

TABLE 444-6 Diagnostic Criteria for Posttraumatic Stress Disorder

A. Exposure to actual or threatened death, serious injury, or sexual violence in one (or more) of the following ways:
   1. Directly experiencing the traumatic event(s).
   2. Witnessing, in person, the event(s) as it occurred to others.
   3. Learning that the traumatic event(s) occurred to a close family member or close friend. In cases of actual or threatened death of a family member or friend, the event(s) must have been violent or accidental.
   4. Experiencing repeated or extreme exposure to aversive details of the traumatic event(s) (e.g., first responders collecting human remains; police officers repeatedly exposed to details of child abuse).

B. Presence of one (or more) of the following intrusion symptoms associated with the traumatic event(s), beginning after the traumatic event(s) occurred:
   1. Recurrent, involuntary, and intrusive distressing memories of the traumatic event(s).
   2. Recurrent distressing dreams in which the content and/or affect of the dream are related to the traumatic event(s).
   3. Dissociative reactions (e.g., flashbacks) in which the individual feels or acts as if the traumatic event(s) were recurring. (Such reactions may occur on a continuum, with the most extreme expression being a complete loss of awareness of present surroundings.)
   4. Intense or prolonged psychological distress at exposure to internal or external cues that symbolize or resemble an aspect of the traumatic event(s).
   5. Marked physiologic reactions to internal or external cues that symbolize or resemble an aspect of the traumatic event(s).

C. Persistent avoidance of stimuli associated with the traumatic event(s), beginning after the traumatic event(s) occurred, as evidenced by one or both of the following:
   1. Avoidance of or efforts to avoid distressing memories, thoughts, or feelings about or closely associated with the traumatic event(s).
   2. Avoidance of or efforts to avoid external reminders (people, places, conversations, activities, objects, situations) that arouse distressing memories, thoughts, or feelings about or closely associated with the traumatic event(s).

D. Negative alterations in cognitions and mood associated with the traumatic event(s), beginning or worsening after the traumatic event(s) occurred as evidenced by two (or more) of the following:
   1. Inability to remember an important aspect of the traumatic event(s) (typically due to dissociative amnesia and not to other factors such as head injury, alcohol, or drugs).
   2. Persistent and exaggerated negative beliefs or expectations about oneself, others, or the world (e.g., "I am bad," "No one can be trusted," "The world is completely dangerous," "My whole nervous system is permanently ruined").
   3. Persistent, distorted cognitions about the cause or consequences of the traumatic event(s) that lead the individual to blame himself/herself or others.
   4. Persistent negative emotional state (e.g., fear, horror, anger, guilt, or shame).
   5. Markedly diminished interest or participation in significant activities.
   6. Feelings of detachment or estrangement from others.
   7. Persistent inability to experience positive emotions (e.g., inability to experience happiness, satisfaction, or loving feelings).

E. Marked alterations in arousal and reactivity associated with the traumatic event(s), beginning or worsening after the traumatic event(s) occurred, as evidenced by two (or more) of the following:
   1. Irritable behavior and angry outbursts (with little or no provocation) typically expressed as verbal or physical aggression toward people or objects.
   2. Reckless or self-destructive behavior.
   3. Hypervigilance.
   4. Exaggerated startle response.
   5. Problems with concentration.
   6. Sleep disturbance (e.g., difficulty falling or staying asleep or restless sleep).

F. Duration of the disturbance (criteria B, C, D, and E) is >1 month.

G. The disturbance causes clinically significant distress or impairment in social, occupational, or other important areas of functioning.

H. The disturbance is not attributable to the physiologic effects of a substance (e.g., medication, alcohol) or another medical condition.

Source: Diagnostic and Statistical Manual of Mental Disorders, 5th ed. Washington, DC, American Psychiatric Association, 2013.

however, requires a more complex approach using drug and behavioral treatments. PTSD is highly correlated with peritraumatic dissociative symptoms and the development of an acute stress disorder at the time of the trauma. The SSRIs (paroxetine and sertraline are FDA approved for PTSD), venlafaxine, nefazadone, and topiramate can all reduce anxiety, symptoms of intrusion, and avoidance behaviors. Hydrocortisone, intranasal oxytocin, and opiates such as morphine, given shortly after the acute stress, may have beneficial effects in preventing the development of PTSD, and adjunctive naltrexone can be effective when comorbid alcoholism is present. Low dose trazodone and mirtazepine, sedating antidepressants, are frequently used at night to help with insomnia. Benzodiazepines and SSRIs, however, should not be given in the early aftermath of trauma. Psychotherapeutic strategies for PTSD help the patient overcome avoidance behaviors and demoralization and master fear of recurrence of the trauma; therapies that encourage the patient to dismantle avoidance behaviors through stepwise focusing on the experience of the traumatic event, such as trauma-focused cognitive-behavioral therapy, exposure therapy, and eye movement desensitization and reprocessing, are the most effective. Debriefing after the traumatic event does not prevent PTSD and may exacerbate symptoms.

OBSESSIVE-COMPULSIVE DISORDER

Clinical Manifestations Obsessive-compulsive disorder (OCD) is characterized by obsessive thoughts and compulsive behaviors that impair everyday functioning. Fears of contamination and germs are common, as are handwashing, counting behaviors, and having to check and recheck such actions as whether a door is locked. The degree to which the disorder is disruptive for the individual varies, but in all cases, obsessive-compulsive activities take up >1 h per day and are undertaken to relieve the anxiety triggered by the core fear. Patients often conceal their symptoms, usually because they are embarrassed by the content of their thoughts or the nature of their actions. Physicians must ask specific questions regarding recurrent thoughts and behaviors, particularly if physical clues such as chafed and reddened hands or patchy hair loss (from repetitive hair pulling, or trichotillomania) are present. Comorbid conditions are common, the most frequent being depression, other anxiety disorders, eating disorders, and tics. OCD has a lifetime prevalence of 2–3% worldwide. Onset is usually gradual, beginning in early adulthood, but childhood onset is not rare. The disorder usually has a waxing and waning course, but some cases may show a steady deterioration in psychosocial functioning.

Etiology and Pathophysiology A genetic contribution to OCD is suggested by twin studies, but no susceptibility gene for OCD has been identified to date. Family studies show an aggregation of OCD with Tourette's disorder, and both are more common in males and in first-born children.

The anatomy of obsessive-compulsive behavior is thought to include the orbital frontal cortex, caudate nucleus, and globus pallidus. The caudate nucleus appears to be involved in the acquisition and maintenance of habit and skill learning, and interventions that are successful in reducing obsessive-compulsive behaviors also decrease metabolic activity measured in the caudate.

TREATMENT

Obsessive-Compulsive Disorder

Clomipramine, fluoxetine, fluvoxamine, and sertraline are approved for the treatment of OCD in adults (fluvoxamine is also approved for children). Clomipramine is a TCA that is often tolerated poorly owing to anticholinergic and sedative side effects at the doses required to treat the illness (25–250 mg/d); its efficacy in OCD is unrelated to its antidepressant activity. Fluoxetine (5–60 mg/d), fluvoxamine (25–300 mg/d), and sertraline (50–150 mg/d) are as effective as clomipramine and have a more benign side effect profile. Only 50–60% of patients with OCD show adequate improvement with pharmacotherapy alone. In treatment-resistant cases, augmentation with other serotonergic agents such as buspirone, or with a neuroleptic or benzodiazepine, may be beneficial, and in severe cases, deep brain stimulation has been found to be effective. When a therapeutic response is achieved, long-duration maintenance therapy is usually indicated.

For many individuals, particularly those with time-consuming compulsions, behavior therapy, and exposure response prevention will result in as much improvement as that afforded by medication. Effective techniques include the gradual increase in exposure to stressful situations, maintenance of a diary to clarify stressors, and homework assignments that substitute new activities for compulsive behaviors.

MOOD DISORDERS

Mood disorders are characterized by a disturbance in the regulation of mood, behavior, and affect. Mood disorders are subdivided into (1) depressive disorders, (2) bipolar disorders, and (3) depression in association with medical illness or alcohol and substance abuse **(Chaps. 445 through 448)**. Major depressive disorder (MDD) is differentiated from bipolar disorder by the absence of a manic or hypomanic episode. The relationship between pure depressive syndromes and bipolar disorders is not well understood; MDD is more frequent in families of bipolar individuals, but the reverse is not true. In the Global Burden of Disease Study conducted by the World Health Organization, unipolar major depression ranked fourth among all diseases in terms of disability-adjusted life-years and was projected to rank second by the year 2020. In the United States, lost productivity directly related to mood disorders has been estimated at $55.1 billion per year.

DEPRESSION IN ASSOCIATION WITH MEDICAL ILLNESS

Depression occurring in the context of medical illness is difficult to evaluate. Depressive symptomatology may reflect the psychological stress of coping with the disease, may be caused by the disease process itself or by the medications used to treat it, or may simply coexist in time with the medical diagnosis.

Virtually every class of *medication* includes some agent that can induce depression. Antihypertensive drugs, anticholesterolemic agents, and antiarrhythmic agents are common triggers of depressive symptoms. Iatrogenic depression should also be considered in patients receiving glucocorticoids, antimicrobials, systemic analgesics, antiparkinsonian medications, and anticonvulsants. To decide whether a causal relationship exists between pharmacologic therapy and a patient's change in mood, it may sometimes be necessary to undertake an empirical trial of an alternative medication.

Between 20 and 30% of *cardiac* patients manifest a depressive disorder; an even higher percentage experience depressive symptomatology when self-reporting scales are used. Depressive symptoms following unstable angina, myocardial infarction, cardiac bypass surgery, or heart transplant impair rehabilitation and are associated with higher rates of mortality and medical morbidity. Depressed patients often show decreased variability in heart rate (an index of reduced parasympathetic nervous system activity), which may predispose individuals to ventricular arrhythmia and increased morbidity. Depression also appears to increase the risk of coronary heart disease, possibly through increased platelet aggregation. TCAs are contraindicated in patients with bundle branch block, and TCA-induced tachycardia is an additional concern in patients with congestive heart failure. SSRIs appear not to induce ECG changes or adverse cardiac events and thus are reasonable first-line drugs for patients at risk for TCA-related complications. SSRIs may interfere with hepatic metabolism of anticoagulants, however, causing increased anticoagulation.

In patients with *cancer*, the mean prevalence of depression is 25%, but depression occurs in 40–50% of patients with cancers of the pancreas or oropharynx. This association is not due to the effect of cachexia alone, as the higher prevalence of depression in patients with pancreatic cancer persists when compared to those with advanced gastric cancer. Initiation of antidepressant medication in cancer patients has been shown to improve quality of life as well as mood. Psychotherapeutic approaches, particularly group therapy, may have some effect on short-term depression, anxiety, and pain symptoms.

Depression occurs frequently in patients with *neurologic disorders*, particularly cerebrovascular disorders, Parkinson's disease, dementia, multiple sclerosis, and traumatic brain injury. One in five patients with left-hemisphere stroke involving the dorsolateral frontal cortex experiences major depression. Late-onset depression in otherwise cognitively normal individuals increases the risk of a subsequent diagnosis of Alzheimer's disease. All classes of antidepressant agents are effective against these depressions, as are, in some cases, stimulant compounds.

The reported prevalence of depression in patients with *diabetes mellitus* varies from 8 to 27%, with the severity of the mood state correlating with the level of hyperglycemia and the presence of diabetic complications. Treatment of depression may be complicated by effects of antidepressive agents on glycemic control. MAOIs can induce hypoglycemia and weight gain, whereas TCAs can produce hyperglycemia and carbohydrate craving. SSRIs and SNRIs, like MAOIs, may reduce fasting plasma glucose, but they are easier to use and may also improve dietary and medication compliance.

Hypothyroidism is frequently associated with features of depression, most commonly depressed mood and memory impairment. Hyperthyroid states may also present in a similar fashion, usually in geriatric populations. Improvement in mood usually follows normalization of thyroid function, but adjunctive antidepressant medication is sometimes required. Patients with subclinical hypothyroidism can also experience symptoms of depression and cognitive difficulty that respond to thyroid replacement.

The lifetime prevalence of depression in *HIV-positive* individuals has been estimated at 22–45%. The relationship between depression and disease progression is multifactorial and likely to involve psychological and social factors, alterations in immune function, and central nervous system (CNS) disease. Chronic hepatitis C infection is also associated with depression, which may worsen with interferon-α treatment.

Some chronic disorders of uncertain etiology, such as chronic fatigue syndrome **(Chap. 442)** and fibromyalgia **(Chap. 366)**, are strongly associated with depression and anxiety; patients may benefit from antidepressant treatment or anticonvulsant agents such as pregabalin.

DEPRESSIVE DISORDERS

Clinical Manifestations Major depression is defined as depressed mood on a daily basis for a minimum duration of 2 weeks **(Table 444-7)**. An episode may be characterized by sadness, indifference,

TABLE 444-7 Criteria for a Major Depressive Episode

A. Five (or more) of the following symptoms have been present during the same 2-week period and represent a change from previous functioning; at least one of the symptoms is either (1) depressed mood or (2) loss of interest or pleasure. (*Note:* Do not include symptoms that are clearly attributable to another medical condition.)
   1. Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad, empty, hopeless) or observation made by others (e.g., appears tearful).
   2. Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation).
   3. Significant weight loss when not dieting or weight gain (e.g., a change of >5% of body weight in a month), or decrease or increase in appetite nearly every day.
   4. Insomnia or hypersomnia nearly every day.
   5. Psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down).
   6. Fatigue or loss of energy nearly every day.
   7. Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick).
   8. Diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others).
   9. Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide.

B. The symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning

C. The episode is not attributable to the physiologic effects of a substance or to another medical condition.

D. The occurrence of the major depressive episode is not better explained by seasonal affective disorder, schizophrenia, schizophreniform disorder, delusional disorder, or other specified and unspecified schizophrenia spectrum and other psychotic disorders.

E. There has never been a manic episode or a hypomanic episode.

Source: Diagnostic and Statistical Manual of Mental Disorders, 5th ed. Washington, DC, American Psychiatric Association, 2013.

apathy, or irritability and is usually associated with changes in sleep patterns, appetite, and weight; motor agitation or retardation; fatigue; impaired concentration and decision-making; feelings of shame or guilt; and thoughts of death or dying. Patients with depression have a profound loss of pleasure in all enjoyable activities, exhibit early morning awakening, feel that the dysphoric mood state is qualitatively different from sadness, and often notice a diurnal variation in mood (worse in morning hours). Patients experiencing bereavement or grief may exhibit many of the same signs and symptoms of major depression, although the emphasis is usually on feelings of emptiness and loss, rather than anhedonia and loss of self-esteem, and the duration is usually limited. In certain cases, however, the diagnosis of major depression may be warranted even in the context of a significant loss.

Approximately 15% of the population experiences a major depressive episode at some point in life, and 6–8% of all outpatients in primary care settings satisfy diagnostic criteria for the disorder. Depression is often undiagnosed, and even more frequently, it is treated inadequately. If a physician suspects the presence of a major depressive episode, the initial task is to determine whether it represents unipolar or bipolar depression or is one of the 10–15% of cases that are secondary to general medical illness or substance abuse. Physicians should also assess the risk of suicide by direct questioning, as patients are often reluctant to verbalize such thoughts without prompting. If specific plans are uncovered or if significant risk factors exist (e.g., a past history of suicide attempts, profound hopelessness, concurrent medical illness, substance abuse, or social isolation), the patient must be referred to a mental health specialist for immediate care. The physician should specifically probe each of these areas in an empathic and hopeful manner, being sensitive to denial and possible minimization of distress. The presence of anxiety, panic, or agitation significantly increases near-term suicidal risk. Approximately 4–5% of all depressed patients will commit suicide; most will have sought help from physicians within 1 month of their deaths.

In some depressed patients, the mood disorder does not appear to be episodic and is not clearly associated with either psychosocial dysfunction or change from the individual's usual experience in life. Persistent depressive disorder (*dysthymic disorder*) consists of a pattern of chronic (at least 2 years), ongoing depressive symptoms that are usually less severe and/or less numerous than those found in major depression, but the functional consequences may be equivalent to or even greater; the two conditions are sometimes difficult to separate and can occur together ("double depression"). Many patients who exhibit a profile of pessimism, disinterest, and low self-esteem respond to antidepressant treatment. Persistent and chronic depressive disorders occur in ~2% of the general population.

Depression is approximately twice as common in women as in men, and the incidence increases with age in both sexes. Twin studies indicate that the liability to major depression of early onset (before age 25) is largely genetic in origin. Negative life events can precipitate and contribute to depression, but genetic factors influence the sensitivity of individuals to these stressful events. In most cases, both biologic and psychosocial factors are involved in the precipitation and unfolding of depressive episodes. The most potent stressors appear to involve death of a relative, assault, or severe marital or relationship problems.

Unipolar depressive disorders usually begin in early adulthood and recur episodically over the course of a lifetime. The best predictor of future risk is the number of past episodes; 50–60% of patients who have a first episode have at least one or two recurrences. Some patients experience multiple episodes that become more severe and frequent over time. The duration of an untreated episode varies greatly, ranging from a few months to ≥1 year. The pattern of recurrence and clinical progression in a developing episode are also variable. Within an individual, the nature of episodes (e.g., specific presenting symptoms, frequency, and duration) may be similar over time. In a minority of patients, a severe depressive episode may progress to a psychotic state; in elderly patients, depressive symptoms may be associated with cognitive deficits mimicking dementia ("pseudodementia"). A seasonal pattern of depression, called *seasonal affective disorder*, may manifest with onset and remission of episodes at predictable times of the year. This disorder is more common in women, whose symptoms are anergy, fatigue, weight gain, hypersomnia, and episodic carbohydrate craving. The prevalence increases with distance from the equator, and improvement may occur by altering light exposure.

Etiology and Pathophysiology Although evidence for genetic transmission of unipolar depression is not as strong as in bipolar disorder, monozygotic twins have a higher concordance rate (46%) than dizygotic siblings (20%), with little support for any effect of a shared family environment. A recent association study of over 5000 Chinese women identified genome wide significant loci near the gene for sirtuin 1(*SIRT1*) and in an intron of the phospholysine phosphohistidine inorganic pyrophosphate phosphatase (*LHPP*) gene.

Neuroendocrine abnormalities that reflect the neurovegetative signs and symptoms of depression include increased cortisol and corticotropin-releasing hormone (CRH) secretion, a decreased inhibitory response of glucocorticoids to dexamethasone, and a blunted response of thyroid-stimulating hormone (TSH) level to infusion of thyroid-releasing hormone (TRH). Antidepressant treatment leads to normalization of these abnormalities. Major depression is also associated with changes in levels of proinflammatory cytokines and neurotrophins, an increase in measures of oxidative stress and cellular aging, telomere shortening, and mitochondrial dysfunction.

Diurnal variations in symptom severity and alterations in circadian rhythmicity of a number of neurochemical and neurohumoral factors suggest that biologic differences may be secondary to a primary defect in regulation of biologic rhythms. Patients with major depression show consistent findings of a decrease in rapid eye movement (REM) sleep

onset (REM latency), an increase in REM density, and, in some subjects, a decrease in stage IV delta slow-wave sleep.

Although antidepressant drugs inhibit neurotransmitter uptake within hours, their therapeutic effects typically emerge over several weeks, implicating adaptive changes in second messenger systems and transcription factors as possible mechanisms of action.

TREATMENT

Depressive Disorders

Treatment planning requires coordination of short-term strategies to induce remission combined with longer term maintenance designed to prevent recurrence. The most effective intervention for achieving remission and preventing relapse is medication, but combined treatment, incorporating psychotherapy to help the patient cope with decreased self-esteem and demoralization, improves outcome **(Fig. 444-1)**. Approximately 40% of primary care patients with depression drop out of treatment and discontinue medication if symptomatic improvement is not noted within a month, unless additional support is provided. Outcome improves with (1) increased intensity and frequency of visits during the first 4–6 weeks of treatment, (2) supplemental educational materials, and (3) psychiatric consultation as indicated. Despite the widespread use of SSRIs and other second-generation antidepressant drugs, there is no convincing evidence that these classes of antidepressants are more efficacious than TCAs. Between 60 and 70% of all depressed patients respond to any drug chosen, if it is given in a sufficient dose for 6–8 weeks.

A rational approach to selecting which antidepressant to use involves matching the patient's preference and medical history with the metabolic and side effect profile of the drug (Tables 444-4 and 444-5). A previous response, or a family history of a positive response, to a specific antidepressant often suggests that that drug be tried first. Before initiating antidepressant therapy, the physician should evaluate the possible contribution of comorbid illnesses and consider their specific treatment. In individuals with suicidal ideation, particular attention should be paid to choosing a drug with low toxicity if taken in overdose. Newer antidepressant drugs are distinctly safer in this regard; nevertheless, the advantages of TCAs have not been completely superseded. The existence of generic equivalents makes TCAs relatively cheap, and for secondary tricyclics, particularly nortriptyline and desipramine, well-defined relationships among dose, plasma level, and therapeutic response exist. The steady-state plasma level achieved for a given drug dose can vary more than tenfold between individuals, and plasma levels may help in interpreting apparent resistance to treatment and/or unexpected drug toxicity. The principal side effects of TCAs are antihistamine (sedation) and anticholinergic (constipation, dry mouth, urinary hesitancy, blurred vision). TCAs are contraindicated in patients with serious cardiovascular risk factors, and overdoses of tricyclic agents can be lethal, with desipramine carrying the greatest risk. It is judicious to prescribe only a 10-day supply when suicide is a risk. Most patients require a daily dose of 150–200 mg of imipramine or amitriptyline or its equivalent to achieve a therapeutic blood level of 150–300 ng/mL and a satisfactory remission; some patients show a partial effect at lower doses. Geriatric patients may require a low starting dose and slow escalation. Ethnic differences in drug metabolism are significant, with Hispanic, Asian, and African-American patients generally requiring lower doses to achieve a comparable blood level.

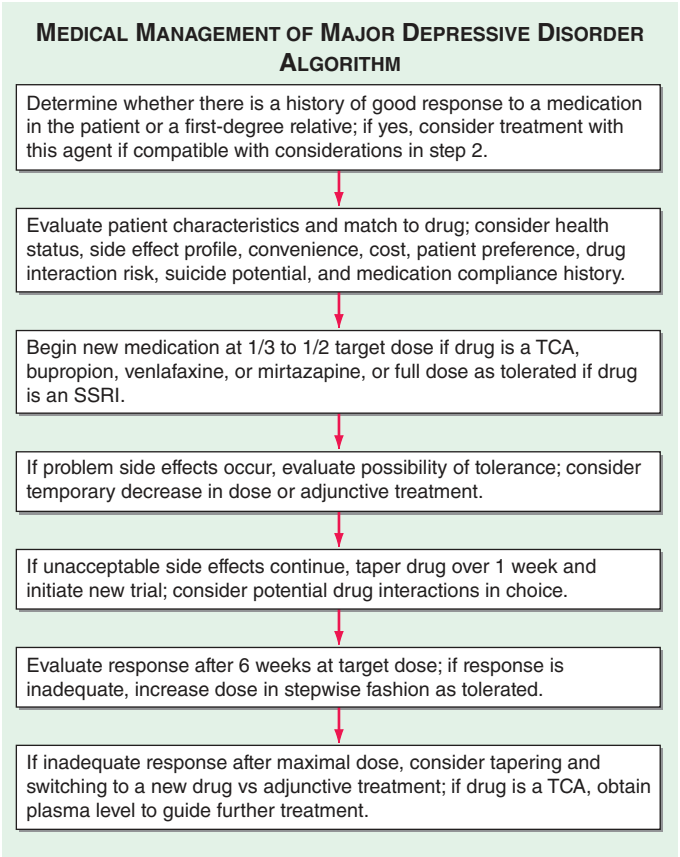


FIGURE 444-1 A guideline for the medical management of major depressive disorder. SSRI, selective serotonin reuptake inhibitor; TCA, tricyclic antidepressant.

Second-generation antidepressants are similar to tricyclics in their effect on neurotransmitter reuptake, although some also have specific actions on catecholamine and indolamine receptors as well. Amoxapine is a dibenzoxazepine derivative that blocks norepinephrine and serotonin reuptake and has a metabolite that shows a degree of dopamine blockade. Long-term use of this drug carries a risk of tardive dyskinesia. Maprotiline is a potent noradrenergic reuptake blocker that has little anticholinergic effect but may produce seizures. Bupropion is a novel antidepressant whose mechanism of action is thought to involve enhancement of noradrenergic function. It has no anticholinergic, sedating, or orthostatic side effects and has a low incidence of sexual side effects. It may, however, be associated with stimulant-like side effects, may lower seizure threshold, and has an exceptionally short half-life, requiring frequent dosing. An extended-release preparation is available.

SSRIs such as fluoxetine, sertraline, paroxetine, citalopram, and escitalopram cause a lower frequency of anticholinergic, sedating, and cardiovascular side effects but a possibly greater incidence of gastrointestinal complaints, sleep impairment, and sexual dysfunction than do TCAs. Akathisia, involving an inner sense of restlessness and anxiety in addition to increased motor activity, may also be more common, particularly during the first week of treatment. One concern is the risk of "serotonin syndrome," which is thought to result from hyperstimulation of brainstem 5-HT_{1A} receptors and characterized by myoclonus, agitation, abdominal cramping, hyperpyrexia, hypertension, and potentially death. Serotonergic agonists taken in combination should be monitored closely for this reason. Considerations such as half-life, compliance, toxicity, and drug-drug interactions may guide the choice of a particular SSRI. Fluoxetine and its principal active metabolite, norfluoxetine, for example, have a combined half-life of almost 7 days, resulting in a delay of 5 weeks before steady-state levels are achieved and a similar delay for complete drug excretion once its use is discontinued; paroxetine appears to incur a greater risk of withdrawal symptoms with abrupt discontinuation. All the SSRIs may impair sexual function, resulting in diminished libido, impotence, or difficulty in achieving orgasm. Sexual dysfunction frequently results in noncompliance and should be asked about specifically. Sexual dysfunction can sometimes be ameliorated by lowering the dose, by instituting weekend drug holidays (two or three times a month), or by treatment with amantadine (100 mg tid), bethanechol (25 mg tid), buspirone (10 mg tid), or bupropion (100–150 mg/d). Paroxetine appears to be more anticholinergic than either fluoxetine or sertraline, and sertraline carries a lower risk of producing an adverse drug interaction than the other

two. Rare side effects of SSRIs include angina due to vasospasm and prolongation of the prothrombin time. Escitalopram is the most specific of currently available SSRIs and appears to have no specific inhibitory effects on the P450 system.

Venlafaxine, desvenlafaxine, duloxetine, vilazodone, vortioxetine, and levomilnacipran block the reuptake of both norepinephrine and serotonin but produce relatively little in the way of traditional tricyclic side effects. Unlike the SSRIs, venlafaxine and vortioxetine have relatively linear dose-response curves. Patients on immediate release venlafaxine should be monitored for a possible increase in diastolic blood pressure, and multiple daily dosing is required because of the drug's short half-life. An extended-release form is available and has a somewhat lower incidence of gastrointestinal side effects. Mirtazapine is a TCA that has a unique spectrum of activity. It increases noradrenergic and serotonergic neurotransmission through a blockade of central α_2-adrenergic receptors and postsynaptic 5-HT_2 and 5-HT_3 receptors. It is also strongly antihistaminic and, as such, may produce sedation. Levomilnacipran is the most noradrenergic of the SNRIs and theoretically may be appropriate for patients with more severe fatigue and anergia.

With the exception of citalopram and escitalopram, each of the SSRIs may inhibit one or more cytochrome P450 enzymes. Depending on the specific isoenzyme involved, the metabolism of a number of concomitantly administered medications can be dramatically affected. Fluoxetine and paroxetine, for example, by inhibiting 2D6, can cause dramatic increases in the blood level of type 1C antiarrhythmics, whereas sertraline, by acting on 3A4, may alter blood levels of carbamazepine or digoxin. Depending on drug specificity for a particular CYP enzyme for its own metabolism, concomitant medications or dietary factors, such as grapefruit juice, may in turn affect the efficacy or toxicity of the SSRI.

The MAOIs are highly effective, particularly in atypical depression, but the risk of hypertensive crisis following intake of tyramine-containing food or sympathomimetic drugs makes them inappropriate as first-line agents. Transdermal selegiline may avert this risk at low dose. Common side effects include orthostatic hypotension, weight gain, insomnia, and sexual dysfunction. MAOIs should not be used concomitantly with SSRIs, because of the risk of serotonin syndrome, or with TCAs, because of possible hyperadrenergic effects.

Electroconvulsive therapy is at least as effective as medication, but its use is reserved for treatment-resistant cases and delusional depressions. Repetitive transcranial magnetic stimulation (rTMS) is approved for treatment-resistant depression and has been shown to have efficacy in several controlled trials. Vagus nerve stimulation (VNS) has also recently been approved for treatment-resistant depression, but its degree of efficacy is controversial. Some meta-analyses of low intensity transcranial current stimulation (tCS) have shown a positive benefit over sham treatment, but whether this is comparable to or synergistic with antidepressant treatment is unclear. In off-label usage, intravenous ketamine or esketamine and intranasal esketamine have been shown to have short-term antidepressant efficacy, often after a single administration, suggesting a possible utility in addressing suicidality. Questions remain, however, about the risk/benefit ratio over the longer term. Lastly, deep brain stimulation of the ventral anterior limb of the internal capsule and of the subcallosal cingulate region have demonstrable efficacy in randomized experimental trials of treatment-resistant depression.

Regardless of the treatment undertaken, the response should be evaluated after ~2 months. Three-quarters of patients show improvement by this time, but if remission is inadequate, the patient should be questioned about compliance, and an increase in medication dose should be considered if side effects are not troublesome. If this approach is unsuccessful, referral to a mental health specialist is advised. Strategies for treatment then include selection of an alternative drug, combinations of antidepressants, and/or adjunctive treatment with other classes of drugs, including lithium, thyroid hormone, l-methylfolate or s-adenylmethionine, atypical antipsychotic agents, and dopamine agonists. In switching to a different monotherapy, other drugs from the same class appear to be as likely to be efficacious as choosing a drug from a different class. A large randomized trial (STAR-D) was unable to show preferential efficacy, but the addition of certain atypical antipsychotic drugs (quetiapine extended-release; aripiprazole; brexpiprazole) has received FDA approval, as has usage of a combined medication, olanzapine and fluoxetine (Symbyax). Patients whose response to an SSRI wanes over time may benefit from the addition of buspirone (10 mg tid) or pindolol (2–5 mg tid) or small amounts of a TCA such as nortriptyline (25 mg bid or tid). Most patients will show some degree of response, but aggressive treatment should be pursued until remission is achieved, and drug treatment should be continued for at least 6–9 months to prevent relapse. In patients who have had two or more episodes of depression, indefinite maintenance treatment should be considered. Pharmacogenomic testing focusing on cytochrome p450 allelic variation may sometimes be helpful in identifying individuals who are poor or rapid metabolizers, but assessing pharmacodynamic gene variants has not been shown to be cost-effective or affect clinical outcomes.

It is essential to educate patients both about depression and the benefits and side effects of medications they are receiving. Advice about stress reduction and cautions that alcohol may exacerbate depressive symptoms and impair drug response are helpful. Patients should be given time to describe their experience, their outlook, and the impact of the depression on them and their families. Occasional empathic silence may be as helpful for the treatment alliance as verbal reassurance. Controlled trials have shown that cognitive-behavioral and interpersonal therapies are effective in improving psychological and social adjustment and that a combined treatment approach is more successful than medication alone for many patients.

BIPOLAR DISORDER

Clinical Manifestations Bipolar disorder is characterized by unpredictable swings in mood from mania (or hypomania) to depression. Some patients suffer only from recurrent attacks of *mania*, which in its pure form is associated with increased psychomotor activity; excessive social extroversion; decreased need for sleep; impulsivity and impaired judgment; and expansive, grandiose, and sometimes irritable mood **(Table 444-8)**. In severe mania, patients may experience delusions and paranoid thinking indistinguishable from schizophrenia. One-half of patients with bipolar disorder present with a mixture of psychomotor agitation and activation with dysphoria, anxiety, and irritability. It may be difficult to distinguish *mixed mania* from *agitated depression*. In some bipolar patients (*bipolar II disorder*), the full criteria for mania are lacking, and the requisite recurrent depressions are separated by periods of mild activation and increased energy (hypomania). In *cyclothymic disorder*, there are numerous hypomanic periods, usually of relatively short duration, alternating with clusters of depressive symptoms that fail, either in severity or duration, to meet the criteria of major depression. The mood fluctuations are chronic and should be present for at least 2 years before the diagnosis is made.

Manic episodes typically emerge over a period of days to weeks, but onset within hours is possible, usually in the early morning hours. An untreated episode of either depression or mania can be as short as several weeks or last as long as 8–12 months, and rare patients have an unremitting chronic course. The term *rapid cycling* is used for patients who have four or more episodes of either depression or mania in a given year. This pattern occurs in 15% of all patients, most of whom are women. In some cases, rapid cycling is linked to an underlying thyroid dysfunction, and in others, it is iatrogenically triggered by prolonged antidepressant treatment. Approximately one-half of patients have sustained difficulties in work performance and psychosocial functioning, with depressive phases being more responsible for impairment than mania.

Bipolar disorder is common, affecting ~1.5% of the population in the United States. Onset is typically between 20 and 30 years of age, but many individuals report premorbid symptoms in late childhood or early adolescence. The prevalence is similar for men and women;

TABLE 444-8 Criteria for a Manic Episode

A. A distinct period of abnormally and persistently elevated, expansive, or irritable mood and abnormally and persistently increased goal-directed activity or energy, lasting at least 1 week and present most of the day, nearly every day (or any duration if hospitalization is necessary).

B. During the period of the mood disturbance and increased energy or activity, three (or more) of the following symptoms (four if the mood is only irritable) are present to a significant degree and represent a noticeable change from usual behavior:

1. Inflated self-esteem or grandiosity.
2. Decreased need for sleep (e.g., feels rested after only 3 h of sleep).
3. More talkative than usual or pressure to keep talking.
4. Flight of ideas or subjective experience that thoughts are racing.
5. Distractibility (i.e., attention too easily drawn to unimportant or irrelevant external stimuli), as reported or observed.
6. Increase in goal-directed activity (either socially, at work or school, or sexually) or psychomotor agitation (i.e., purposeless non-goal-directed activity).
7. Excessive involvement in activities that have a high potential for painful consequences (e.g., engaging in unrestrained buying sprees, sexual indiscretions, or foolish business investments).

C. The mood disturbance is sufficiently severe to cause marked impairment in social or occupational functioning or to necessitate hospitalization to prevent harm to self or others, or there are psychotic features.

D. The episode is not attributable to the physiologic effects of a substance (e.g., a drug of abuse, a medication, or other treatment) or another medical condition.

Source: Diagnostic and Statistical Manual of Mental Disorders, 5th ed. Washington, DC, American Psychiatric Association, 2013.

women are likely to have more depressive and men more manic episodes over a lifetime.

Differential Diagnosis The differential diagnosis of mania includes secondary mania induced by stimulant or sympathomimetic drugs, hyperthyroidism, AIDS, and neurologic disorders such as Huntington's or Wilson's disease and cerebrovascular accidents. Comorbidity with alcohol and substance abuse is common, either because of poor judgment and increased impulsivity or because of an attempt to self-treat the underlying mood symptoms and sleep disturbances.

Etiology and Pathophysiology Genetic predisposition to bipolar disorder is evident from family studies; the concordance rate for monozygotic twins approaches 80%. A number of risk genes that have been identified to date overlap with those conveying risk for other psychiatric disorders, such as schizophrenia and autism, implying some degree of shared pathophysiology. Replicated loci include the alpha subunit of the L-type calcium channel (*CACNA1C*), teneurin transmembrane protein 4 (*ODZ4*), ankryn 3 (*ANK3*), neurocan (NCAN), and tetratricopeptide repeat and ankyrin repeat containing 1 (*TRANK1*). No clear biomarkers have been identified, but there is evidence for circadian rhythm dysregulation and oxidative stress, mitochondrial, and endoplasmic reticulum abnormalities.

TREATMENT

Bipolar Disorder

(Table 444-9) Lithium carbonate is the mainstay of treatment in bipolar disorder, although sodium valproate and carbamazepine, as well as a number of second-generation antipsychotic agents (aripiprazole, asenapine, olanzapine, quetiapine, risperidone, ziprasidone), also have FDA approval for the treatment of acute mania. Oxcarbazepine is not FDA approved, but appears to enjoy carbamazepine's spectrum of efficacy. The response rate to lithium carbonate is 70–80% in acute mania, with beneficial effects appearing in 1–2 weeks. Lithium also has a prophylactic effect in prevention of recurrent mania and, to a lesser extent, in the prevention of recurrent depression, which is more difficult to treat than unipolar depression. A simple cation, lithium is rapidly absorbed from the gastrointestinal tract and remains unbound to plasma or tissue proteins. Some 95% of a given dose is excreted unchanged through the kidneys within 24 h.

TABLE 444-9 Clinical Pharmacology of Mood Stabilizers

AGENT AND DOSING	SIDE EFFECTS AND OTHER EFFECTS
Lithium	**Common Side Effects**
Starting dose: 300 mg bid or tid Therapeutic blood level: 0.8–1.2 meq/L	Nausea/anorexia/diarrhea, fine tremor, thirst, polyuria, fatigue, weight gain, acne, folliculitis, neutrophilia, hypothyroidism Blood level is increased by thiazides, tetracyclines, and NSAIDs Blood level is decreased by bronchodilators, verapamil, and carbonic anhydrase inhibitors *Rare side effects:* Neurotoxicity, renal toxicity, hypercalcemia, ECG changes
Valproic Acid	**Common Side Effects**
Starting dose: 250 mg tid Therapeutic blood level: 50–125 μg/mL	Nausea/anorexia, weight gain, sedation, tremor, rash, alopecia Inhibits hepatic metabolism of other medications *Rare side effects:* Pancreatitis, hepatotoxicity, Stevens-Johnson syndrome
Carbamazepine/ Oxcarbazepine	**Common Side Effects**
Starting dose: 200 mg bid for carbamazepine, 150 mg bid for oxcarbazepine Therapeutic blood level: 4–12 μg/mL for carbamazepine	Nausea/anorexia, sedation, rash, dizziness/ataxia Carbamazepine, but not oxcarbazepine, induces hepatic metabolism of other medications *Rare side effects:* Hyponatremia, agranulocytosis, Stevens-Johnson syndrome
Lamotrigine	**Common Side Effects**
Starting dose: 25 mg/d	Rash, dizziness, headache, tremor, sedation, nausea *Rare side effect*: Stevens-Johnson syndrome

Abbreviations: ECG, electrocardiogram; NSAIDs, nonsteroidal anti-inflammatory drugs.

Serious side effects from lithium are rare, but minor complaints such as gastrointestinal discomfort, nausea, diarrhea, polyuria, weight gain, skin eruptions, alopecia, and edema are common. Over time, urine-concentrating ability may be decreased, but significant nephrotoxicity is relatively rare. Lithium exerts an antithyroid effect by interfering with the synthesis and release of thyroid hormones. More serious side effects include tremor, poor concentration and memory, ataxia, dysarthria, and incoordination. There is suggestive, but not conclusive, evidence that lithium is teratogenic, inducing cardiac malformations in the first trimester.

In the treatment of acute mania, lithium is initiated at 300 mg bid or tid, and the dose is then increased by 300 mg every 2–3 days to achieve blood levels of 0.8–1.2 meq/L. Because the therapeutic effect of lithium may not appear until after 7–10 days of treatment, adjunctive usage of lorazepam (1–2 mg every 4 h) or clonazepam (0.5–1 mg every 4 h) may be beneficial to control agitation. Antipsychotics are indicated in patients with severe agitation who respond only partially to benzodiazepines. Patients using lithium should be monitored closely, since the blood levels required to achieve a therapeutic benefit are close to those associated with toxicity.

Valproic acid may be more effective than lithium for patients who experience rapid cycling (i.e., more than four episodes a year) or who present with a mixed or dysphoric mania. Tremor and weight gain are the most common side effects; hepatotoxicity and pancreatitis are rare toxicities.

The recurrent nature of bipolar mood disorder necessitates maintenance treatment. A sustained blood lithium level of at least 0.8 meq/L is important for optimal prophylaxis and has been shown to reduce the risk of suicide, a finding not yet apparent for other mood stabilizers. Combinations of mood stabilizers together or with atypical antipsychotic drugs are sometimes required to maintain mood stability. Quetiapine extended release, olanzapine,

risperidone, and lamotrigine have been approved for maintenance treatment as sole agents, in combination with lithium and with aripiprazole and ziprasidone as adjunctive drugs. Lurasidone, olanzapine/fluoxetine, and quetiapine are also approved to treat acute depressive episodes in bipolar disorder. Compliance is frequently an issue and often requires enlistment and education of concerned family members. Efforts to identify and modify psychosocial factors that may trigger episodes are important, as is an emphasis on lifestyle regularity (social rhythm therapy). Mobile apps for smartphones that alert the individual and clinician to changes in activity and speech are proving useful in early detection of behavioral change and in delivering clinical interventions and education. Antidepressant medications are sometimes required for the treatment of severe breakthrough depressions, but their use should generally be avoided during maintenance treatment because of the risk of precipitating mania or accelerating the cycle frequency. Alternative off-label agents for bipolar depression include pramipexole, modafenil, omega-3 fatty acids, and *N*-acetyl cysteine; interventions such as ECT, light therapy and rTMS may also be effective. Loss of efficacy over time may be observed with any of the mood-stabilizing agents. In such situations, an alternative agent or combination therapy is usually helpful.

SOMATIC SYMPTOM DISORDER

Many patients presenting in general medical practice, perhaps as many as 5–7%, will experience a somatic symptom(s) as particularly distressing and preoccupying, to the point that it comes to dominate their thoughts, feelings, and beliefs and interferes to a varying degree with everyday functioning. Although the absence of a medical explanation for these complaints was historically emphasized as a diagnostic element, it has been recognized that the patient's interpretation and elaboration of the experience is the critical defining factor and that patients with well-established medical causation may qualify for the diagnosis. Multiple complaints are typical, but severe single symptoms can occur as well. Comorbidity with depressive and anxiety disorders is common and may affect the severity of the experience and its functional consequences. Personality factors may be a significant risk factor, as may a low level of educational or socioeconomic status or a history of recent stressful life events. Cultural factors are relevant as well and should be incorporated into the evaluation. Individuals who have persistent preoccupations about having or acquiring a serious illness, but who do not have a specific somatic complaint, may qualify for a related diagnosis—illness anxiety disorder. The diagnosis of conversion disorder (functional neurologic symptom disorder) is used to specifically identify those individuals whose somatic complaints involve one or more symptoms of altered voluntary motor or sensory function that cannot be medically explained and that causes significant distress or impairment or requires medical evaluation.

In *factitious illnesses*, the patient consciously and voluntarily produces physical symptoms of illness. The term *Munchausen's syndrome* is reserved for individuals with particularly dramatic, chronic, or severe factitious illness. In true factitious illness, the sick role itself is gratifying. A variety of signs, symptoms, and diseases have been either simulated or caused by factitious behavior, the most common including chronic diarrhea, fever of unknown origin, intestinal bleeding or hematuria, seizures, and hypoglycemia. Factitious disorder is usually not diagnosed until 5–10 years after its onset, and it can produce significant social and medical costs. In *malingering*, the fabrication derives from a desire for some external reward such as a narcotic medication or disability reimbursement.

TREATMENT

Somatic Symptom Disorder and Related Disorders

Patients with somatic symptom disorder are frequently subjected to many diagnostic tests and exploratory surgeries in an attempt to find their "real" illness. Such an approach is doomed to failure and does not address the core issue. Successful treatment is best achieved through behavior modification, in which access to the physician is tightly regulated and adjusted to provide a sustained and predictable level of support that is less clearly contingent on the patient's level of presenting distress. Visits can be brief and should not be associated with a need for a diagnostic or treatment action. Although the literature is limited, some patients may benefit from antidepressant treatment.

Any attempt to confront the patient usually creates a sense of humiliation and causes the patient to abandon treatment from that caregiver. A better strategy is to introduce psychological causation as one of a number of possible explanations in the differential diagnoses that are discussed. Without directly linking psychotherapeutic intervention to the diagnosis, the patient can be offered a face-saving means by which the pathologic relationship with the health care system can be examined and alternative approaches to life stressors developed. Specific medical treatments also may be indicated and effective in treating some of the functional consequences of conversion disorder.

FEEDING AND EATING DISORDERS

CLINICAL MANIFESTATIONS

Feeding and eating disorders constitute a group of conditions in which there is a persistent disturbance of eating or associated behaviors that significantly impair an individual's physical health or psychosocial functioning. In DSM-5 the described categories (with the exception of pica) are defined to be mutually exclusive in a given episode, based on the understanding that although they are phenotypically similar in some ways, they differ in course, prognosis, and effective treatment interventions. Compared with DSM-IV-TR, three disorders (i.e., avoidant/restrictive food intake disorder, rumination disorder, pica) that were previously classified as disorders of infancy or childhood have been grouped together with the disorders of anorexia and bulimia nervosa. Binge-eating disorder is also now included as a formal diagnosis; the intent of each of these modifications is to encourage clinicians to be more specific in their codification of eating and feeding pathology.

PICA

Pica is diagnosed when the individual, aged >2, eats one or more nonnutritive, nonfood substances for a month or more and requires medical attention as a result. There is usually no specific aversion to food in general but a preferential choice to ingest substances such as clay, starch, soap, paper, or ash. The diagnosis requires the exclusion of specific culturally approved practices and has not been commonly found to be caused by a specific nutritional deficiency. Onset is most common in childhood but the disorder can occur in association with other major psychiatric conditions in adults. An association with pregnancy has been observed, but the condition is only diagnosed when medical risks are increased by the behavior.

RUMINATION DISORDER

In this condition, individuals who have no demonstrable associated gastrointestinal or other medical condition repeatedly regurgitate their food after eating and then either rechew or swallow it or spit it out. The behavior typically occurs on a daily basis and must persist for at least 1 month. Weight loss and malnutrition are common sequelae, and individuals may attempt to conceal their behavior, either by covering their mouth or through social avoidance while eating. In infancy, the onset is typically between 3 and 12 months, and the behavior may remit spontaneously, although in some it appears to be recurrent.

AVOIDANT/RESTRICTIVE FOOD INTAKE DISORDER

The cardinal feature of this disorder is avoidance or restriction of food intake, usually stemming from a lack of interest in or distaste of food and associated with weight loss, nutritional deficiency, dependency on nutritional supplementation, or marked impairment in psychosocial

functioning, either alone or in combination. Culturally approved practices, such as fasting or a lack of available food, must be excluded as possible causes. The disorder is distinguished from anorexia nervosa by the presence of emotional factors, such as a fear of gaining weight and distortion of body image in the latter condition. Onset is usually in infancy or early childhood, but avoidant behaviors may persist into adulthood. The disorder is equally prevalent in males and females and is frequently comorbid with anxiety and cognitive and attention-deficit disorders and situations of familial stress. Developmental delay and functional deficits may be significant if the disorder is long-standing and unrecognized.

ANOREXIA NERVOSA

Individuals are diagnosed with anorexia nervosa if they restrict their caloric intake to a degree that their body weight deviates significantly from age, gender, health, and developmental norms and if they also exhibit a fear of gaining weight and an associated disturbance in body image. The condition is further characterized by differentiating those who achieve their weight loss predominantly through restricting intake or by excessive exercise (restricting type) from those who engage in recurrent binge eating and/or subsequent purging, self-induced vomiting, and usage of enemas, laxatives, or diuretics (binge-eating/purging type). Such subtyping is more state than trait specific, as individuals may transition from one profile to the other over time. Determination of whether an individual satisfies the primary criterion of significant low weight is complex and must be individualized, using all available historical information and comparison of body habitus to international body mass norms and guidelines.

Individuals with anorexia nervosa frequently lack insight into their condition and are in denial about possible medical consequences; they often are not comforted by their achieved weight loss and persist in their behaviors despite having met previously self-designated weight goals. Recent research has identified alterations in the circuitry of reward sensitivity and executive function in anorexia and implicated disturbances in frontal cortex and anterior insula regulation of interoceptive awareness of satiety and hunger. Neurochemical findings, including the role of ghrelin, remain controversial.

Onset is most common in adolescence, although onset in later life can occur. Many more females than males are affected, with a lifetime prevalence in women of up to 4%. The disorder appears most prevalent in postindustrialized and urbanized countries and is frequently comorbid with preexisting anxiety disorders. The medical consequences of prolonged anorexia nervosa are multisystemic and can be life-threatening in severe presentations. Changes in blood chemistry include leukopenia with lymphocytosis, elevations in blood urea nitrogen, and metabolic alkalosis and hypokalemia when purging is present. History and physical examination may reveal amenorrhea in females, skin abnormalities (petechiae, lanugo hair, dryness), and signs of hypometabolic function, including hypotension, hypothermia, and sinus bradycardia. Endocrine effects include hypogonadism, growth hormone resistance, and hypercortisolemia. Osteoporosis is a longer-term concern.

The course of the disorder is variable, with some individuals recovering after a single episode, while others exhibit recurrent episodes or a chronic course. Untreated anorexia has a mortality of 5.1/1000, the highest among psychiatric conditions. Maudsley Anorexia Nervosa Treatment for Adults (MANTRA) and eating disorder focused cognitive behavior therapy have proven to be effective therapies, with strict behavioral contingencies used when weight loss becomes critical. No pharmacologic intervention has proven to be specifically beneficial, but comorbid depression and anxiety should be treated. Weight gain should be undertaken gradually with a goal of 0.5–1 pound per week to prevent refeeding syndrome. Most individuals are able to achieve remission within 5 years of the original diagnosis.

BULIMIA NERVOSA

Bulimia nervosa describes individuals who engage in recurrent and frequent (at least once a week for 3 months) periods of binge eating and who then resort to compensatory behaviors, such as self-induced purging, enemas, use of laxatives, or excessive exercise to avoid weight gain. Binge eating itself is defined as excessive food intake in a prescribed period of time, usually <2 h. As in anorexia nervosa, disturbances in body image occur and promote the behavior, but unlike in anorexia, individuals are of normal weight or even somewhat overweight. Subjects typically describe a loss of control and express shame about their actions, and often relate that their episodes are triggered by feelings of negative self-esteem or social stresses. The lifetime prevalence in women is ~2%, with a 10:1 female-to-male ratio. The disorder typically begins in adolescence and may be persistent over a number of years. Transition to anorexia occurs in only 10–15% of cases. Many of the medical risks associated with bulimia nervosa parallel those of anorexia nervosa and are a direct consequence of purging, including fluid and electrolyte disturbances and conduction abnormalities. Physical examination often results in no specific findings, but dental erosion and parotid gland enlargement may be present. Effective treatment approaches include SSRI antidepressants, usually in combination with cognitive-behavioral, emotion regulation, or interpersonal-based psychotherapies.

BINGE-EATING DISORDER

Binge-eating disorder is distinguished from bulimia nervosa by the absence of compensatory behaviors to prevent weight gain after an episode and by a lack of effort to restrict weight gain between episodes. Other features are similar, including distress over the behavior and the experience of loss of control, resulting in eating more rapidly or in greater amounts than intended or eating when not hungry. The 12-month prevalence in females is 1.6%, with a much lower female-to-male ratio than bulimia nervosa. Little is known about the course of the disorder, given its recent categorization, but its prognosis is markedly better than for other eating disorders, both in terms of its natural course and response to treatment. Transition to other eating disorder conditions is thought to be rare.

PERSONALITY DISORDERS

CLINICAL MANIFESTATIONS

Personality disorders are characteristic patterns of thinking, feeling, and interpersonal behavior that are relatively inflexible and cause significant functional impairment or subjective distress for the individual. The observed behaviors are not secondary to another mental disorder, nor are they precipitated by substance abuse or a general medical condition. This distinction is often difficult to make in clinical practice, because personality change may be the first sign of serious neurologic, endocrine, or other medical illness. Patients with frontal lobe tumors, for example, can present with changes in motivation and personality while the results of the neurologic examination remain within normal limits. Individuals with personality disorders are often regarded as "difficult patients" in clinical medical practice because they are seen as excessively demanding and/or unwilling to follow recommended treatment plans. Although DSM-5 portrays personality disorders as qualitatively distinct categories, there is an alternative and emerging perspective that personality characteristics vary as a continuum between normal functioning and formal mental disorder, the essential features being moderate or greater impairment in self/interpersonal functioning and one or more pathological personality traits.

Personality disorders have been grouped into three overlapping clusters. *Cluster A* includes paranoid, schizoid, and schizotypal personality disorders. It includes individuals who are odd and eccentric and who maintain an emotional distance from others. Individuals have a restricted emotional range and remain socially isolated. Patients with schizotypal personality disorder frequently have unusual perceptual experiences and express magical beliefs about the external world. The essential feature of paranoid personality disorder is a pervasive mistrust and suspiciousness of others to an extent that is unjustified by available evidence. *Cluster B* disorders include antisocial, borderline, histrionic, and narcissistic types and describe individuals whose behavior is impulsive, excessively emotional, and erratic. *Cluster C*

incorporates avoidant, dependent, and obsessive-compulsive personality types; enduring traits are anxiety and fear. The boundaries between cluster types are to some extent artificial, and many patients who meet criteria for one personality disorder also meet criteria for aspects of another. The risk of a comorbid major mental disorder is increased in patients who qualify for a diagnosis of personality disorder.

ETIOLOGY AND PATHOPHYSIOLOGY

Genetic studies have increasingly suggested a genetic contribution to the development of personality disorders. One study of 106,000 subjects identified 9 loci significantly linked to aspects of neuroticism.

TREATMENT

Personality Disorders

Dialectical behavior therapy (DBT) is a cognitive-behavioral approach that focuses on behavioral change while providing acceptance, compassion, and validation of the patient. Several randomized trials have demonstrated the efficacy of DBT in the treatment of personality disorders. Antidepressant medications and low-dose antipsychotic drugs have some efficacy in cluster A personality disorders, whereas anticonvulsant mood-stabilizing agents and MAOIs may be considered for patients with cluster B diagnoses who show marked mood reactivity, behavioral dyscontrol, and/or rejection hypersensitivity. Anxious or fearful cluster C patients often respond to medications used for axis I anxiety disorders (see above). It is important that the physician and the patient have reasonable expectations vis-à-vis the possible benefit of any medication used and its side effects. Improvement may be subtle and observable only over time.

SCHIZOPHRENIA

CLINICAL MANIFESTATIONS

Schizophrenia is a heterogeneous syndrome characterized by perturbations of language, perception, thinking, social activity, affect, and volition. There are no pathognomonic features. The syndrome commonly begins in late adolescence, has an insidious (and less commonly acute) onset, and, often, a poor outcome, progressing from social withdrawal and perceptual distortions to recurrent delusions and hallucinations. Patients may present with positive symptoms (such as conceptual disorganization, delusions, or hallucinations) or negative symptoms (loss of function, anhedonia, decreased emotional expression, impaired concentration, and diminished social engagement) and must have at least two of these for a 1-month period and continuous signs for at least 6 months to meet formal diagnostic criteria. Disorganized thinking or speech and grossly disorganized motor behavior, including catatonia, may also be present. As individuals age, positive psychotic symptoms tend to attenuate, and some measure of social and occupational function may be regained. "Negative" symptoms predominate in one-third of the schizophrenic population and are associated with a poor long-term outcome and a poor response to drug treatment. However, marked variability in the course and individual character of symptoms is typical.

The term *schizophreniform disorder* describes patients who meet the symptom requirements but not the duration requirements for schizophrenia, and *schizoaffective disorder* is used for those who manifest symptoms of schizophrenia and independent periods of mood disturbance. The terms *schizotypal* and *schizoid* refer to specific personality disorders and are discussed in that section. The diagnosis of delusional disorder is used for individuals who have delusions of various content for at least 1 month but who otherwise do not meet criteria for schizophrenia. Patients who experience a sudden onset of a brief (<1 month) alteration in thought processing, characterized by delusions, hallucinations, disorganized speech, or gross motor behavior, are most appropriately designated as having a brief psychotic disorder. Catatonia is recognized as a nonspecific syndrome that can occur as a consequence of other severe psychiatric/medical disorders and is diagnosed by the documentation of three or more of a cluster of motor and behavioral symptoms, including stupor, cataplexy, mutism, waxy flexibility, and stereotypy, among others. Prognosis depends not on symptom severity but on the response to antipsychotic medication. A permanent remission without recurrence does occasionally occur. About 10% of schizophrenic patients commit suicide.

Schizophrenia is present in 0.85% of individuals worldwide, with a lifetime prevalence of ~1–1.5%. An estimated 300,000 episodes of acute schizophrenia occur annually in the United States, resulting in direct and indirect costs of $62.7 billion.

DIFFERENTIAL DIAGNOSIS

The diagnosis is principally one of exclusion, requiring the absence of significant associated mood symptoms, any relevant medical condition, and substance abuse. Drug reactions that cause hallucinations, paranoia, confusion, or bizarre behavior may be dose-related or idiosyncratic; parkinsonian medications, clonidine, quinacrine, and procaine derivatives are the most common prescription medications associated with these symptoms. Drug causes should be ruled out in any case of newly emergent psychosis. The general neurologic examination in patients with schizophrenia is usually normal, but motor rigidity, tremor, and dyskinesias are noted in one-quarter of untreated patients.

EPIDEMIOLOGY AND PATHOPHYSIOLOGY

Epidemiologic surveys identify several risk factors for schizophrenia, including genetic susceptibility, early developmental insults, winter birth, and increasing parental age. Genetic factors are involved in at least a subset of individuals who develop schizophrenia. Schizophrenia is observed in ~6.6% of all first-degree relatives of an affected proband. If both parents are affected, the risk for offspring is 40%. The concordance rate for monozygotic twins is 50%, compared to 10% for dizygotic twins. Schizophrenia-prone families are also at risk for other psychiatric disorders, including schizoaffective disorder and *schizotypal* and *schizoid personality disorders*, the latter terms designating individuals who show a lifetime pattern of social and interpersonal deficits characterized by an inability to form close interpersonal relationships, eccentric behavior, and mild perceptual distortions. Large scale genomewide association studies have identified >100 small effect risk loci and a few larger effect copy number variants, along with epigenetic effects. Pathways identified include ones involved in immunity, inflammation, and cell signaling.

TREATMENT

Schizophrenia

Antipsychotic agents **(Table 444-10)** are the cornerstone of acute and maintenance treatment of schizophrenia and are effective in the treatment of hallucinations, delusions, and thought disorders, regardless of etiology. The mechanism of action involves, at least in part, binding to dopamine D_2/D_3 receptors in the ventral striatum; the clinical potencies of traditional antipsychotic drugs parallel their affinities for the D_2 receptor, and even the newer "atypical" agents exert some degree of D_2 receptor blockade. All neuroleptics induce expression of the immediate-early gene c-*fos* in the nucleus accumbens, a dopaminergic site connecting prefrontal and limbic cortices. The clinical efficacy of newer atypical neuroleptics, however, may involve *N*-methyl-D-aspartate (NMDA) receptor blockade, α_1- and α_2-noradrenergic activity, altering the relationship between 5-HT_2 and D_2 receptor activity, and faster dissociation of D_2 binding and effects on neuroplasticity.

Conventional neuroleptics differ in their potency and side effect profile. Older agents, such as chlorpromazine and thioridazine, are more sedating and anticholinergic and more likely to cause orthostatic hypotension, whereas higher potency antipsychotics, such as haloperidol, perphenazine, and thiothixene, are more likely to induce extrapyramidal side effects. The model "atypical" antipsychotic agent is clozapine, a dibenzodiazepine that has a greater

TABLE 444-10 Antipsychotic Agents

NAME	USUAL PO DAILY DOSE, mg	SIDE EFFECTS	SEDATION	COMMENTS
First-Generation Antipsychotics				
Low potency				
Chlorpromazine (Thorazine)	100–1000	Anticholinergic effects; orthostasis; photosensitivity; cholestasis; QT prolongation	+++	EPSEs usually not prominent; can cause anticholinergic delirium in elderly patients
Thioridazine (Mellaril)	100–600			
Midpotency				
Trifluoperazine (Stelazine)	2–50	Fewer anticholinergic side effects	++	Well tolerated by most patients
Perphenazine (Trilafon)	4–64	Fewer EPSEs than with higher potency agents	++	
Loxapine (Loxitane)	30–100	Frequent EPSEs	++	
Molindone (Moban)	30–100	Frequent EPSEs	0	Little weight gain
High potency				
Haloperidol (Haldol)	5–20	No anticholinergic side effects; EPSEs often prominent	0/+	Often prescribed in doses that are too high; long-acting injectable forms of haloperidol and fluphenazine available
Fluphenazine (Prolixin)	1–20	Frequent EPSEs	0/+	
Thiothixene (Navane)	2–50	Frequent EPSEs	0/+	
Second-Generation Antipsychotics				
Clozapine (Clozaril)	150–600	Agranulocytosis (1%); weight gain; seizures; drooling; hyperthermia	+ +	Requires weekly WBC count for first 6 months, then biweekly if stable
Risperidone (Risperdal)	2–8	Orthostasis	+	Requires slow titration; EPSEs observed with doses >6 mg qd
Olanzapine (Zyprexa)	10–30	Weight gain	++	Mild prolactin elevation
Quetiapine (Seroquel)	350–800	Sedation; weight gain; anxiety	+++	Bid dosing
Ziprasidone (Geodon)	120–200	Orthostatic hypotension	+/++	Minimal weight gain; increases QT interval
Aripiprazole (Abilify)	10–30	Nausea, anxiety, insomnia	0/+	Mixed agonist/antagonist; ER available
Paliperidone (Invega)	3–12	Restlessness, EPSEs, increased prolactin, headache	+	Active metabolite of risperidone
Iloperidone (Fanapt)	12–24	Dizziness, hypotension	0/+	Requires dose titration; long acting injectable available
Asenapine (Saphris)	10–20	Dizziness, anxiety, EPSEs, minimal weight gain	++	Sublingual tablets; bid dosing
Lurasidone (Latuda)	40–80	Nausea, EPSEs	++	Uses CYP3A4
Brexpiprazole (Rexulti)	1–4	Anxiety, dizziness, fatigue	++	CYP3A4 and 2D6 interactions
Pimavanserin (Nuplazid)	34	Edema, confusion, sedation	++	Approved for Parkinson's disease psychosis
Cariprazine (Vraylar)	1.5–6	EPSEs, vomiting	++	Preferential D3 receptor affinity

Abbreviations: EPSEs, extrapyramidal side effects; WBC, white blood cell.

potency in blocking the 5-HT_2 than the D_2 receptor and a much higher affinity for the D_4 than the D_2 receptor. Its principal disadvantage is a risk of blood dyscrasias. Paliperidone is a recently approved agent that is a metabolite of risperidone and shares many of its properties. Unlike other antipsychotics, clozapine does not cause a rise in prolactin level. Approximately 30% of patients who do not benefit from conventional antipsychotic agents will have a better response to this drug, which also has a demonstrated superiority to other antipsychotic agents in preventing suicide; however, its side effect profile makes it most appropriate for treatment-resistant cases. Risperidone, a benzisoxazole derivative, is more potent at 5-HT_2 than D_2 receptor sites, like clozapine, but it also exerts significant α_2 antagonism, a property that may contribute to its perceived ability to improve mood and increase motor activity. Risperidone is not as effective as clozapine in treatment-resistant cases but does not carry a risk of blood dyscrasias. Olanzapine is similar neurochemically to clozapine but has a significant risk of inducing weight gain. Quetiapine is distinct in having a weak D_2 effect but potent α_1 and histamine blockade. Ziprasidone causes minimal weight gain and is unlikely to increase prolactin but may increase QT prolongation. Aripiprazole also has little risk of weight gain or prolactin increase but may increase anxiety, nausea, and insomnia as a result of its partial agonist properties. Asenapine is associated with minimal weight gain and anticholinergic effect but may have a higher than expected risk of extrapyramidal symptoms.

Antipsychotic agents are effective in 70% of patients presenting with a first episode. Improvement may be observed within hours or days, but full remission usually requires 6–8 weeks. The choice of agent depends principally on the side effect profile and cost of treatment or on a past personal or family history of a favorable response to the drug in question. Atypical agents appear to be more effective in treating negative symptoms and improving cognitive function. An equivalent treatment response can usually be achieved with relatively low doses of any drug selected (i.e., 4–6 mg/d of haloperidol, 10–15 mg of olanzapine, or 4–6 mg/d of risperidone). Doses in this range result in >80% D_2 receptor blockade, and there is little evidence that higher doses increase either the rapidity or degree of response. Maintenance treatment requires careful attention to the possibility of relapse and monitoring for the development of a movement disorder. Intermittent drug treatment is less effective than regular dosing, but gradual dose reduction is likely to improve social functioning in many schizophrenic patients who have been maintained at high doses. If medications are completely discontinued, however, the relapse rate is 60% within 6 months. Long-acting injectable preparations (risperidone, paliperidone, olanzapine, aripiprazole) are considered when noncompliance with oral therapy leads to relapses but should not be considered interchangeable, because the agents differ in their indications, injection intervals and sites/volumes, and possible adverse reactions, among other factors. In treatment-resistant patients, a transition to clozapine usually results in rapid improvement, but a prolonged delay in response in some cases necessitates a 6- to 9-month trial for maximal benefit to occur.

Antipsychotic medications can cause a broad range of side effects, including lethargy, weight gain, postural hypotension, constipation, and dry mouth. Extrapyramidal symptoms such as dystonia,

akathisia, and akinesia are also frequent with first-generation agents and may contribute to poor adherence if not specifically addressed. Anticholinergic and parkinsonian symptoms respond well to trihexyphenidyl, 2 mg bid, or benztropine mesylate, 1–2 mg bid. Akathisia may respond to beta blockers. In rare cases, more serious and occasionally life-threatening side effects may emerge, including hyperprolactinemia, ventricular arrhythmias, gastrointestinal obstruction, retinal pigmentation, obstructive jaundice, and neuroleptic malignant syndrome (characterized by hyperthermia, autonomic dysfunction, muscular rigidity, and elevated creatine phosphokinase levels). The most serious adverse effects of clozapine are agranulocytosis, which has an incidence of 1%, and induction of seizures, which has an incidence of 10%. Weekly white blood cell counts are required, particularly during the first 3 months of treatment.

The risk of type 2 diabetes mellitus appears to be increased in schizophrenia, and second-generation agents as a group produce greater adverse effects on glucose regulation, independent of effects on obesity, than traditional agents. Clozapine, olanzapine, and quetiapine seem more likely to cause hyperglycemia, weight gain, and hypertriglyceridemia than other atypical antipsychotic drugs. Close monitoring of plasma glucose and lipid levels are indicated with the use of these agents.

A serious side effect of long-term use of first-generation and, to a lesser extent, second-generation antipsychotic agents is tardive dyskinesia, characterized by repetitive, involuntary, and potentially irreversible movements of the tongue and lips (bucco-linguo-masticatory triad) and, in approximately half of cases, choreoathetosis. Tardive dyskinesia has an incidence of 2–4% per year of exposure and a prevalence of 20% in chronically treated patients. The prevalence increases with age, total dose, and duration of drug administration and may involve formation of free radicals and perhaps mitochondrial energy failure. Valbenazine, a vesicular monoamine transporter 2 inhibitor that depletes presynaptic dopamine, has recently received FDA approval for treatment of tardive dyskinesia.

The CATIE study, a large-scale investigation of the effectiveness of antipsychotic agents in "real-world" patients, revealed a high rate of discontinuation of treatment >18 months. Olanzapine showed greater effectiveness than quetiapine, risperidone, perphenazine, or ziprasidone but also a higher discontinuation rate due to weight gain and metabolic effects. Surprisingly, perphenazine, a first-generation agent, showed little evidence of inferiority to newer drugs.

Drug treatment of schizophrenia is by itself insufficient. Educational efforts directed toward families and relevant community resources have proved to be necessary to maintain stability and optimize outcome. A treatment model using social cognition interventions and involving a multidisciplinary case-management team that seeks out and closely follows the patient in the community has proved particularly effective. Attempts to prevent schizophrenia through early identification and treatment (both psychosocial and psychopharmacologic) of high-risk children and adolescents are currently being evaluated.

ASSESSMENT AND EVALUATION OF VIOLENCE

Primary care physicians may encounter situations in which family, domestic, or societal violence is discovered or suspected. Such an awareness can carry legal and moral obligations; many state laws mandate reporting of child, spousal, and elder abuse. Physicians are frequently the first point of contact for both victim and abuser. Approximately 2 million older Americans and 1.5 million U.S. children are thought to experience some form of physical maltreatment each year. Spousal abuse is thought to be even more prevalent. An interview study of 24,000 women in 10 countries found a lifetime prevalence of physical or sexual violence that ranged from 15 to 71%; these individuals are more likely to suffer from depression, anxiety, and substance abuse and to have attempted suicide. In addition, abused individuals frequently express low self-esteem, vague somatic symptomatology, social isolation, and a passive feeling of loss of control. Although it is essential to treat these elements in the victim, the first obligation is to ensure that the perpetrator has taken responsibility for preventing any further violence. Substance abuse and/or dependence and serious mental illness in the abuser may contribute to the risk of harm and require direct intervention. Depending on the situation, law enforcement agencies, community resources such as support groups and shelters, and individual and family counseling can be appropriate components of a treatment plan. A safety plan should be formulated with the victim, in addition to providing information about abuse, its likelihood of recurrence, and its tendency to increase in severity and frequency. Antianxiety and antidepressant medications may sometimes be useful in treating the acute symptoms, but only if independent evidence for an appropriate psychiatric diagnosis exists.

FURTHER READING

Combs G, Oshman L: Pearls for working with people who have personality disorder diagnoses. Prim Care 43:263, 2017.

Craske MG, Stein MB: Anxiety. Lancet 388:3048, 2016.

Gelauff J, Stone J: Prognosis of functional neurologic disorders. Handb Clin Neurol 139:523, 2016.

Gitlin M, Miklowitz DJ: The difficult lives of individuals with bipolar disorder: A review of functional outcomes and their implications for treatment. J Affec Disord 209:147, 2017.

Kok RM, Reynolds CF: Management of depression in older adults: A review. JAMA 317:2114, 2017.

Maselko J: Social epidemiology and global mental health: Expanding the evidence from high-income to low- and middle-income countries. Curr Epidemiol Rep 3:166, 2017

Morin L, Franck N: Rehabilitation interventions to promote recovery from schizophrenia: A systematic review. Front Psychiatry 8:100, 2017.

National Guideline Alliance (UK): Eating Disorders: Recognition and treatment. National Institute for Health and Care Excellence (NICE) May 2017.

Ortiz-Orendain J et al: Antipsychotic combinations for schizophrenia. Cochrane Database Syst Rev 28:6, 2017.

Shalev A et al: Post-traumatic stress disorder. N Engl J Med 376:2459, 2017.

445 Alcohol and Alcohol Use Disorders

Marc A. Schuckit

Alcohol (beverage ethanol) has diverse and widespread effects on the body, and impacts directly or indirectly on almost every neurochemical system in the brain. A large majority of patients in most clinical settings consume alcohol, with the highest proportions of drinkers of at least modest levels of alcohol seen in more educated and affluent patient groups. At even relatively low doses, this drug can exacerbate most medical problems and affect medications metabolized in the liver, and at higher doses can temporarily mimic many medical (e.g., diabetes) and psychiatric (e.g., depression) conditions. The lifetime risk for repetitive serious alcohol problems (e.g., alcohol use disorders as described below) in patients is about 20% for men and 10% for women, regardless of a person's education or income. Although low doses of alcohol might have healthful benefits, greater than three standard drinks per day enhances the risk for cancer and vascular disease, and alcohol use disorders decrease the life span by about 10 years. Unfortunately, most clinicians have had only limited training regarding identifying and

treating alcohol-related disorders. This chapter presents a brief overview of clinically useful information about alcohol use and associated problems.

PHARMACOLOGY AND NUTRITIONAL IMPACT OF ETHANOL

Ethanol blood levels are expressed as milligrams or grams of ethanol per deciliter (e.g., 100 mg/dL = 0.10 g/dL), with values of ~0.02 g/dL resulting from the ingestion of one typical drink. In round figures, a standard drink is 10–12 g, as seen in 340 mL (12 oz) of beer, 115 mL (4 oz) of nonfortified wine, and 43 mL (1.5 oz) (a shot) of 80-proof beverage (e.g., whisky); 0.5 L (1 pint) of 80-proof beverage contains ~160 g of ethanol (about 16 standard drinks), and 750 mL of wine contains ~60 g of ethanol. These beverages also have additional components (*congeners*) that affect the drink's taste and might contribute to adverse effects on the body. Congeners include methanol, butanol, acetaldehyde, histamine, tannins, iron, and lead. Alcohol acutely decreases neuronal activity and has similar behavioral effects and cross-tolerance with other depressants, including benzodiazepines and barbiturates.

Alcohol is absorbed from mucous membranes of the mouth and esophagus (in small amounts), from the stomach and large bowel (in modest amounts), and from the proximal portion of the small intestine (the major site). The rate of absorption is increased by rapid gastric emptying (as seen with carbonation); by the absence of proteins, fats, or carbohydrates (which interfere with absorption); and by dilution to a modest percentage of ethanol (maximum at ~20% by volume).

Between 2% (at low blood alcohol concentrations) and 10% (at high blood alcohol concentrations) of ethanol is excreted directly through the lungs, urine, or sweat, but most is metabolized to acetaldehyde, primarily in the liver. The most important pathway occurs in the cell cytosol where alcohol dehydrogenase (ADH) produces acetaldehyde, which is then rapidly destroyed by aldehyde dehydrogenase (ALDH) in the cytosol and mitochondria **(Fig. 445-1)**. A second pathway occurs in the microsomes of the smooth endoplasmic reticulum (the microsomal ethanol-oxidizing system or MEOS) that is responsible for ≥10% of ethanol oxidation at high blood alcohol concentrations.

Although a drink contains ~300 kJ, or 70–100 kcal, these are devoid of minerals, proteins, and vitamins. In addition, alcohol interferes with absorption of vitamins in the small intestine and decreases their storage in the liver with modest effects on folate (folacin or folic acid), pyridoxine (B_6), thiamine (B_1), nicotinic acid (niacin, B_3), and vitamin A.

Heavy drinking in a fasting, healthy individual can produce transient hypoglycemia within 6–36 h, secondary to the acute actions of ethanol that decrease gluconeogenesis. This can result in temporary abnormal glucose tolerance tests (with a resulting erroneous diagnosis of diabetes mellitus) until the heavy drinker has abstained for 2–4 weeks. Alcohol ketoacidosis, probably reflecting a decrease in fatty acid oxidation coupled with poor diet or persistent vomiting, can be misdiagnosed as diabetic ketosis. With alcohol-related ketoacidosis, patients show an increase in serum ketones along with a mild increase in glucose but a large anion gap, a mild to moderate increase in serum lactate, and a β-hydroxybutyrate/lactate ratio of between 2:1 and 9:1 (with normal being 1:1).

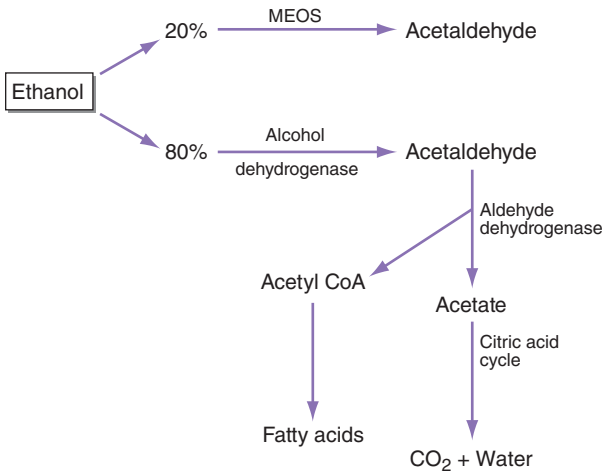


FIGURE 445-1 The metabolism of alcohol. CoA, coenzyme A; MEOS, microsomal ethanol oxidizing system.

In the brain, alcohol affects almost all neurotransmitter systems, with acute effects that are often the opposite of those seen following desistance after a period of heavy drinking. The most prominent acute actions relate to boosting γ-aminobutyric acid (GABA) activity, especially at $GABA_A$ receptors. Enhancement of this complex chloride channel system contributes to anticonvulsant, sleep-inducing, antianxiety, and muscle relaxation effects of all GABA-boosting drugs. Acutely administered alcohol produces a release of GABA, and continued use increases density of $GABA_A$ receptors, whereas alcohol withdrawal states are characterized by decreases in GABA-related activity. Equally important is the ability of acute alcohol to inhibit postsynaptic *N*-methyl-D-aspartate (NMDA) excitatory glutamate receptors, whereas chronic drinking and desistance are associated with an upregulation of these excitatory receptor subunits. The relationships between greater GABA and diminished NMDA receptor activity during acute intoxication and diminished GABA with enhanced NMDA actions during alcohol withdrawal explain much of intoxication and withdrawal phenomena.

As with all pleasurable activities, alcohol acutely increases dopamine levels in the ventral tegmentum and related brain regions, and this effect plays an important role in continued alcohol use, craving, and relapse. The changes in dopamine pathways are also linked to increases in "stress hormones," including cortisol and adrenocorticotropic hormone (ACTH) during intoxication and in the context of the stresses of withdrawal. Such alterations are likely to contribute to both feelings of reward during intoxication and depression during falling blood alcohol concentrations. Also closely linked to alterations in dopamine (especially in the nucleus accumbens) are alcohol-induced changes in opioid receptors, with acute alcohol causing release of β-endorphins.

Additional neurochemical changes include increases in synaptic levels of serotonin during acute intoxication and subsequent upregulation of serotonin receptors. Acute increases in nicotinic acetylcholine systems contribute to the impact of alcohol in the ventral tegmental region, which occurs in concert with enhanced dopamine activity. In the same regions, alcohol impacts on cannabinol receptors, with resulting release of dopamine, GABA, and glutamate as well as subsequent effects on brain reward circuits.

BEHAVIORAL EFFECTS, TOLERANCE, AND WITHDRAWAL

The acute effects of a drug depend on the dose, the rate of increase in plasma, the concomitant presence of other drugs, and past experience with the agent. "Legal intoxication" with alcohol in most states is based on a blood alcohol concentration of 0.08 g/dL, some states are considering lowering acceptable levels to <.05 g/dL, and levels of 0.04 are cited for pilots in the United States and automobile drivers in some other countries. However, behavioral, psychomotor, and cognitive changes are seen at 0.02–0.04 g/dL (i.e., after one to two drinks) **(Table 445-1)**. Deep but disturbed sleep can be seen at .15 g/dL in individuals who have not developed tolerance, and death can occur with levels between

TABLE 445-1 Effects of Blood Alcohol Levels in the Absence of Tolerance

BLOOD LEVEL, g/dL	USUAL EFFECT
0.02	Decreased inhibitions, a slight feeling of intoxication
0.08	Decrease in complex cognitive functions and motor performance
0.20	Obvious slurred speech, motor incoordination, irritability, and poor judgment
0.30	Light coma and depressed vital signs
0.40	Death

0.30 and 0.40 g/dL. Beverage alcohol is probably responsible for more overdose deaths than any other drug.

Repeated use of alcohol contributes to the need for a greater number of standard drinks to produce effects originally observed with fewer drinks (acquired tolerance), a phenomenon involving at least three compensatory mechanisms. (1) After 1–2 weeks of daily drinking, *metabolic* or *pharmacokinetic tolerance* can be seen, with up to 30% increases in the rate of hepatic ethanol metabolism. This alteration disappears almost as rapidly as it develops. (2) *Cellular or pharmacodynamic tolerance* develops through neurochemical changes that maintain relatively normal physiologic functioning despite the presence of alcohol. Subsequent decreases in blood levels contribute to symptoms of withdrawal. (3) Individuals learn to adapt their behavior so that they can function better than expected under the influence of the drug (*learned* or *behavioral tolerance*).

The cellular changes caused by chronic ethanol exposure may not resolve for several weeks or longer following cessation of drinking. Rapid decreases in blood alcohol levels before that time can produce a withdrawal syndrome, which is most intense during the first 5 days, but with some symptoms (e.g., disturbed sleep and anxiety) lasting up to 4–6 months as part of a "protracted withdrawal" syndrome.

THE EFFECTS OF ETHANOL ON ORGAN SYSTEMS

Relatively low doses of alcohol (one or two drinks per day) may have potential beneficial effects of increasing high-density lipoprotein cholesterol and decreasing aggregation of platelets, with a resulting possible decrease in risk for occlusive coronary disease and embolic strokes. Red wine has additional potential health-promoting qualities at relatively low doses due to flavinols and related substances. Such modest drinking might also decrease the risk for vascular dementia and, possibly, Alzheimer's disease. However, any potential healthful effects disappear with the regular consumption of three or more drinks per day, and knowledge about the deleterious effects of alcohol can both help the physician to identify patients with alcohol use disorders and to supply them with information that might help motivate changes in behavior.

NERVOUS SYSTEM

Approximately 35% of drinkers overall, including as many as 50% of drinking college students and a much higher proportion of individuals with alcohol use disorders, ever experience a *blackout*. This is an episode of temporary anterograde amnesia, in which the person was awake but forgot all (en bloc blackouts at blood alcohol levels >.20 mg/dL) or part (fragmentary blackouts at >.12 mg/dL) of what occurred during a drinking period.

Another common problem, one seen after as few as one or two drinks shortly before bedtime, is disturbed sleep. Although alcohol might initially help a person fall asleep, it disrupts sleep throughout the rest of the night. The stages of sleep are altered, and times spent in rapid eye movement (REM) and deep sleep early in the night are reduced. Alcohol relaxes muscles in the pharynx, which can cause snoring and exacerbate sleep apnea; symptoms of the latter occur in 75% of men with alcohol use disorders aged ≥60 years. Patients may also experience prominent and sometimes disturbing dreams later in the night. All these sleep impairments can contribute to relapses to drinking in persons with alcohol use disorders.

Another common consequence of alcohol use even at relatively low alcohol levels is impaired judgment and coordination, which increases the risk of injuries. In the United States, ~40% of drinkers have at some time driven while intoxicated. Heavy drinking can also be associated with headache, thirst, nausea, vomiting, and fatigue the following day, a hangover syndrome that is responsible for much missed work and school time and temporary cognitive deficits.

Chronic high alcohol doses cause *peripheral neuropathy* in ~10% of individuals with alcohol use disorders: similar to diabetes, patients experience bilateral limb numbness, tingling, and paresthesias, all of which are more pronounced distally. Approximately 1% of those with alcohol use disorders develop *cerebellar degeneration* or *atrophy*, producing a syndrome of progressive unsteady stance and gait often accompanied by mild nystagmus; neuroimaging studies reveal atrophy of the cerebellar vermis. Perhaps as few as 1 in 500 individuals with alcohol use disorders develop full *Wernicke's* (ophthalmoparesis, ataxia, and encephalopathy) and *Korsakoff's* (severe retrograde and anterograde amnesia) *syndromes*, although a higher proportion has one or more neuropathologic findings related to these conditions. These result from low levels of thiamine, especially in predisposed individuals with transketolase deficiencies. Repeated heavy drinking can contribute to *cognitive problems* and temporary memory impairment lasting for weeks to months after abstinence. Brain atrophy, evident as ventricular enlargement and widened cortical sulci on magnetic resonance imaging (MRI) and computed tomography (CT) scans, occurs in ~50% of individuals with long-term alcohol use disorders; these changes are usually reversible if abstinence is maintained. Adolescents may be especially vulnerable to alcohol-related brain changes. There is no single "alcoholic dementia" syndrome; rather, this label describes patients who have irreversible cognitive changes (possibly from diverse causes) in the context of chronic alcohol use disorders.

Psychiatric Comorbidity As many as two-thirds of individuals with alcohol use disorders meet criteria for another independent or temporary substance-induced psychiatric syndrome as defined in the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5) of the American Psychiatric Association **(Chap. 444)**. A substantial proportion of those with independent psychiatric conditions (i.e., not just temporary symptoms only seen during intoxication or withdrawal) relate to a preexisting antisocial personality disorder (ASPD) manifesting as severe impulsivity and disinhibition that contribute to both alcohol and drug use disorders. The lifetime ASPD risk is 3% in males, and ≥80% of such individuals demonstrate alcohol- and/or drug-related conditions. Another common psychiatric comorbidity occurs with problems regarding other substances of abuse. The remainder of individuals with alcohol use disorders who have an independent psychiatric syndrome relate to preexisting conditions such as schizophrenia or manic-depressive disease or anxiety syndromes such as panic disorder. The comorbidities of alcohol use disorders with independent psychiatric disorders might represent an overlap in genetic vulnerabilities, impaired judgment regarding the use of alcohol from the independent psychiatric condition, or an attempt to use alcohol to alleviate symptoms of the disorder or side effects of medications.

Many alcohol-related psychiatric syndromes can be seen *temporarily* during heavy drinking and subsequent withdrawal. These alcohol-induced conditions include an intense *sadness* lasting for days to weeks in the midst of heavy drinking seen in 40% of individuals with alcohol use disorders, which tends to disappear over several weeks of abstinence (alcohol-induced mood disorder); 10–30% have temporary severe *anxiety*, often beginning during alcohol withdrawal, which can persist for a month or more after cessation of drinking (alcohol-induced anxiety disorder); and 3–5% have auditory *hallucinations* and/or paranoid delusions while they are otherwise alert and oriented (*alcohol-induced psychotic disorder*).

Treatment of all forms of alcohol-induced psychopathology includes helping patients achieve abstinence and offering supportive care, as well as reassurance and "talk therapy" such as cognitive-behavioral approaches. However, with the exception of short-term antipsychotic medications for substance-induced psychoses, substance-induced psychiatric conditions only rarely require medications. Recovery is likely within several days to 4 weeks of abstinence. Conversely, because alcohol-induced conditions are temporary and do not indicate a need for long-term pharmacotherapy, a history of heavy alcohol intake is an important part of the workup for any patient who presents with any of these psychiatric symptoms.

THE GASTROINTESTINAL SYSTEM

Esophagus and Stomach Alcohol can cause inflammation of the esophagus and stomach causing epigastric distress and gastrointestinal bleeding, making alcohol one of the most common causes of

hemorrhagic gastritis. Violent vomiting can produce severe bleeding through a Mallory-Weiss lesion, a longitudinal tear in the mucosa at the gastroesophageal junction.

Pancreas and Liver The incidence of acute pancreatitis (~25 per 1000 per year) is almost threefold higher in individuals with alcohol use disorders than in the general population, accounting for an estimated 10% or more of the total cases. Alcohol impairs gluconeogenesis in the liver, resulting in a fall in the amount of glucose produced from glycogen, increased lactate production, and decreased oxidation of fatty acids. These contribute to an increase in fat accumulation in liver cells. In healthy individuals, these changes are reversible, but with repeated exposure to ethanol, especially daily heavy drinking, more severe changes in the liver occur, including alcohol-induced hepatitis, perivenular sclerosis, and cirrhosis, with the latter observed in an estimated 15% of individuals with alcohol use disorders **(Chap. 335)**. Perhaps through an enhanced vulnerability to infections, individuals with alcohol use disorders have an elevated rate of hepatitis C, and drinking in the context of that disease is associated with more severe liver deterioration.

CANCER

As few as 1.5 drinks per day increases a woman's risk of breast cancer 1.4-fold. For both sexes, four drinks per day increases the risk for oral and esophageal cancers approximately threefold and rectal cancers by a factor of 1.5; seven to eight or more drinks per day produces an approximately fivefold increased risk for many other cancers. These consequences may result directly from cancer-promoting effects of alcohol and acetaldehyde or indirectly by interfering with immune homeostasis.

HEMATOPOIETIC SYSTEM

Ethanol causes an increase in red blood cell size (mean corpuscular volume [MCV]), which reflects its effects on stem cells. If heavy drinking is accompanied by folic acid deficiency, there can also be hypersegmented neutrophils, reticulocytopenia, and a hyperplastic bone marrow; if malnutrition is present, sideroblastic changes can be observed. Chronic heavy drinking can decrease production of white blood cells, decrease granulocyte mobility and adherence, and impair delayed-hypersensitivity responses to novel antigens (with a possible false-negative tuberculin skin test). Associated immune deficiencies can contribute to vulnerability toward infections, including hepatitis and HIV, and interfere with their treatment. Finally, many individuals with alcohol use disorders have mild thrombocytopenia, which usually resolves within a week of abstinence unless there is hepatic cirrhosis or congestive splenomegaly.

CARDIOVASCULAR SYSTEM

Acutely, ethanol decreases myocardial contractility and causes peripheral vasodilation, with a resulting mild decrease in blood pressure and a compensatory increase in cardiac output. Exercise-induced increases in cardiac oxygen consumption are higher after alcohol intake. These acute effects have little clinical significance for the average healthy drinker but can be problematic when persisting cardiac disease is present.

The consumption of three or more drinks per day results in a dose-dependent increase in blood pressure, which returns to normal within weeks of abstinence. Thus, heavy drinking is an important factor in mild to moderate hypertension. Chronic heavy drinkers also have a sixfold increased risk for coronary artery disease, related, in part, to increased low-density lipoprotein cholesterol, and carry an increased risk for cardiomyopathy through direct effects of alcohol on heart muscle. Symptoms of the latter include unexplained arrhythmias in the presence of left ventricular impairment, heart failure, hypocontractility of heart muscle, and dilation of all four heart chambers with associated potential mural thrombi and mitral valve regurgitation. Atrial or ventricular arrhythmias, especially paroxysmal tachycardia, can also occur temporarily after heavy drinking in individuals showing no other evidence of heart disease—a syndrome known as the "holiday heart."

GENITOURINARY SYSTEM CHANGES, SEXUAL FUNCTIONING, AND FETAL DEVELOPMENT

Heavy drinking in adolescence can affect normal sexual development and reproductive onset. At any age, modest ethanol doses (e.g., blood alcohol concentrations of 0.06 g/dL) can increase sexual drive but also decrease erectile capacity in men. Even in the absence of liver impairment, a significant minority of chronic heavy drinking men show irreversible testicular atrophy with shrinkage of the seminiferous tubules, decreases in ejaculate volume, and a lower sperm count **(Chap. 384)**.

The repeated ingestion of high doses of ethanol by women can result in amenorrhea, a decrease in ovarian size, absence of corpora lutea with associated infertility, and an increased risk of spontaneous abortion. Drinking during pregnancy results in the rapid placental transfer of both ethanol and acetaldehyde, which may contribute to a range of consequences known as fetal alcohol spectrum disorder (FASD). One severe result is the *fetal alcohol syndrome* (FAS), seen in ~5% of children born to heavy-drinking mothers, which can include any of the following: facial changes with epicanthal eye folds; poorly formed ear concha; small teeth with faulty enamel; cardiac atrial or ventricular septal defects; an aberrant palmar crease and limitation in joint movement; and microcephaly with intellectual impairment. Less pervasive FASD conditions include combinations of low birth weight, a lower intelligence quotient (IQ), hyperactive behavior, and some modest cognitive deficits. The amount of ethanol required and the time of vulnerability during pregnancy have not been defined, making it advisable for pregnant women to abstain from alcohol completely.

OTHER EFFECTS

Between one-half and two-thirds of individuals with alcohol use disorders have skeletal muscle weakness caused by acute *alcoholic myopathy*, a condition that improves but which might not fully remit with abstinence. Effects of repeated heavy drinking on the *skeletal system* include changes in calcium metabolism, lower bone density, and decreased growth in the epiphyses, leading to an increased risk for fractures and osteonecrosis of the femoral head. *Hormonal changes* include an increase in cortisol levels, which can remain elevated during heavy drinking; inhibition of vasopressin secretion at rising blood alcohol concentrations and enhanced secretion at falling blood alcohol concentrations (with the final result that most individuals with alcohol use disorders are likely to be slightly overhydrated); a modest and reversible decrease in serum thyroxine (T_4); and a more marked decrease in serum triiodothyronine (T_3). Hormone irregularities may disappear after a month or more of abstinence.

ALCOHOL USE DISORDERS

Because many drinkers occasionally imbibe to excess, temporary alcohol-related problems are common, especially in the late teens to the late twenties. However, repeated problems in multiple life areas can indicate an alcohol use disorder as defined in DSM-5.

DEFINITIONS AND EPIDEMIOLOGY

An *alcohol use disorder* (aka alcoholism or alcohol dependence in prior diagnostic manuals) is defined as repeated alcohol-related difficulties in at least 2 of 11 life areas that cluster together in the same 12-month period **(Table 445-2)**. Ten of the 11 items in DSM-5 (published in 2013) were taken directly from the 7 dependence and 4 abuse criteria in DSM-IV (most recently revised in 2000), after deleting legal problems and adding craving. Severity of an alcohol use disorder is based on the number of items endorsed: mild is two or three items; moderate is four or five; and severe is six or more of the criterion items. The 2013 diagnostic approach is similar enough to DSM-IV that the following descriptions of associated phenomena are still accurate.

The lifetime risk for an alcohol use disorder in most Western countries is about 10–15% for men and 5–8% for women; higher rates are seen in individuals who seek help from health-care deliverers. Between 2001 and 2013, the proportion of the United States population with a current (i.e., past 12-month) alcohol use disorder increased by 49% with increases of almost 100% in women, African Americans, and individuals aged ≥45. Rates are similar in the United States, Canada, Germany, Australia, and the United Kingdom tend to be lower in

TABLE 445-2 ***Diagnostic and Statistical Manual of Mental Disorders,* Fifth Edition, Classification of Alcohol Use Disorder (AUD)**

Criteria
Two or more of the following items occurring in the same 12-month period must be endorsed for the diagnosis of an alcohol use disorder[a]:
Drinking resulting in recurrent failure to fulfill role obligations
Recurrent drinking in hazardous situations
Continued drinking despite alcohol-related social or interpersonal problems
Tolerance
Withdrawal, or substance use for relief/avoidance of withdrawal
Drinking in larger amounts or for longer than intended
Persistent desire/unsuccessful attempts to stop or reduce drinking
Great deal of time spent obtaining, using, or recovering from alcohol
Important activities given up/reduced because of drinking
Continued drinking despite knowledge of physical or psychological problems caused by alcohol
Alcohol craving

[a]*Mild AUD*: 2–3 criteria required; *Moderate AUD*: 4–5 items endorsed; *severe AUD:* 6 or more items endorsed.

most Mediterranean countries, such as Italy, Greece, and Israel, and may be higher in Ireland, France, Eastern Europe (e.g., Russia), and Scandinavia. An even higher lifetime prevalence has been reported for most native cultures, including Native Americans, Eskimos, Maori groups, and aboriginal tribes of Australia. These differences in prevalence reflect both cultural and genetic influences, as described below. In Western countries, the typical person with an alcohol use disorder is more often a blue- or white-collar worker or homemaker. The lifetime risk for this disorder among physicians is similar to that of the general population.

GENETICS

Approximately 60% of the risk for alcohol use disorders is attributed to genes, as indicated by the fourfold higher risk in children with an alcohol use disorder parent (even if adopted early in life and raised by nonalcoholics) and a higher risk in identical twins compared to fraternal twins of affected individuals. The genetic variations operate primarily through intermediate characteristics that subsequently combine with environmental influences to alter the risk for heavy drinking and alcohol problems. These include genes relating to a high risk for all substance use disorders that operate through impulsivity, schizophrenia, and bipolar disorder. Another characteristic, an intense skin flushing response when drinking, decreases the risk for only alcohol use disorders through gene variations for several alcohol-metabolizing enzymes, especially ALDH (a mutation only seen in Asians), and to a lesser extent, variations in ADH.

An additional genetically influenced characteristic that increases the risk for heavy drinking, a low level of response to alcohol, can be seen very early in the drinking career and before acquired tolerance or alcohol used disorders develop. The low response per drink operates, in part, through variations in genes relating to calcium and potassium channels, GABA, nicotinic, dopamine, and serotonin systems. Follow-up studies have demonstrated that this need for higher doses of alcohol to achieve effects predicts future heavy drinking, alcohol problems, and alcohol use disorders. The impact of a low response to alcohol on adverse drinking outcomes is partially mediated by a range of environmental and attitudinal influences, including the selection of heavier-drinking friends, more positive expectations of the effects of high doses of alcohol, and using alcohol to cope with stress.

NATURAL HISTORY

Although the average age of the first drink (~15 years) is similar in individuals who do and do not go on to develop alcohol use disorders, an earlier onset of regular drinking and drunkenness, especially in the context of conduct problems, is associated with a higher risk for later alcohol-related diagnoses. By the mid-twenties, most nonalcoholic men and women begin to moderate their drinking (perhaps learning from negative consequences), whereas those with alcohol use disorders are likely to escalate their drinking despite difficulties. The first major life problem from alcohol often appears in the late teens to early twenties, and a pattern of multiple alcohol difficulties by the midtwenties. Once established, the course is likely to include exacerbations and remissions, with little difficulty in temporarily stopping or controlling alcohol use when problems develop, but without help, desistance usually gives way to escalations in alcohol intake and subsequent problems. Following treatment, between half and two-thirds of those with alcohol use disorders maintain abstinence for at least a year, and often permanently. Even without formal treatment or self-help groups, there is at least a 20% chance of spontaneous remission with long-term abstinence. However, should the individual continue to drink heavily, the life span is shortened by ~10 years on average, with the leading causes of early death being enhanced rates of heart disease, cancer, accidents, and suicide.

TREATMENT

The approach to treating alcohol-related conditions is relatively straightforward: (1) recognize that at least 20% of patients have an alcohol use disorder; (2) learn how to identify and treat acute alcohol-related conditions (e.g., severe intoxication); (3) know how to help patients begin to address their alcohol problems; and (4) know how to treat alcohol withdrawal symptoms and to appropriately refer patients for additional help.

IDENTIFICATION OF PATIENTS WITH ALCOHOL USE DISORDERS

Even in affluent locales, the ~20% of patients who have an alcohol use disorder can be identified by asking questions about *alcohol problems* and noting laboratory test results that can reflect regular consumption of six to eight or more drinks per day. The two blood tests with ≥60% sensitivity and specificity for heavy alcohol consumption are γ-glutamyl transferase (GGT) (>35 U) and carbohydrate-deficient transferrin (CDT) (>20 U/L or >2.6%); the combination of the two tests is likely to be more accurate than either alone. The values for these serologic markers are likely to return toward normal within several weeks of abstinence. Other useful blood tests include high-normal MCVs (≥91 μm^3) and serum uric acid (>416 mol/L, or 7 mg/dL).

The diagnosis of an alcohol use disorder ultimately rests on the documentation of a pattern of repeated difficulties associated with alcohol (Table 445-2). Thus, in screening, it is important to probe for marital or job problems, legal difficulties, histories of accidents, medical problems, evidence of tolerance, and so on, and then attempt to relate these issues to use of alcohol. Some standardized questionnaires can be helpful, including the 10-item Alcohol Use Disorders Identification Test (AUDIT) **(Table 445-3)**, but these are only screening tools, and a face-to-face interview is still required for a meaningful diagnosis.

TREATMENT

Alcohol-Related Conditions

ACUTE INTOXICATION

The first priority in treating severe intoxication is to assess vital signs and manage respiratory depression, cardiac arrhythmias, and blood pressure instability, if present. The possibility of intoxication with other drugs should be considered by obtaining, if needed, toxicology screens for other central nervous system (CNS) depressants such as benzodiazepines and for opioids. Aggressive behavior should be handled by offering reassurance but also by considering a possible show of force with an intervention team. If the aggressive behavior continues, relatively low doses of a short-acting benzodiazepine such as lorazepam (e.g., 1–2 mg PO or IV) may be used and can be repeated as needed, but care must be taken not to destabilize vital signs or worsen confusion. An alternative approach is to use an antipsychotic medication (e.g., 0.5–5 mg of haloperidol PO or IM every 4–8 h as needed, or olanzapine 2.5–10 mg IM repeated at 2 and 6 h, if needed).

TABLE 445-3 The Alcohol Use Disorders Identification Test (AUDIT)[a]

ITEM	5-POINT SCALE (LEAST TO MOST)
1. How often do you have a drink containing alcohol?	Never (0) to 4+ per week (4)
2. How many drinks containing alcohol do you have on a typical day?	1 or 2 (0) to 10+ (4)
3. How often do you have six or more drinks on one occasion?	Never (0) to daily or almost daily (4)
4. How often during the last year have you found that you were not able to stop drinking once you had started?	Never (0) to daily or almost daily (4)
5. How often during the last year have you failed to do what was normally expected from you because of drinking?	Never (0) to daily or almost daily (4)
6. How often during the last year have you needed a first drink in the morning to get yourself going after a heavy drinking session?	Never (0) to daily or almost daily (4)
7. How often during the last year have you had a feeling of guilt or remorse after drinking?	Never (0) to daily or almost daily (4)
8. How often during the last year have you been unable to remember what happened the night before because you had been drinking?	Never (0) to daily or almost daily (4)
9. Have you or someone else been injured as a result of your drinking?	No (0) to yes, during the last year (4)
10. Has a relative, friend, doctor, or other health worker been concerned about your drinking or suggested that you should cut down?	No (0) to yes, during the last year (4)

[a]The AUDIT is scored by simply summing the values associated with the endorsed response. A score ≥8 may indicate harmful alcohol use.

INTERVENTION

There are two main elements to highlighting the need for compliance with treatment in a person with an alcohol use disorder: motivational interviewing and brief interventions. During motivational interviewing, the clinician helps the patient to think through the assets (e.g., comfort in social situations) and liabilities (e.g., health- and interpersonal-related problems) of the current pattern of drinking. The patient's responses are key, and the clinician should listen empathetically, helping patients to weigh options and encouraging them to take responsibility for needed changes. Patients should be reminded that only they can decide to avoid the consequences that will occur without changes in drinking. The process of the similar approach, brief intervention, has been summarized by the acronym FRAMES: *F*eedback to the patient; *R*esponsibility to be taken by the patient; *A*dvice, rather than orders, on what needs to be done; *M*enus of options that might be considered; *E*mpathy for understanding the patient's thoughts and feelings; and *S*elf-efficacy, i.e., offering support for the capacity of the patient to make changes.

Once the patient begins to consider change, the discussions can focus more on the consequences of high alcohol consumption, suggested approaches to stopping drinking, and help in recognizing and avoiding situations likely to lead to heavy drinking. Both motivational interviewing and brief interventions can be carried out in 15-min sessions, but because patients often do not change behavior immediately, multiple meetings are often required to explore the problem and possible options, discuss optimal treatments, and explain the benefits of abstinence.

ALCOHOL WITHDRAWAL

If the patient agrees to stop drinking, sudden decreases in alcohol intake can produce withdrawal symptoms, most of which are the opposite of those produced by intoxication. Features include tremor of the hands (shakes); agitation and anxiety; autonomic nervous system overactivity including an increase in pulse, respiratory rate, sweating, and body temperature; and insomnia. These symptoms usually begin within 5–10 h of decreasing ethanol intake, peak on day 2 or 3, and improve by day 4 or 5, although mild levels of these problems may persist for 4–6 months as a protracted abstinence syndrome.

About 2% of individuals with alcohol use disorders experience a withdrawal seizure, with the risk increasing in the context of older age, concomitant medical problems, misuse of additional drugs, and higher alcohol quantities. The same risk factors also contribute to the even lower rate of withdrawal delirium, also known as *delirium tremens* (DTs), where the withdrawal includes delirium (mental confusion, agitation, and fluctuating levels of consciousness) associated with a tremor and autonomic overactivity (e.g., marked increases in pulse, blood pressure, and respirations). The risks for seizures and DTs can be diminished by identifying and treating any underlying medical conditions early in the course of withdrawal.

Thus, the first step in dealing with possible withdrawal phenomena is a thorough physical examination in all very heavy drinkers who are considering abstinence. This includes a search for evidence of liver failure, gastrointestinal bleeding, cardiac arrhythmias, infection, and glucose or electrolyte imbalances. It is also important to offer adequate nutrition and oral multiple B vitamins, including 50–100 mg of oral thiamine daily for a week or more. Because most patients with alcohol use disorders who enter withdrawal are either normally hydrated or mildly overhydrated, IV fluids should be avoided unless there is a relevant medical problem or significant recent bleeding, vomiting, or diarrhea.

The next step is to recognize that because withdrawal symptoms reflect the rapid removal of a CNS depressant, alcohol, the symptoms can be controlled by administering any depressant in doses that decrease symptoms (e.g., a rapid pulse and tremor) and then tapering the dose over 3–5 days. Although most depressants are effective, benzodiazepines **(Chap. 444)** have the most supportive data for use in this situation, highest margin of safety and lowest cost and are, therefore, the preferred class of drugs. Short-half-life benzodiazepines can be considered for patients with serious liver impairment or evidence of significant brain damage, but they must be given every 4 h to avoid abrupt blood-level fluctuations that may increase the risk for seizures. Therefore, most clinicians use drugs with longer half-lives (e.g., chlordiazepoxide), adjusting the dose if signs of withdrawal escalate, and withholding the drug if the patient is sleeping or has orthostatic hypotension. The average patient requires 25–50 mg of chlordiazepoxide or 10 mg of diazepam given PO every 4–6 h on the first day, with doses then decreased to zero over the next 5 days. Although alcohol withdrawal can be treated in a hospital, patients in good physical condition who demonstrate mild signs of withdrawal despite low blood alcohol concentrations and who have no prior history of DTs or withdrawal seizures can be considered for outpatient detoxification. For the next 4 or 5 days, these patients should receive only 1 or 2 days of medications at a time, and return daily for evaluation of vital signs. They can be hospitalized if signs and symptoms of withdrawal markedly escalate.

Treatment of patients with DTs can be challenging, and the condition is likely to run a course of 3–5 days regardless of the therapy used. However, conditions that meet the criteria for DTs outlined above represent medical emergencies that carry an estimated mortality as high as 5%, and treatment is best carried out in an intensive care unit by well-trained clinicians who closely monitor vital signs. Medications can include high-dose benzodiazepines (e.g., as much as 800 mg/d of chlordiazepoxide has been reported), or, for those who do not respond to that regimen, closely monitored doses of propofol or dexmedetomidine. The focus of care is to identify and correct medical problems and to control behavior and prevent injuries. We do not recommend the use of antipsychotic medications in the treatment of alcohol withdrawal symptoms; although antipsychotics are less likely than benzodiazepines to exacerbate confusion, they may increase the risk of seizures.

Generalized withdrawal seizures rarely require more than the administration of an adequate dose of benzodiazepines. There is little evidence that anticonvulsants such as phenytoin or gabapentin are more effective than benzodiazepines for alcohol-withdrawal seizures, and the risk of seizures has usually passed by the time effective drug levels are reached. The rare patient with status epilepticus must be treated aggressively **(Chap. 418)**.

HELPING INDIVIDUALS WITH ALCOHOL USE DISORDERS TO STOP DRINKING: THE REHABILITATION PHASE

An Overview After completing alcoholic rehabilitation, ≥60% of individuals with alcohol use disorders, especially highly functioning patients, maintain abstinence for at least a year; many also achieve long-term sobriety. The core components of the rehabilitation phase of treatment include cognitive-behavioral approaches to help patients recognize the need to change, while working with them to alter their behaviors to enhance compliance. A key step is to optimize motivation toward abstinence through education of patients and their significant others about alcohol use disorders and their likely course over time. After years of heavy drinking, many patients also require vocational or avocational counseling to help to structure their days, and all patients should try self-help groups such as Alcoholics Anonymous (AA) to assist them in developing a sober peer group and to learn how to deal with life's stresses while remaining sober. *Relapse prevention education* helps patients identify situations in which a return to drinking is likely (e.g., spending time with heavily drinking friends or stopping in a bar to meet friends but planning to only have a nonalcoholic beverage), formulate ways to avoid the risky situation and if not possible to mitigate the risks to which they are exposed. It is also important to develop coping strategies that increase the chances of a return to abstinence quickly after an episode of drinking.

Although many patients can be treated as outpatients, more intense interventions are more effective, and some individuals with alcohol use disorders do not respond to AA or outpatient groups. Whatever the setting, ongoing contact with outpatient treatment staff should be maintained for at least 6 months and preferably for a year after abstinence. Counseling focuses on areas of improved functioning in the absence of alcohol (i.e., why it is a good idea to continue abstinence), helping patients to manage free time without alcohol, encouraging them to develop a nondrinking peer group, and discussions of ways to handle stress without drinking.

The physician serves an important role in identifying the alcohol problem, diagnosing and treating associated medical and independent or substance-induced psychiatric syndromes, overseeing detoxification, referring the patient to outpatient or inpatient rehabilitation programs, providing counseling, and, if appropriate, selecting which (if any) medication might be needed. For insomnia, patients should be reassured that troubled sleep is temporary after alcohol withdrawal and will begin to improve over subsequent weeks. They should be taught the elements of "sleep hygiene" including maintaining consistent schedules for bedtime and awakening, avoiding exercising or eating large meals before bedtime, and keeping the bedroom cool, dark, and quiet at night **(Chap. 27)**. Depressant sleep medications are not the optimal approach for this type of insomnia that often continues for several weeks or months. Patients are likely to develop rebound insomnia when the depressant dose is decreased or stopped. The rebound increases the chance they will increase the dose and potentially develop problems controlling the prescribed depressant drug. Sedating antidepressants (e.g., trazodone) should not be used because they interfere with cognitive functioning the next morning and disturb the normal sleep architecture, but occasional use of over-the-counter sleeping medications (sedating antihistamines) can be considered. An additional problem, anxiety symptoms, can be addressed by increasing patients' insights into the temporary nature of the symptoms and helping them develop strategies to achieve relaxation by using forms of cognitive therapy.

Medications for the Alcohol Rehabilitation Treatment Phase Several medications have modest benefits when used in the first 6–12 months of recovery. The opioid antagonist, naltrexone, may shorten subsequent relapses, whether used in the oral form (50–150 mg/d) or as a once-per-month 380-mg injection. By blocking opioid receptors, naltrexone decreases activity in the dopamine-rich ventral tegmental reward system and decreases the feeling of pleasure if alcohol is imbibed. A second medication, acamprosate (Campral) (~2 g/d divided into three oral doses), has similar modest effects. Acamprosate inhibits NMDA receptors, decreasing mild symptoms of protracted withdrawal. Several trials of combined naltrexone and acamprosate have reported that the combination is well tolerated and the efficacy might be superior to either drug alone, although not all studies agree.

It is more difficult to establish the asset-to-liability ratio of a third drug, disulfiram, an ALDH inhibitor, used clinically at doses of 250 mg/d. This drug produces vomiting and autonomic nervous system instability in the presence of alcohol as a result of rapidly rising blood levels of acetaldehyde. This reaction can be dangerous, especially for patients with heart disease, stroke, diabetes mellitus, or hypertension. The drug itself carries potential risks of temporary depressive or psychotic symptoms, peripheral neuropathy, and liver damage. Disulfiram is best given under supervision by someone (such as a spouse), especially during high-risk drinking situations (such as the Christmas holidays). Additional drugs under investigation include another opioid antagonist nalmefene, the nicotinic receptor agonist varenicline, the serotonin antagonist ondansetron, the α-adrenergic agonist prazosin, the $GABA_B$ receptor agonist baclofen, the anticonvulsant topiramate, and cannabinol receptor antagonists. At present, there are insufficient data to determine the asset-to-liability ratio for these medications in treating alcohol use disorders and, therefore, few data yet offer solid support for their routine use in clinical settings.

GLOBAL CONSIDERATIONS

As described above, rates of alcohol use disorders differ across sex, age, ethnicity, and country. There are also differences across countries regarding the definition of a standard drink (e.g., 10–12 g of ethanol in the United States and 8 g in the United Kingdom) and the definition of being legally drunk. The preferred alcoholic beverage also varies across groups, even within countries. That said, regardless of sex, ethnicity, or country, the actual drug in the drink is still ethanol, and the risks for problems, course of alcohol use disorders, and approaches to treatment are similar across the world.

FURTHER READING

Grant BF et al: Prevalence of 12-month alcohol use, high-risk drinking, and DSM-IV alcohol use disorder in the United States, 2001-2002 to 2012-2013: Results from the National Epidemiologic Survey on Alcohol and Related Conditions. JAMA Psychiatry 74:911, 2017.

Heikkinen N et al: Alcohol consumption during adolescence is associated with reduced grey matter volumes. Addiction 112:604, 2017.

Litten RZ et al: Potential medications for the treatment of alcohol use disorder: An evaluation of clinical efficacy and safety. Subst Abuse 37:286, 2016.

Reilly MT et al: Genetic studies of alcohol dependence in the context of the addiction cycle. Neuropsychopharmacology 122:3, 2017.

Schuckit MA: Recognition and management of withdrawal delirium (delirium tremens). N Engl J Med 371:2109, 2014.

Wetherill RR, Fromme K: Alcohol-induced blackouts: A review of recent clinical research with practical implications and recommendations for future studies. Alcohol Clin Exp Res 40:922, 2016.

446 Opioid-Related Disorders

Thomas R. Kosten, Colin N. Haile

Opioid analgesics have been used since at least 300 B.C. Nepenthe (Greek "free from sorrow") helped the hero of the *Odyssey*, but widespread opium smoking in China and the Near East has caused harm for centuries. Since the first chemical isolation of opium and codeine 200 years ago, a wide range of synthetic opioids have been developed, and opioid receptors were cloned in the 1990s. Two of the most important adverse effects of all these agents are the development of opioid use disorder and overdose. Prescription opioids are primarily used for pain management, but due to ease of availability individuals procure and misuse these drugs with dire consequences. In 2015, for example, 3.8 million individuals in the United States were current misusers of pain relievers. More concerning, during 2015 >20,000 overdose deaths involved opioids with an additional 12,990 overdose deaths related to heroin alone. These numbers continue to increase and have accelerated due to mixing high potency fentanyl derivatives with heroin. The accelerating death rates are partially because reversal of fentanyl overdoses can require several-fold larger doses of naloxone than the doses in the intranasal devices used for nonmedical street resuscitations. Indeed, according to the most recent World Drug Report, opioid misuse causes the greatest global burden of morbidity and mortality; disease transmission; increased health care, crime, and law enforcement costs; and less tangible costs of family distress and lost productivity.

The terms "dependence" and "addiction" are no longer used to describe substance use disorders. Opioid-related disorders encompass opioid use disorder, opioid intoxication, and opioid withdrawal. The diagnosis of opioid use disorder as defined in the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (DSM-5) requires the repeated use of the opiate while producing problems in two or more areas in a 12-month period. The areas include tolerance, withdrawal, use of greater amounts of opioids than intended, craving, and use despite adverse consequences. This new definition of opioid use disorder, reducing the criteria for diagnosis from three problem areas to two, is not expected to change the rates of these disorders because most individuals using these substances meet more than three criteria.

A striking recent aspect of illicit opioid use has been its marked increase as the gateway to illicit drugs in the United States. Since 2007, prescription opiates have surpassed marijuana as the most common illicit drug that adolescents initially use, although overall rates of opioid use are far lower than marijuana. The most commonly used opioids are diverted prescriptions for oxycodone and hydrocodone, followed by heroin and morphine, and—among health professionals—meperidine and fentanyl. Heroin is metabolized into 6-monoacetylmorphine and morphine thus acting as a prodrug that more readily penetrates the brain and is converted rapidly to morphine in the body. Two opioid maintenance treatment agents—methadone and buprenorphine—are also misused, but at substantially lower rates, and the partial opioid agonists such as butorphanol, tramadol, and pentazocine are misused even less frequently. Because the chemistry and general pharmacology of these agents are covered in major pharmacology texts, this chapter focuses on the neurobiology and pharmacology relevant to opioid use disorder and its treatments. Although the neurobiology of misuse involves all four of the known opioid receptors—mu, kappa, delta, and nociceptin/orphanin—this discussion focuses on the mu receptor targeted by most of the clinically used opioids.

NEUROBIOLOGY

The neurobiology of opioids and their effects not only include opioid receptors, but also downstream intracellular messenger systems and ion channels that the receptors regulate. The different functional activities of opioid receptors are summarized in Table 446-1. Abuse liability of opioids is primarily associated with the mu receptor. All opioid receptors are G protein–linked and coupled to the cyclic adenosine monophosphate (cAMP) second messenger system and to G protein–coupled, inwardly rectifying potassium channels (GIRKs). Opioids activate GIRKs, increasing permeability to potassium ions to cause hyperpolarization, which inhibits the production of action potentials. Thus, opioids inhibit the activity of diverse and widely distributed neuronal types. The major effects of opioids, such as analgesia, sedation, and drug reinforcement, are produced through this inhibition of neurons that belong to specific brain pathways.

TABLE 446-1 Actions of Opioid Receptors

RECEPTOR TYPE	ACTIONS
Mu (μ) (e.g., morphine, buprenorphine)	Analgesia, reinforcement euphoria, cough and appetite suppression, decreased respirations, decreased GI motility, sedation, hormone changes, dopamine and acetylcholine release
Kappa (κ) (e.g., butorphanol)	Dysphoria, decreased GI motility, decreased appetite, decreased respiration, psychotic symptoms, sedation, diuresis, and analgesia
Delta (δ) (e.g., etorphine)	Analgesia, euphoria, physical dependence, hormone changes, appetite suppression, and dopamine release
Nociceptin/orphanin (e.g., buprenorphine)	Analgesia, appetite, anxiety, tolerance to opioids, hypotension, decreased GI motility, and 5-HT and NE release

Abbreviations: 5-HT, serotonin; GI, gastrointestinal; NE, norepinephrine.

Many opioid actions are related to the specific neuroanatomic locations of mu receptors. Reinforcing and euphoric effects of opioids relate primarily to activation of the mesolimbic dopaminergic pathway from the ventral tegmental area (VTA) to the nucleus accumbens (NAc), where opioids increase synaptic levels of dopamine. This increase is due to inhibition of GABAergic neurons that inhibit both the activity of neurons within the VTA and the NAc. The positive subjective effects of opioid drugs also include mu receptor desensitization and internalization, potentially related to stimulation of beta-arrestin signaling pathways. However, the "high" only occurs when the *rate of change* in dopamine is fast. Large, rapidly administered doses of opioids block γ-aminobutyric acid (GABA) inhibition and produce a burst of VTA dopamine neuron activity that is associated with a "high" in commonly misused substances. Therefore, routes of administration that slowly increase opioid blood and brain levels, such as oral and transdermal routes, are effective for analgesia and sedation but do not produce an opioid "high" that follows smoking and intravenous routes. Other acute effects such as analgesia and respiratory depression involve opioid receptors located in other brain areas such as the locus coeruleus (LC).

Opioid tolerance and withdrawal are chronic effects related to the cAMP-protein kinase A (PKA)-cAMP response-element binding protein (CREB) intracellular cascade (Fig. 446-1). These effects are also reflective of genetic risk factors for developing opioid use disorder, with estimates of up to 50% of the risk due to polygenic inheritance. Specific functional polymorphisms in the mu opiate receptor gene appear to be associated with this risk for opioid misuse, including one producing a threefold increase in this receptor's affinity for opiates and the endogenous ligand β-endorphin. Epigenetic methylation changes also occur on DNA in the region of the mu receptor gene in individuals with opioid use disorder, inhibiting gene transcription. This molecular cascade links acute intoxication and sedation to opioid tolerance and withdrawal mediated by the LC. Noradrenergic (NE) neurons in the LC mediate activation of the cortical hemispheres. When large opioid doses saturate and activate all of its mu receptors, action potentials cease. When this direct inhibitory effect is sustained over weeks and months of opioid use, a secondary set of adaptive changes occur that lead to tolerance and withdrawal symptoms (Fig. 446-1). Withdrawal symptoms reflect, in part, overactivity of NE neurons in the LC. This molecular model of NE neuronal activation during withdrawal has had important treatment implications, such as the use of the alpha-2 agonist clonidine to treat opioid withdrawal. Other contributors to withdrawal include deficits within the dopamine reward system.

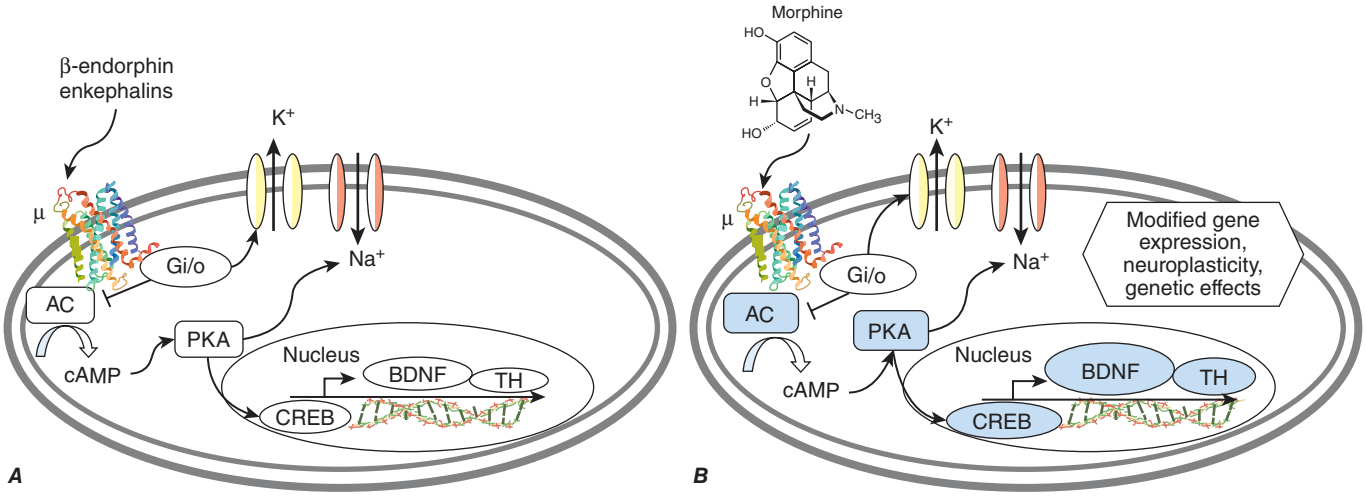


FIGURE 446-1 Normal mu-receptor activation by endogenous opioids inhibits the cyclic adenosine monophosphate (cAMP)-protein kinase A (PKA)-cAMP response-element binding protein (CREB) cascade in noradrenergic (NE) neurons within the locus coeruleus (***A***) through inhibitory Gi/o protein influence on adenylyl cyclase (AC). Similarly, acute exposure to opioids (e.g., morphine) inhibits this system, whereas chronic exposure to opiates (***B***) leads to upregulation of the cAMP pathway in an attempt to oppose opioid-induced inhibitory influence. Upregulation of this system is involved in opioid tolerance, and when the opioid is removed, unopposed NE neurotransmission is involved in opioid withdrawal. Upregulated PKA phosphorylates CREB, initiating the expression of various genes such as tyrosine hydroxylase (*TH*) and brain-derived neurotrophic factor (*BDNF*). *BDNF* is implicated in long-term neuroplastic changes in response to chronic opioids.

PHARMACOLOGY

Tolerance and withdrawal commonly occur with chronic daily use, developing as quickly as 6–8 weeks depending on dose concentration and dosing frequency. Tolerance appears to be primarily a pharmacodynamic rather than pharmacokinetic effect, with relatively limited induction of cytochrome P450 or other liver enzymes. The metabolism of opioids occurs in the liver, primarily through the cytochrome P450 systems of 2D6 and 3A4. They then are conjugated to glucuronic acid and excreted in small amounts in feces. The plasma half-lives generally range from 2.5 to 3 h for morphine and >22 h for methadone. The shortest half-lives of several minutes are for fentanyl-related opioids and the longest are for buprenorphine and its active metabolites, which can block opioid withdrawal for up to 3 days after a single dose. Tolerance to opioids leads to the need for increasing amounts of drugs to sustain the desired euphoric effects—as well as to avoid the discomfort of withdrawal. This combination has the expected consequence of strongly reinforcing misuse once it has started. Methadone taken chronically at maintenance doses is stored in the liver, which may reduce the occurrence of withdrawal between daily doses. The role of endogenous opioid peptides in tolerance and withdrawal is uncertain.

The clinical features of opioid misuse are tied to route of administration and rapidity of the drug reaching the brain. Intravenous and smoked administration rapidly produces high drug concentrations in the brain. This produces a "rush," followed by euphoria, a feeling of tranquility, and sleepiness ("the nod"). Heroin produces effects that last 3–5 h, and several doses a day are required to forestall manifestations of withdrawal in chronic users. Symptoms of opioid withdrawal begin 8–10 h after the last dose; lacrimation, rhinorrhea, yawning, and sweating appear first. Restless sleep followed by weakness, chills, gooseflesh ("cold turkey"), nausea and vomiting, muscle aches, and involuntary movements ("kicking the habit"), hyperpnea, hyperthermia, and hypertension occur in later stages of the withdrawal syndrome. The acute course of withdrawal may last 7–10 days. A secondary phase of protracted abstinence lasts for 26–30 weeks and is characterized by hypotension, bradycardia, hypothermia, mydriasis, and decreased responsiveness of the respiratory center to carbon dioxide.

Besides the brain effects of opioids on sedation and euphoria and the combined brain and peripheral nervous system effects on analgesia, a wide range of other organs can be affected. The release of several pituitary hormones is inhibited, including corticotropin-releasing factor (CRF) and luteinizing hormone, which reduces levels of cortisol and sex hormones and can lead to impaired stress responses and reduced libido. An increase in prolactin also contributes to the reduced sex drive in males. Two other hormones affected are thyrotropin, which is reduced, and growth hormone, which is increased. Respiratory depression results from opioid-induced insensitivity of brainstem neurons to increases in carbon dioxide, and in patients with pulmonary disease, this can result in clinically significant complications. In overdoses, aspiration pneumonia is commonly due to loss of the gag reflex. Opioids reduce gut motility, which is helpful for treating diarrhea, but can lead to nausea, constipation, and anorexia with weight loss. Deaths occurred in early methadone maintenance programs due to severe constipation and toxic megacolon. Opioids such as methadone may prolong QT intervals and lead to sudden death in some patients. Orthostatic hypotension may occur due to histamine release and peripheral blood vessel dilation, which is an opioid effect usefully applied to managing acute myocardial infarction. During opioid maintenance, interactions with other medications are of concern; these include inducers of the cytochrome P450 system (usually CYP3A4) such as rifampin and carbamazepine.

Heroin users in particular tend to use opioids intravenously and are likely to be polydrug users, also using alcohol, sedatives, cannabinoids, and stimulants. None of these other drugs are substitutes for opioids, but they have desired additive effects. Therefore, one needs to be sure that the person undergoing a withdrawal reaction is not also withdrawing from alcohol or sedatives, which might be more dangerous and more difficult to manage.

Intravenous opioid use carries with it the risk of serious complications. The common sharing of hypodermic syringes can lead to infections with hepatitis B and HIV/AIDS, among others. Bacterial infections can lead to septic complications such as meningitis, osteomyelitis, and abscesses in various organs. Off-target effects of opioids synthesized in illicit drug labs can lead to serious toxicity. For example, attempts to illicitly manufacture meperidine in the 1980s resulted in the production of a highly specific neurotoxin, MPTP, which produced parkinsonism in users **(Chap. 427)**.

Lethal overdose is a relatively common complication of opioid use disorder. Rapid recognition and treatment with naloxone, a highly specific reversal agent that is relatively free of complications, is essential. The diagnosis is based on recognition of characteristic signs and symptoms, including shallow and slow respirations, pupillary miosis (mydriasis does not occur until significant brain anoxia supervenes), bradycardia, hypothermia, and stupor or coma. Blood or urine toxicology studies can confirm a suspected diagnosis, but immediate management must be based on clinical criteria. If naloxone is not administered, progression to respiratory and cardiovascular collapse leading to death

occurs. At autopsy, cerebral edema and sometimes frothy pulmonary edema are generally found. Opioids generally do not produce seizures except for unusual cases of polydrug use with the opioid meperidine, with high doses of tramadol, or in the newborn.

TREATMENT

Opioid Overdose

Beyond the acute treatment of opioid overdose with naloxone, clinicians have two general treatment options: opioid maintenance or detoxification. Opioid agonist and partial agonist medications are commonly used for both maintenance and detoxification purposes. Alpha-2-adrenergic agonists are primarily used for detoxification. Antagonists are used to accelerate detoxification and then continued after detoxification to prevent relapse. Only the residential medication-free programs have had success that comes close to matching that of the medication-based programs. Success of the various treatment approaches is assessed as retention in treatment and reduced opioid and other drug use; secondary outcomes, such as reduced HIV risk behaviors, crime, psychiatric symptoms, and medical comorbidity, also indicate successful treatment.

Stopping opioid use is much easier than preventing relapse. Long-term relapse prevention for individuals with opioid use disorder requires combined pharmacologic and psychosocial approaches. Chronic users tend to prefer pharmacologic approaches; those with shorter histories of drug use are more amenable to detoxification and psychosocial interventions.

OPIOID OVERDOSE

Managing overdose requires naloxone and support of vital functions, including intubation if needed **(Table 446-2)**. If the overdose is due to buprenorphine, then naloxone might be required at total doses of ≥10 mg, but primary buprenorphine overdose is nearly impossible because this agent is a partial opioid agonist, meaning that as the dose of buprenorphine is increased it has greater opioid antagonist than agonist activity. Thus, a 0.2-mg buprenorphine dose leads to analgesia and sedation, while a hundred times greater 20-mg dose produces profound opioid antagonism, precipitating opioid withdrawal in a person who had opioid use disorder on morphine or methadone. It is important to recognize that the goal is to reverse the respiratory depression and not to administer so much naloxone that it precipitates opiate withdrawal. Because naloxone only lasts a few hours and most opioids last considerably longer, an IV naloxone drip with close monitoring is frequently employed to provide a continuous level of antagonism for 24–72 h depending on the opioid used in the overdose (e.g., morphine vs methadone). Whenever naloxone has only a limited effect, other sedative drugs that produce significant overdoses must be considered. The most common are benzodiazepines, which have produced overdoses and deaths in combination with buprenorphine. A specific antagonist for benzodiazepines—flumazenil at 0.2 mg/min—can be given to a maximum of 3 g/h, but it may precipitate seizures and increase intracranial pressure. Like naloxone, administration for a prolonged period is usually required because most benzodiazepines remain active for considerably longer than flumazenil. Support of vital functions may include oxygen and positive-pressure breathing, IV fluids, pressor agents for hypotension, and cardiac monitoring to detect QT prolongation, which might require specific treatment. Activated charcoal and gastric lavage may be helpful for oral ingestions, but intubation will be needed if the patient is stuporous.

TABLE 446-2 Management of Opioid Overdose

Establish airway. Intubation and mechanical ventilation may be necessary.
Naloxone 0.4–2.0 mg (IV, IM, or endotracheal tube). Onset of action with IV is ~1–2 min.
Repeat doses of naloxone if needed to restore adequate respiration or a continuous infusion of naloxone can be used.
One-half to two-thirds of the initial naloxone dose that reversed the respiratory depression is administered on an hourly basis (note: naloxone dosing is not necessary if the patient has been intubated).

OPIOID WITHDRAWAL

The principles of detoxification are the same for all drugs: to substitute a longer-acting, orally active, pharmacologically equivalent medication for the substance being used, stabilize the patient on that medication, and then gradually withdraw the substituted medication. Methadone and buprenorphine are the two medications used to treat opioid use disorder. Clonidine, a centrally acting sympatholytic agent, has also been used for detoxification in the United States. By reducing central sympathetic outflow, clonidine mitigates many of the signs of sympathetic overactivity but typically requires augmentation with other agents. Clonidine has no narcotic action and is not addictive. Lofexidine, a clonidine analogue with less hypotensive effect, is not yet approved in the United States.

Methadone for Detoxification Dose-tapering regimens for detoxification using methadone range from 2 to 3 weeks to as long as 180 days, but this approach is controversial given the relative effectiveness of methadone maintenance and the low success rates of detoxification. Unfortunately, the vast majority of patients tend to relapse to heroin or other opioids during or after the detoxification period, indicative of the chronic and relapsing nature of opioid use disorder.

Buprenorphine for Detoxification Buprenorphine does not appear to lead to better outcomes than methadone but is superior to clonidine in reducing symptoms of withdrawal, retaining patients in a withdrawal protocol, and in completing treatment.

Alpha-2-Adrenergic Agonists for Detoxification Several alpha-2-adrenergic agonists have relieved opioid withdrawal by suppressing brain NE hyperactivity. Clonidine relieves some signs and symptoms of opioid withdrawal such as lacrimation, rhinorrhea, muscle pain, joint pain, restlessness, and gastrointestinal symptoms. Related agents are lofexidine, guanfacine, and guanabenz acetate. Lofexidine can be dosed up to ~2 mg/d and appears to be associated with fewer adverse effects. Clonidine or lofexidine is typically administered orally, in three or four doses per day, with dizziness, sedation, lethargy, and dry mouth as the primary adverse side effects. Outpatient-managed withdrawal will require close follow-up often with naltrexone maintenance to prevent relapse.

Rapid and Ultrarapid Opioid Detoxification The opioid antagonist naltrexone typically combined with an alpha-2-adrenergic agonist has been purported to shorten the duration of withdrawal without significantly increasing patient discomfort. Completion rates using naltrexone and clonidine range from 75 to 81% compared to 40 to 65% for methadone or clonidine alone. Ultrarapid opioid detoxification is an extension of this approach using anesthetics but is highly controversial due to the medical risks and mortality associated with it.

Opioid Agonist Medications for Maintenance Methadone maintenance substitutes a once-daily oral opioid dose for three- to four-times daily heroin. Methadone saturates the opioid receptors and, by inducing a high level of opioid tolerance, blocks the euphoria from additional opioids. Buprenorphine, a partial opioid agonist, also can be given once daily at sublingual doses of 4–32 mg daily, and in contrast to methadone, it can be given in an office-based primary care setting.

METHADONE MAINTENANCE Methadone's slow onset of action when taken orally, long elimination half-life (24–36 h), and production of cross-tolerance at doses from 80 to 150 mg are the basis for its efficacy in treatment retention and reductions in IV drug use, criminal activity, and HIV risk behaviors and mortality. Methadone can prolong the QT interval at rates as high as 16% above the rates in nonmethadone-maintained, drug-injecting patients, but it has been used safely in the treatment of opioid use disorder for 40 years.

BUPRENORPHINE MAINTENANCE While France and Australia have had sublingual buprenorphine maintenance since 1996, it was first approved by the U.S. Food and Drug Administration (FDA) in 2002 as

a Schedule III drug for managing opioid use disorder. Unlike the full agonist methadone, buprenorphine is a partial agonist of mu-opioid receptors with a slow onset and long duration of action. Its partial agonism reduces the risk of unintentional overdose but limits its efficacy to patients who need the equivalent of only 60–70 mg of methadone, and many patients in methadone maintenance require higher doses up to 150 mg daily. Buprenorphine is combined with naloxone at a 4:1 ratio in order to reduce its abuse liability. Because of pediatric exposures and diversion of buprenorphine to illicit use, a new formulation, using mucosal films rather than sublingual pills that were crushed and snorted, is now marketed. A subcutaneous buprenorphine implant that lasts up to 6 months has recently been approved by the FDA as a formulation improvement to prevent pediatric exposures and illicit diversion and to enhance compliance.

In the United States, the ability of primary care physicians to prescribe buprenorphine for opioid use disorder represents an important opportunity to improve access and quality of treatment as well as reduce social harm. Europe, Asia, and Australia have found reduced opioid-related deaths and drug-injection-related medical morbidity with buprenorphine available in primary care. Retention in office-based buprenorphine treatment has been high as 70% at 6-month follow-ups.

Opioid Antagonist Medications The rationale for using narcotic antagonist therapy is that blocking the action of self-administered opioids should eventually extinguish the habit, but this therapy is poorly accepted by patients. Naltrexone, a long-acting orally active pure opioid antagonist, can be given three times a week at doses of 100–150 mg. Because it is an antagonist, the patient must first be detoxified from opioids before starting naltrexone. It is safe even when taken chronically for years, is associated with few side effects (headache, nausea, and abdominal pain), and can be given to patients infected with hepatitis B or C without producing hepatotoxicity. However, most providers refrain from prescribing naltrexone if liver function tests are three times above normal levels. Naltrexone maintenance combined with psychosocial therapy is effective in reducing heroin use, but medication adherence is low. Depot injection formulations lasting up to 4 weeks markedly improve adherence, retention, and drug use. Subcutaneous naltrexone implants in Russia, China, and Australia have doubled treatment retention and reduced relapse to half that of oral naltrexone. In the United States, a depot naltrexone formulation is available for monthly use and maintains blood levels equivalent to 25 mg of daily oral use.

Medication-Free Treatment Most opioid users enter medication-free treatments in inpatient, residential, or outpatient settings, but 1- to 5-year outcomes are very poor compared to pharmacotherapy except for residential settings lasting 6–18 months. The residential programs require full immersion in a regimented system with progressively increasing levels of independence and responsibility within a controlled community of fellow drug users. These medication-free programs, as well as the pharmacotherapy programs, also include counseling and behavioral treatments designed to teach interpersonal and cognitive skills for coping with stress and for avoiding situations leading to easy access to drugs or to craving. Relapse is prevented by having the individual very gradually reintroduced to greater responsibilities and to the working environment outside of the protected therapeutic community.

PREVENTION

Preventing the development of opioid use disorder represents a critically important challenge for physicians. Opioid prescriptions are the most common source of drugs accessed by adolescents who begin a pattern of illicit drug use. The major sources of these drugs are family members, not drug dealers or the Internet. Pain management involves providing sufficient opioids to relieve the pain over as short a period of time as the pain warrants (Chap. 10). The patient then needs to dispose of any remaining opioids, not save them in the medicine cabinet, because this behavior leads to diversion by adolescents. Finally, physicians should never prescribe opioids for themselves.

FURTHER READING

Brady KT et al: Prescription opioid misuse, abuse, and treatment in the United States: An update. Am J Psychiatry 173:18, 2016.

D'Onofrio G et al: Emergency department-initiated buprenorphine/naloxone treatment for opioid dependence: A randomized clinical trial. JAMA 313:1636, 2015.

Dugosh K et al: A systematic review on the use of psychosocial interventions in conjunction with medications for the treatment of opioid addiction. J Addict Med 10:93, 2016.

Trang T et al: Pain and poppies: The good, the bad, and the ugly of opioid analgesics. J Neurosci 35:13879, 2015.

447 Cocaine and Other Commonly Used Drugs

Karran A. Phillips, Antonello Bonci

The use of cocaine and other psychostimulants reflects a complex interaction between the pharmacology of the drug, the personality and expectations of the user, and the environmental context in which the drug is used. Polydrug use involving the concurrent use of several drugs with different pharmacologic effects is increasingly common. Sometimes one drug is used to enhance the effects of another, as with the combined use of cocaine and nicotine, or cocaine and heroin in methadone-maintained patients. Some forms of polydrug use, such as the combined use of IV heroin and cocaine, are especially dangerous and account for many hospital emergency room visits. Chronic cocaine and psychostimulant use may cause a number of adverse health consequences and may exacerbate preexisting disorders such as hypertension and cardiac disease. In addition, the combined use of two or more drugs may accentuate medical complications associated with use of one drug. Chronic drug use is often associated with immune system dysfunction and increased vulnerability to infections, including risk for HIV infection. The concurrent use of cocaine and opiates ("speedball") is frequently associated with needle sharing by people using drugs intravenously. People who use IV drugs represent the largest single group of individuals with HIV infection in several major metropolitan areas in the United States as well as in many parts of Europe and Asia.

Stimulants and hallucinogens have been used to induce euphoria and alter consciousness for centuries. Cocaine and marijuana are two of the most commonly used drugs today. Synthetic variations of marijuana and a variety of hallucinogens have become popular recently, and new drugs are continually being developed. This chapter describes the subjective and adverse medical effects of cocaine, other psychostimulants including methamphetamine, 3,4-methylenedioxymethamphetamine (MDMA), and cathinones; as well as hallucinogens such as phencyclidine (PCP), D-lysergic acid diethylamide (LSD), salvia divinorum; and marijuana and the synthetic cannabinoids. Some options for medical management of severe adverse effects are also described.

PSYCHOSTIMULANTS

PHARMACOKINETICS/DYNAMICS, NEUROBIOLOGY, AND EPIDEMIOLOGY

Cocaine Cocaine is a powerful stimulant drug made from the cocoa plant. It has local anesthetic, vasoconstrictor, and stimulant properties. Cocaine is a Schedule II drug, which means that it has high potential for abuse but can be administered by a physician for legitimate medical uses, such as local anesthesia for some eye, ear, and throat surgeries.

Cocaine comes in a variety of forms, the most commonly used being the hydrochloride salt, sulfate, and a base. The salt is an acidic, water-soluble powder with a high melting point, used by snorting or sniffing intranasally or by dissolving it in water and injecting it

intravenously. When used intranasally the bioavailability of cocaine is about 60 per percent. Cocaine sulfate ("paste") has a melting point of almost 200°C, so it has limited use, but is sometimes smoked with tobacco. The base form can be freebase or crystallized as crack. Cocaine freebase is made by adding a strong base to an aqueous solution of cocaine and extracting the alkaline freebase precipitate. It has a melting point of 98°C and can be vaporized and inhaled. Freebase cocaine can also be crystallized and sold as crack or rock, which is also smoked or inhaled. Street dealers often dilute (or "cut") cocaine with nonpsychoactive substances such as cornstarch, talcum powder, flour, or baking soda, or adulterate it with other substances with similar effects (like procaine or amphetamine) to increase their profits.

Given the extensive pulmonary vasculature, smoked or inhaled cocaine reaches the brain very quickly and produces a rapid and intense (yet transient) high, which enhances its addictive potential. Cocaine binds to the dopamine (DA) transporter and blocks DA reuptake, which increases synaptic levels of the monoamine neurotransmitters DA, norepinephrine (NE), and serotonin, in both the central nervous system (CNS) and the peripheral nervous system (PNS). Use of cocaine, like other drugs of abuse, induces long-term changes in the brain. Animal studies have shown adaptations in neurons that release the excitatory neurotransmitter glutamate after cocaine exposure.

According to the National Survey on Drug Use and Health (NSDUH), in 2015 about 1.9 million people (~0.7% of the population) were current users of cocaine, including about 394,000 current users of crack (0.1% of the population in the United States). There were 53,000 adolescents aged 12–17 (0.2% of adolescents) who were current users of cocaine in 2015. About 896,000 people aged ≥12 (0.3% of the population) in 2015 had a cocaine use disorder in the past year. The Drug Abuse Warning Network (DAWN) reported that in 2011 there were 505,224 cocaine-related emergency department (ED) visits, or about 162 ED visits per 100,000 of the U.S. population.

Methamphetamine Methamphetamine is a stimulant drug usually used as a white, bitter-tasting powder or a pill. Crystal methamphetamine is a form of the drug that looks like glass fragments or shiny, bluish-white rocks. It can be inhaled/smoked, swallowed (pill), snorted, or injected after being dissolved in water or alcohol. When smoked, methamphetamine exhibits 90.3% bioavailability, compared to 67.2% for oral ingestion. Methamphetamine exists in two stereoisomers, the L- and D-forms. D-Methamphetamine, or the dextrorotatory enantiomer, is a more powerful psychostimulant, with 3–5 times the CNS activity as compared to L-methamphetamine. Methamphetamine is a cationic lipophilic molecule which stimulates the release, and partially blocks the reuptake, of newly synthesized catecholamines in the CNS. Methamphetamine has a similar structure to the DA, NE, serotonin, and vesicular monoamine transporters and reverses their endogenous function, resulting in release of monoamines from storage vesicles into the synapse. Methamphetamine also attenuates the metabolism of monoamines by inhibiting monoamine oxidase.

Methamphetamine is more potent and more efficacious than amphetamine, resulting in much higher concentrations of synaptic DA and more toxic effects on nerve terminals. Outside the medical context, methamphetamine's pharmacokinetics and low cost often result in a chronic and continuous, high dose self-administered use pattern.

According to the NSDUH, ~897,000 people (0.3% of the population) aged ≥12 were current users of methamphetamine in 2015. Meanwhile, about 13,000 adolescents (0.1%) aged 12–17 were current methamphetamine users in 2015. There were also 51.3 ED visits, per 100,000 of the population, related to illicit stimulants (predominately amphetamines and methamphetamine) in 2011.

MDMA and Cathinones MDMA is an illegal drug that has stimulant and psychedelic effects. With MDMA use, individuals experience increased physical and mental energy, distortions in time and perception, emotional warmth, empathy toward others, a general sense of well-being, decreased anxiety, and an enhanced enjoyment of tactile experience. MDMA is usually taken orally in a tablet, capsule, or liquid form, and its effects last ~3–6 h. MDMA alters brain chemistry by binding to serotonin transporters and increasing the release of serotonin, NE, and DA. Research in animals has shown that MDMA in moderate to high doses can cause loss of serotonin-containing nerve endings and permanent damage. MDMA is a Schedule I drug, along with other substances with no proven therapeutic value. MDMA is currently in clinical trials as a possible treatment for posttraumatic stress disorder and anxiety in terminally ill patients, and for social anxiety in autistic adults.

Adulteration of MDMA tablets with methamphetamine, ketamine, caffeine, the over-the-counter cough suppressant dextromethorphan (DXM), the diet drug ephedrine, and cocaine is common. MDMA is rarely used alone and is often mixed with other substances, such as alcohol and marijuana, making the scope of its use difficult to ascertain. The Monitoring the Future study estimated that, in 2016, the lifetime prevalence of MDMA use among eighth graders was 1.7%, tenth graders was 2.8%, and twelfth graders was 4.9%, with the most use among 18–25 year olds.

Cathinone is an alkaloid psychostimulant found in the khât (Catha edulis) plant, which grows at high altitudes in East Africa and the Middle East. The actions and effects of khât are like those of the amphetamines, and misusers are at increased risk for acute myocardial infarction and stroke, due to inotropic and chronotropic effects on the heart, vasospasm of coronary arteries, and catecholamine-induced platelet aggregation.

Prescribed Psychostimulants Methylphenidate, amphetamine, and methamphetamine are psychostimulants approved in the United States for treatment of attention-deficit hyperactivity disorder (ADHD), weight control, and narcolepsy. Phenylpropanolamine, a psychostimulant used primarily for weight control, was found to be related to hemorrhagic stroke in women and removed from the market in 2005. These drugs deserve mention here, as there has been increased use of nonprescribed amphetamines or methylphenidate as a study aid among college students, and an energy and productivity booster for so-called "supermoms." According to the NSDUH, of the 7.7 million people, aged ≥12, who had a past year stimulant use disorder (SUD) related to their use of illicit drugs, 0.4 million misused prescription stimulants.

CLINICAL MANIFESTATIONS

Psychostimulants produce the same acute CNS effects: euphoria, increased energy/decreased fatigue, reduced need for sleep, decreased appetite, decreased distractibility, increased self-confidence and alertness, increased libido, and prolonged orgasm, independent of the specific psychostimulant or route of administration. Peripheral effects may include tremor, diaphoresis, hypertonia, tachypnea, hyperreflexia, and hyperthermia. Many of the effects are biphasic; for example, low doses improve psychomotor performance, while higher doses may cause tremors or convulsions. α-adrenergically mediated cardiovascular effects are also biphasic, with low doses resulting in increased vagal tone and decreased heart rate, and high doses causing increased heart rate and blood pressure. Psychostimulant use can result in restlessness, irritability, and insomnia and, at higher doses, suspiciousness, repetitive stereotyped behaviors, and bruxism. Endocrine effects may include impotence, gynecomastia, menstrual function disruptions, and persistent hyperprolactinemia **(Table 447-1)**.

Overdose presents as sympathetic nervous system overactivity with psychomotor agitation, hypertension, tachycardia, headache, and mydriasis, and can lead to convulsions, cerebral hemorrhage or infarction, cardiac arrhythmias or ischemia, respiratory failure, or rhabdomyolysis. It is a medical emergency; treatment is largely symptomatic and should occur in an intensive care or telemetry unit. Inhalation of crack cocaine that is vaporized at high temperatures can cause airway burns, bronchospasm and other symptoms of pulmonary disease. MDMA has also been shown to raise body temperature and can occasionally result in liver, kidney, or heart failure, or even death.

Psychostimulants are often used with other drugs, including opioids and alcohol, whose CNS-depressant effects tend to attenuate psychostimulant-induced CNS stimulation. These combinations often have additive deleterious effects, increasing the risk of morbidity and mortality. An example of this risk is the use of cocaine with alcohol, which results in the metabolite, cocaethylene. Cocaethylene's effects on the cardiovascular system are additive to that of cocaine's effects, resulting in intensified pathophysiologic consequences.

TABLE 447-1 Complications of Cocaine Use

Cardiovascular	Acute • Arterial vasoconstriction • Thrombosis • Tachycardia • Hypertension • Increased myocardial oxygen demand • Increased vascular shearing forces • Coronary vasoconstriction • Cardiac ischemia • Left ventricular dysfunction/heart failure (high blood concentrations) • Supraventricular and ventricular dysrhythmias • Aortic dissection/rupture Chronic • Accelerated atherogenesis • Left ventricular hypertrophy • Dilated cardiomyopathy
Central and Peripheral Nervous	• Hyperthermia • Psychomotor agitation • Tremor • Hyperreflexia • Hypertonia • Headache • Seizures • Coma • Intracranial hemorrhage • Focal neurologic symptoms
Pulmonary	• Angioedema (inhaled) • Pharyngeal burns (inhaled) • Pneumothorax • Pneumomediastinum • Pneumopericardium • Reversible airway disease exacerbations • Bronchospasm • Shortness of breath (“crack lung”) • Tachypnea • Pulmonary infarction
Gastrointestinal	• Perforated ulcers • Ischemic colitis • Bowel infarction • Impaction (body packing) • Hepatic enzyme elevation
Renal	• Metabolic acidosis • Renal infarction • Rhabdomyolysis
Endocrine	• Impotence • Gynecomastia • Menstrual function disruptions • Hyperprolactinemia
Other	• Diaphoresis • Irritability • Insomnia • Bruxism • Stereotypy • Splenic infarction • Acute angle-closure glaucoma • Vasospasm of the retinal vessels (unilateral or bilateral vision loss) • Mydriasis • Madarosis • Abruptio placentae

Adulteration of psychostimulants, particularly cocaine, with other drugs is common and can have additional potential health consequences. Levamisole, an anthelminthic and immunomodulator used primarily in veterinary medicine, has been found in cocaine and can cause agranulocytosis, leukoencephalopathy, and cutaneous vasculitis, which has resulted in cutaneous necrosis. Clenbuterol, a sympathomimetic amine used clinically as a bronchodilator, has also been found in cocaine and can result in tachycardia, hyperglycemia, palpitations, and hypokalemia. Studies in Europe have found that in addition to levamisole some of the most common adulterants in cocaine include: phenacetin, lidocaine, caffeine, diltiazem, hydroxyzine, procaine, tetracaine, paracetamol, creatine, and benzocaine.

Withdrawal from psychostimulants often includes hypersomnia, increased appetite, and depressed mood. Acute withdrawal typically lasts 7–10 days, but residual symptoms, possibly associated with neurotoxicity, may persist for several months. Psychostimulant withdrawal is not thought to be a driver of ongoing use. Debate remains as to whether, in psychostimulant withdrawal, symptoms decline monotonically or occur in discrete phases, getting worse before they get better. Most current theories of psychostimulant addiction emphasize the primary role of conditioned craving, which can persist long after physiological withdrawal has abated.

Injection of psychostimulants places people at increased risk of contracting infectious diseases from exposure to blood or other bodily fluids, such as HIV and hepatitis B and C. Psychostimulant use can also increase risk for infection by causing altered judgment and decision-making, leading to risky behaviors, such as unprotected sex. There is some evidence that psychostimulant use may worsen the progression of HIV/AIDS via increased injury to nerve cells exacerbating cognitive problems.

SCREENING AND DIAGNOSIS

The *Diagnostic and Statistical Manual of Psychiatric Disorders*, 5th edition (*DSM*-5) defines a SUD as a pattern of use of amphetamine-type substances, cocaine, or other stimulants leading to clinically significant impairment or distress, as manifested by at least two of the following 11 problems within a 12-month period: taking larger amounts, or over a longer period of time, than intended; persistent desire or unsuccessful efforts to cut down or control; a great deal of time spent in activities necessary to obtain, use, or recover; craving; use resulting in failure to fulfill major role obligations; continued use, despite recurrent social or interpersonal problems; giving up social, occupational, or recreational activities; recurrent use in physically hazardous situations; continued use despite persistent or recurrent physical or psychological problems; tolerance; and withdrawal symptoms, or avoidance of withdrawal symptoms, by continued use.

TREATMENT

Psychostimulants

COCAINE ACUTE INTOXICATION

As with all emergency situations the first task is to ensure a patent airway, breathing, and circulation. With cocaine use, succinylcholine is relatively contraindicated in rapid sequence intubation; consider rocuronium (1 mg/kg IV) or another nondepolarizing agent as an alternative. If psychomotor agitation occurs, rule out hypoglycemia and hypoxemia first, and then administer benzodiazepines (e.g., diazepam 10 mg IV and then 5–10 mg IV every 3–5 min until agitation controlled). Benzodiazepines are usually sufficient to address cardiovascular side effects. Severe or symptomatic hypertension can be treated with phentolamine, nitroglycerin, or nitroprusside. Hyperthermic patients should be cooled within ≤30 min with the goal to achieve a core body temperature of <39°C (102°F). Evaluation of chest pain in someone using cocaine should include an electrocardiogram, chest radiograph, and biomarkers to exclude myocardial infarction. The treatment approach is similar to noncocaine-induced chest pain, however, it is recommended that whenever possible beta blockers not be used in people who use cocaine. The concern

arises from the potential unopposed alpha-adrenergic stimulation that results from beta blockade possibly causing coronary arterial vasoconstriction, ischemia, and infarction and also limited data supporting the benefit of beta blockers in cocaine-related cardiovascular complications. If beta blockers are to be given, it is suggested that mixed alpha/beta blockers, e.g., labetalol and carvedilol, be used rather than nonselective beta blockers, and only in situations where the benefits outweigh the risks. Because many instances of cocaine-related mortality have been associated with concurrent use of other illicit drugs (particularly heroin), the physician must be prepared to institute effective emergency treatment for multiple drug toxicities.

COCAINE USE DISORDERS

Treatment of cocaine use disorders requires the combined efforts of primary care physicians, psychiatrists, and psychosocial care providers. Early abstinence from cocaine use is often complicated by symptoms of depression and guilt, insomnia, and anorexia, which may be as severe as those observed in major affective disorders and can last for months and even years after use has stopped.

Behavioral therapies, including cognitive-behavioral therapy (CBT), the community reinforcement approach (CRA), contingency management (CM; providing rewards to patients who remain substance free), motivational enhancement therapy (MET), combinations of these, and others remain the mainstay of treatment for stimulant use disorders and show modest benefit. These behavioral therapies are designed to help modify the patient's thinking, expectancies, and behaviors, and to increase life-coping skills, with behavioral interventions to support long-term, drug-free recovery.

There are no U.S. Food and Drug Administration (FDA)-approved medications for psychostimulant addiction. Current research includes several neurotransmitter-based strategies, including DA agonist-, serotonin-, γ-aminobutyric acid (GABA)-, and glutamate-based approaches. Other therapies being studied for the treatment of psychostimulant use disorder include: acamprosate (possibly via a role in Ca^{2+} supply), galantamine (reversible acetylcholine esterase inhibitor, which may strengthen impulse control, as well as cognitive and social abilities depleted by long-term psychostimulant use), naltrexone (opiate receptor antagonist), doxazosin (alpha-adrenergic antagonist), and varenicline (partial agonist at the α4β2 nicotinic acetylcholine receptors and DA neurotransmission enhancer). Vaccines for cocaine and methamphetamine use disorders are also being developed. Finally, recent preliminary studies have brought attention to the use of brain stimulation techniques such as transcranial magnetic stimulation (TMS), theta-burst stimulation (TBS), and transcranial direct current stimulation (tDCS) to treat psychostimulant use disorders, although further studies are warranted.

HALLUCINOGENS

Hallucinogens are a diverse group of drugs causing alteration of thoughts, feelings, sensations, and perceptions. Their use in religious and spiritual rituals goes back centuries. Hallucinogens can be found naturally in plants and mushrooms, or can be human-made. They include: ayahuasca (a tea made from Amazonian plants containing *dimethyltryptamine* (DMT), the primary mind-altering ingredient); DMT (aka Dimitri, can also be synthesized in a lab); *LSD* (clear or white odorless material made from lysergic acid found in rye and other grain fungus); peyote (mescaline, derived from a small, spineless cactus or made synthetically); and *4-phosphoryloxy-N,N-dimethyltryptamine* (psilocybin, comes from certain South and North American mushrooms).

A subgroup of hallucinogens produces the added sensation of feeling out of control or disconnected from one's body or surroundings, these include: DXM (an over-the-counter cough suppressant, when used in high doses); ketamine (a human and veterinary anesthetic); PCP (cyclohexylamine derivative and dissociative anesthetic); and *Salvia divinorum* (salvia, a Mexican, Central, and South American plant).

Hallucinogens are used in a wide variety of ways, including smoking, snorting, and transmucosally. Except for salvia, whose effects last 30 min, the onset of action of hallucinogens is within 20–90 min and the duration of action can be as long as 6–12 h. Hallucinogens disrupt brain chemistry, specifically the neurotransmitters serotonin and glutamate. Effects on the serotonin system can disturb mood, sensory perception, sleep, appetite, body temperature, sexual behavior, and muscle control. Glutamate system effects include perturbations in pain perception, responses to the environment, emotion, and learning and memory.

According to the NSDUH, in 2015, an estimated 1.2 million (0.5%) of people aged ≥12 reported current hallucinogen use. The highest rates were among young adults aged 18–25, with 1.8% (636,000) young adults reporting current hallucinogen use.

Clinical manifestations of hallucinogen use include: hallucinations, intensified feelings, heightened sensory experiences, and time perturbations. Additional physiologic responses include: nausea, increased heart rate, blood pressure, respiratory rate, or body temperature, loss of appetite, xerostomia, sleep problems, synesthesia, impaired coordination, and hyperhidrosis. "Bad trips" (negative experiences with hallucinogen use) can include panic, paranoia, and psychosis, and may persist for up to 24 h. Such experiences are best treated with supportive reassurance. There is some evidence that chronic effects of hallucinogen use can occur, including persistent psychosis, memory loss, anxiety, depression, and flashbacks.

There are currently no FDA-approved medications for the treatment of hallucinogen addiction. Research on behavioral treatments for hallucinogen addiction is underway.

MARIJUANA

Marijuana policies in several states in the United States have legalized marijuana for medical and/or recreational use. Marijuana refers to the dried leaves, flowers, stems, and seeds from the hemp plant, *Cannabis sativa*. There are >480 natural components found within the *Cannabis sativa* plant, of which 66 have been classified as "cannabinoids"; chemicals unique to the plant. The degree of psychological activity allows for the differentiation of the cannabinoids. Three classes of cannabinoids, the cannabigerols (CBGs), cannabichromenes (CBCs), and cannabidiols (CBDs) are not known to have psychological effects. The psychologically active cannabinoids include: tetrahydrocannabinols (THC), cannabinol (CBN), and cannabinodiol (CBDL), among other cannabinoids. Delta-9-tetrahydrocannabinol (THC) is the main psychoactive chemical, responsible for most of the intoxicating effects. Stronger forms of marijuana include sinsemilla (from specially tended female plants) and concentrated resins, including honey-like *hash oil*, waxy *budder*, and hard amber-like *shatter*.

When smoked, marijuana is quickly absorbed from the lungs into the blood and then sequestered in tissues and metabolized by the liver. Marijuana can also be baked into foods (edibles) and eaten with a resulting slower onset of action of 30–60 min. Cannabinoid receptors (CB_1 and CB_2) have been identified in the CNS (cerebral cortex, basal ganglia, and hippocampus) and PNS, as well as on T and B lymphocytes. Endogenous cannabinoids (such as anandamide) as well as exogenous cannabinoids (THC) bind to the CB_1 receptors. Cannabinoid effects occur in the limbic system, affecting memory, cognition and psychomotor performance, and the mesolimbic pathway, impacting the reward pathway and areas of pain perception. Effects include: altered senses, altered sense of time, laughter, changes in mood, psychomotor retardation, difficulty with thinking and problem-solving, and impaired memory.

Marijuana is the most commonly used illicit drug in the United States, with 22.2 million (8.3%) current marijuana users aged ≥12 (i.e., users in the past 30 days). In 2015, >11 million young adults, ages 18–25, used marijuana in the past year, and 19.8% used in the past month. Of the 7.7 million people aged ≥12 who had a past year SUD related to their use of illicit drugs, 4.0 million had a past year disorder related to their use of marijuana. In 2015, 2.6% of adolescents aged 12–17, 5.1% of young adults aged 18–25, and 0.8% of adults aged ≥26 had a marijuana use disorder in the past year. Emergency room visits involving marijuana have increased, which may be due to increased

THC levels in marijuana over the past few decades resulting in a greater chance of a harmful reaction.

Acute intoxication brings with it a perceived sense of relaxation and mild euphoria, accompanied by some degree of impairment in memory, concentration, judgment, and perceptual and psychomotor function, as well as anxiety, paranoia, and rarely, psychosis. As with all psychoactive compounds, the experience changes depending on the individual's environment and state of mind at the time of use. Physical signs of marijuana use include conjunctival injection and tachycardia. Adverse physical effects of marijuana include: respiratory problems due to inhaled pulmonary irritants and lower birth weights in pregnancy.

Chronic marijuana use may also have adverse psychological effects, which may not be permanent, such as impaired concentration and learning, insomnia, and worsening symptoms in schizophrenia. Upon cessation, or cutting back, there is evidence of a withdrawal syndrome consisting of irritability, insomnia, anorexia, anxiety, and craving. Individuals who begin marijuana use before age 17, while the brain is still developing, may be more prone to cognitive deficits, and may be at higher risk for polydrug addiction in the future.

There are no current medications to treat marijuana use disorder. Behavioral therapies (CBT, CM, MET) and symptomatic treatment of withdrawal, for example selective serotonin reuptake inhibitors (SSRIs) to treat related anxiety, may be effective. Preliminary studies and small clinical trials with zolpidem (sleep aid), buspirone (antianxiety/antistress medication), and gabapentin (antiepileptic) have been promising. Other agents being studied include *N*-acetylcysteine; fatty acid amid hydrolase (FAAH) inhibitors, which may reduce withdrawal by inhibiting the breakdown of endocannabinoids; and allosteric modulators that interact with cannabinoid receptors to inhibit THC's rewarding effects.

Therapeutic use of marijuana includes as an antiemetic in chemotherapy, appetite promoter in AIDS, intraocular pressure reducer in glaucoma, and spasticity reducer in multiple sclerosis and other neurologic disorders.

EMERGING DRUGS

With the aid of the Internet, and some basic over-the-counter (and other) ingredients, the rise of the "kitchen chemist" is upon us. The production of new psychoactive substances (NPSs), such as synthetic cathinones (bath salts) and synthetic cannabinoids (spice), is on the rise and has resulted in the use of unregulated psychoactive substances that are intended to copy the effects of more expensive illegal drugs, such as methamphetamine and cocaine.

Synthetic cathinones (bath salts) are human-made drugs that are chemically like khât, and are often stronger and more dangerous than the natural product. They usually take the form of a white or brown crystal-like powder, packaged in small plastic or foil bundles labeled "not for human consumption," or as "plant food," "jewelry cleaner," or "phone screen cleaner," and sold online and in drug paraphernalia stores. The popular nickname Molly (slang for "molecular") often refers to the purported "pure" crystalline powder form of MDMA, usually sold in capsules. However, people who purchase powder or capsules sold as Molly often actually get other drugs, such as synthetic cathinones. The uncertainty of what is actual in these synthetic products, whose components might change from batch to batch, makes them even more dangerous as anyone using them is unaware of what they contain and how their bodies will react.

The three most common synthetic cathinones are mephedrone, methylone, and MDPV (*3,4-methylenedioxypyrovalerone)*. With oral ingestion, these drugs have an onset of action from 15 to 45 min, and a duration of action that varies from 2 to 7 h. A recent study found that MDPV affects the brain in a manner similar to cocaine, but is at least 10 times more powerful. MDPV is the most common synthetic cathinone found in the blood and urine of patients admitted to EDs after taking "bath salts." High doses, or chronic use, of synthetic cathinones can lead to dangerous medical consequences, including psychosis, violent behaviors, tachycardia, hyperthermia, and even death.

Synthetic cannabinoids refer to a growing number of human-made psychoactive chemicals that are either sprayed on dried, shredded plant material so they can be smoked (herbal incense), or sold as liquids to be vaporized and inhaled in e-cigarettes and other devices (liquid incense). Synthetic cannabinoids act on the same brain cell receptors as THC, the psychoactive ingredient in marijuana, and with use people report elevated mood, relaxation, altered perception, and symptoms of psychosis, including extreme anxiety, confusion, paranoia, and hallucinations.

Overdose with synthetic cannabinoids can result in tachycardia, vomiting, violent behavior, and suicidal thoughts. Elevations in blood pressure due to vasoconstriction can impair blood flow to the heart, brain, kidney, liver, and other vital organs. Withdrawal symptoms include: headaches, anxiety, depression, and irritability. Behavioral and pharmacologic therapies for treatment of synthetic cannabinoid addiction have not yet been tested.

The ability to synthesize addictive and dangerous drugs relatively simply and rapidly, changing just a few molecules, yet retaining the effects, has allowed many of these emerging drugs to outpace the attempt to regulate them, resulting in a developing global public health concern.

GLOBAL CONSIDERATIONS

Cannabis remains the most commonly used drug globally, with an estimated 183 million people having used the drug in 2014, while amphetamines are the second most commonly used drug. Overall global trends show the use of cannabis has remained stable over the past 3 years; however, in subregions of North America and Western and Central Europe, cannabis use has increased. Cocaine use had remained stable until 2010 when it also began to rise, driven by an increase in cocaine use in South America. The use of amphetamines appears to be stable; however, drug use data may be an underestimate particularly in subregions in East and South-East Asia, where information is sparse. Globally, men are three times more likely than women to use cannabis, cocaine, or amphetamines, whereas women are more likely than men to participate in the nonmedical use of opioids and tranquilizers. Opioids, cocaine, amphetamines, and cannabis together accounted for almost 12 million life years lost due to premature death or disability in 2013 according to the United Nation's World Drug Report 2016. Stigma and marginalization makes treatment of drug use disorders difficult and hinders sustainable development incorporating gender equality and the empowerment of women and girls. Drug use further corrodes environmental and economic well-being as well as the ability to develop and sustain safe communities.

FUTURE DIRECTIONS

Despite their prevalence and public health impact, psychostimulant, hallucinogen, and marijuana use disorders have no FDA-approved treatment medications. While behavioral therapies such as CBT, CM, and MET have been shown effective in psychostimulant use disorders, further research needs to be done regarding their utility for hallucinogen and marijuana use disorders. Furthermore, based upon experience with opioid and alcohol use disorders, it is likely that the most efficacious treatments will employ a combination of behavioral and pharmacological therapy.

Additionally, new approaches that utilize emerging technologies have considerable potential for future treatment of psychostimulant use disorders. These include neurostimulation/neuromodulation (TMS, TBS, tDCS), optogenetic techniques (use of light to control neurons that have been genetically modified to express light-sensitive ion channels), wearable biosensors, and mobile technology, including ecological and geographical momentary assessment (EMA/GMA) as well as real-time interventions delivered via smartphone or other mobile devices.

FURTHER READING

Koob GF, Volkow ND: The neurobiology of addiction: A neurocircuitry analysis. Lancet Psychiatry 3:760, 2016.

Volkow ND et al: Neurobiologic advances from the brain disease model of addiction. N Engl J Med 374:363, 2016.

WEBSITES

American Society of Addiction Medicine: *https://www.asam.org/public-resources*

National Institute on Drug Abuse: *https://www.drugabuse.gov/drugs-abuse*

World Health Organization: *http://www.who.int/substance_abuse/en/*

448 Nicotine Addiction

David M. Burns

The use of tobacco leaf to create and satisfy nicotine addiction was introduced to Columbus by Native Americans and spread rapidly to Europe. Use of tobacco as cigarettes, however, only became popular in the twentieth century and so is a modern phenomenon, as is the epidemic of disease caused by this form of tobacco use.

Nicotine is the principal constituent of tobacco responsible for its addictive character, but other smoke constituents and behavioral associations contribute to the strength of the addiction. Addicted smokers regulate their nicotine intake by adjusting the frequency and intensity of their tobacco use both to obtain the desired psychoactive effects and avoid withdrawal.

Unburned cured tobacco used orally contains nicotine, carcinogens, and other toxicants capable of causing gum disease, oral and pancreatic cancers, and an increase in the risk of heart disease. When tobacco is burned, the resultant smoke contains, in addition to nicotine, >7000 other compounds that result from volatilization, pyrolysis, and pyrosynthesis of tobacco leaf and various chemical additives used in making different tobacco products. Tobacco smoke is composed of a fine aerosol and a vapor phase; the aerosol is of a size range that results in deposition in the airways and alveolar surfaces of the lungs. The aggregate of particulate matter, after subtracting nicotine and moisture, is referred to as tar.

The alkaline pH of smoke from blends of tobacco used for pipes and cigars allows sufficient absorption of nicotine across the oral mucosa to satisfy the smoker's need for this drug. Therefore, those who smoke pipes and cigars exclusively tend not to inhale the smoke into the lung, confining the toxic and carcinogenic exposure (and the increased rates of disease) largely to the upper airway. The acidic pH of smoke generated by the tobacco used in cigarettes dramatically reduces absorption of nicotine in the mouth, necessitating inhalation of the smoke into the larger surface of the lungs in order to absorb quantities of nicotine sufficient to satisfy the smoker's addiction. The shift to using tobacco as cigarettes, with resultant increased deposition and absorption of smoke in the lung, has created the epidemic of heart disease, lung disease, and lung cancer that dominates the current disease manifestations of tobacco use.

Several genes have been associated with nicotine addiction. Some reduce the clearance of nicotine, and others have been associated with an increased likelihood of becoming dependent on tobacco and other drugs as well as a higher incidence of depression. It is likely that genetic susceptibility can influence the probability that adolescent experimentation with tobacco will lead to addiction as an adult. Rates of smoking cessation have increased, and rates of nicotine addiction have decreased dramatically, since the mid-1950s, suggesting that factors other than genetics are more important influences for tobacco use.

Adult cigarette smoking prevalence has declined to about 15% in the United States, with only 11.4% of the population smoking every day. However, the rapidly rising smoking rate observed in the developing world is of concern, and the World Health Organization Framework Convention on Tobacco Control is encouraging effective tobacco control approaches in these countries with the hope of preventing a future epidemic of tobacco-related illness.

DISEASE MANIFESTATIONS OF CIGARETTE SMOKING

More than 480,000 individuals die prematurely each year in the United States from cigarette use; this represents almost one of every five deaths in the United States. Approximately 40% of cigarette smokers will die prematurely due to cigarette smoking unless they are able to quit.

The major diseases caused by cigarette smoking are listed in Table 448-1. The ratio of smoking-related disease rates in smokers compared to never smokers (relative risk) increases with advancing age for most cancers and for chronic obstructive pulmonary disease (COPD). However, relative risk declines with advancing age for cardiovascular diseases due to the increasing contribution of other risk factors to cardiovascular disease occurrence as age advances. Nevertheless, even for cardiovascular disease, the absolute difference in mortality rate between smokers and never smokers, called excess death rate, continues to increase with advancing age, as one would expect from a process of cumulative injury.

CARDIOVASCULAR DISEASES

Cigarette smokers are more likely than nonsmokers to develop both large-vessel atherosclerosis and small-vessel disease. Approximately 90% of peripheral vascular disease in the nondiabetic population can be attributed to cigarette smoking, as can ~50% of aortic aneurysms. In contrast, 24% of coronary artery disease and ~11% of ischemic and hemorrhagic strokes are caused by cigarette smoking. There is a multiplicative interaction between cigarette smoking and other cardiac risk factors such that the increment in risk produced by smoking among individuals with hypertension or elevated serum lipids is substantially greater than the increment in risk produced by smoking for individuals without these risk factors.

In addition to its role in promoting atherosclerosis, cigarette smoking also increases the likelihood of myocardial infarction and sudden

TABLE 448-1 Relative Risks for Current Smokers of Cigarettes

AGE	35–44	45–64	65–74	≥75
Males				
Lung cancer	14.33	19.03	28.29	22.51
Coronary heart disease	3.88	2.99	2.76	1.98
Cerebrovascular disease	2.17	1.48	1.23	1.12
Other vascular diseases			7.25	4.93
Chronic obstructive pulmonary disease (COPD)			29.69	23.01
All causes	2.55	2.97	3.02	2.40
Females				
Lung cancer	13.30	18.95	23.65	23.08
Other tobacco-related cancers	1.28	2.08	2.06	1.93
Coronary heart disease	4.98	3.25	3.29	2.25
Cerebrovascular disease	2.27	1.70	1.24	1.10
Other vascular diseases			6.81	5.77
COPD			38.89	20.96
All causes	1.79	2.63	2.87	2.47

Relative Risks for Selected Other Cancers		
Other cancers	Male	Female
Larynx	14.6	13
Lip, oral cavity, pharynx	10.9	5.1
Esophagus	6.8	7.8
Bladder	3.3	2.2
Kidney	2.7	1.3
Pancreas	2.3	2.3
Stomach	2	1.4
Liver	1.7	1.7
Colorectal	1.2	1.2
Cervix		1.6
Acute myeloid leukemia	1.4	1.4

cardiac death by promoting platelet aggregation and vascular occlusion. Reversal of these effects on coagulation may explain the rapid benefit of smoking cessation for a new coronary event demonstrable among those who have survived a first myocardial infarction. This effect may also explain the substantially higher rates of graft occlusion among continuing smokers following vascular bypass surgery for cardiac or peripheral vascular disease.

Cessation of cigarette smoking reduces the risk of a second coronary event within 6–12 months. Rates of first myocardial infarction and death from coronary heart disease decline within 2–4 years following cessation among those with no prior cardiovascular history. After 15 years of abstinence, the risk of a new myocardial infarction or death from coronary heart disease in former smokers is similar to that for those who have never smoked.

CANCER

Tobacco smoking causes cancer of the lung; lip; oral cavity; naso-, oro-, and hypopharynx; nasal cavity and paranasal sinuses; larynx; esophagus; stomach; pancreas; liver (hepatocellular); colon and rectum; kidney (body and pelvis); ureter; urinary bladder; uterine cervix; and acute myeloid leukemia. There is evidence suggesting that cigarette smoking may play a role in increasing the risk of breast cancer, particularly for premenopausal women. There does not appear to be a causal link between cigarette smoking and cancer of the endometrium, and there is a lower risk of uterine cancer among postmenopausal women who smoke. The risks of cancer increase with the increasing number of cigarettes smoked per day and with increasing duration of smoking. Additionally, there are synergistic interactions between cigarette smoking and alcohol use for cancer of the oral cavity and esophagus. Several occupational exposures synergistically increase lung cancer risk among cigarette smokers, most notably occupational asbestos and radon exposure.

Cessation of cigarette smoking reduces the risk of developing cancer relative to continuing smoking after about 4 years of abstinence, but even 20 years after cessation there is a modest persistent increased risk of developing lung cancer.

RESPIRATORY DISEASE

Cigarette smoking is responsible for 80% of COPD. Within 1–2 years of beginning to smoke regularly, many young smokers will develop inflammatory changes in their small airways, although lung function measures of these changes do not predict development of chronic airflow obstruction. Pathophysiologic changes in the lungs manifest and progress over longer durations of smoking proportional to smoking intensity and duration. Chronic mucous hyperplasia of the larger airways results in a chronic productive cough in as many as 80% of smokers >60 years of age. Chronic inflammation and narrowing of the small airways, and/or enzymatic digestion of alveolar walls resulting in pulmonary emphysema, can reduce expiratory airflow sufficiently to produce clinical symptoms of respiratory limitation in ~15–25% of smokers.

Changes in the small airways of young smokers will reverse after 1–2 years of cessation. There is also a small increase in measures of expiratory airflow following cessation among many individuals who have developed chronic airflow obstruction, but the major change following cessation is a slowing of the rate of decline in lung function with advancing age rather than a return of lung function toward normal.

PREGNANCY

Cigarette smoking is associated with several maternal complications of pregnancy: premature rupture of membranes, abruptio placentae, and placenta previa; there is also a small increase in the risk of spontaneous abortion among smokers. Infants of smoking mothers are more likely to experience preterm delivery, have a higher perinatal mortality rate, be small for their gestational age, and have higher rates of infant respiratory distress syndrome. They are more likely to die of sudden infant death syndrome and appear to have a developmental lag for at least the first several years of life.

OTHER CONDITIONS

Smoking delays healing of peptic ulcers; increases the risk of developing periodontal disease, diabetes, active tuberculosis, rheumatoid arthritis, osteoporosis, senile cataracts, and neovascular and atrophic forms of macular degeneration; and results in premature menopause, wrinkling of the skin, gallstones and cholecystitis in women, and male impotence. Patients who continue to smoke during treatment for cancer with chemotherapy or radiation have poorer outcomes and reduced survival.

ENVIRONMENTAL TOBACCO SMOKE

Long-term exposure to environmental tobacco smoke increases the risk of lung cancer and coronary artery disease among nonsmokers. It also increases the incidence of respiratory infections, chronic otitis media, and asthma in children and causes exacerbation of asthma in children. Some evidence suggests that environmental tobacco smoke exposure may increase the risk of premenopausal breast cancer.

PHARMACOLOGIC INTERACTIONS

Cigarette smoking may interact with a variety of other drugs (Table 448-2). Cigarette smoking induces the cytochrome P450 system, which may alter the metabolic clearance of drugs such as warfarin. This may result in inadequate serum levels in smokers as outpatients when the dosage is established in the hospital under nonsmoking conditions. Correspondingly, serum levels may rise when smokers are hospitalized

TABLE 448-2 Interactions of Smoking and Prescription Drugs

DRUG	INTERACTION
Cardiovascular and Pulmonary Drugs	
β blockers	Reduced lowering of heart rate and blood pressure
Flecainide	Increased first-pass clearance
Heparin	Faster clearance
Lidocaine	Increased first-pass clearance
Mexiletine	Increased first-pass clearance
Propranolol	Increased first-pass clearance
Theophylline	Faster metabolic clearance
Verapamil	Increased clearance
Warfarin	Increased metabolism lowers serum levels
Neuropsychiatric Drugs	
Amitriptyline	Increased clearance
Benzodiazepines	Less sedation
Clomipramine, Imipramine	Decreased serum concentrations
Chlorpromazine	Decreased serum concentrations
Clozapine	Decreased serum concentrations
Duloxetine	Decreased serum concentrations
Fluphenazine	Decreased serum concentrations
Fluvoxamine	Decreased serum concentrations
Haloperidol	Decreased serum concentrations
Naratriptan	Increased clearance
Olanzapine	Faster clearance
Trazodone	Decreased serum concentrations
Anticancer Drugs	
Erlotinib	Increased clearance, higher response rate, and improved survival in non-smokers
Gefitinib	Higher response rate and improved survival in non-smokers
Irinotecan	Increased clearance
Other Drugs	
Dextropropoxyphene	Less analgesia
Estrogens (oral)	Increased hepatic clearance
Fentanyl	Increased clearance
Insulin	Delayed absorption due to skin vasoconstriction
Rivastigmine	Increased clearance

and not allowed to smoke. Smokers may also have higher first-pass clearance for drugs such as lidocaine, and the stimulant effects of nicotine may reduce the effect of benzodiazepines or beta blockers.

OTHER FORMS OF TOBACCO USE

Other major forms of tobacco use are moist snuff deposited between the cheek and gum, chewing tobacco, pipes and cigars, and recently bidi (tobacco wrapped in tendu or temburni leaf; commonly used in India), clove cigarettes, and water pipes. Oral tobacco use leads to gum disease and can result in oral and pancreatic cancer as well as heart disease. There are dramatic differences in the risks evident for products used in Africa or Asia as compared to those in the United States and Europe.

The risk of upper airway cancers is similar among cigarette, pipe, and cigar smokers, whereas those who have smoked only pipes and cigars have a much lower risk of lung cancer, heart disease, and COPD. Cigarette smokers who switch to pipes or cigars do tend to inhale the smoke, increasing their risk.

A resurgence of cigar, bidi, and water pipe use among adolescents of both genders has raised concerns that these older forms of tobacco use are once again causing a public health problem.

ELECTRONIC CIGARETTES

Use of electronic nicotine delivery systems, often called e-cigarettes, is a new behavior where both the products and patterns of use are rapidly evolving. These devices electronically heat a solution, which may or may not contain nicotine, to produce a vapor that can be inhaled. Smaller devices resembling cigarettes often deliver less nicotine than combustible cigarettes, whereas the larger devices with substantial tanks for the nicotine solution can deliver amounts of nicotine similar to combustible cigarettes. Absent the identification of new toxicants or new disease risks, the existing evidence on toxicant exposure establishes that smokers who use e-cigarettes exclusively will receive less toxicant exposure than combustible cigarette smokers and, with sustained exclusive use, would be expected to experience less disease risk.

Addicted cigarette smokers can derive sufficient nicotine from e-cigarettes to satisfy their addiction and some will persist in e-cigarette use for multiple months. Given the newness of this behavior, it is not clear whether smokers who switch to exclusive e-cigarette use will persist in that behavior, quit nicotine use entirely, or relapse back to smoking over the multiyear period needed to alter disease risks.

Patterns of e-cigarette use indicate that only small percentages of adults use e-cigarettes. The highest prevalence of e-cigarette use is among adolescents and young adults where the prevalence of e-cigarette use substantially exceeds that of cigarettes. It is currently uncertain whether the high prevalence of e-cigarette use among youth will translate into nicotine addiction and combustible cigarette use as they age. Nevertheless, the prevalence of combustible cigarette use among adolescents and young adults has continued to decline over recent years, even with widespread uptake of e-cigarettes.

LOWER TAR AND NICOTINE CIGARETTES

Filtered cigarettes with lower machine-measured yields of tar and nicotine commonly use ventilation holes in the filters and other engineering designs to artificially lower the machine measurements. Smokers compensate for the lowered nicotine delivery resulting from these design changes by changing the manner in which they puff on the cigarette or the number of cigarettes smoked per day, and actual tar and nicotine deliveries are not reduced with use of these products negating any reduction in disease risks with their use. In addition, the amount of carcinogenic tobacco-specific nitrosamines in the tobacco used in cigarettes has increased over time.

Cigarette design changes that reduce machine-measured tar and nicotine also lead to deeper inhalation of the smoke. Presentation of more carcinogenic smoke to the alveolar portions of the lung increases the risk of adenocarcinoma of the lung. The increased adenocarcinoma risk produces a substantively greater overall risk for lung cancer among current smokers compared with smokers of cigarettes manufactured prior to the 1960s. This increased risk may also be present for COPD. It is the changes in cigarette design and composition of cigarettes over the past six decades that caused the dramatic rise in rates of adenocarcinoma of the lung observed over the past half century. There has been no increase in risk of lung cancer or adenocarcinoma of the lung over the same period among never smokers.

CESSATION

The process of stopping smoking is commonly a cyclical one, with the smoker sometimes making multiple attempts to quit and failing before finally being successful. Approximately 70–80% of smokers would like to quit smoking. More than one-half of current smokers attempted to quit in the last year, but only 6% quit for 6 months, and only 3% remain abstinent for 2 years. Clinician-based smoking interventions should repeatedly encourage smokers to try to quit, and to use different forms of cessation assistance with each new cessation attempt, rather than focusing exclusively on immediate cessation at the time of the first visit.

Advice from a physician to quit smoking, particularly at the time of an acute illness, is a powerful trigger for cessation attempts, with up to half of patients who are advised to quit making a cessation effort. Other triggers that may be enhanced by timely physician advice to quit include increases in the tax on cigarettes, media campaigns, and changes in rules to restrict smoking in the workplace.

PHYSICIAN INTERVENTION (TABLE 448-3)

All patients should be asked whether they smoke, how much they smoke, how long they have smoked, their past experience with quitting, and whether they are currently interested in quitting. The number of cigarettes smoked per day and smoking within 30 min of waking are useful measures of the intensity of nicotine addiction. Even those who are not interested in quitting should be encouraged and motivated to quit; provided a clear, strong, and personalized message by the clinician that smoking is an important health concern; and offered assistance if they become interested in quitting in the future. Many of those not currently expressing an interest in quitting may nevertheless make an attempt to quit in the subsequent year. For those interested in quitting, a quit date should be negotiated, usually not the day of the

TABLE 448-3 Clinical Practice Guidelines

Physician Actions
Ask: Systematically identify all tobacco users at every visit
Advise: Strongly urge all smokers to quit
Identify smokers willing to quit
Assist the patient in quitting
Arrange follow-up contact
Effective Pharmacologic Interventions[a]
First-line therapies
Nicotine gum (1.5)
Nicotine patch (1.9)
Nicotine nasal inhaler (2.3)
Nicotine oral inhaler (2.1)
Nicotine lozenge (2 mg 2.0, 4 mg 2.8)
Bupropion (2.0)
Varenicline (3.1)
Second-line therapies
Clonidine (2.1)
Nortriptyline (1.8)
Other Effective Interventions[a]
Physician or other medical personnel counseling (10 min) (1.84)
Intensive group smoking cessation programs (at least 4–7 sessions of 20- to 30-min duration lasting at least 2 and preferably 8 weeks) (1.3)
Intensive individual counseling (1.7)
Clinic-based smoking status identification system (3.1)
Telephone counseling (1.6)

[a]Numerical value following the intervention is the multiple for cessation success compared to no intervention.

visit but within the next few weeks, and a follow-up contact by office staff around the time of the quit date should be provided. There is a relationship between the amount of assistance a patient is willing to accept and the success of the cessation attempt.

There are a variety of nicotine-replacement products, including over-the-counter nicotine patches, gum, and lozenges, as well as nicotine nasal and oral inhalers available by prescription. These products can be used for up to 3–6 months, and some products are formulated to allow a gradual step-down in dosage with increasing duration of abstinence. Antidepressants such as bupropion (300 mg in divided doses for up to 6 months) have also been shown to be effective, as has varenicline, a partial agonist for the nicotinic acetylcholine receptor (initial dose 0.5 mg daily increasing to 1 mg twice daily at day 8; treatment duration up to 6 months). Combined use of nicotine-replacement therapy (NRT) and antidepressants as well as the use of gum or lozenges for acute cravings in patients using patches can increase cessation outcomes. Pretreatment with antidepressants or varenicline is recommended for 1–2 weeks prior to the quit date. Pretreatment with nicotine-replacement products is also useful prior to a cessation date. Longer duration of nicotine replacement as a maintenance therapy for those who are unsuccessful in quitting with a shorter duration of use is a useful strategy. NRT is provided in different dosages, with higher doses being recommended for more intense smokers. Clonidine or the tricyclic antidepressant nortriptyline should be reserved for patients who have failed on first-line pharmacologic treatment or who are unable to use other therapies. Antidepressants are more effective among smokers with a history of depression symptoms.

Current recommendations are to offer pharmacologic treatment, usually with nicotine patches or varenicline, to all who will accept it and to provide counseling and other support as a part of the cessation attempt. Cessation advice alone by a physician or his or her staff is likely to increase success compared with no intervention; a more comprehensive approach with advice, pharmacologic assistance, and counseling can increase cessation success nearly threefold.

For adult addicted smokers, switching to exclusive use of e-cigarettes, but not dual use with combusted cigarettes, may have a role in promoting cessation, particularly for those unlikely to try to quit with other proven cessation modalities. However, it is not yet clear whether e-cigarette use results in higher cessation outcomes than other pharmacological approaches in the context of physician-based smoking interventions.

Incorporation of cessation assistance into a practice requires a change of the care delivery infrastructure. Simple changes include (1) adding questions about smoking and interest in cessation on patient-intake questionnaires, (2) asking patients whether they smoke as part of the initial vital sign measurements made by office staff, (3) listing smoking as a problem in the medical record, and (4) automating follow-up contact with the patient on the quit date. These changes are essential to institutionalizing smoking intervention within the practice setting; without this institutionalization, the best intentions of physicians to intervene with their patients who smoke are often lost in the time crush of a busy practice.

PREVENTION

Approximately 85% of individuals who become cigarette smokers initiate the behavior during adolescence. Factors that promote adolescent initiation are parental or older-sibling cigarette smoking, tobacco advertising and promotional activities, the availability of cigarettes, and the social acceptability of smoking. The need for an enhanced self-image and to imitate adult behavior is greatest for those adolescents who have the least external validation of their self-worth, which may explain in part the enormous differences in adolescent smoking prevalence by socioeconomic and school performance strata.

Prevention of smoking initiation must begin early, preferably in the elementary school years. Physicians who treat adolescents should be sensitive to the prevalence of this problem even in the preteen population. Physicians should ask all adolescents whether they have experimented with tobacco or currently use tobacco, reinforce the fact that most adolescents and adults do not smoke, and explain that all forms of tobacco are both addictive and harmful.

FURTHER READING

El Dib R et al: Electronic nicotine delivery systems and/or electronic non-nicotine delivery systems for tobacco smoking cessation or reduction: A systematic review and meta-analysis. BMJ Open 23;7:e012680, 2017.

Sohn HS et al: Evidence supporting the need for considering the effects of smoking on drug disposition and effectiveness in medication practices: A systematic narrative review. Int J Clin Pharmacol Ther 53:621, 2015.

U.S. Department of Health and Human Services: *The Health Consequences of Smoking: 50 Years of Progress. A Report of the Surgeon General*. Treatment for Tobacco Use and Dependence; Appendix 14.5 Smoking Cessation Medications. Atlanta, GA, 2014. Available from *https://www.surgeongeneral.gov/library/reports/50-years-of-progress/50-years-of-progress-by-section.html* only available electronically. Accessed March 23, 2017.

U.S. Department of Health and Human Services: *The Health Consequences of Smoking: 50 Years of Progress. A Report of the Surgeon General*. Atlanta, GA, 2014. Available from *https://www.surgeongeneral.gov/library/reports/50-years-of-progress/index.html*. Accessed March 23, 2017.

449 Heavy Metal Poisoning

Howard Hu

Metals pose a significant threat to health through low-level environmental as well as occupational exposures. One indication of their importance relative to other potential hazards is their ranking by the U.S. Agency for Toxic Substances and Disease Registry, which maintains an updated list of all hazards present in toxic waste sites according to their prevalence and the severity of their toxicity. The first, second, third, and seventh hazards on the list are heavy metals: lead, mercury, arsenic, and cadmium, respectively (*http://www.atsdr.cdc.gov/spl/*). Specific information pertaining to each of these metals, including sources and metabolism, toxic effects produced, diagnosis, and the appropriate treatment for poisoning, is summarized in **Table 449-1**.

Metals are inhaled primarily as dusts and fumes (the latter defined as tiny particles generated by combustion). Metal poisoning can also result from exposure to vapors (e.g., mercury vapor in creating dental amalgams). When metals are ingested in contaminated food or drink or by hand-to-mouth activity (implicated especially in children), their gastrointestinal absorption varies greatly with the specific chemical form of the metal and the nutritional status of the host. Once a metal is absorbed, blood is the main medium for its transport, with the precise kinetics dependent on diffusibility, protein binding, rates of biotransformation, availability of intracellular ligands, and other factors. Some organs (e.g., bone, liver, and kidney) sequester metals in relatively high concentrations for years. Most metals are excreted through renal clearance and gastrointestinal excretion; some proportion is also excreted through salivation, perspiration, exhalation, lactation, skin exfoliation, and loss of hair and nails. The intrinsic stability of metals facilitates tracing and measurement in biologic material, although the clinical significance of the levels measured is not always clear.

Some metals, such as copper and selenium, are essential to normal metabolic function as trace elements **(Chap. 326)** but are toxic at high levels of exposure. Others, such as lead and mercury, are xenobiotic and theoretically are capable of exerting toxic effects at any level of exposure. Indeed, much research is currently focused on the contribution of low-level xenobiotic metal exposure to chronic diseases and to subtle changes in health that may have significant public health consequences. Genetic factors, such as polymorphisms that encode for variant enzymes with altered properties in terms of metal binding, transport, and effects, also may modify the impact of metals on health and thereby account, at least in part, for individual susceptibility to metal effects.

The most important component of treatment for metal toxicity is the termination of exposure. *Chelating agents* are used to bind metals into stable cyclic compounds with relatively low toxicity and to enhance their excretion. The principal chelating agents are dimercaprol (British anti-Lewisite [BAL]), ethylenediamine tetraacetic acid (EDTA), succimer (dimercaptosuccinic acid [DMSA]), and penicillamine; their specific use depends on the metal involved and the clinical circumstances. Activated charcoal does not bind metals and thus is of limited usefulness in cases of acute metal ingestion.

In addition to the information provided in Table 449-1, several other aspects of exposure, toxicity, or management are worthy of discussion with respect to the four most hazardous toxicants (arsenic, cadmium, lead, and mercury).

Arsenic, even at moderate levels of exposure, has been clearly linked with increased risks for cancer of the skin, bladder, renal pelvis, ureter, kidney, liver, and lung. These risks appear to be modified by smoking, folate and selenium status, genetic traits (such as ability to methylate arsenic), and other factors. Recent studies in community-based populations have generated strong evidence that arsenic exposure is a risk factor for increased coronary heart disease and stroke, lung function impairment, acute respiratory tract infections, respiratory symptoms, and non-malignant lung disease mortality. Evidence is also emerging that low-level arsenic may cause neurodevelopmental delays in children and diabetes.

Serious *cadmium* poisoning from the contamination of food and water by mining effluents in Japan contributed to the 1946 outbreak of "itai-itai" ("ouch-ouch") disease, so named because of cadmium-induced bone toxicity that led to painful bone fractures. Modest exposures from environmental contamination have been associated in some studies with a lower bone density, a higher incidence of fractures, and a faster decline in height in both men and women, effects that may be related to cadmium's calciuric effect on the kidney. There is some evidence for synergy between the adverse impacts of cadmium and lead on kidney function. Environmental exposures have also been linked to lower lung function (even after adjusting for smoking cigarettes, which contain cadmium) as well as increased risk of cardiovascular disease and mortality, stroke, and heart failure. Several studies have also raised concerns that cadmium may be carcinogenic and contribute to elevated risks of prostate, breast, and pancreatic cancer. Overall, this growing body of research indicates that cadmium exposure may be contributing significantly to morbidity and mortality rates in the general population.

Advances in our understanding of *lead* toxicity have recently benefited by the development of K x-ray fluorescence (KXRF) instruments for making safe in vivo measurements of lead levels in bone (which, in turn, reflect cumulative exposure over many years, as opposed to blood lead levels, which mostly reflect recent exposure). Higher bone lead levels measured by KXRF have been linked to increased risk of hypertension and accelerated declines in cognition in both men and women from an urban population. Upon reviewing these studies in conjunction with other epidemiologic and toxicologic studies, a recent federal expert panel concluded that the impact of lead exposure on hypertension and cognition in adults was causal. Prospective studies have also demonstrated that higher bone lead levels are a major risk factor for increased cardiovascular morbidity and mortality rates in both community-based and occupational-exposed populations. Lead exposure at community levels has also been recently associated with increased risks of hearing loss, Parkinson's disease, and amyotrophic lateral sclerosis. With respect to pregnancy-associated risks, high maternal bone lead levels were found to predict lower birth weight, head circumference, birth length, and neurodevelopmental performance in offspring by age 2 years. Offspring have also been shown to have higher blood pressures at age 7–14 years, an age range at which higher blood pressures are known to predict an elevated risk of developing hypertension. In a randomized trial, calcium supplementation (1200 mg daily) was found to significantly reduce the mobilization of lead from maternal bone into blood during pregnancy.

The toxicity of low-level organic *mercury* exposure (as manifested by neurobehavioral performance) is of increasing concern based on studies of the offspring of mothers who ingested mercury-contaminated fish. With respect to whether the consumption of fish by women during pregnancy is good or bad for offspring neurodevelopment, balancing the trade-offs of the beneficial effects of the omega-3-fatty acids (FAs) in fish versus the adverse effects of mercury contamination in fish has led to some confusion and inconsistency in public health recommendations. Overall, it would appear that it would be best for pregnant women to either limit fish consumption to those species known to be low in mercury contamination but high in omega-3-FAs (such as sardines or mackerel) or to avoid fish and obtain omega-3-FAs through supplements or other dietary sources. Accumulated evidence has not supported the contention that ethyl mercury, used as a preservative in multiuse vaccines administered in early childhood, has played a significant role in causing neurodevelopmental problems such as autism.

TABLE 449-1 Heavy Metals

MAIN SOURCES	METABOLISM	TOXICITY	DIAGNOSIS	TREATMENT
Arsenic				
Smelting and microelectronics industries; wood preservatives, pesticides, herbicides, fungicides; contaminant of deep-water wells; folk remedies; and coal; incineration of these products.	Organic arsenic (arsenobetaine, arsenocholine) is ingested in seafood and fish, but is nontoxic; inorganic arsenic is readily absorbed (lung and GI); sequesters in liver, spleen, kidneys, lungs, and GI tract; residues persist in skin, hair, and nails; biomethylation results in detoxification, but this process saturates.	Acute arsenic poisoning results in necrosis of intestinal mucosa with hemorrhagic gastroenteritis, fluid loss, hypotension, delayed cardiomyopathy, acute tubular necrosis, and hemolysis. Chronic arsenic exposure causes diabetes, vasospasm, peripheral vascular insufficiency and gangrene, peripheral neuropathy, and cancer of skin, lung, liver (angiosarcoma), bladder, and kidney. Lethal dose: 120–200 mg (adults); 2 mg/kg (children).	Nausea, vomiting, diarrhea, abdominal pain, delirium, coma, seizures; garlicky odor on breath; hyperkeratosis, hyperpigmentation, exfoliative dermatitis, and Mees' lines (transverse white striae of the fingernails); sensory and motor polyneuritis, distal weakness. Radiopaque sign on abdominal x-ray; ECG–QRS broadening, QT prolongation, ST depression, T-wave flattening; 24-h urinary arsenic >67 μmol/d or 50 μg/d; (no seafood × 24 h); if recent exposure, serum arsenic >0.9 μmol/L (7 μg/dL). High arsenic in hair or nails.	If acute ingestion, ipecac to induce vomiting, gastric lavage, activated charcoal with a cathartic. Supportive care in ICU. Dimercaprol 3–5 mg/kg IM q4h × 2 days; q6h × 1 day, then q12h × 10 days; alternative: oral succimer.
Cadmium				
Metal-plating, pigment, smelting, battery, and plastics industries; tobacco; incineration of these products; ingestion of food that concentrates cadmium (grains, cereals).	Absorbed through ingestion or inhalation; bound by metallothionein, filtered at the glomerulus, but reabsorbed by proximal tubules (thus, poorly excreted). Biologic half-life: 10–30 y. Binds cellular sulfhydryl groups, competes with zinc, calcium for binding sites. Concentrates in liver and kidneys.	Acute cadmium inhalation causes pneumonitis after 4–24 h; acute ingestion causes gastroenteritis. Chronic exposure causes anosmia, yellowing of teeth, emphysema, minor LFT elevations, microcytic hypochromic anemia unresponsive to iron therapy, proteinuria, increased urinary β_2-microglobulin, calciuria, leading to chronic renal failure, osteomalacia, and fractures. Possible risks of cardiovascular disease and cancer.	With inhalation: pleuritic chest pain, dyspnea, cyanosis, fever, tachycardia, nausea, noncardiogenic pulmonary edema. With ingestion: nausea, vomiting, cramps, diarrhea. Bone pain, fractures with osteomalacia. If recent exposure, serum cadmium >500 nmol/L (5 μg/dL). Urinary cadmium >100 nmol/L (10 μg/g creatinine) and/or urinary β_2-microglobulin >750 μg/g creatinine (but urinary β_2-microglobulin also increased in other renal diseases such as pyelonephritis).	There is no effective treatment for cadmium poisoning (chelation not useful; dimercaprol can exacerbate nephrotoxicity). Avoidance of further exposure, supportive therapy, vitamin D for osteomalacia.
Lead				
Manufacturing of auto batteries, lead crystal, ceramics, fishing weights, etc.; demolition or sanding of lead-painted houses, bridges; stained glass–making, plumbing, soldering; environmental exposure to paint chips, house dust (in homes built <1975), firing ranges (from bullet dust), food or water from improperly glazed ceramics, lead pipes; contaminated herbal remedies, candies; exposure to the combustion of leaded fuels.	Absorbed through ingestion or inhalation; organic lead (e.g., tetraethyl lead) absorbed dermally. In blood, 95–99% sequestered in RBCs—thus, must measure lead in whole blood (not serum). Distributed widely in soft tissue, with half-life ~30 days; 15% of dose sequestered in bone with half-life of >20 years. Excreted mostly in urine, but also appears in other fluids including breast milk. Interferes with mitochondrial oxidative phosphorylation, ATPases, calcium-dependent messengers; enhances oxidation and cell apoptosis.	Acute exposure with blood lead levels (BPb) of >60–80 μg/dL can cause impaired neurotransmission and neuronal cell death (with central and peripheral nervous system effects); impaired hematopoiesis and renal tubular dysfunction. At higher levels of exposure (e.g., BPb >80–120 μg/dL), acute encephalopathy with convulsions, coma, and death may occur. Subclinical exposures in children (BPb 25–60 μg/dL) are associated with anemia; mental retardation; and deficits in language, motor function, balance, hearing, behavior, and school performance. Impairment of IQ appears to occur at even lower levels of exposure with no measurable threshold above the limit of detection in most assays of 1 μg/dL. In adults, chronic subclinical exposures (BPb >40 μg/dL) are associated with an increased risk of anemia, demyelinating peripheral neuropathy (mainly motor), impairments of reaction time and hearing, accelerated declines in cognition, hypertension, ECG conduction delays, higher risk of cardiovascular disease and death, interstitial nephritis and chronic renal failure, diminished sperm counts, and spontaneous abortions.	Abdominal pain, irritability, lethargy, anorexia, anemia, Fanconi's syndrome, pyuria, azotemia in children with blood lead level (BPb) >80 μg/dL; may also see epiphyseal plate "lead lines" on long bone x-rays. Convulsions, coma at BPb >120 μg/dL. Noticeable neurodevelopmental delays at BPb of 40–80 μg/dL; may also see symptoms associated with higher BPb levels. Screening of all U.S. children when they begin to crawl (~6 months) is recommended by the CDC; source identification and intervention is begun if the BPb >10 μg/dL. In adults, acute exposure causes similar symptoms as in children as well as headaches, arthralgias, myalgias, depression, impaired short-term memory, loss of libido. Physical examination may reveal a "lead line" at the gingiva-tooth border, pallor, wrist drop, and cognitive dysfunction (e.g., declines on the mini-mental state exam); lab tests may reveal a normocytic, normochromic anemia, basophilic stippling, an elevated blood protoporphyrin level (free erythrocyte or zinc), and motor delays on nerve conduction. U.S. OSHA requires regular testing of lead-exposed workers with removal if BPb >40 μg/dL. New guidelines have been proposed recommending that BPb be maintained at <10 μg/dL, removal of workers if BPb >20 μg/dL, and monitoring of cumulative exposure parameters.	Identification and correction of exposure sources is critical. In some U.S. states, screening and reporting to local health boards of children with BPb >10 μg/dL and workers with BPb >40 μg/dL is required. In the highly exposed individual with symptoms, chelation is recommended with oral DMSA (succimer); if acutely toxic, hospitalization and IV or IM chelation with ethylenediaminetetraacetic acid calcium disodium (CaEDTA) may be required, with the addition of dimercaprol to prevent worsening of encephalopathy. It is uncertain whether children with asymptomatic lead exposure (e.g., BPb 20–40 μg/dL) benefit from chelation; a recent randomized trial showed no benefit. Correction of dietary deficiencies in iron, calcium, magnesium, and zinc will lower lead absorption and may also improve toxicity. Vitamin C is a weak but natural chelating agent. Calcium supplements (1200 mg at bedtime) have been shown to lower blood lead levels in pregnant women.

(*Continued*)

TABLE 449-1 Heavy Metals (Continued)

MAIN SOURCES	METABOLISM	TOXICITY	DIAGNOSIS	TREATMENT
Mercury				
Metallic, mercurous, and mercuric mercury (Hg, Hg^{+}, Hg^{2+}) exposures occur in some chemical, metal-processing, electrical-equipment, automotive industries; they are also in thermometers, dental amalgams, batteries. Mercury is dispersed by waste incineration. Environmental bacteria convert inorganic to organic mercury, which then bioconcentrates up the aquatic food chain to contaminate tuna, swordfish, and other pelagic fish.	Elemental mercury (Hg) is not well absorbed; however, it will volatilize into highly absorbable vapor. Inorganic mercury is absorbed through the gut and skin. Organic mercury is well absorbed through inhalation and ingestion. Elemental and organic mercury cross the blood-brain barrier and placenta. Mercury is excreted in urine and feces and has a half-life in blood of ~60 days; however, deposits will remain in the kidney and brain for years. Exposure to mercury stimulates the kidney to produce metallothionein, which provides some detoxification benefit. Mercury binds sulfhydryl groups and interferes with a wide variety of critical enzymatic processes.	Acute inhalation of Hg vapor causes pneumonitis and noncardiogenic pulmonary edema leading to death, CNS symptoms, and polyneuropathy. Chronic high exposure causes CNS toxicity (mercurial *erethism*; see Diagnosis); lower exposures impair renal function, motor speed, memory, coordination. Acute ingestion of inorganic mercury causes gastroenteritis, the nephritic syndrome, or acute renal failure, hypertension, tachycardia, and cardiovascular collapse, with death at a dose of 10–42 mg/kg. Ingestion of organic mercury causes gastroenteritis, arrhythmias, and lesions in the basal ganglia, gray matter, and cerebellum at doses >1.7 mg/kg. High exposure during pregnancy causes derangement of fetal neuronal migration resulting in severe mental retardation. Mild exposures during pregnancy (from fish consumption) are associated with declines in neurobehavioral performance in offspring. Dimethylmercury, a compound only found in research labs, is "supertoxic"—a few drops of exposure via skin absorption or inhaled vapor can cause severe cerebellar degeneration and death.	Chronic exposure to metallic mercury vapor produces a characteristic intention tremor and mercurial *erethism*: excitability, memory loss, insomnia, timidity, and delirium ("mad as a hatter"). On neurobehavioral tests: decreased motor speed, visual scanning, verbal and visual memory, visuomotor coordination. Children exposed to mercury in any form may develop *acrodynia* ("pink disease"): flushing, itching, swelling, tachycardia, hypertension, excessive salivation or perspiration, irritability, weakness, morbilliform rashes, desquamation of palms and soles. Toxicity from elemental or inorganic mercury exposure begins when blood levels >180 nmol/L (3.6 µg/dL) and urine levels >0.7 µmol/L (15 µg/dL). Exposures that ended years ago may result in a >20-µg increase in 24-h urine after a 2-g dose of succimer. Organic mercury exposure is best measured by levels in blood (if recent) or hair (if chronic); CNS toxicity in children may derive from fetal exposures associated with maternal hair Hg >30 nmol/g (6 µg/g).	Treat acute ingestion of mercuric salts with induced emesis or gastric lavage and polythiol resins (to bind mercury in the GI tract). Chelate with dimercaprol (up to 24 mg/kg per day IM in divided doses), DMSA (succimer), or penicillamine, with 5-day courses separated by several days of rest. If renal failure occurs, treat with peritoneal dialysis, hemodialysis, or extracorporeal regional complexing hemodialysis and succimer. Chronic inorganic mercury poisoning is best treated with *N*-acetyl penicillamine.

Abbreviations: ATPase, adenosine triphosphatase; BPb, blood lead; CDC, Centers for Disease Control and Prevention; CNS, central nervous system; DMSA, dimercaptosuccinic acid; ECG, electrocardiogram; GI, gastrointestinal; ICU, intensive care unit; IQ, intelligence quotient; LFT, liver function tests; OSHA, Occupational Safety and Health Administration; RBC, red blood cell.

With regard to adults, there is conflicting evidence as to whether mercury exposure is associated with increased risk of hypertension and cardiovascular disease. At this point, conclusions cannot be drawn. Mercury exposure may also be associated with perturbations in markers of autoimmunity. The clinical significance of these findings remains unclear.

Heavy metals pose risks to health that are especially burdensome in selected parts of the world. For example, *arsenic* exposure from natural contamination of shallow tube wells inserted for drinking water is a major environmental problem for millions of residents in parts of Bangladesh and Western India. Contamination was formerly considered only a problem with deep wells; however, the geology of this region allows most residents only a few alternatives for potable drinking water. The combustion of leaded gasoline with resulting contamination of air and soil with *lead oxide* remains a problem in some countries of Central Asia, Southeast Asia, Africa, and the Middle East. Populations living in the Arctic have been shown to have particularly high exposures to *mercury* due to long-range transport patterns that concentrate mercury in the polar regions, as well as the traditional dependence of Arctic peoples on the consumption of fish and other wildlife that bioconcentrate methylmercury.

A few additional metals deserve brief mention but are not covered in Table 449-1 because of the relative rarity of their being clinically encountered or the uncertainty regarding their potential toxicities. *Aluminum* contributes to the encephalopathy in patients with severe renal disease, who are undergoing dialysis **(Chap. 403)**. High levels of aluminum are found in the neurofibrillary tangles in the cerebral cortex and hippocampus of patients with Alzheimer's disease, as well as in the drinking water and soil of areas with an unusually high incidence of Alzheimer's. The experimental and epidemiologic evidence for the aluminum–Alzheimer's disease link remains relatively weak, however, and it cannot be concluded that aluminum is a causal agent or a contributing factor in neurodegenerative disease. Hexavalent chromium is corrosive and sensitizing. Workers in the chromate and chrome pigment production industries have consistently had a greater risk of lung cancer. The introduction of *cobalt* chloride as a fortifier in beer led to outbreaks of fatal cardiomyopathy among heavy consumers. Occupational exposure (e.g., of miners, dry-battery manufacturers, and arc welders) to *manganese* (Mn) can cause a parkinsonian syndrome within 1–2 years, including gait disorders; postural instability; a masked, expressionless face; tremor; and psychiatric symptoms. With the introduction of methylcyclopentadienyl manganese tricarbonyl (MMT) as a gasoline additive, there is concern for the toxic potential of environmental manganese exposure. For example, a recent study found a high prevalence of parkinsonian disorders in a community with risks proportionate to estimated manganese exposures emitted by local ferroalloy industries. Epidemiologic studies have also suggested that manganese may interfere with early childhood neurodevelopment in ways similar to that of lead. Manganese toxicity is clearly associated with dopaminergic dysfunction and its toxicity is likely influenced by age, gender, ethnicity, genetics, and preexisting medical conditions. *Nickel* exposure induces an allergic response, and inhalation of nickel compounds with low aqueous solubility (e.g., nickel subsulfide and nickel oxide) in occupational settings is associated with an increased risk of lung cancer. Overexposure to selenium may cause local irritation of the respiratory system and eyes, gastrointestinal irritation, liver inflammation, loss of hair, depigmentation, and peripheral nerve damage. Workers exposed to certain organic forms of *tin* (particularly

trimethyl and triethyl derivatives) have developed psychomotor disturbances, including tremor, convulsions, hallucinations, and psychotic behavior.

Thallium, which is a component of some insecticides, metal alloys, and fireworks, is absorbed through the skin as well as by ingestion and inhalation. Severe poisoning follows a single ingested dose of >1 g or >8 mg/kg. Nausea and vomiting, abdominal pain, and hematemesis precede confusion, psychosis, organic brain syndrome, and coma. Thallium is radiopaque. Induced emesis or gastric lavage is indicated within 4–6 h of acute ingestion; Prussian blue prevents absorption and is given orally at 250 mg/kg in divided doses. Unlike other types of metal poisoning, thallium poisoning may be less severe when activated charcoal is used to interrupt its enterohepatic circulation. Other measures include forced diuresis, treatment with potassium chloride (which promotes renal excretion of thallium), and peritoneal dialysis.

Chelation Therapy The Trial to Assess Chelation Therapy (TACT), a multi-center double-blind, placebo-controlled, prospective randomized trial funded by NIH of 1708 patients aged ≥50 years who had experienced a myocardial infarction (MI), recently found that a protocol of repeated intravenous chelation with disodium EDTA, compared with placebo, modestly but significantly reduced the risk of adverse cardiovascular outcomes, many of which were revascularization procedures. The effect was particularly pronounced among those with concurrent diabetes. However, the trial did not include rigorous measures of lead exposure or any selection criteria based on lead exposure; thus, even though chelation reduces metal burdens, which have been associated with adverse cardiovascular effects (especially lead), it remains unclear whether the beneficial effects result from a reduction in metal burden. In view of the risks of side effects associated with chelation, by themselves, the results are not sufficient to support the routine use of chelation therapy for treatment of patients either who have had an MI or who have had low-level lead exposure. A follow-up trial with rigorous measures of metals exposure is planned.

FURTHER READING

Alamolhodaei NS et al: Arsenic cardiotoxicity: An overview. Environ Toxicol Pharmacol 40:1005, 2015.

Aneni EC et al: Chronic toxic metal exposure and cardiovascular disease: Mechanisms of risk and emerging role of chelation therapy. Curr Atheroscler Rep 18:81, 2016.

Cheng X et al: Cadmium exposure and risk of any fracture: A PRISMA-compliant systematic review and meta-analysis. Medicine (Baltimore) 95:e2932, 2016.

Gidlow DA: Lead toxicity. Occup Med (Lond) 65:348, 2015.

Kim KH et al: A review on the distribution of Hg in the environment and its human health impacts. J Hazard Mater 306:376, 2016.

Lucchini RG, Hashim D: Tremor secondary to neurotoxic exposure: Mercury, lead, solvents, pesticides. Handb Clin Neurol 131:241, 2015.

O'Neal SL, Zheng W: Manganese toxicity upon overexposure: A decade in review. Curr Environ Health Rep 2:315, 2015.

450 Poisoning and Drug Overdose

Mark B. Mycyk

Poisoning refers to the development of dose-related adverse effects following exposure to chemicals, drugs, or other xenobiotics. To paraphrase Paracelsus, the dose makes the poison. Although most poisons have predictable dose-related effects, individual responses to a given dose may vary because of genetic polymorphism, enzymatic induction or inhibition in the presence of other xenobiotics, or acquired tolerance. Poisoning may be local (e.g., skin, eyes, or lungs) or systemic depending on the route of exposure, the chemical and physical properties of the poison, and its mechanism of action. The severity and reversibility of poisoning also depend on the functional reserve of the individual or target organ, which is influenced by age and preexisting disease.

EPIDEMIOLOGY

More than 5 million poison exposures occur in the United States each year. Most are acute, are accidental (unintentional), involve a single agent, occur in the home, result in minor or no toxicity, and involve children <6 years of age. Pharmaceuticals are involved in 47% of exposures and in 84% of serious or fatal poisonings. In the last decade, the rate of injury-related deaths from poisoning has overtaken the rate of deaths related to motor-vehicle crashes in the United States. According to the Centers for Disease Control (CDC), twice as many Americans died from drug overdoses in 2014 compared to 2000. Although prescription opioids have appropriately received attention as a major reason for the increased number of poisoning deaths, the availability of other pharmaceuticals and rapid proliferation of novel drugs of abuse also contribute to the increasing death rate. In many parts of the United States, where these issues are particularly prevalent, there are efforts to develop better prescription drug databases and enhanced training for health care professionals in pain management and the use of opiates. Unintentional exposures can result from the improper use of chemicals at work or play; label misreading; product mislabeling; mistaken identification of unlabeled chemicals; uninformed self-medication; and dosing errors by nurses, pharmacists, physicians, parents, and the elderly. Excluding the recreational use of ethanol, attempted suicide (deliberate self-harm) is the most common reported reason for intentional poisoning. Recreational use of prescribed and over-the-counter drugs for psychotropic or euphoric effects (*abuse*) or excessive self-dosing (*misuse*) is increasingly common and may also result in unintentional self-poisoning.

About 20–25% of exposures require bedside health-professional evaluation, and 5% of all exposures require hospitalization. Poisonings account for 5–10% of all ambulance transports, emergency department visits, and intensive care unit admissions. Hospital admissions related to poisoning are also associated with longer lengths of stay and increase the utilization of resources such as radiography and other laboratory services. Up to 30% of psychiatric admissions are prompted by attempted suicide via overdosage. Overall, the mortality rate is low: <1% of all poisoning exposures. It is significantly higher (1–2%) among hospitalized patients with intentional (suicidal) overdose or complications from drugs of abuse, who account for the majority of serious poisonings. Acetaminophen is the pharmaceutical agent most often implicated in fatal poisoning. Overall, carbon monoxide is the leading cause of death from poisoning, but this prominence is not reflected in hospital or poison center statistics because patients with such poisoning are typically dead when discovered and are referred directly to medical examiners.

DIAGNOSIS

Although poisoning can mimic other illnesses, the correct diagnosis can usually be established by the history, physical examination, routine and toxicologic laboratory evaluations, and characteristic clinical course.

HISTORY

The *history* should include the time, route, duration, and circumstances (location, surrounding events, and intent) of exposure; the name and amount of each drug, chemical, or ingredient involved; the time of onset, nature, and severity of symptoms; the time and type of first-aid measures provided; and the medical and psychiatric history.

In many cases the patient is confused, comatose, unaware of an exposure, or unable or unwilling to admit to one. Suspicious circumstances include unexplained sudden illness in a previously healthy person or a group of healthy people; a history of psychiatric problems (particularly depression); recent changes in health, economic status, or social relationships; and onset of illness during work with chemicals

or after ingestion of food, drink (especially ethanol), or medications. When patients become ill soon after arriving from a foreign country or being arrested for criminal activity, "body packing" or "body stuffing" (ingesting or concealing illicit drugs in a body cavity) should be suspected. Relevant information may be available from family, friends, paramedics, police, pharmacists, physicians, and employers, who should be questioned regarding the patient's habits, hobbies, behavioral changes, available medications, and antecedent events. Patients need to be asked explicitly about their prescribed medications and recreational drug use. Drugs previously considered "illicit" such as cannabinoids are now legal in many places and prescribed for therapeutic purposes. A search of clothes, belongings, and place of discovery may reveal a suicide note or a container of drugs or chemicals. Without a clear history in a patient clinically suspected to be poisoned, all medications available anywhere in the patient's home or belongings should be considered as possible agents, including medications for pets. The imprint code on pills and the label on chemical products may be used to identify the ingredients and potential toxicity of a suspected poison by consulting a reference text, a computerized database, the manufacturer, or a regional poison information center (800-222-1222). Occupational exposures require review of any available safety data sheet (SDS) from the worksite. Because of increasing globalization from travel and internet consumerism, unfamiliar poisonings may result in local emergency department evaluation. Pharmaceuticals, industrial chemicals, or drugs of abuse from foreign countries may be identified with the assistance of a regional poison center or via the World Wide Web.

PHYSICAL EXAMINATION AND CLINICAL COURSE

The *physical examination* should focus initially on vital signs, the cardiopulmonary system, and neurologic status. The neurologic examination should include documentation of neuromuscular abnormalities such as dyskinesia, dystonia, fasciculations, myoclonus, rigidity, and tremors. The patient should also be examined for evidence of trauma and underlying illnesses. Focal neurologic findings are uncommon in poisoning, and their presence should prompt evaluation for a structural central nervous system (CNS) lesion. Examination of the eyes (for nystagmus and pupil size and reactivity), abdomen (for bowel activity and bladder size), and skin (for burns, bullae, color, warmth, moisture, pressure sores, and puncture marks) may reveal findings of diagnostic value. When the history is unclear, all orifices should be examined for the presence of chemical burns and drug packets. The odor of breath or vomitus and the color of nails, skin, or urine may provide important diagnostic clues.

The diagnosis of poisoning in cases of unknown etiology primarily relies on pattern recognition. The first step is to assess the pulse, blood pressure, respiratory rate, temperature, and neurologic status and to characterize the overall physiologic state as stimulated, depressed, discordant, or normal **(Table 450-1)**. Obtaining a complete set of vital signs and reassessing them frequently are critical. Measuring core temperature is especially important, even in difficult or combative patients, since temperature elevation is the most reliable prognosticator of poor outcome in poisoning from stimulants (e.g., cocaine) or drug withdrawal (e.g., alcohol or GHB). The next step is to consider the underlying causes of the physiologic state and to attempt to identify a pathophysiologic pattern or toxic syndrome (*toxidrome*) based on the observed findings. Assessing the severity of physiologic derangements **(Table 450-2)** is useful in this regard and also for monitoring the clinical course and response to treatment. The final step is to attempt to identify the particular agent involved by looking for unique or relatively poison-specific physical or ancillary test abnormalities. Distinguishing among toxidromes on the basis of the physiologic state is summarized next.

The Stimulated Physiologic State Increased pulse, blood pressure, respiratory rate, temperature, and neuromuscular activity characterize the *stimulated* physiologic state, which can reflect sympathetic, anticholinergic, or hallucinogen poisoning or drug withdrawal (Table 450-1). Other features are noted in (Table 450-2). Mydriasis, a characteristic feature of all stimulants, is most marked in anticholinergic poisoning since pupillary reactivity relies on muscarinic control. In sympathetic poisoning (e.g., due to cocaine), pupils are also enlarged, but some reactivity to light remains. The anticholinergic toxidrome is also distinguished by hot, dry, flushed skin; decreased bowel sounds; and urinary retention. Other stimulant syndromes increase sympathetic activity and cause diaphoresis, pallor, and increased bowel activity with varying degrees of nausea, vomiting, abnormal distress, and occasionally diarrhea. The absolute and relative degree of vital-sign changes and neuromuscular hyperactivity can help distinguish among stimulant toxidromes. Since sympathetics stimulate the peripheral nervous system more directly than do hallucinogens or drug withdrawal, markedly increased vital signs and organ ischemia suggest sympathetic poisoning. Findings helpful in suggesting the particular drug or class causing physiologic stimulation include reflex bradycardia from selective α-adrenergic stimulants (e.g., decongestants), hypotension from selective β-adrenergic stimulants (e.g., asthma therapeutics), limb ischemia from ergot alkaloids, rotatory nystagmus from phencyclidine and ketamine (the only physiologic stimulants that cause this finding), and delayed cardiac conduction from high doses of cocaine and some anticholinergic agents (e.g., antihistamines, cyclic antidepressants, and antipsychotics). Seizures suggest a sympathetic etiology, an anticholinergic agent with membrane-active properties (e.g., cyclic antidepressants, phenothiazines), or a withdrawal syndrome. Close attention to core temperature is critical in patients with grade 4 physiologic stimulation (Table 450-2).

The Depressed Physiologic State Decreased pulse, blood pressure, respiratory rate, temperature, and neuromuscular activity are indicative of the *depressed* physiologic state caused by "functional" sympatholytics (agents that decrease cardiac function and vascular tone as well as sympathetic activity), cholinergic (muscarinic and nicotinic) agents, opioids, and sedative-hypnotic γ-aminobutyric acid (GABA)-ergic agents (Tables 450-1 and 450-2). Miosis is also common and is most pronounced in opioid and cholinergic poisoning. Miosis is distinguished from other depressant syndromes by muscarinic and nicotinic signs and symptoms (Table 450-1). Pronounced cardiovascular depression in the absence of significant CNS depression suggests a direct or peripherally acting sympatholytic. In contrast, in opioid and sedative-hypnotic poisoning, vital-sign changes are secondary to depression of CNS cardiovascular and respiratory centers (or consequent hypoxemia), and significant abnormalities in these parameters do not occur until there is a marked decrease in the level of consciousness (grade 3 or 4 physiologic depression; [Table 450-2]). Other clues that suggest the cause of physiologic depression include cardiac arrhythmias and conduction disturbances (due to antiarrhythmics, β-adrenergic antagonists, calcium channel blockers, digitalis glycosides, propoxyphene, and cyclic antidepressants), mydriasis (due to tricyclic antidepressants, some antiarrhythmics, meperidine, and diphenoxylate-atropine [Lomotil]), nystagmus (due to sedative-hypnotics), and seizures (due to cholinergic agents, propoxyphene, and cyclic antidepressants).

The Discordant Physiologic State The *discordant* physiologic state is characterized by mixed vital-sign and neuromuscular abnormalities, as observed in poisoning by asphyxiants, CNS syndromes, membrane-active agents, and anion-gap metabolic acidosis (AGMA) inducers (Table 450-1). In these conditions, manifestations of physiologic stimulation and physiologic depression occur together or at different times during the clinical course. For example, membrane-active agents can cause simultaneous coma, seizures, hypotension, and tachyarrhythmias. Alternatively, vital signs may be normal while the patient has an altered mental status or is obviously sick or clearly symptomatic. Early, pronounced vital-sign and mental-status changes suggest asphyxiant or membrane-active agent poisoning; the lack of such abnormalities suggests an AGMA inducer; and marked neuromuscular dysfunction without significant vital-sign abnormalities

TABLE 450-1 Differential Diagnosis of Poisoning Based on Physiologic State

STIMULATED	DEPRESSED	DISCORDANT	NORMAL
Sympathetics	Sympatholytics	Asphyxiants	Nontoxic exposure
Sympathomimetics	α_1-Adrenergic antagonists	Cytochrome oxidase inhibitors	Psychogenic illness
Ergot alkaloids	α_2-Adrenergic agonists	Inert gases	"Toxic time-bombs"
Methylxanthines	ACE inhibitors	Irritant gases	Slow absorption
Monoamine oxidase inhibitors	Angiotensin receptor blockers	Methemoglobin inducers	Anticholinergics
Thyroid hormones	Antipsychotics	Oxidative phosphorylation inhibitors	Carbamazepine
Anticholinergics	β-Adrenergic blockers	AGMA inducers	Concretion formers
Antihistamines	Calcium channel blockers	Alcohol (ketoacidosis)	Extended-release phenytoin sodium capsules (Dilantin Kapseals)
Antiparkinsonian agents	Cardiac glycosides	Ethylene glycol	Drug packets
Antipsychotics	Cyclic antidepressants	Iron	Enteric-coated pills
Antispasmodics	Cholinergics	Methanol	Diphenoxylate-atropine (Lomotil)
Belladonna alkaloids	Acetylcholinesterase inhibitors	Other alcohols	Opioids
Cyclic antidepressants	Muscarinic agonists	Salicylate	Salicylates
Mushrooms and plants	Nicotinic agonists	Toluene	Sustained-release pills
Hallucinogens	Opioids	CNS syndromes	Valproate
Cannabinoids (marijuana)	Analgesics	Extrapyramidal reactions	Slow distribution
LSD and analogues	GI antispasmodics	Hydrocarbon inhalation	Cardiac glycosides
Mescaline and analogues	Heroin	Isoniazid	Lithium
Mushrooms	Sedative-hypnotics	Lithium	Metals
Phencyclidine and analogues	Alcohols	Neuroleptic malignant syndrome	Salicylate
Withdrawal syndromes	Anticonvulsants	Serotonin syndrome	Valproate
Barbiturates	Barbiturates	Strychnine	Toxic metabolite
Benzodiazepines	Benzodiazepines	Membrane-active agents	Acetaminophen
Ethanol	GABA precursors	Amantadine	Carbon tetrachloride
GHB products	Muscle relaxants	Antiarrhythmics	Cyanogenic glycosides
Opioids	Other agents	Antihistamines	Ethylene glycol
Sedative-hypnotics	GHB products	Antipsychotics	Methanol
Sympatholytics		Carbamazepine	Methemoglobin inducers
		Cyclic antidepressants	Mushroom toxins
		Local anesthetics	Organophosphate insecticides
		Opioids (some)	Paraquat
		Quinoline antimalarials	Metabolism disruptors
			Antineoplastic agents
			Antiviral agents
			Colchicine
			Hypoglycemic agents
			Immunosuppressive agents
			MAO inhibitors
			Metals
			Other oral anticoagulants
			Salicylate
			Warfarin

Abbreviations: ACE, angiotensin-converting enzyme; AGMA, anion-gap metabolic acidosis; CNS, central nervous system; GABA, γ-aminobutyric acid; GHB, γ-hydroxybutyrate; GI, gastrointestinal; LSD, lysergic acid diethylamide; MAO, monoamine oxidase.

suggests a CNS syndrome. The *discordant* physiologic state may also be evident in patients poisoned with multiple agents.

The Normal Physiologic State A *normal* physiologic status and physical examination may be due to a nontoxic exposure, psychogenic illness, or poisoning by "toxic time-bombs": agents that are slowly absorbed, are slowly distributed to their sites of action, require metabolic activation, or disrupt metabolic processes (Table 450-1). Because so many medications have now been reformulated into a once-a-day preparations for the patient's convenience and adherence, toxic time-bombs are increasingly common. Diagnosing a nontoxic exposure requires that the identity of the exposure agent be known or that a toxic time-bomb exposure be excluded and the time since exposure exceed the longest known or predicted interval between exposure and peak toxicity. Psychogenic illness (fear of being poisoned, mass hysteria) may also follow a nontoxic exposure and should be considered when symptoms are inconsistent with exposure history. Anxiety reactions resulting from a nontoxic exposure can cause mild physiologic stimulation (Table 450-2) and be indistinguishable from toxicologic causes without ancillary testing or a suitable period of observation.

LABORATORY ASSESSMENT

Laboratory assessment may be helpful in the differential diagnosis. Increased anion gap metabolic acidosis (AGMA) is most common in advanced methanol, ethylene glycol, and salicylate intoxication but can occur with any poisoning that results in hepatic, renal, or respiratory failure; seizures; or shock. The serum lactate concentration is more commonly low (less than the anion gap) in the former and high (nearly equal to the anion gap) in the latter. An abnormally low anion gap can

TABLE 450-2 Severity of Physiologic Stimulation and Depression in Poisoning and Drug Withdrawal

Physiologic Stimulation	
Grade 1	Anxious, irritable, tremulous; vital signs normal; diaphoresis, flushing or pallor, mydriasis, and hyperreflexia sometimes present
Grade 2	Agitated; may have confusion or hallucinations but can converse and follow commands; vital signs mildly to moderately increased
Grade 3	Delirious; unintelligible speech, uncontrollable motor hyperactivity; moderately to markedly increased vital signs; tachyarrhythmias possible
Grade 4	Coma, seizures, cardiovascular collapse
Physiologic Depression	
Grade 1	Awake, lethargic, or sleeping but arousable by voice or tactile stimulation; able to converse and follow commands; may be confused
Grade 2	Responds to pain but not voice; can vocalize but not converse; spontaneous motor activity present; brainstem reflexes intact
Grade 3	Unresponsive to pain; spontaneous motor activity absent; brainstem reflexes depressed; motor tone, respirations, and temperature decreased
Grade 4	Unresponsive to pain; flaccid paralysis; brainstem reflexes and respirations absent; cardiovascular vital signs decreased

be due to elevated blood levels of bromide, calcium, iodine, lithium, or magnesium. An increased osmolal gap—a difference of >10 mmol/L between serum osmolality (measured by freezing-point depression) and osmolality calculated from serum sodium, glucose, and blood urea nitrogen levels—suggests the presence of a low-molecular-weight solute such as acetone; an alcohol (benzyl, ethanol, isopropanol, methanol); a glycol (diethylene, ethylene, propylene); ether (ethyl, glycol); or an "unmeasured" cation (calcium, magnesium) or sugar (glycerol, mannitol, sorbitol). Ketosis suggests acetone, isopropyl alcohol, salicylate poisoning, or alcoholic ketoacidosis. Hypoglycemia may be due to poisoning with β-adrenergic blockers, ethanol, insulin, oral hypoglycemic agents, quinine, and salicylates, whereas hyperglycemia can occur in poisoning with acetone, β-adrenergic agonists, caffeine, calcium channel blockers, iron, theophylline, or *N*-3-pyridylmethyl-*N*′-*p*-nitrophenylurea (PNU [Vacor]). Hypokalemia can be caused by barium, β-adrenergic agonists, caffeine, diuretics, theophylline, or toluene; hyperkalemia suggests poisoning with an α-adrenergic agonist, a β-adrenergic blocker, cardiac glycosides, or fluoride. Hypocalcemia may be seen in ethylene glycol, fluoride, and oxalate poisoning. PT and INR are useful for risk stratification in cases of warfarin or rodenticide poisoning, but are not to be relied on when evaluating overdose or complications from new anticoagulant pharmaceuticals (e.g., dabigatran).

The *electrocardiogram* (ECG) can be useful for rapid diagnostic purposes. Bradycardia and atrioventricular block may occur in patients poisoned by α-adrenergic agonists, antiarrhythmic agents, beta blockers, calcium channel blockers, cholinergic agents (carbamate and organophosphate insecticides), cardiac glycosides, lithium, or tricyclic antidepressants. QRS- and QT-interval prolongation may be caused by hyperkalemia, various antidepressants, and other membrane-active drugs (Table 450-1). Ventricular tachyarrhythmias may be seen in poisoning with cardiac glycosides, fluorides, membrane-active drugs, methylxanthines, sympathomimetics, antidepressants, and agents that cause hyperkalemia or potentiate the effects of endogenous catecholamines (e.g., chloral hydrate, aliphatic and halogenated hydrocarbons).

Radiologic studies may occasionally be useful. Pulmonary edema (adult respiratory distress syndrome [ARDS]) can be caused by poisoning with carbon monoxide, cyanide, an opioid, paraquat, phencyclidine, a sedative-hypnotic, or salicylate; by inhalation of irritant gases, fumes, or vapors (acids and alkali, ammonia, aldehydes, chlorine, hydrogen sulfide, isocyanates, metal oxides, mercury, phosgene, polymers); or by prolonged anoxia, hyperthermia, or shock. Aspiration pneumonia is common in patients with coma, seizures, and petroleum distillate aspiration. Chest x-ray is useful for identifying complications from metal fume fever or elemental mercury. The presence of radiopaque densities on abdominal x-rays or abdominal CT scan suggests the ingestion of calcium salts, chloral hydrate, chlorinated hydrocarbons, heavy metals, illicit drug packets, iodinated compounds, potassium salts, enteric-coated tablets, or salicylates.

Toxicologic analysis of urine and blood (and occasionally of gastric contents and chemical samples) can sometimes confirm or rule out suspected poisoning. Interpretation of laboratory data requires knowledge of the qualitative and quantitative tests used for screening and confirmation (enzyme-multiplied, fluorescence polarization, and radio-immunoassays; colorimetric and fluorometric assays; thin-layer, gas-liquid, or high-performance liquid chromatography; gas chromatography; mass spectrometry), their sensitivity (limit of detection) and specificity, the preferred biologic specimen for analysis, and the optimal time of specimen sampling. Personal communication with the hospital laboratory is essential to an understanding of institutional testing capabilities and limitations.

Rapid *qualitative* hospital-based urine tests for drugs of abuse are only screening tests that cannot confirm the exact identity of the detected substance and should not be considered diagnostic or used for forensic purposes: False-positive and false-negative results are common. A positive screen may result from other pharmaceuticals that interfere with laboratory analysis (e.g., fluoroquinolones commonly cause "false-positive" opiate screens). Confirmatory testing with gas chromatography/mass spectrometry can be requested, but it often takes weeks to obtain a reported result. A negative screening result may mean that the responsible substance is not detectable by the test used or that its concentration is too low for detection at the time of sampling. For instance, recent new drugs of abuse that often result in emergency department evaluation for unexpected complications, such as synthetic cannabinoids (spice), cathinones (bath salts), and opiate substitutes (kratom), are not detectable by hospital-based tests. In cases where a drug concentration is too low to be detected early during clinical evaluation, repeating the test at a later time may yield a positive result. Patients symptomatic from drugs of abuse often require immediate management based on the history, physical examination, and observed toxidrome without laboratory confirmation (e.g., apnea from opioid intoxication). When the patient is asymptomatic or when the clinical picture is consistent with the reported history, qualitative screening is neither clinically useful nor cost-effective. Thus, qualitative drug screens are of greatest value for the evaluation of patients with severe or unexplained toxicities, such as coma, seizures, cardiovascular instability, metabolic or respiratory acidosis, and nonsinus cardiac rhythms. In contrast to qualitative drug screens, *quantitative* serum tests are useful for evaluation of patients poisoned with acetaminophen **(Chap. 333)**, alcohols (including ethylene glycol and methanol), anticonvulsants, barbiturates, digoxin, heavy metals, iron, lithium, salicylate, and theophylline as well as for the presence of carboxyhemoglobin and methemoglobin. The serum concentration in these cases guides clinical management, and results are often available within an hour.

The *response to antidotes* is sometimes useful for diagnostic purposes. Resolution of altered mental status and abnormal vital signs within minutes of IV administration of dextrose, naloxone, or flumazenil is virtually diagnostic of hypoglycemia, opioid poisoning, and benzodiazepine intoxication, respectively. The prompt reversal of dystonic (extrapyramidal) signs and symptoms following an IV dose of benztropine or diphenhydramine confirms a drug etiology. Although complete reversal of both central and peripheral manifestations of anticholinergic poisoning by physostigmine is diagnostic of this condition, physostigmine may cause some arousal in patients with CNS depression of any etiology.

TREATMENT

Poisoning and Drug Overdose

GENERAL PRINCIPLES

Treatment goals include support of vital signs, prevention of further poison absorption (decontamination), enhancement of poison elimination, administration of specific antidotes, and prevention of reexposure **(Table 450-3)**. Specific treatment depends on the identity of the poison, the route and amount of exposure, the time of presentation relative to the time of exposure, and the severity of poisoning. Knowledge of the offending agents' pharmacokinetics and pharmacodynamics is essential.

During the *pretoxic phase*, prior to the onset of poisoning, decontamination is the highest priority, and treatment is based solely on the history. The maximal potential toxicity based on the greatest possible exposure should be assumed. Since decontamination is more effective when accomplished soon after exposure and when the patient is asymptomatic, the initial history and physical examination should be focused and brief. It is also advisable to establish IV access and initiate cardiac monitoring, particularly in patients with potentially serious ingestions or unclear histories.

When an accurate history is not obtainable and a poison causing delayed toxicity (i.e., a toxic time-bomb) or irreversible damage is suspected, blood and urine should be sent for appropriate toxicologic screening and quantitative analysis. During poison absorption and distribution, blood levels may be greater than those in tissue and may not correlate with toxicity. However, high blood levels of agents whose metabolites are more toxic than the parent compound (acetaminophen, ethylene glycol, or methanol) may indicate the need for additional interventions (antidotes, dialysis). Most patients who remain asymptomatic or who become asymptomatic 6 h after ingestion are unlikely to develop subsequent toxicity and can be discharged safely. Longer observation will be necessary for patients who have ingested toxic time-bombs.

TABLE 450-3 Fundamentals of Poisoning Management

Supportive Care	
Airway protection	Treatment of seizures
Oxygenation/ventilation	Correction of temperature abnormalities
Treatment of arrhythmias	Correction of metabolic derangements
Hemodynamic support	Prevention of secondary complications
Prevention of Further Poison Absorption	
Gastrointestinal decontamination	Decontamination of other sites
Gastric lavage	Eye decontamination
Activated charcoal	Skin decontamination
Whole-bowel irrigation	Body cavity evacuation
Dilution	
Endoscopic/surgical removal	
Enhancement of Poison Elimination	
Multiple-dose activated charcoal administration	Extracorporeal removal
Alteration of urinary pH	Hemodialysis
Chelation	Hemoperfusion
	Hemofiltration
	Plasmapheresis
	Exchange transfusion
	Hyperbaric oxygenation
Administration of Antidotes	
Neutralization by antibodies	Metabolic antagonism
Neutralization by chemical binding	Physiologic antagonism
Prevention of Reexposure	
Adult education	Notification of regulatory agencies
Child-proofing	Psychiatric referral

During the *toxic phase*—the interval between the onset of poisoning and its peak effects—management is based primarily on clinical and laboratory findings. *Effects after an overdose usually begin sooner, peak later, and last longer than they do after a therapeutic dose*. A drug's published pharmacokinetic profile in standard references such as the *Physician's Desk Reference* (PDR) is usually different from its toxicokinetic profile in overdose. Resuscitation and stabilization are the first priority. Symptomatic patients should have an IV line placed and should undergo oxygen saturation determination, cardiac monitoring, and continuous observation. Baseline laboratory, ECG, and x-ray evaluation may also be appropriate. Intravenous glucose (unless the serum level is documented to be normal), naloxone, and thiamine should be considered in patients with altered mental status, particularly those with coma or seizures. Decontamination should also be considered, but it is less likely to be effective during this phase than during the pretoxic phase.

Measures that enhance poison elimination may shorten the duration and severity of the toxic phase. However, they are not without risk, which must be weighed against the potential benefit. Diagnostic certainty (usually via laboratory confirmation) is generally a prerequisite. Intestinal (gut) dialysis with repetitive doses of activated charcoal (see "Multiple-Dose Activated Charcoal," later) can enhance the elimination of selected poisons such as theophylline or carbamazepine. Urinary alkalinization may enhance the elimination of salicylates and a few other poisons. Chelation therapy can enhance the elimination of selected metals. Extracorporeal elimination methods are effective for many poisons, but their expense and risk make their use reasonable only in patients who would otherwise have an unfavorable outcome.

During the *resolution phase* of poisoning, supportive care and monitoring should continue until clinical, laboratory, and ECG abnormalities have resolved. Since chemicals are eliminated sooner from the blood than from tissues, blood levels are usually lower than tissue levels during this phase and again may not correlate with toxicity. This discrepancy applies particularly when extracorporeal elimination procedures are used. Redistribution from tissues may cause a rebound increase in the blood level after termination of these procedures (e.g., lithium). When a metabolite is responsible for toxic effects, continued treatment may be necessary in the absence of clinical toxicity or abnormal laboratory studies.

SUPPORTIVE CARE

The goal of supportive therapy is to maintain physiologic homeostasis until detoxification is accomplished and to prevent and treat secondary complications such as aspiration, bedsores, cerebral and pulmonary edema, pneumonia, rhabdomyolysis, renal failure, sepsis, thromboembolic disease, coagulopathy, and generalized organ dysfunction due to hypoxemia or shock.

Admission to an intensive care unit is indicated for the following: patients with severe poisoning (coma, respiratory depression, hypotension, cardiac conduction abnormalities, cardiac arrhythmias, hypothermia or hyperthermia, seizures); those needing close monitoring, antidotes, or enhanced elimination therapy; those showing progressive clinical deterioration; and those with significant underlying medical problems. Patients with mild to moderate toxicity can be managed on a general medical service, on an intermediate care unit, or in an emergency department observation area, depending on the anticipated duration and level of monitoring needed (intermittent clinical observation versus continuous clinical, cardiac, and respiratory monitoring). Patients who have attempted suicide require continuous observation and measures to prevent self-injury until they are no longer suicidal.

Respiratory Care Endotracheal intubation for protection against the aspiration of gastrointestinal contents is of paramount importance in patients with CNS depression or seizures as this complication can

increase morbidity and mortality rates. Mechanical ventilation may be necessary for patients with respiratory depression or hypoxemia and for facilitation of therapeutic sedation or paralysis of patients in order to prevent or treat hyperthermia, acidosis, and rhabdomyolysis associated with neuromuscular hyperactivity. Since clinical assessment of respiratory function can be inaccurate, the need for oxygenation and ventilation is best determined by continuous pulse oximetry or arterial blood-gas analysis. The gag reflex is not a reliable indicator of the need for intubation. A patient with CNS depression may maintain airway patency while being stimulated but not if left alone. Drug-induced pulmonary edema is usually noncardiac rather than cardiac in origin, although profound CNS depression and cardiac conduction abnormalities suggest the latter. Measurement of pulmonary artery pressure may be necessary to establish the cause and direct appropriate therapy. Extracorporeal measures (membrane oxygenation, ECMO, venoarterial perfusion, cardiopulmonary bypass) and partial liquid (perfluorocarbon) ventilation may be appropriate for severe but reversible respiratory failure.

Cardiovascular Therapy Maintenance of normal tissue perfusion is critical for complete recovery to occur once the offending agent has been eliminated. Focused bedside echocardiography or measurement of CVP may help prioritize therapeutic strategies. If hypotension is unresponsive to volume expansion and appropriate goal-directed antidotal therapy, treatment with norepinephrine, epinephrine, or high-dose dopamine may be necessary. Intraaortic balloon pump counterpulsation and venoarterial or cardiopulmonary perfusion techniques should be considered for severe but reversible cardiac failure. For patients with a return of spontaneous circulation after resuscitative treatment for cardiopulmonary arrest secondary to poisoning, therapeutic hypothermia should be used according to protocol. Bradyarrhythmias associated with hypotension generally should be treated as described in **Chaps. 239 and 240**. Glucagon, calcium, and high-dose insulin with dextrose may be effective in beta blocker and calcium channel blocker poisoning. Antibody therapy may be indicated for cardiac glycoside poisoning.

Supraventricular tachycardia associated with hypertension and CNS excitation is almost always due to agents that cause generalized physiologic excitation (Table 450–1). Most cases are mild or moderate in severity and require only observation or nonspecific sedation with a benzodiazepine. In severe cases or those associated with hemodynamic instability, chest pain, or ECG evidence of ischemia, specific therapy is indicated. When the etiology is sympathetic hyperactivity, treatment with a benzodiazepine should be prioritized. Further treatment with a combined alpha and beta blocker (labetalol), a calcium channel blocker (verapamil or diltiazem), or a combination of a beta blocker and a vasodilator (esmolol and nitroprusside) may be considered for cases refractory to high doses of benzodiazepines only when adequate sedation has been achieved but cardiac conduction or blood pressure abnormalities persist. Treatment with an α-adrenergic antagonist (phentolamine) alone may sometimes be appropriate. If the cause is anticholinergic poisoning, physostigmine alone can be effective. Supraventricular tachycardia without hypertension is generally secondary to vasodilation or hypovolemia and responds to fluid administration.

For ventricular tachyarrhythmias due to tricyclic antidepressants and other membrane-active agents (Table 450-1), sodium bicarbonate is indicated, whereas class IA, IC, and III antiarrhythmic agents are contraindicated because of similar electrophysiologic effects. Although lidocaine and phenytoin are historically safe for ventricular tachyarrhythmias of any etiology, sodium bicarbonate should be considered first for any ventricular arrhythmia suspected to have a toxicologic etiology. Intravenous lipid emulsion therapy has shown benefit for treatment of arrhythmias and hemodynamic instability from various membrane-active agents. Beta blockers can be hazardous if the arrhythmia is due to sympathetic hyperactivity. Magnesium sulfate and overdrive pacing (by isoproterenol or a pacemaker) may be useful in patients with torsades des pointes and prolonged QT intervals. Magnesium and anti-digoxin antibodies should be considered in patients with severe cardiac glycoside poisoning. Invasive (esophageal or intracardiac) ECG recording may be necessary to determine the origin (ventricular or supraventricular) of wide-complex tachycardias **(Chap. 241)**. If the patient is hemodynamically stable, however, it is reasonable to simply observe him or her rather than to administer another potentially proarrhythmic agent. Arrhythmias may be resistant to drug therapy until underlying acid-base, electrolyte, oxygenation, and temperature derangements are corrected.

Central Nervous System Therapies Neuromuscular hyperactivity and seizures can lead to hyperthermia, lactic acidosis, and rhabdomyolysis and should be treated aggressively. Seizures caused by excessive stimulation of catecholamine receptors (sympathomimetic or hallucinogen poisoning and drug withdrawal) or decreased activity of GABA (isoniazid poisoning) or glycine (strychnine poisoning) receptors are best treated with agents that enhance GABA activity, such as benzodiazepine or barbiturates. Since benzodiazepines and barbiturates act by slightly different mechanisms (the former increases the frequency via allosteric modulation at the receptor and the latter directly increases the duration of chloride channel opening in response to GABA), therapy with both may be effective when neither is effective alone. Seizures caused by isoniazid, which inhibits the synthesis of GABA at several steps by interfering with the cofactor pyridoxine (vitamin B_6), may require high doses of supplemental pyridoxine. Seizures resulting from membrane destabilization (beta blocker or cyclic antidepressant poisoning) require GABA enhancers (benzodiazepines first, barbiturates second). Phenytoin is contraindicated in toxicologic seizures: Animal and human data demonstrate worse outcomes after phenytoin loading, especially in theophylline overdose. For poisons with central dopaminergic effects (methamphetamine, phencyclidine) manifested by psychotic behavior, a dopamine receptor antagonist, such as haloperidol or ziprasidone, may be useful. In anticholinergic and cyanide poisoning, specific antidotal therapy may be necessary. The treatment of seizures secondary to cerebral ischemia or edema or to metabolic abnormalities should include correction of the underlying cause. Neuromuscular paralysis is indicated in refractory cases. Electroencephalographic monitoring and continuing treatment of seizures are necessary to prevent permanent neurologic damage. Serotonergic receptor overstimulation in serotonin syndrome may be treated with cyproheptadine.

Other Measures Temperature extremes, metabolic abnormalities, hepatic and renal dysfunction, and secondary complications should be treated by standard therapies.

PREVENTION OF POISON ABSORPTION

Gastrointestinal Decontamination Whether or not to perform gastrointestinal decontamination and which procedure to use depends on the time since ingestion; the existing and predicted toxicity of the ingestant; the availability, efficacy, and contraindications of the procedure; and the nature, severity, and risk of complications. The efficacy of all decontamination procedures decreases with time, and data are insufficient to support or exclude a beneficial effect when they are used >1 h after ingestion. The average time from ingestion to presentation for treatment is >1 h for children and >3 h for adults. Most patients will recover from poisoning uneventfully with good supportive care alone, but complications of gastrointestinal decontamination, particularly aspiration, can prolong this process. Hence, gastrointestinal decontamination should be performed selectively, not routinely, in the management of overdose patients. It is clearly unnecessary when predicted toxicity is minimal or the time of expected maximal toxicity has passed without significant effect.

Activated charcoal has comparable or greater efficacy; has fewer contraindications and complications; and is less aversive and invasive than ipecac or gastric lavage. Thus it is the preferred method of gastrointestinal decontamination in most situations. Activated charcoal suspension (in water) is given orally via a cup, straw, or small-bore nasogastric tube. The generally recommended dose is 1 g/kg body weight because of its dosing convenience, although in vitro and in vivo studies have demonstrated that charcoal adsorbs ≥90% of most substances when given in an amount equal to 10 times the weight of the substance. Palatability may be increased by adding a sweetener (sorbitol) or a flavoring agent (cherry, chocolate, or cola syrup) to the suspension. Charcoal adsorbs ingested poisons within the gut lumen, allowing the charcoal-toxin complex to be evacuated with stool. Charged (ionized) chemicals such as mineral acids, alkalis, and highly dissociated salts of cyanide, fluoride, iron, lithium, and other inorganic compounds are not well adsorbed by charcoal. In studies with animals and human volunteers, charcoal decreases the absorption of ingestants by an average of 73% when given within 5 min of ingestant administration, 51% when given at 30 min, and 36% when given at 60 min. For this reason, charcoal given before hospital arrival increases the potential clinical benefit. Side effects of charcoal include nausea, vomiting, and diarrhea or constipation. Charcoal may also prevent the absorption of orally administered therapeutic agents. Complications include mechanical obstruction of the airway, aspiration, vomiting, and bowel obstruction and infarction caused by inspissated charcoal. Charcoal is not recommended for patients who have ingested corrosives because it obscures endoscopy.

Gastric lavage should be considered for life-threatening poisons that cannot be treated effectively with other decontamination, elimination, or antidotal therapies (e.g., colchicine). Gastric lavage is performed by sequentially administering and aspirating ~5 mL of fluid per kilogram of body weight through a no. 40 French orogastric tube (no. 28 French tube for children). Except in infants, for whom normal saline is recommended, tap water is acceptable. The patient should be placed in Trendelenburg and left lateral decubitus positions to prevent aspiration (even if an endotracheal tube is in place). Lavage decreases ingestant absorption by an average of 52% if performed within 5 min of ingestion administration, 26% if performed at 30 min, and 16% if performed at 60 min. Significant amounts of ingested drug are recovered from <10% of patients. Aspiration is a common complication (occurring in up to 10% of patients), especially when lavage is performed improperly. Serious complications (esophageal and gastric perforation, tube misplacement in the trachea) occur in ~1% of patients. For this reason, the physician should personally insert the lavage tube and confirm its placement, and the patient must be cooperative during the procedure. Gastric lavage is contraindicated in corrosive or petroleum distillate ingestions because of the respective risks of gastroesophageal perforation and aspiration pneumonitis. It is also contraindicated in patients with a compromised unprotected airway and those at risk for hemorrhage or perforation due to esophageal or gastric pathology or recent surgery. Finally, gastric lavage is absolutely contraindicated in combative patients or those who refuse, as most published complications involve patient resistance to the procedure.

Syrup of ipecac, an emetogenic agent that was once the substance most commonly used for decontamination, no longer has a role in poisoning management. Even the American Academy of Pediatrics—traditionally the strongest proponent of ipecac—issued a policy statement in 2003 recommending that ipecac should no longer be used in poisoning treatment. Chronic ipecac use (by patients with anorexia nervosa or bulimia) has been reported to cause electrolyte and fluid abnormalities, cardiac toxicity, and myopathy.

Whole-bowel irrigation is performed by administering a bowel-cleansing solution containing electrolytes and polyethylene glycol (Golytely, Colyte) orally or by gastric tube at a rate of 2 L/h (0.5 L/h in children) until rectal effluent is clear. The patient must be in a sitting position. Although data are limited, whole-bowel irrigation appears to be as effective as other decontamination procedures in volunteer studies. It is most appropriate for those who have ingested foreign bodies, packets of illicit drugs, and agents that are poorly adsorbed by charcoal (e.g., heavy metals). This procedure is contraindicated in patients with bowel obstruction, ileus, hemodynamic instability, and compromised unprotected airways.

Cathartics are salts (disodium phosphate, magnesium citrate and sulfate, sodium sulfate) or saccharides (mannitol, sorbitol) that historically have been given with activated charcoal to promote the rectal evacuation of gastrointestinal contents. However, no animal, volunteer, or clinical data have ever demonstrated any decontamination benefit from cathartics. Abdominal cramps, nausea, and occasional vomiting are side effects. Complications of repeated dosing include severe electrolyte disturbances and excessive diarrhea. Cathartics are contraindicated in patients who have ingested corrosives and in those with preexisting diarrhea. Magnesium-containing cathartics should not be used in patients with renal failure.

Dilution (i.e., drinking water, another clear liquid, or milk at a volume of 5 mL/kg of body weight) is recommended only after the ingestion of corrosives (acids, alkali). It may increase the dissolution rate (and hence absorption) of capsules, tablets, and other solid ingestants and should *not* be used in these circumstances.

Endoscopic or surgical removal of poisons may be useful in rare situations, such as ingestion of a potentially toxic foreign body that fails to transit the gastrointestinal tract, a potentially lethal amount of a heavy metal (arsenic, iron, mercury, thallium), or agents that have coalesced into gastric concretions or bezoars (heavy metals, lithium, salicylates, sustained-release preparations). Patients who become toxic from cocaine due to its leakage from ingested drug packets require immediate surgical intervention.

Decontamination of Other Sites Immediate, copious flushing with water, saline, or another available clear, drinkable liquid is the initial treatment for topical exposures (exceptions include alkali metals, calcium oxide, phosphorus). Saline is preferred for eye irrigation. A triple wash (water, soap, water) may be best for dermal decontamination. Inhalational exposures should be treated initially with fresh air or supplemental oxygen. The removal of liquids from body cavities such as the vagina or rectum is best accomplished by irrigation. Solids (drug packets, pills) should be removed manually, preferably under direct visualization.

ENHANCEMENT OF POISON ELIMINATION

Although the elimination of most poisons can be accelerated by therapeutic interventions, the pharmacokinetic efficacy (removal of drug at a rate greater than that accomplished by intrinsic elimination) and clinical benefit (shortened duration of toxicity or improved outcome) of such interventions are often more theoretical than proven. Accordingly, the decision to use such measures should be based on the actual or predicted toxicity and the potential efficacy, cost, and risks of therapy.

Multiple-Dose Activated Charcoal Repetitive oral dosing with charcoal can enhance the elimination of previously absorbed substances by binding them within the gut as they are excreted in the bile, are secreted by gastrointestinal cells, or passively diffuse into the gut lumen (*reverse absorption* or *enterocapillary exsorption*). Doses of 0.5–1 g/kg of body weight every 2–4 h, adjusted downward to avoid regurgitation in patients with decreased gastrointestinal motility, are generally recommended. Pharmacokinetic efficacy approaches that of hemodialysis for some agents (e.g., phenobarbital, theophylline). Multiple-dose therapy should be considered only for selected agents (theophylline, phenobarbital, carbamazepine, dapsone, quinine). Complications include intestinal obstruction, pseudo-obstruction, and nonocclusive intestinal infarction in patients with decreased gut motility. Because of electrolyte and fluid shifts, sorbitol and other cathartics are absolutely contraindicated when multiple doses of activated charcoal are administered.

Urinary Alkalinization Ion trapping via alteration of urine pH may prevent the renal reabsorption of poisons that undergo excretion by glomerular filtration and active tubular secretion. Since membranes are more permeable to non-ionized molecules than to their ionized counterparts, acidic (low-pK_a) poisons are ionized and trapped in alkaline urine, whereas basic ones become ionized and trapped in acid urine. Urinary alkalinization (producing a urine pH ≥7.5 and a urine output of 3–6 mL/kg of body weight per hour by the addition of sodium bicarbonate to an IV solution) enhances the excretion of chlorophenoxyacetic acid herbicides, chlorpropamide, diflunisal, fluoride, methotrexate, phenobarbital, sulfonamides, and salicylates. Contraindications include congestive heart failure, renal failure, and cerebral edema. Acid-base, fluid, and electrolyte parameters should be monitored carefully. Although acid diuresis may make theoretical sense for some overdoses (amphetamines), it is *never* indicated and is potentially harmful.

Extracorporeal Removal Hemodialysis, charcoal or resin hemoperfusion, hemofiltration, plasmapheresis, and exchange transfusion are capable of removing any toxin from the bloodstream. Agents most amenable to enhanced elimination by dialysis have low molecular mass (<500 Da), high water solubility, low protein binding, small volumes of distribution (<1 L/kg of body weight), prolonged elimination (long half-life), and high dialysis clearance relative to total-body clearance. Molecular weight, water solubility, and protein binding do not limit the efficacy of the other forms of extracorporeal removal.

Dialysis should be considered in cases of severe poisoning due to carbamazepine, ethylene glycol, isopropyl alcohol, lithium, methanol, theophylline, salicylates, and valproate. Although hemoperfusion may be more effective in removing some of these poisons, it does not correct associated acid-base and electrolyte abnormalities, and most hospitals no longer have hemoperfusion cartridges readily available. Fortunately, recent advances in hemodialysis technology make it as effective as hemoperfusion for removing poisons such as caffeine, carbamazepine, and theophylline. Both techniques require central venous access and systemic anticoagulation and may result in transient hypotension. Hemoperfusion may also cause hemolysis, hypocalcemia, and thrombocytopenia. Peritoneal dialysis and exchange transfusion are less effective but may be used when other procedures are unavailable, contraindicated, or technically difficult (e.g., in infants). Exchange transfusion may be indicated in the treatment of severe arsine- or sodium chlorate–induced hemolysis, methemoglobinemia, and sulfhemoglobinemia. Although hemofiltration can enhance elimination of aminoglycosides, vancomycin, and metal-chelate complexes, the roles of hemofiltration and plasmapheresis in the treatment of poisoning are not yet defined.

Candidates for extracorporeal removal therapies include patients with severe toxicity whose condition deteriorates despite aggressive supportive therapy; those with potentially prolonged, irreversible, or fatal toxicity; those with dangerous blood levels of toxins; those who lack the capacity for self-detoxification because of liver or renal failure; and those with a serious underlying illness or complication that will adversely affect recovery.

Other Techniques The elimination of heavy metals can be enhanced by chelation, and the removal of carbon monoxide can be accelerated by hyperbaric oxygenation.

ADMINISTRATION OF ANTIDOTES

Antidotes counteract the effects of poisons by neutralizing them (e.g., antibody-antigen reactions, chelation, chemical binding) or by antagonizing their physiologic effects (e.g., activation of opposing nervous system activity, provision of a competitive metabolic or receptor substrate). Poisons or conditions with specific antidotes include acetaminophen, anticholinergic agents, anticoagulants, benzodiazepines, beta blockers, calcium channel blockers, carbon monoxide, cardiac glycosides, cholinergic agents, cyanide, drug-induced dystonic reactions, ethylene glycol, fluoride, heavy metals, hypoglycemic agents, isoniazid, membrane-active agents, methemoglobinemia, opioids, sympathomimetics, and a variety of envenomations. Intravenous lipid emulsion has been shown to be a successful antidote for poisoning from various anesthetics and membrane-active agents (e.g., cyclic antidepressants), but the exact mechanism of benefit is still under investigation. Antidotes can significantly reduce morbidity and mortality rates but are potentially toxic if used for inappropriate reasons. Since their safe use requires correct identification of a specific poisoning or syndrome, details of antidotal therapy are discussed with the conditions for which they are indicated **(Table 450-4)**.

PREVENTION OF REEXPOSURE

Poisoning is a preventable illness. Unfortunately, some adults and children are poison-prone, and recurrences are common. Unintentional polypharmacy poisoning has become especially common among adults with developmental delays, among the growing population of geriatric patients who are prescribed a large number of medications, and among adolescents and young adults experimenting with pharmaceuticals for recreational euphoria. Adults with unintentional exposures should be instructed regarding the safe use of medications and chemicals (according to labeling instructions). Confused patients may need assistance with the administration of their medications. Errors in dosing by health care providers may require educational efforts. Patients should be advised to avoid circumstances that result in chemical exposure or poisoning. Appropriate agencies and health departments (e.g., Occupational Health and Safety Administration [OSHA]) should be notified in cases of environmental or workplace exposure. The best approach to young children and patients with intentional overdose (deliberate self-harm or attempted suicide) is to limit their access to poisons. In households where children live or visit, alcoholic beverages, medications, household products (automotive, cleaning, fuel, pet-care, and toiletry products), inedible plants, and vitamins should be kept out of reach or in locked or child-proof cabinets. Depressed or psychotic patients should undergo psychiatric assessment, disposition, and follow-up. They should be given prescriptions for a limited supply of drugs with a limited number of refills and should be monitored for compliance and response to therapy.

SPECIFIC TOXIC SYNDROMES AND POISONINGS

Table 450-4 summarizes the pathophysiology, clinical features, and treatment of toxidromes and poisonings that are common, produce life-threatening toxicity, or require unique therapeutic interventions. In all cases, treatment should include attention to the general principles discussed above and, in particular, supportive care. Poisonings not covered in this chapter are discussed in specialized texts.

Alcohol, cocaine, hallucinogen, and opioid poisoning and alcohol and opioid withdrawal are discussed in Chaps. 445–447; nicotine addiction is discussed in Chap. 448; acetaminophen poisoning is discussed in Chap. 333; the neuroleptic malignant syndrome is discussed in Chap. 427; and heavy metal poisoning is discussed in Chap. 449.

GLOBAL CONSIDERATIONS

Risks of poisoning in the United States and throughout the world are in transition. Patterns of travel, immigration, and internet consumerism should always be considered in patients suspected of poisoning or overdose without a clear etiology. Immigrants into various countries may have underlying poisoning from various metals from work or the environment where they previously lived; herbal remedies, food products, and cosmetics imported from overseas may be contaminated with metals, toxic plants, or other pharmaceutical contaminants; and new drugs of abuse that originate in one part of the world quickly circulate due to the ease afforded by the internet. Expanding the history at the time of evaluation, recruiting the assistance of global health specialists, and ordering expanded laboratory panels may be indicated.

TABLE 450-4 Pathophysiologic Features and Treatment of Specific Toxic Syndromes and Poisonings

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Stimulated				
Sympathetics[a]				
Sympathomimetics	α_1-Adrenergic agonists (decongestants): phenylephrine, phenylpropanolamine β_2-Adrenergic agonists (bronchodilators): albuterol, terbutaline Nonspecific adrenergic agonists: amphetamines, cocaine, ephedrine	Stimulation of central and peripheral sympathetic receptors directly or indirectly (by promoting release or inhibiting reuptake of norepinephrine and sometimes dopamine)	Physiologic stimulation (Table 450-2). Reflex bradycardia can occur with selective α_1 agonists; β agonists can cause hypotension and hypokalemia.	Phentolamine, a nonselective α_1-adrenergic receptor antagonist, for severe hypertension due to α_1-adrenergic agonists; propranolol, a nonselective β blocker, for hypotension and tachycardia due to β_2 agonists; *either* labetalol, a β blocker with α-blocking activity, *or* phentolamine with esmolol, metoprolol, or another cardioselective β blocker for hypertension with tachycardia due to nonselective agents (β blockers, if used alone, can exacerbate hypertension and vasospasm due to unopposed α stimulation.); benzodiazepines; propofol
Ergot alkaloids	Ergotamine, methysergide, bromocriptine, pergolide	Stimulation and inhibition of serotonergic and α-adrenergic receptors; stimulation of dopamine receptors	Physiologic stimulation (Table 450-2); formication; vasospasm with limb (isolated or generalized), myocardial, and cerebral ischemia progressing to gangrene or infarction. Hypotension, bradycardia, and involuntary movements can also occur.	Nitroprusside or nitroglycerine for severe vasospasm; prazosin (an α_1 blocker), captopril, nifedipine, and cyproheptadine (a serotonin receptor antagonist) for mild-to-moderate limb ischemia; dopamine receptor antagonists (antipsychotics) for hallucinations and movement disorders
Methylxanthines	Caffeine, theophylline	Inhibition of adenosine synthesis and adenosine receptor antagonism; stimulation of epinephrine and norepinephrine release; inhibition of phosphodiesterase resulting in increased intracellular cyclic adenosine and guanosine monophosphate	Physiologic stimulation (Table 450-2); pronounced gastrointestinal symptoms and β agonist effects (see above). Toxicity occurs at lower drug levels in chronic poisoning than in acute poisoning.	Propranolol, a nonselective β blocker, for tachycardia with hypotension; any β blocker for supraventricular or ventricular tachycardia without hypotension; elimination enhanced by multiple-dose charcoal, hemoperfusion, and hemodialysis. Indications for hemoperfusion or hemodialysis include unstable vital signs, seizures, and a theophylline level of 80–100 μg/mL after an acute overdose and 40–60 μg/mL with chronic exposure.
Monoamine oxidase inhibitors	Phenelzine, tranylcypromine, selegiline	Inhibition of monoamine oxidase resulting in impaired metabolism of endogenous catecholamines and exogenous sympathomimetic agents	Delayed or slowly progressive physiologic stimulation (Table 450-2); terminal hypotension and bradycardia in severe cases	Short-acting agents (e.g., nitroprusside, esmolol) for severe hypertension and tachycardia; direct-acting sympathomimetics (e.g., norepinephrine, epinephrine) for hypotension and bradycardia
Anticholinergics				
Antihistamines	Diphenhydramine, doxylamine, pyrilamine	Inhibition of central and postganglionic parasympathetic muscarinic cholinergic receptors. At high doses, amantadine, diphenhydramine, orphenadrine, phenothiazines, and tricyclic antidepressants have additional nonanticholinergic activity (see below).	Physiologic stimulation (Table 450-2); dry skin and mucous membranes, decreased bowel sounds, flushing, and urinary retention; myoclonus and picking activity. Central effects may occur without significant autonomic dysfunction.	Physostigmine, an acetylcholinesterase inhibitor (see below), for delirium, hallucinations, and neuromuscular hyperactivity. Contraindications include asthma and non-anticholinergic cardiovascular toxicity (e.g., cardiac conduction abnormalities, hypotension, and ventricular arrhythmias).
Antipsychotics	Chlorpromazine, olanzapine, quetiapine, thioridazine	Inhibition of α-adrenergic, dopaminergic, histaminergic, muscarinic, and serotonergic receptors. Some agents also inhibit sodium, potassium, and calcium channels.	Physiologic depression (Table 450-2), miosis, anticholinergic effects (see above), extrapyramidal reactions (see below), tachycardia	Sodium bicarbonate for ventricular tachydysrhythmias associated with QRS prolongation; magnesium, isoproterenol, and overdrive pacing for torsades des pointes. Avoid class IA, IC, and III antiarrhythmics.
Belladonna alkaloids	Atropine, hyoscyamine, scopolamine	Inhibition of central and postganglionic parasympathetic muscarinic cholinergic receptors	Physiologic stimulation (Table 450-2); dry skin and mucous membranes, decreased bowel sounds, flushing, and urinary retention; myoclonus and picking activity. Central effects may occur without significant autonomic dysfunction.	Physostigmine, an acetylcholinesterase inhibitor (see below), for delirium, hallucinations, and neuromuscular hyperactivity. Contraindications include asthma and non-anticholinergic cardiovascular toxicity (e.g., cardiac conduction abnormalities, hypotension, and ventricular arrhythmias).

(*Continued*)

TABLE 450-4 Pathophysiologic Features and Treatment of Specific Toxic Syndromes and Poisonings (Continued)

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Cyclic antidepressants	Amitriptyline, doxepin, imipramine	Inhibition of α-adrenergic, dopaminergic, GABA-ergic, histaminergic, muscarinic, and serotonergic receptors; inhibition of sodium channels (see membrane-active agents); inhibition of norepinephrine and serotonin reuptake	Physiologic depression (Table 450-2), seizures, tachycardia, cardiac conduction delays (increased PR, QRS, JT, and QT intervals; terminal QRS right-axis deviation) with aberrancy and ventricular tachydysrhythmias; anticholinergic toxidrome (see above)	Hypertonic sodium bicarbonate (or hypertonic saline) for ventricular tachydysrhythmias associated with QRS prolongation. Use of phenytoin is controversial. Avoid class IA, IC, and III antiarrhythmics. IV emulsion therapy may be beneficial in some cases.
Mushrooms and plants	*Amanita muscaria* and *A. pantherina*, henbane, jimson weed, nightshade	Inhibition of central and postganglionic parasympathetic muscarinic cholinergic receptors	Physiologic stimulation (Table 450-2); dry skin and mucous membranes, decreased bowel sounds, flushing, and urinary retention; myoclonus and picking activity. Central effects may occur without significant autonomic dysfunction.	Physostigmine, an acetylcholinesterase inhibitor (see below), for delirium, hallucinations, and neuromuscular hyperactivity. Contraindications include asthma and nonanticholinergic cardiovascular toxicity (e.g., cardiac conduction abnormalities, hypotension, and ventricular arrhythmias).
Depressed				
Sympatholytics				
α_2-Adrenergic agonists	Clonidine, guanabenz, tetrahydrozoline and other imidazoline decongestants, tizanidine and other imidazoline muscle relaxants	Stimulation of α_2-adrenergic receptors leading to inhibition of CNS sympathetic outflow. Activity at nonadrenergic imidazoline binding sites also contributes to CNS effects.	Physiologic depression (Table 450-2), miosis. Transient initial hypertension may be seen.	Dopamine and norepinephrine for hypotension; atropine for symptomatic bradycardia; naloxone for CNS depression (inconsistently effective)
Antipsychotics	Chlorpromazine, clozapine, haloperidol, risperidone, thioridazine	Inhibition of α-adrenergic, dopaminergic, histaminergic, muscarinic, and serotonergic receptors. Some agents also inhibit sodium, potassium, and calcium channels.	Physiologic depression (Table 450-2), miosis, anticholinergic effects (see above), extrapyramidal reactions (see below), tachycardia. Cardiac conduction delays (increased PR, QRS, JT, and QT intervals) with ventricular tachydysrhythmias, including torsades des pointes, can sometimes develop.	Sodium bicarbonate for ventricular tachydysrhythmias associated with QRS prolongation; magnesium, isoproterenol, and overdrive pacing for torsades des pointes. Avoid class IA, IC, and III antiarrhythmics.
β-Adrenergic blockers	Cardioselective (β_1) blockers: atenolol, esmolol, metoprolol Nonselective (β_1 and β_2) blockers: nadolol, propranolol, timolol Partial β agonists: acebutolol, pindolol α_1 Antagonists: carvedilol, labetalol Membrane-active agents: acebutolol, propranolol, sotalol	Inhibition of β-adrenergic receptors (class II antiarrhythmic effect). Some agents have activity at additional receptors or have membrane effects (see below).	Physiologic depression (Table 450-2), atrioventricular block, hypoglycemia, hyperkalemia, seizures. Partial agonists can cause hypertension and tachycardia. Sotalol can cause increased QT interval and ventricular tachydysrhythmias. Onset may be delayed after sotalol and sustained-release formulation overdose.	Glucagon for hypotension and symptomatic bradycardia. Atropine, isoproterenol, dopamine, dobutamine, epinephrine, and norepinephrine may sometimes be effective. High-dose insulin (with glucose and potassium to maintain euglycemia and normokalemia), electrical pacing, and mechanical cardiovascular support for refractory cases.
Calcium channel blockers	Diltiazem, nifedipine and other dihydropyridine derivatives, verapamil	Inhibition of slow (type L) cardiovascular calcium channels (class IV antiarrhythmic effect)	Physiologic depression (Table 450-2), atrioventricular block, organ ischemia and infarction, hyperglycemia, seizures. Hypotension is usually due to decreased vascular resistance rather than to decreased cardiac output. Onset may be delayed for ≥12 h after overdose of sustained-release formulations.	Calcium and glucagon for hypotension and symptomatic bradycardia. Dopamine, epinephrine, norepinephrine, atropine, and isoproterenol are less often effective but can be used adjunctively. High-dose insulin (with glucose and potassium to maintain euglycemia and normokalemia), IV lipid emulsion therapy, electrical pacing, and mechanical cardiovascular support for refractory cases.
Cardiac glycosides	Digoxin, endogenous cardioactive steroids, foxglove and other plants, toad skin secretions (Bufonidae spp.)	Inhibition of cardiac Na^+, K^+-ATPase membrane pump	Physiologic depression (Table 450-2); gastrointestinal, psychiatric, and visual symptoms; atrioventricular block with or without concomitant supraventricular tachyarrhythmia; ventricular tachyarrhythmias; hyperkalemia in acute poisoning. Toxicity occurs at lower drug levels in chronic poisoning than in acute poisoning.	Digoxin-specific antibody fragments for hemodynamically compromising dysrhythmias, Mobitz II or third-degree atrioventricular block, hyperkalemia (>5.5 meq/L; in acute poisoning only). Temporizing measures include atropine, dopamine, epinephrine, and external cardiac pacing for bradydysrhythmias and magnesium, lidocaine, or phenytoin, for ventricular tachydysrhythmias. Internal cardiac pacing and cardioversion can increase ventricular irritability and should be reserved for refractory cases.

(*Continued*)

TABLE 450-4 Pathophysiologic Features and Treatment of Specific Toxic Syndromes and Poisonings (Continued)

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Depressed (Cont.)				
Sympatholytics (Cont.)				
Cyclic antidepressants	Amitriptyline, doxepin, imipramine	Inhibition of α-adrenergic, dopaminergic, GABA-ergic, histaminergic, muscarinic, and serotonergic receptors; inhibition of sodium channels (see membrane-active agents); inhibition of norepinephrine and serotonin reuptake	Physiologic depression (Table 450-2), seizures, tachycardia, cardiac conduction delays (increased PR, QRS, JT, and QT intervals; terminal QRS right-axis deviation) with aberrancy and ventricular tachydysrhythmias; anticholinergic toxidrome (see above)	Hypertonic sodium bicarbonate (or hypertonic saline) for ventricular tachydysrhythmias associated with QRS prolongation. Use of phenytoin is controversial. Avoid class IA, IC, and III antiarrhythmics. IV emulsion therapy may be beneficial in some cases.
Cholinergics				
Acetylcholin-esterase inhibitors	Carbamate insecticides (aldicarb, carbaryl, propoxur) and medicinals (neostigmine, physostigmine, tacrine); nerve gases (sarin, soman, tabun, VX); organophosphate insecticides (diazinon, chlorpyrifos-ethyl, malathion)	Inhibition of acetylcholinesterase leading to increased synaptic acetylcholine at muscarinic and nicotinic cholinergic receptor sites	Physiologic depression (Table 450-2). Muscarinic signs and symptoms: seizures, excessive secretions (lacrimation, salivation, bronchorrhea and wheezing, diaphoresis), and increased bowel and bladder activity with nausea, vomiting, diarrhea, abdominal cramps, and incontinence of feces and urine. Nicotinic signs and symptoms: hypertension, tachycardia, muscle cramps, fasciculations, weakness, and paralysis. Death is usually due to respiratory failure. Cholinesterase activity in plasma and red cells is <50% of normal in acetylcholinesterase inhibitor poisoning.	Atropine for muscarinic signs and symptoms; 2-PAM, a cholinesterase reactivator, for nicotinic signs and symptoms due to organophosphates, nerve gases, or an unknown anticholinesterase
Muscarinic agonists	Bethanechol, mushrooms (*Boletus, Clitocybe, Inocybe* spp.), pilocarpine	Stimulation of CNS and postganglionic parasympathetic cholinergic (muscarinic) receptors		
Nicotinic agonists	Lobeline, nicotine (tobacco)	Stimulation of preganglionic sympathetic and parasympathetic and striated muscle (neuromuscular junction) cholinergic (nicotine) receptors		
Sedative-hypnotics[b]				
Anticonvulsants	Carbamazepine, ethosuximide, felbamate, gabapentin, lamotrigine, levetiracetam, oxcarbazepine, phenytoin, tiagabine, topiramate, valproate, zonisamide	Potentiation of the inhibitory effects of GABA by binding to the neuronal GABA–A chloride channel receptor complex and increasing the frequency or duration of chloride channel opening in response to GABA stimulation. Baclofen and, to some extent, GHB act at the GABA–B receptor complex. Meprobamate, its metabolite carisoprodol, felbamate, and orphenadrine antagonize NDMA excitatory receptors. Ethosuximide, valproate, and zonisamide decrease conduction through T-type calcium channels. Valproate decreases GABA degradation, and tiagabine blocks GABA reuptake. Carbamazepine, lamotrigine, oxcarbazepine, phenytoin, topiramate, valproate, and zonisamide slow the rate of recovery of inactivated sodium channels. Some agents also have α_2 agonist, anticholinergic, and sodium channel–blocking activity (see above and below).	Physiologic depression (Table 450-2), nystagmus. Delayed absorption can occur with carbamazepine, phenytoin, and valproate. Myoclonus, seizures, hypertension, and tachyarrhythmias can occur with baclofen, carbamazepine, and orphenadrine.	Benzodiazepines, barbiturates, or propofol for seizures.
Barbiturates	Short-acting: butabarbital, pentobarbital, secobarbital Long-acting: phenobarbital, primidone			Hemodialysis and hemoperfusion may be indicated for severe poisoning by some agents (see "Extracorporeal Removal," in text).
Benzodiazepines	Ultrashort-acting: estazolam, midazolam, temazepam, triazolam Short-acting: alprazolam, flunitrazepam, lorazepam, oxazepam Long-acting: chlordiazepoxide, clonazepam, diazepam, flurazepam Pharmacologically related agents: zaleplon, zolpidem		Tachyarrhythmias can also occur with chloral hydrate. AGMA, hypernatremia, hyperosmolality, hyperammonemia, chemical hepatitis, and hypoglycemia can be seen in valproate poisoning. Carbamazepine and oxcarbazepine may produce hyponatremia from SIADH.	See above and below for treatment of anticholinergic and sodium channel (membrane)–blocking effects.
GABA precursors	γ-Hydroxybutyrate (sodium oxybate; GHB), γ-butyrolactone (GBL), 1,4-butanediol		Some agents can cause anticholinergic and sodium channel (membrane) blocking effects (see above and below).	

(*Continued*)

TABLE 450-4 Pathophysiologic Features and Treatment of Specific Toxic Syndromes and Poisonings (Continued)

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Muscle relaxants	Baclofen, carisoprodol, cyclobenzaprine, etomidate, metaxalone, methocarbamol, orphenadrine, propofol, tizanidine and other imidazoline muscle relaxants	Baclofen acts at GABA-B receptor complex; Stimulation of α2-adrenergic receptors inhibits CNS sympathetic outflow. Activity at nonadrenergic imidazoline binding sites also contributes to CNS effects. The others have centrally-acting and various other unknown mechanisms of action	Physiologic depression (Table 450-2)	Goal-directed supportive care; benzodiazepines and barbiturates for seizures
Other agents	Chloral hydrate, ethchlorvynol, glutethimide, meprobamate, methaqualone, methyprylon			
Discordant				
Asphyxiants				
Cytochrome oxidase inhibitors	Cyanide, hydrogen sulfide	Inhibition of mitochondrial cytochrome oxidase, with consequent blockage of electron transport and oxidative metabolism. Carbon monoxide also binds to hemoglobin and myoglobin and prevents oxygen binding, transport, and tissue uptake. (Binding to hemoglobin shifts the oxygen dissociation curve to the left.)	Signs and symptoms of hypoxemia with initial physiologic stimulation and subsequent depression (Table 450-2); lactic acidosis; normal Po_2 and calculated oxygen saturation but decreased oxygen saturation by co-oximetry. (That measured by pulse oximetry is falsely elevated but is less than normal and less than the calculated value.) Headache and nausea are common with carbon monoxide. Sudden collapse may occur with cyanide and hydrogen sulfide exposure. A bitter almond breath odor may be noted with cyanide ingestion, and hydrogen sulfide smells like rotten eggs.	High-dose oxygen; IV hydroxocobalamin or IV sodium nitrite and sodium thiosulfate (Lilly cyanide antidote kit) for coma, metabolic acidosis, and cardiovascular dysfunction in cyanide poisoning or victims from a fire
Methemoglobin inducers	Aniline derivatives, dapsone, local anesthetics, nitrates, nitrites, nitrogen oxides, nitro- and nitrosohydrocarbons, phenazopyridine, primaquine-type antimalarials, sulfonamides	Oxidation of hemoglobin iron from ferrous (Fe^{2+}) to ferric (Fe^{3+}) state prevents oxygen binding, transport, and tissue uptake. (Methemoglobinemia shifts oxygen dissociation curve to the left.) Oxidation of hemoglobin protein causes hemoglobin precipitation and hemolytic anemia (manifesting as Heinz bodies and "bite cells" on peripheral-blood smear).	Signs and symptoms of hypoxemia with initial physiologic stimulation and subsequent depression (Table 450-2), gray-brown cyanosis unresponsive to oxygen at methemoglobin fractions >15–20%, headache, lactic acidosis (at methemoglobin fractions >45%), normal Po_2 and calculated oxygen saturation but decreased oxygen saturation and increased methemoglobin fraction by co-oximetry (Oxygen saturation by pulse oximetry may be falsely increased or decreased but is less than normal and less than the calculated value.)	High-dose oxygen; IV methylene blue for methemoglobin fraction >30%, symptomatic hypoxemia, or ischemia (contraindicated in G6PD deficiency); exchange transfusion and hyperbaric oxygen for severe or refractory cases
AGMA inducers	Ethylene glycol	Ethylene glycol causes CNS depression and increased serum osmolality. Metabolites (primarily glycolic acid) cause AGMA, CNS depression, and renal failure. Precipitation of oxalic acid metabolite as calcium salt in tissues and urine results in hypocalcemia, tissue edema, and crystalluria.	Initial ethanol-like intoxication, nausea, vomiting, increased osmolar gap, calcium oxylate crystalluria; delayed AGMA, back pain, renal failure; coma, seizures, hypotension, ARDS in severe cases	Sodium bicarbonate to correct acidemia; thiamine, folinic acid, magnesium, and high-dose pyridoxine to facilitate metabolism; ethanol or fomepizole for AGMA, crystalluria or renal dysfunction, ethylene glycol level >3 mmol/L (20 mg/dL), and ethanol-like intoxication or increased osmolal gap if level not readily obtainable; hemodialysis for persistent AGMA, lack of clinical improvement, and renal dysfunction; hemodialysis also useful for enhancing ethylene glycol elimination and shortening duration of treatment when ethylene glycol level is >8 mmol/L (50 mg/dL).

(*Continued*)

TABLE 450-4 Pathophysiologic Features and Treatment of Specific Toxic Syndromes and Poisonings (Continued)

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Discordant (Cont.)				
Asphyxiants (Cont.)				
AGMA inducers	Iron	Hydration of ferric (Fe^{3+}) ion generates H^+. Non-transferrin-bound iron catalyzes formation of free radicals that cause mitochondrial injury, lipid peroxidation, increased capillary permeability, vasodilation, and organ toxicity.	Initial nausea, vomiting, abdominal pain, diarrhea; AGMA, cardiovascular and CNS depression, hepatitis, coagulopathy, and seizures in severe cases. Radiopaque iron tablets may be seen on abdominal x-ray.	Whole-bowel irrigation for large ingestions; endoscopy and gastrostomy if clinical toxicity and large number of tablets are still visible on x-ray; IV hydration; sodium bicarbonate for acidemia; IV deferoxamine for systemic toxicity, iron level >90 μmol/L (500 μg/dL)
	Methanol	Methanol causes ethanol-like CNS depression and increased serum osmolality. Formic acid metabolite causes AGMA and retinal toxicity.	Initial ethanol-like intoxication, nausea, vomiting, increased osmolar gap; delayed AGMA, visual (clouding, spots, blindness) and retinal (edema, hyperemia) abnormalities; coma, seizures, cardiovascular depression in severe cases; possible pancreatitis	Gastric aspiration for recent ingestion; sodium bicarbonate to correct acidemia; high-dose folinic acid or folate to facilitate metabolism; ethanol or fomepizole for AGMA, visual symptoms, methanol level >6 mmol/L (20 mg/dL), and ethanol-like intoxication or increased osmolal gap if level not readily obtainable; hemodialysis for persistent AGMA, lack of clinical improvement, and renal dysfunction; hemodialysis also useful for enhancing methanol elimination and shortening duration of treatment when methanol level is >15 mmol/L (50 mg/dL)
	Salicylate	Increased sensitivity of CNS respiratory center to changes in and stimulates respiration. Uncoupling of oxidative phosphorylation, inhibition of Krebs cycle enzymes, and stimulation of carbohydrate and lipid metabolism generate unmeasured endogenous anions and cause AGMA.	Initial nausea, vomiting, hyperventilation, alkalemia, alkaluria; subsequent alkalemia with both respiratory alkalosis and AGMA and paradoxical aciduria; late acidemia with CNS and respiratory depression; cerebral and pulmonary edema in severe cases. Hypoglycemia, hypocalcemia, hypokalemia, and seizures can occur.	IV hydration and supplemental glucose; sodium bicarbonate to correct acidemia; urinary alkalinization for systemic toxicity; hemodialysis for coma, cerebral edema, seizures, pulmonary edema, renal failure, progressive acid-base disturbances or clinical toxicity, salicylate level >7 mmol/L (100 mg/dL) following acute overdose
CNS syndromes				
Extrapyramidal reactions	Antipsychotics (see above), some cyclic antidepressants and antihistamines	Decreased CNS dopaminergic activity with relative excess of cholinergic activity	Akathisia, dystonia, parkinsonism	Oral or parenteral anticholinergic agent such as benztropine or diphenhydramine
Isoniazid		Interference with activation and supply of pyridoxal-5-phosphate, a cofactor for glutamic acid decarboxylase, which converts glutamic acid to GABA, results in decreased levels of this inhibitory CNS neurotransmitter; complexation with and depletion of pyridoxine itself; inhibition of nicotine adenine dinucleotide–dependent lactate and hydroxybutyrate dehydrogenases, resulting in substrate accumulation	Nausea, vomiting, agitation, confusion; coma, respiratory depression, seizures, lactic and ketoacidosis in severe cases	High-dose IV pyridoxine (vitamin B_6) for agitation, confusion, coma, and seizures; diazepam or barbiturates for seizures
Lithium		Interference with cell membrane ion transport, adenylate cyclase and Na^+, K^+-ATPase activity, and neurotransmitter release	Nausea, vomiting, diarrhea, ataxia, choreoathetosis, encephalopathy, hyperreflexia, myoclonus, nystagmus, nephrogenic diabetes insipidus, falsely elevated serum chloride with low anion gap, tachycardia; coma, seizures, arrhythmias, hyperthermia, and prolonged or permanent encephalopathy and movement disorders in severe cases; delayed onset after acute overdose, particularly with delayed-release formulations. Toxicity occurs at lower drug levels in chronic poisoning than in acute poisoning.	Whole-bowel irrigation for large ingestions; IV hydration; hemodialysis for coma, seizures, encephalopathy or neuromuscular dysfunction (severe, progressive, or persistent), peak lithium level >4 meq/L following acute overdose

(*Continued*)

TABLE 450-4 Pathophysiologic Features and Treatment of Specific Toxic Syndromes and Poisonings (Continued)

PHYSIOLOGIC CONDITION, CAUSES	EXAMPLES	MECHANISM OF ACTION	CLINICAL FEATURES	SPECIFIC TREATMENTS
Serotonin syndrome	Amphetamines, cocaine, dextromethorphan, meperidine, MAO inhibitors, selective serotonin (5-HT) reuptake inhibitors, tricyclic antidepressants, tramadol, triptans, tryptophan	Promotion of serotonin release, inhibition of serotonin reuptake, or direct stimulation of CNS and peripheral serotonin receptors (primarily 5-HT-1a and 5-HT-2), alone or in combination	Altered mental status (agitation, confusion, mutism, coma, seizures), neuromuscular hyperactivity (hyperreflexia, myoclonus, rigidity, tremors), and autonomic dysfunction (abdominal pain, diarrhea, diaphoresis, fever, flushing, labile hypertension, mydriasis, tearing, salivation, tachycardia). Complications include hyperthermia, lactic acidosis, rhabdomyolysis, and multisystem organ failure.	Discontinue the offending agent(s); the serotonin receptor antagonist cyproheptadine may be helpful in severe cases.
Membrane-active agent	Amantadine, antiarrhythmics (class I and III agents; some β blockers), antipsychotics (see above), antihistamines (particularly diphenhydramine), carbamazepine, local anesthetics (including cocaine), opioids (meperidine, propoxyphene), orphenadrine, quinoline antimalarials (chloroquine, hydroxychloroquine, quinine), cyclic antidepressants (see above)	Blockade of fast sodium membrane channels prolongs phase 0 (depolarization) of the cardiac action potential, which prolongs QRS duration and promotes reentrant (monomorphic) ventricular tachycardia. Class Ia, Ic, and III antiarrhythmics also block potassium channels during phases 2 and 3 (repolarization) of the action potential, prolonging the JT interval and promoting early after-depolarizations and polymorphic (torsades des pointes) ventricular tachycardia. Similar effects on neuronal membrane channels cause CNS dysfunction. Some agents also block α-adrenergic and cholinergic receptors or have opioid effects (see above and **Chap. 446**).	QRS and JT prolongation (or both) with hypotension, ventricular tachyarrhythmias, CNS depression, seizures; anticholinergic effects with amantadine, antihistamines, carbamazepine, disopyramide, antipsychotics, and cyclic antidepressants (see above); opioid effects with meperidine and propoxyphene **(see Chap. 446)**; cinchonism (hearing loss, tinnitus, nausea, vomiting, vertigo, ataxia, headache, flushing, diaphoresis), and blindness with quinoline antimalarials	Hypertonic sodium bicarbonate (or hypertonic saline) for cardiac conduction delays and monomorphic ventricular tachycardia; lidocaine for monomorphic ventricular tachycardia (except when due to class Ib antiarrhythmics); magnesium, isoproterenol, and overdrive pacing for polymorphic ventricular tachycardia; physostigmine for anticholinergic effects (see above); naloxone for opioid effects **(see Chap. 446)**; extracorporeal removal for some agents (see text).

[a]See above and **Chap. 447**. [b]See above and **Chap. 446**.

Abbreviations: AGMA, anion-gap metabolic acidosis; ARDS, adult respiratory distress syndrome; CNS, central nervous system; GABA, γ-aminobutyric acid; GBL, γ-butyrolactone; GHB, γ-hydroxybutyrate; G6PD, glucose-6-phosphate dehydrogenase; MAO, monoamine oxidase; NDMA, *N*-methyl-D-aspartate; 2-PAM, pralidoxime; SIADH, syndrome of inappropriate antidiuretic hormone secretion.

FURTHER READING

DART RC et al: Expert consensus guidelines for stocking of antidotes in hospitals that provide emergency care. Ann Emerg Med 54:386, 2009.

HOFFMAN RS et al (eds): *Goldfrank's Toxicologic Emergencies*, 10th ed. New York, McGraw-Hill, 2015.

MAZER-AMIRSHAHI M et al: Trends in Emergency Department Resource Utilization for Poisoning-Related Visits, 2003–2011. J Med Toxicol 12:248, 2016.

MOWRY JB et al: 2015 Annual report of the American Association of Poison Control Centers' National Poison Data System (NPDS): 33rd Annual Report. Clin Toxicol 54:924, 2016.

THOMPSON TM et al: The general approach to the poisoned patient. Dis Mon 60:509, 2014.

WELKER K, MYCYK MB: Pharmacology in the geriatric patient. Emerg Med Clin North Am 34:469, 2016.

451 Disorders Caused by Venomous Snakebites and Marine Animal Exposures

Charles Lei, Kirsten B. Hornbeak, Paul S. Auerbach, Robert L. Norris

This chapter outlines general principles for the evaluation and management of victims of envenomation and poisoning by venomous snakes and marine animals. Because the incidence of serious bites and stings is relatively low in developed nations, there is a paucity of relevant clinical research; as a result, therapeutic decision-making often is based on anecdotal information.

VENOMOUS SNAKEBITE

EPIDEMIOLOGY

The venomous snakes of the world belong to the families Viperidae (subfamily Viperinae: Old World vipers; subfamily Crotalinae: New World and Asian pit vipers), Elapidae (including cobras, coral snakes, sea snakes, kraits, and all Australian venomous snakes), Lamprophiidae (subfamily Atractaspidinae: burrowing asps), and Colubridae (a large family in which most species are nonvenomous

and only a few are dangerously toxic to humans). Most snakebites occur in developing countries with temperate and tropical climates in which populations subsist on agriculture and fishing. Recent estimates indicate that somewhere between 1.2 million and 5.5 million snakebites occur worldwide each year, with 421,000–1,841,000 envenomations and 20,000–94,000 deaths. Such wide-ranging estimates reflect the challenges of collecting accurate data in the regions most affected by venomous snakes; many victims do not seek medical attention, and reporting and record keeping are generally poor.

SNAKE ANATOMY/IDENTIFICATION

The typical snake venom delivery apparatus consists of bilateral venom glands situated behind the eyes and connected by ducts to hollow anterior maxillary fangs. In viperids, these fangs are long and highly mobile; they are retracted against the roof of the mouth when the snake is at rest and brought to an upright position for a strike. In elapids, the fangs are shorter and relatively fixed in an erect position. Approximately 20% of pit viper bites and higher percentages of other snakebites (up to 75% for sea snakes) are "dry" bites; i.e., no venom is released. Significant envenomation probably occurs in ~50% of all venomous snakebites.

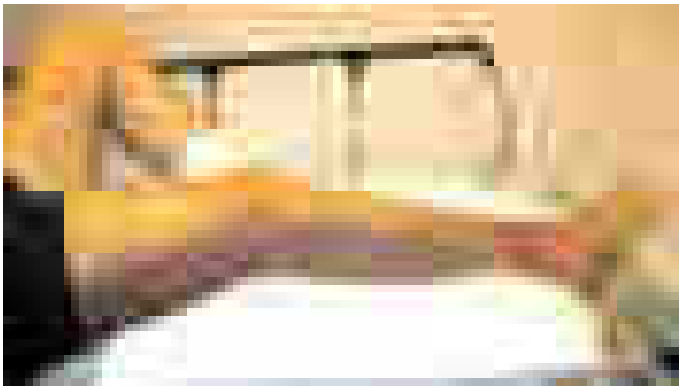
Differentiation of venomous from nonvenomous snake species can be difficult. Viperids are characterized by somewhat triangular heads (a feature shared with many harmless snakes), elliptical pupils (also seen in some nonvenomous snakes, such as boas and pythons), enlarged maxillary fangs, and, in pit vipers, heat-sensing organs (pits) on each side of the head that assist with locating prey and aiming strikes. Rattlesnakes possess a series of interlocking hollow keratin plates (the rattle) on the tip of the tail that emits a buzzing sound when the snake rapidly vibrates its tail; this sound serves as a warning signal to perceived threats. Identifying venomous snakes by color pattern is notoriously misleading, as many harmless snakes have color patterns that closely mimic those of venomous snakes found in the same region.

VENOMS AND CLINICAL MANIFESTATIONS

Snake venoms are highly variable and complex mixtures of enzymes, polypeptides, glycoproteins, and other constituents. Among the deleterious components are proteolytic enzymes that cause local tissue necrosis, affect the coagulation pathway at various steps, and impair organ function. Hemorrhagins cause vascular leakage, resulting in fluid shifts and spontaneous local and systemic bleeding. Hyaluronidases promote the spread of venom through connective tissue. Myocardial depressant factors reduce cardiac output, and bradykinins cause vasodilation and hypotension. Neurotoxins act at various sites of the neuromuscular junction to block transmission and cause muscle paralysis. Most snake venoms have multisystem effects on their victims.

After a venomous snakebite, the time to symptom onset and clinical presentation can be quite variable and depends on the species involved, the anatomic location of the bite, and the amount of venom injected. Envenomations by most viperids and some elapids with necrotizing venoms cause progressive local pain, soft-tissue swelling, and ecchymosis **(Fig. 451-1)**. Hemorrhagic or serum-filled vesicles and bullae may develop at the bite site over a period of hours to days. In serious bites, tissue loss can be significant **(Figs. 451-2 and 451-3)**. Systemic findings are extremely variable and can include generalized fatigue, nausea, changes in taste, mouth numbness, tachycardia or bradycardia, hypotension, muscle fasciculations, pulmonary edema, renal dysfunction, and spontaneous hemorrhage from essentially any anatomic site. Envenomations by neurotoxic elapids, such as kraits (*Bungarus* species), many Australian elapids (e.g., death adders [*Acanthophis* species] and tiger snakes [*Notechis* species]), some cobras (*Naja* species), some viperids (e.g., the South American rattlesnake [*Crotalus durissus*], and certain Indian Russell's vipers [*Daboia russelii*]), cause neurologic dysfunction. Early findings may consist of nausea and vomiting, headache, paresthesias or numbness, and altered mental status. Victims may develop cranial nerve abnormalities (e.g., ptosis, difficulty swallowing), followed by peripheral motor weakness. Severe envenomation may result in muscle paralysis, including the muscles of respiration, and lead to death from respiratory failure and aspiration. Sea snake

A



B

FIGURE 451-1 Northern Pacific rattlesnake (*Crotalus oreganus oreganus*) envenomations. *A.* Moderately severe envenomation. Note edema and early ecchymosis 2 h after a bite to the finger. ***B.*** Severe envenomation. Note extensive ecchymosis 5 days after a bite to the ankle. *(Courtesy of Robert Norris, with permission.)*

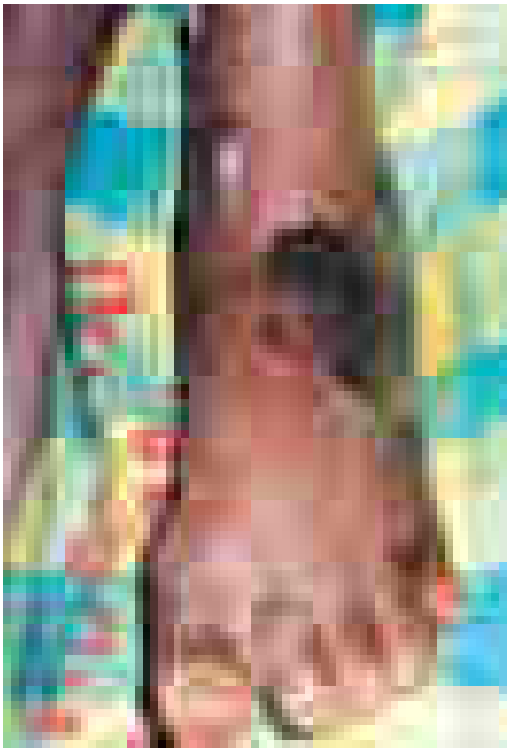


FIGURE 451-2 Early stages of severe, full-thickness necrosis 5 days after a Russell's viper (*Daboia russelii*) bite in southwestern India. *(Courtesy of Robert Norris, with permission.)*

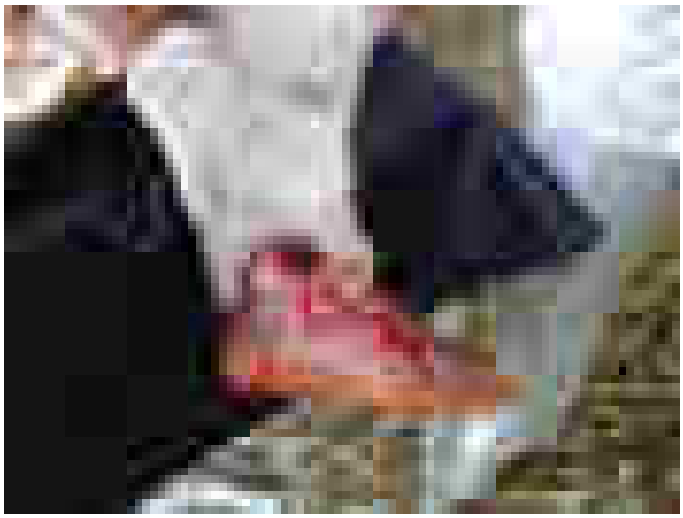


FIGURE 451-3 Severe necrosis 10 days after a pit viper bite in a young child in Colombia. *(Courtesy of Jay Stanka, with permission.)*

envenomation results in local pain (variable), generalized myalgias, trismus, rhabdomyolysis, and progressive flaccid paralysis; these manifestations can be delayed for several hours.

TREATMENT

Venomous Snakebite

FIELD MANAGEMENT

The most important aspect of prehospital care of a person bitten by a venomous snake is rapid transport to a medical facility equipped to provide supportive care (airway, breathing, and circulation) and antivenom therapy. Any jewelry or tight-fitting clothing near the bite should be removed to avoid constriction from anticipated soft-tissue swelling. Without a delay in transport, the wound should be cleaned with soap and running water and covered with a sterile dressing. It is reasonable to apply a splint to the bitten extremity to limit movement and lessen bleeding. If possible, the extremity should be maintained in a neutral position of comfort at approximately heart level. Attempting to capture and transport the offending snake, alive or dead, is not advised; instead, digital photographs taken from a safe distance may assist with snake identification and treatment decisions.

Most of the first-aid measures recommended in the past are of little benefit, and some actually worsen outcome. Incising and/or applying suction to the bite site should be avoided, as these measures exacerbate local tissue damage, increase the risk of infection, and have not been shown to be effective. Similarly ineffective and potentially harmful are the application of poultices, ice, and electric shock. Venom sequestration devices (e.g., lympho-occlusive bandages or tourniquets) are not advised, as they may intensify local tissue damage by restricting the spread of potentially necrotizing venom. Tourniquet use can result in loss of function, ischemia, and limb amputation, even in the absence of envenomation. In developing countries, indigenous people should be encouraged to seek immediate treatment at a medical facility equipped with antivenom instead of consulting traditional healers and thus incurring significant delays in reaching appropriate care.

Elapid venoms that are primarily neurotoxic and have no significant effects on local tissue may be localized by pressure-immobilization, a technique in which the entire bitten limb is immediately wrapped with a bandage (e.g., crepe or elastic) and then immobilized. For this technique to be effective, the wrap pressure must be precise (40–70 mmHg in upper-extremity application and 55–70 mmHg in lower-extremity application) and the victim must be carried out of the field because walking generates muscle-pumping activity that—regardless of the anatomic site of the bite—will disperse venom into the systemic circulation. Pressure-immobilization should be used only in cases in which the offending snake is reliably identified and known to be primarily neurotoxic, the rescuer is skilled in pressure-wrap application, the necessary supplies are readily available, and the victim can be fully immobilized and carried to medical care—a rare combination of conditions, particularly in the regions of the world where such bites are most common.

HOSPITAL MANAGEMENT

Initial hospital management should focus on the victim's airway, breathing, and circulation. Patients with bites to the face or neck may require early endotracheal intubation to prevent loss of airway patency caused by rapid soft-tissue swelling. Vital signs, cardiac rhythm, oxygen saturation, and urine output should be closely monitored. Two large-bore IV lines should be established in unaffected extremities. Because of the potential for coagulopathy, venipuncture attempts should be minimized and noncompressible sites (e.g., a subclavian vein) avoided. Early hypotension is due to pooling of blood in the pulmonary and splanchnic vascular beds. Later, systemic bleeding, hemolysis, and loss of intravascular volume into the soft tissues may play important roles. Fluid resuscitation with isotonic saline (20–40 mL/kg IV) should be initiated if there is any evidence of hemodynamic instability, and a trial of 5% albumin (10–20 mL/kg IV) may be undertaken if the response to saline infusion is inadequate. Vasopressors (e.g., norepinephrine, dopamine) should be added only if venom-induced shock persists after aggressive volume resuscitation and antivenom administration (see below). Invasive hemodynamic monitoring (central venous and/or continuous arterial pressures) can be helpful in such cases, although gaining central vascular access is risky if coagulopathy has developed.

A thorough history (including the time of the bite and any symptoms of envenomation) should be obtained and a complete physical examination performed. Bandages or wraps applied in the field should be removed once IV access has been obtained, with cognizance that the release of such ligatures may result in hypotension or dysrhythmias when stagnant acidotic blood containing venom is released into the systemic circulation. To objectively evaluate the progression of local envenomation, the leading edge of swelling, ecchymosis, and tenderness should be marked and limb circumference should be measured every 15 min until the local tissue effects have stabilized. During this period of observation, the bitten extremity should be positioned at approximately heart level. Victims of neurotoxic envenomation should be monitored closely for evidence of cranial nerve dysfunction (e.g., ptosis) that may precede more overt signs of impending airway compromise (e.g., difficulty swallowing, respiratory insufficiency) necessitating endotracheal intubation and mechanical ventilation.

Blood should be drawn for laboratory evaluation as soon as possible. Important studies include a complete blood count to determine the degree of hemorrhage or hemolysis and to detect thrombocytopenia; blood type and cross-matching; assessment of renal and hepatic function; coagulation studies to diagnose consumptive coagulopathy; measurement of creatine kinase for suspected rhabdomyolysis; and testing of urine for blood or myoglobin. In developing regions, the 20-min whole-blood clotting test can be used to reliably diagnose coagulopathy. A few milliliters of fresh blood are placed in a new, clean, plain glass receptacle (e.g., a test tube) and left undisturbed for 20 min. The tube is then tipped once to 45°. If the blood is still liquid and a clot has not formed, coagulopathy is present. Arterial blood gas studies, electrocardiography, and chest radiography may be helpful in severe envenomations or when there is significant comorbidity. Any arterial puncture in the setting of coagulopathy requires great caution and must be performed at an anatomic site amenable to direct-pressure tamponade. After antivenom therapy (see below), laboratory values should be rechecked every 6 h until clinical stability is achieved. If initial laboratory values are normal, the complete blood count and coagulation studies should be repeated every hour until it is clear that no systemic envenomation has occurred.

The mainstay of treatment of a venomous snakebite resulting in significant envenomation is prompt administration of specific antivenom. Antivenoms are produced by injecting animals (generally horses or sheep) with venoms from medically important snakes. Once the stock animals develop antibodies to the venoms, their serum is harvested and the antibodies are isolated for antivenom preparation. The goal of antivenom administration is to allow antibodies (or antibody fragments) to bind and deactivate circulating venom components before they can attach to target tissues and cause deleterious effects. Antivenoms may be monospecific (directed against a particular snake species) or polyspecific (covering several species in a geographic region) but rarely offer cross-protection against snake species other than those used in their production unless the species are known to have homologous venoms. Thus, antivenom selection must be specific for the offending snake; if the antivenom chosen does not contain antibodies to that snake's venom components, it will provide no benefit and may lead to unnecessary complications (see below). In the United States, assistance in finding appropriate antivenom can be obtained from a regional poison-control center, which can be reached by telephone 24 h a day at (800) 222-1222.

For victims of bites by viperids or cytotoxic elapids, indications for antivenom administration include significant progressive local findings (e.g., soft-tissue swelling that crosses a joint, involves more than half the bitten limb, or is rapidly spreading; extensive blistering or bruising; severe pain) and any evidence of systemic envenomation (systemic symptoms or signs, laboratory abnormalities). Caution must be used in determining the significance of isolated pain or soft-tissue swelling after the bite of an unidentified snake because the saliva of some relatively harmless species can cause mild discomfort or edema at the bite site; in such bites, antivenoms are useless and potentially harmful. Antivenoms have limited efficacy in preventing local tissue damage caused by necrotizing venoms, as venom components bind to local tissues very quickly, before antivenom administration can be initiated. Nevertheless, antivenom should be administered as soon as the need for it is identified to limit further tissue damage and systemic effects. Antivenom administration after bites by neurotoxic elapids is indicated at the first sign of any evidence of neurotoxicity (e.g., cranial nerve dysfunction, peripheral neuropathy). In general, antivenom is effective only in reversing active venom toxicity; it is of no benefit in reversing effects that have already been established (e.g., renal failure, established paralysis) and that will improve only with time and other supportive therapies.

Specific comments related to the management of venomous snakebites in the United States and Canada appear in **Table 451-1**. The package insert for the selected antivenom should be consulted regarding species covered, method of administration, starting dose, and need (if any) for re-dosing. The mobile app SnakeBite911 ER, developed by the manufacturers of the Crotalidae Polyvalent Immune Fab (CroFab) (Ovine) antivenom, offers a treatment algorithm for pit viper envenomations, has photos and detailed descriptions of North American pit viper species, and can help locate nearby hospitals. Whenever possible, it is advisable for health

TABLE 451-1 Management of Venomous Snakebite in the United States and Canada[a]

Pit Viper Bites: Rattlesnakes (*Crotalus* and *Sistrurus* spp.), Cottonmouth Water Moccasins (*Agkistrodon piscivorus*), and Copperheads (*Agkistrodon contortrix*)

- Stabilize airway, breathing, and circulation.
- Institute monitoring (vital signs, cardiac rhythm, and oxygen saturation).
- Establish two large-bore IV lines.
  - If the patient is hypotensive, administer a normal saline bolus (20–40 mL/kg IV).
  - If hypotension persists, consider 5% albumin (10–20 mL/kg IV).
- Take thorough history and perform complete physical examination.
- Identify offending snake if possible.
- Measure and record circumference of bitten extremity q15min until swelling has stabilized.
- Order laboratory studies (CBC, blood type and cross-matching, metabolic panel, PT/INR/PTT, fibrinogen level, FDP, CK, urinalysis).
  - If normal, repeat CBC and coagulation studies every hour until it is clear that no systemic envenomation has occurred.
  - If abnormal, repeat 6 h after antivenom administration (see below).
- Determine severity of envenomation.
  - None: fang marks only ("dry" bite)
  - Mild: local findings only (e.g., pain, ecchymosis, nonprogressive swelling)
  - Moderate: swelling that is clearly progressing, systemic symptoms or signs, and/or laboratory abnormalities
  - Severe: neurologic dysfunction, respiratory distress, and/or cardiovascular instability/shock
- Contact regional poison-control center.
- Locate and administer antivenom as indicated: Crotalidae Polyvalent Immune Fab (CroFab) (Ovine) (BTG International Inc., West Conshohocken, PA).
  - Starting dose
    - Based on severity of envenomation
      - None or mild: none
      - Moderate: 4–6 vials
      - Severe: 6 vials
    - Dilute reconstituted vials in 250 mL of normal saline.
    - Infuse IV over 1 h (with medical provider in close attendance).
      - Start at rate of 20–50 mL/h for first 10 min.
      - If there is no allergic reaction, increase rate to 250 mL/h.
  - If there is an acute reaction to antivenom:
    - Stop infusion.
    - Treat with standard doses of epinephrine (IM or IV; latter route only in setting of severe hypotension), antihistamines (IV), and glucocorticoids (IV).
    - When reaction is controlled, restart antivenom as soon as possible (may further dilute in larger volume of normal saline).
  - Monitor clinical status over 1 h.
    - Stabilized or improved: Admit to hospital.
    - Progressing or unimproved: Repeat starting dose. Continue this pattern until patient's condition is stabilized or improved.

(*Continued*)

TABLE 451-1 Management of Venomous Snakebite in the United States and Canada[a] (Continued)

- Blood products are rarely needed; if required, they should be given only after antivenom administration.
- Provide tetanus immunization as needed.
- Prophylactic antibiotics are unnecessary unless prehospital care included incision or mouth suction.
- Pain management: Administer acetaminophen and/or opioids as needed; avoid salicylates and nonsteroidal anti-inflammatory agents.
- Admit victim to hospital. (If there is no evidence of envenomation, monitor for 8 h before discharge.)
  - Give additional CroFab (2 vials q6h for 3 additional doses, with close monitoring).
  - Monitor for evidence of rising intracompartmental pressures (see text).
  - Provide wound care (see text).
  - Start physical therapy (see text).
- At discharge, warn patient of possible recurrent coagulopathy and symptoms/signs of serum sickness.

Coral *Snakebites* (*Micrurus* spp. and *Micruroides euryxanthus*)

- Stabilize airway, breathing, and circulation.
- Institute monitoring (vital signs, cardiac rhythm, and oxygen saturation).
- Establish one large-bore IV line and initiate normal saline infusion.
- Take thorough history and perform complete physical examination.
- Identify offending snake if possible.
- Laboratory studies are unlikely to be helpful.
- Contact regional poison-control center.
- Locate and administer antivenom as indicated: Antivenin (*Micrurus fulvius*) (Equine) (commonly referred to as North American Coral Snake Antivenin; Wyeth Pharmaceuticals, New York, NY).[b]
  - Refer to antivenom package insert.
  - Dilute 3–5 reconstituted vials in 250 mL of normal saline.
  - Infuse IV over 1 h (with medical provider in close attendance).
  - If signs of envenomation progress despite initial antivenom, repeat the starting dose; up to 10 vials total may be required.
- If there is an acute adverse reaction to antivenom:
  - Stop infusion.
  - Treat with standard doses of epinephrine (IM or IV; latter route only in setting of severe hypotension), antihistamines (IV), and glucocorticoids (IV).
  - When reaction is controlled, restart antivenom as soon as possible (may further dilute in larger volume of normal saline).
- If there is any evidence of neurologic dysfunction (e.g., any cranial nerve abnormalities such as ptosis):
  - Administer trial of acetylcholinesterase inhibitors (see Table 451-2).
  - With any evidence of difficulty swallowing or breathing, proceed with endotracheal intubation and ventilatory support (may be required for days or weeks).
- Provide tetanus immunization as needed.
- Prophylactic antibiotics are unnecessary unless prehospital care included incision or mouth suction.
- Admit victim to hospital (intensive care unit) even if there is no evidence of envenomation; monitor for at least 24 h.

[a]These recommendations are specific to the care of victims of venomous snakebites in the United States and Canada and should not be applied to bites in other regions of the world. [b]At the time of this writing, a single lot of antivenom remained, with an extended expiration date of January 31, 2018.

Abbreviations: CBC, complete blood count; CK, creatine kinase; FDP, fibrin degradation products; PT/INR/PTT, prothrombin time/international normalized ratio/partial thromboplastin time.

care providers to seek advice from experts in snakebite management regarding indications for and dosing of antivenom.

Antivenom should be administered only by the IV route, and the infusion should be started slowly, with the treating clinician at the bedside ready to immediately intervene at the first signs of an acute adverse reaction. In the absence of an adverse reaction, the rate of infusion can be increased gradually until the full starting dose has been administered (over a total period of ~1 h). Further antivenom may be necessary if the patient's acute clinical condition worsens or fails to stabilize or if venom effects that were initially controlled recur. The decision to administer further antivenom to a stabilized patient should be based on clinical evidence of the persistent circulation of unbound venom components. For viperid bites, antivenom administration generally should be continued until the victim shows definite improvement (e.g., reduced pain, stabilized vital signs, restored coagulation). Neurotoxicity from elapid bites may be more difficult to reverse with antivenom. Once neurotoxicity is established and endotracheal intubation is required, further doses of antivenom are unlikely to be beneficial. In such cases, the victim must be maintained on mechanical ventilation until recovery, which may take days or weeks.

Adverse reactions to antivenom administration include early (anaphylaxis) and late (serum sickness) hypersensitivity reactions. Clinical manifestations of early hypersensitivity range from tachycardia, rigors, vomiting, and urticaria to more serious dyspnea, laryngeal edema, bronchospasm, and hypotension. Skin testing for potential early hypersensitivity, although recommended by some antivenom manufacturers, is neither sensitive nor specific and is of no benefit. Worldwide, the quality of antivenoms is highly variable; rates of acute anaphylactic reactions to some of these products exceed 50%. There is some evidence supporting routine pretreatment with low-dose SC epinephrine (0.25 mg of 1:1000 aqueous solution) to prevent acute anaphylactic reactions after antivenom infusion; although widely practiced, prophylactic use of antihistamines and glucocorticoids has not proved beneficial. Modest expansion of the patient's intravascular volume with crystalloids may blunt episodes of acute hypotension during antivenom infusion. Epinephrine and airway equipment should always be immediately available. If the patient develops an anaphylactic reaction to antivenom, the infusion should be stopped temporarily and the reaction treated immediately with IM epinephrine (0.01 mg/kg up to 0.5 mg), an IV antihistamine (e.g., diphenhydramine, 1 mg/kg up to 50 mg), and a glucocorticoid (e.g., hydrocortisone, 2 mg/kg up to 100 mg). Once the reaction has been controlled, if the severity of the envenomation warrants additional antivenom, the dose should be diluted further in isotonic saline and restarted as soon as possible. In rare cases of refractory hypotension, a concomitant IV infusion of epinephrine may be initiated and titrated to clinical effect while antivenom is administered. The

patient must be monitored very closely during such therapy, preferably in an intensive care setting. Serum sickness typically develops 1–2 weeks after antivenom administration and may present as myalgias, arthralgias, fever, chills, urticaria, lymphadenopathy, and renal or neurologic dysfunction. Treatment for serum sickness consists of systemic glucocorticoids (e.g., oral prednisone, 1–2 mg/kg daily) until all symptoms have resolved, with a subsequent taper over 1–2 weeks. Oral antihistamines and analgesics may provide additional relief of symptoms.

Blood products are rarely necessary in the management of an envenomated patient. The venoms of many snake species can deplete coagulation factors and cause a decrease in platelet count or hematocrit. Nevertheless, these components usually rebound within hours after administration of adequate antivenom. If the need for blood products is thought to be great (e.g., a dangerously low platelet count in a hemorrhaging patient), these products should be given only after adequate antivenom administration to avoid fueling ongoing consumptive coagulopathy.

Rhabdomyolysis should be managed in standard fashion. Victims who develop acute renal failure should be evaluated by a nephrologist and referred for hemodialysis or peritoneal dialysis as needed. Such renal failure, which usually is due to acute tubular necrosis, is frequently reversible. If bilateral cortical necrosis occurs, however, the prognosis for renal recovery is less favorable, and long-term dialysis with possible renal transplantation may be necessary.

Most snake envenomations involve subcutaneous deposition of venom. On occasion, however, venom can be injected more deeply into muscle compartments, particularly if the offending snake was large and the bite occurred on the lower leg, forearm, or hand. Intramuscular swelling of the affected extremity may be accompanied by severe pain, decreased strength, altered sensation, cyanosis, and apparent pulselessness—signs suggesting a muscle compartment syndrome. If there is clinical concern that subfascial muscle edema may be impeding tissue perfusion, intracompartmental pressures should be measured by a minimally invasive technique (e.g., with a wick catheter or digital readout device). If the intracompartmental pressure is high (>30–40 mmHg), the extremity should be kept elevated while antivenom is administered. A dose of IV mannitol (1 g/kg) may be given in an effort to reduce muscle edema if the patient is hemodynamically stable. If the intracompartmental pressure remains elevated after 1 h of such therapy, a surgical consultation should be obtained for possible fasciotomy. Although evidence from animal studies suggests that fasciotomy may actually worsen myonecrosis, compartmental decompression may still be necessary to preserve neurologic function. Fortunately, the incidence of compartment syndrome is very low after snakebite, with fasciotomies required in <1% of cases. Nevertheless, vigilance is essential.

Acetylcholinesterase inhibitors (e.g., edrophonium and neostigmine) may promote neurologic improvement in patients bitten by snakes with postsynaptic neurotoxins. Snakebite victims with objective evidence of neurologic dysfunction should receive a test dose of acetylcholinesterase inhibitors, as outlined in **Table 451-2**. If they exhibit improvement, additional doses of long-acting neostigmine can be administered as needed. Close monitoring is required to prevent aspiration if repetitive dosing of neostigmine is used in an attempt to obviate endotracheal intubation. Acetylcholinesterase inhibitors are not a substitute for administration of appropriate antivenom when available.

Care of the bite wound includes simple cleansing with soap and water; application of a dry, sterile dressing; and splinting of the affected extremity with padding between the digits. Once antivenom therapy has been initiated, the extremity should be elevated above heart level to reduce swelling. Patients should receive tetanus immunization as appropriate. Prophylactic antibiotics are generally unnecessary after bites by North American snakes because the incidence of secondary infection is low. In some regions, secondary bacterial infection is more common and the consequences are serious; in these regions, prophylactic treatment with broad-spectrum antibiotics may be appropriate. Antibiotics may also be considered if misguided first-aid efforts have included incision or mouth suction of the bite site. Pain control should be achieved with acetaminophen or opioid analgesics. Salicylates and nonsteroidal anti-inflammatory agents should be avoided because of their effects on blood clotting.

Wound care in the days after the bite should include careful aseptic debridement of clearly necrotic tissue once coagulopathy has been fully reversed. Intact serum-filled vesicles or hemorrhagic blebs should be left undisturbed. If ruptured, they should be debrided with sterile technique. Any debridement of damaged muscle should be conservative because there is evidence that such muscle may recover to a significant degree after antivenom therapy.

Physical therapy should be started as soon as possible to assist the patient in returning to a functional state. The incidence of long-term loss of function (e.g., reduced range of motion, impaired sensory function) is unclear but is probably quite high (>30%), particularly after viperid bites.

Any patient with signs of envenomation should be observed in the hospital for at least 24 h. In North America, a patient with an apparently "dry" viperid bite should be closely monitored for at least 8 h before discharge, as significant toxicity occasionally develops after a delay of several hours. The onset of systemic symptoms commonly is delayed for a number of hours after bites by certain elapids (including coral snakes, *Micrurus* species), some non–North American viperids (e.g., the hump-nosed pit viper [*Hypnale hypnale*]), and sea snakes. Patients bitten by these snakes should be observed in the hospital for at least 24 h. Admission to an intensive care setting is advised for patients with progressive clinical findings despite initial antivenom administration; those bitten in the head, neck, or other high-risk sites; and those who develop an acute hypersensitivity reaction to antivenom.

At hospital discharge, victims of venomous snakebites should be warned about symptoms and signs of wound infection, antivenom-related serum sickness, and potential long-term sequelae, such as pituitary insufficiency from Russell's viper (*D. russelii*) bites. If coagulopathy developed in the acute stages of envenomation, it can recur during the first 2–3 weeks after the bite; in such cases, victims should be warned to avoid elective surgery or activities posing a high risk of trauma during this period. Outpatient analgesic treatment, wound management, and physical therapy should be provided.

TABLE 451-2 Use of Acetylcholinesterase Inhibitors in Envenomations by Neurotoxic Snakes and Cone Snails

1. Patients with clear, objective evidence of neurotoxicity (e.g., ptosis or inability to maintain upward gaze) should receive a test dose of edrophonium (if available) or neostigmine.
   a. Pretreat with atropine: 0.6 mg IV (children, 0.02 mg/kg with a minimum of 0.1 mg)
   b. Treat with:
      Edrophonium: 10 mg IV (children, 0.25 mg/kg)
      or
      Neostigmine: 0.02 mg/kg IV or IM (children, 0.04 mg/kg)
2. If objective improvement is evident after 30 min, treat with:
   a. Neostigmine: 0.5 mg IV, IM, or SC (children, 0.01 mg/kg) every hour as needed
   b. Atropine: 0.6 mg as IV continuous infusion over 8 h (children, 0.02 mg/kg over 8 h)
3. Closely monitor the airway and perform endotracheal intubation as needed.

MORBIDITY AND MORTALITY

The overall mortality rates for victims of venomous snakebites are low in regions with rapid access to medical care and appropriate antivenoms. In the United States, for example, the mortality rate is <1% for victims who receive antivenom. Eastern and western diamondback rattlesnakes (*Crotalus adamanteus* and *Crotalus atrox*, respectively) are responsible for the majority of snakebite deaths in the United States. Snakes responsible

for large numbers of deaths in other countries include cobras (*Naja* species), carpet and saw-scaled vipers (*Echis* species), Russell's vipers (*D. russelii*), large African vipers (*Bitis* species), lancehead pit vipers (*Bothrops* species), and tropical rattlesnakes (*C. durissus*).

The incidence of morbidity—defined as permanent functional loss in a bitten extremity—is difficult to estimate but is substantial. Morbidity may be due to muscle, nerve, or vascular injury or to scar contracture. Such morbidity can have devastating consequences for victims in the developing world when they lose the ability to work and provide for their families. In the United States, functional loss tends to be more common and severe after rattlesnake bites than after bites by copperheads (*Agkistrodon contortrix*) or water moccasins (*Agkistrodon piscivorus*).

GLOBAL CONSIDERATIONS

In many developing countries where snakebites are common, limited access to medical care and antivenoms contributes to high rates of morbidity and mortality. Often, the available antivenoms are inappropriate and ineffective against the venoms of medically important indigenous snakes. In those regions, further research is necessary to determine the actual impact of venomous snakebites and the specific antivenoms needed in terms of both quantity and spectrum of coverage. Without accurate statistics, it is difficult to persuade antivenom manufacturers to begin and sustain production of appropriate antisera in developing nations. There is evidence that antivenoms can be produced by much more cost-effective methods than those currently being employed. Just as important as getting the correct antivenoms into underserved regions is the need to educate populations about snakebite prevention and train medical care providers in proper management approaches. Local protocols written with significant input from experienced providers in the region of concern should be developed and distributed. Appropriate antivenoms must be available at the most likely first point of medical contact for patients (e.g., primary health centers) in order to minimize the common practice of referring victims to more distant, higher levels of care for the initiation of antivenom therapy. Those who care for snakebite victims in these often-remote clinics must have the skills and confidence required to begin antivenom treatment (and to treat possible reactions) as soon as possible when needed.

MARINE ENVENOMATIONS

Much of the management of envenomation by marine animals is supportive in nature. A subset of envenomations can threaten life or limb. Specific antivenom can be used when appropriate.

INVERTEBRATES

Cnidarians Cnidarians, such as hydroids, fire coral, jellyfish, Portuguese men-of-war, and sea anemones, produce specialized living stinging organelles called *cnidocysts* (a term that encompasses nematocysts, ptychocysts, and spirocysts). In the stinging process, cnidocysts are released and discharged upon mechanosensory stimulation. The venom from these organisms contains bioactive substances, such as tetramine, 5-hydroxytryptamine, histamine, serotonin, and high-molecular-weight toxins, all of which can, among other effects, change the permeability of cells to ions. Victims usually report immediate prickling or burning, pruritus, paresthesias, and painful throbbing with radiation. The skin becomes reddened, darkened, edematous, and blistered and may show signs of superficial necrosis. A legion of neurologic, cardiovascular, respiratory, rheumatologic, gastrointestinal, renal, and ocular symptoms have been described, especially following stings from anemones, *Physalia* species, and scyphozoans. Anaphylaxis is possible. Hundreds of deaths have been reported, many of them caused by *Chironex fleckeri, Stomolophus nomurai, Physalia physalis*, and *Chiropsalmus quadrumanus*. Irukandji syndrome, associated with the Australian jellyfish *Carukia barnesi* and other species, is a potentially fatal condition that most commonly is characterized by hypertension; severe back, chest, and abdominal pain; nausea and vomiting; headache; sweating; and, in the most serious cases, myocardial troponin leak, pulmonary edema, and ultimately hypotension. This syndrome is thought to be mediated, at least in part, by the release of endogenous catecholamines followed by cytokines and nitric oxide.

Envenomations by different cnidarians (typified by jellyfish) may respond differently to similar topical therapies; thus, the recommendations in this chapter must be tailored to local species and clinical practices. During stabilization, the skin should be decontaminated immediately with a generous application of saline. Vinegar (5% acetic acid) is an all-purpose agent that appears to be useful for relieving pain caused by a large number of species. If unavailable, rubbing alcohol (40–70% isopropyl alcohol) can sometimes be used, although some data show it may worsen nematocyst discharge in particular species.

For the sting of the venomous box jellyfish (*C. fleckeri*), vinegar should be used, followed by local application of heat (up to 45°C/113°F) by immersion in hot water. Hot-water application is first-line treatment for mild to moderate *Physalia utriculus* (bluebottle jellyfish) stings. If hot water is unavailable, commercial (chemical) cold packs or ice packs applied over a thin dry cloth or plastic membrane are effective in alleviating bluebottle jellyfish stings. In general, rubbing leads to further stinging by adherent cnidocysts and should be avoided.

After decontamination, topical application of an anesthetic ointment (e.g., lidocaine, benzocaine), an antihistamine (e.g., diphenhydramine), or a glucocorticoid (e.g., hydrocortisone) may be helpful for symptom control. Persistent severe pain after decontamination may be treated with IV or oral opioid analgesics. Muscle spasms may respond to IV diazepam (2–5 mg, titrated upward as necessary). An antivenom is available from Commonwealth Serum Laboratories for stings from the box jellyfish found in Australian and Indo-Pacific waters. However, despite its reported clinical efficacy, some in vitro studies suggest that this antivenom cannot bind venom rapidly enough to account for its effects. Until further notice, current recommendations for its use apply (see "Sources of Antivenoms and Other Assistance," below). Treatment of Irukandji syndrome may require administration of opioid analgesics and aggressive treatment of hypertension (e.g., phentolamine, 5 mg IV). All victims with systemic reactions should be observed for at least 6–8 h for rebound from any therapy, and all elderly patients should be assessed for cardiac arrhythmias. Patients may suffer post-inflammatory hyperpigmentation and persistent cutaneous hypersensitivity in areas of skin contact.

Safe Sea, a "jellyfish-safe" sunblock (*www.nidaria.com*) applied to the skin before an individual enters the water, inactivates the recognition and discharge mechanisms of nematocysts; it has been tested successfully against a number of marine stingers and may prevent or diminish the effects of coelenterate stings. Whenever possible, a dive skin or wet suit should be worn when entering ocean waters.

Sea Sponges Many sponges produce crinotoxins. As a result, touching a sea sponge may result in allergic contact dermatitis. Irritant dermatitis may result if the sponge's small spicules of silica or calcium carbonate penetrate the skin. It is impossible to distinguish between the allergic and spicule reactions, so the treatment is the same for both. Afflicted skin should be gently dried and adhesive tape, a commercial facial peel, or a thin layer of rubber cement used to remove embedded spicules. Vinegar should then be applied immediately, with repeated application for 10–30 min three or four times a day thereafter. Rubbing alcohol may be used if vinegar is unavailable. After spicule removal and skin decontamination, glucocorticoid or antihistamine cream may be applied to the skin. Severe vesiculation should be treated with a 2-week tapering course of systemic glucocorticoids. Mild reactions subside in 3–7 days, while involvement of large areas of the skin may result in systemic symptoms of fever, dizziness, nausea, muscle cramps, and formication.

Annelid Worms Annelid worms (bristleworms) possess rows of soft, cactus-like spines capable of inflicting painful stings. Contact results in symptoms similar to those of nematocyst envenomation. Without treatment, pain usually subsides over several hours, but inflammation may persist for up to a week **(Fig. 451-4)**. Victims should resist the urge to scratch because scratching may fracture retrievable spines. Visible bristles should be removed with forceps and adhesive tape or a commercial facial peel; alternatively, a thin layer of rubber

FIGURE 451-4 Rash on the hand of a diver from the spines of a bristleworm. *(Courtesy of Paul Auerbach, with permission.)*

FIGURE 451-5 Spines on the crown-of-thorns sea star (*Acanthaster planci*). *(Courtesy of Paul Auerbach, with permission.)*

cement can be used to entrap and then peel away the spines. Use of vinegar or rubbing alcohol or a brief application of lidocaine may provide additional relief. Local inflammation should be treated with topical or systemic glucocorticoids.

Sea Urchins Venomous sea urchins possess either hollow, venom-filled calcified spines or triple-jawed, globiferous pedicellariae with venom glands. Venom may also be found within the integumentary sheath on the external spine surface of certain species. The venom contains toxic components, including steroid glycosides, hemolysins, proteases, serotonin, and cholinergic substances. Contact with either venom apparatus produces immediate and intensely painful stings. Spines that enter a joint can cause synovitis that may progress to arthritis if the spines remain in or near the joint. If multiple spines penetrate the skin, the patient may develop systemic symptoms, including nausea, vomiting, numbness, muscular paralysis, and respiratory distress. A delayed hypersensitivity reaction 7–10 days after resolution of primary symptoms has been described.

The affected part should be immersed immediately in hot water to tolerance (up to 45°C/113°F). Pedicellariae should be removed by shaving so that envenomation cannot continue. Accessible embedded spines should be removed with care as they may fracture and leave remnants lodged in the victim. Residual dye from the surface of a spine remaining after the spine's removal may mimic a retained spine but is otherwise of no consequence. Soft-tissue radiography, ultrasonography, or MRI can confirm the presence of retained spines, which may warrant referral for surgical removal if the spines are near vital structures (e.g., joints, neurovascular bundles). Retained spines can cause the formation of granulomas that are amenable to excision or to intralesional injection with triamcinolone hexacetonide (5 mg/mL). Chronic granulomatous arthritis of the proximal interphalangeal joints has been treated with synovectomy and removal of granulation tissue. Eosinophilic pneumonia and local and diffuse neuropathies have been observed separately after penetration by multiple spines of the black sea urchin (presumed *Diadema* species). The pathophysiologies of these phenomena have not been determined.

Starfish The crown-of-thorns *Acanthaster planci* produces venom in glandular tissue underneath the epidermis, which is released via its spiny surfaces **(Fig. 451-5)**. Skin puncture causes pain, bleeding, and local edema, usually with remission over 30–180 min. Multiple punctures may result in reactions such as local muscle paralysis; retained fragments may cause granulomatous lesions and synovitis. There has also been a case report of elevated liver enzymes after *A. planci* envenomation. Envenomated persons benefit from acute immersion therapy in hot water, local anesthesia, wound cleansing, imaging, and possible exploration to remove spines and foreign material.

Sea Cucumbers Sea cucumbers produce holothurin (a cantharidin-like liquid toxin) in their body walls. This toxin is concentrated in the tentacular organs that are projected when the animal is threatened. Underwater, holothurin induces minimal contact dermatitis in the skin but can cause significant corneal and conjunctival irritation should ocular contact occur. A severe reaction can lead to blindness. Skin should be detoxified with vinegar or isopropyl alcohol. The eye should be anesthetized with 1–2 drops of 0.5% proparacaine and irrigated copiously with normal saline, with subsequent slit-lamp examination to identify corneal defects.

Cone Snails Cone snails use a detachable dart-like tooth to inject conotoxins into victims, inducing paralysis. Punctures result in small, painful wounds followed by local ischemia, cyanosis, and numbness. Syncope, dysphagia, dysarthria, ptosis, blurred vision, and pruritus also have been documented. Some envenomations induce paralysis leading to respiratory failure, coma, and death. There is no antivenom for treatment. Pressure immobilization (see "Octopuses," below), hot-water soaks, and local anesthetics have been used empirically with success. The wound should be inspected for a foreign body. Edrophonium has been recommended as therapy for paralysis if an edrophonium (Tensilon) test is positive (see Table 451-2).

Octopuses Serious envenomations and deaths have followed bites of Australian blue-ringed octopuses (*Octopus maculosa* and *Octopus lunulata*). Although these animals rarely exceed 20 cm in length, their salivary venom contains a potent neurotoxin (maculotoxin) that inhibits peripheral-nerve transmission by blocking sodium conductance. Oral and facial numbness develop within several minutes of a serious envenomation and rapidly progress to total flaccid paralysis, including failure of respiratory muscles. Immediately after envenomation, a circumferential pressure-immobilization dressing 15 cm wide should be applied over a gauze pad (~7 × 7 × 2 cm) that has been placed directly over the sting. The dressing should be applied at venous-lymphatic pressure, with the preservation of distal arterial pulses. The limb should then be splinted. Once the victim has been transported to the nearest medical facility, the bandage can be released. There is no antidote and treatment is supportive. Patients with respiratory failure may need to be mechanically ventilated. The victim may remain awake although completely paralyzed, so analgesia and sedation should be provided as needed. Even with serious envenomations, significant recovery often takes place within 4–10 h, although complete recovery may require 2–4 days. Sequelae are uncommon unless related to hypoxia.

VERTEBRATES

As for all penetrating injuries, first-aid care should be undertaken. In addition, consideration must be given to local wound infection by marine *Vibrio* species and freshwater *Aeromonas hydrophila* as well as other "aquatic bacteria," particularly if spines and needles remain embedded.

Stingrays A stingray injury is both an envenomation and a traumatic wound. Thoracic and cardiac penetration, major vessel laceration, and compartment syndrome have all been observed. The venom, which contains serotonin, 5′-nucleotidase, and phosphodiesterase, causes immediate, intense pain that peaks at 30–60 min and may last

up to 48 h. The wound often becomes ischemic in appearance and heals poorly, with adjacent soft-tissue swelling and prolonged disability. Systemic effects include weakness, diaphoresis, nausea, vomiting, diarrhea, dysrhythmias, syncope, hypotension, muscle cramps, fasciculations, paralysis, and (in rare cases) death. Because of differences in the toxins present on the tissues covering the stingers, freshwater stingrays may cause more severe injuries than marine stingrays.

Scorpionfish The designation *scorpionfish* encompasses members of the family Scorpaenidae and includes not only scorpionfish but also lionfish and stonefish. A complex venom with neuromuscular toxicity is delivered through 12 or 13 dorsal, 2 pelvic, and 3 anal spines. In general, the sting of a stonefish is regarded as the most serious (severe to life-threatening); that of the scorpionfish is of intermediate seriousness; and that of the lionfish is the least serious. Like that of a stingray, the sting of a scorpionfish is immediately and intensely painful. Pain from a stonefish envenomation may last for days and can be sufficiently intense to cause delirium. Systemic manifestations of scorpionfish stings are similar to those of stingray envenomations but may be more pronounced, particularly in the case of a stonefish sting, which may cause severe local tissue necrosis in addition to vital organ failure. The rare deaths that follow stonefish envenomation usually occur within 6–8 h. A commercially available stonefish antivenom from Commonwealth Serum Laboratories (see "Sources of Antivenoms and Other Assistance," below) can be used in cases of severe envenomation.

Other Fish Three species of marine catfish—*Plotosus lineatus* (oriental catfish), *Bagre marinus* (sail catfish), and *Galeichthys felis* (common sea catfish)—as well as several species of freshwater catfish are capable of stinging humans. Venom is delivered through a single dorsal spine and two pectoral spines. Clinically, a catfish sting is comparable to that of a stingray, although marine catfish envenomations are generally more severe than those of their freshwater counterparts. Surgeonfish (doctorfish, tang), weeverfish, ratfish, and horned venomous sharks also have the capacity to envenomate humans.

Platypuses The platypus is a venomous mammal. The male has a keratinous spur on each hind limb; the spur is connected to a venom gland within the upper thigh. Skin puncture causes soft-tissue edema and pain that may last for days or weeks. Care is supportive, and hot-water therapy does not appear to benefit the victim.

TREATMENT

Marine Vertebrate Stings

The stings of all marine vertebrates are treated in a similar fashion. Except for stonefish and serious scorpionfish envenomations, no antivenom is available. The affected part should be immersed immediately in non-scalding hot water (45°C/113°F) for 30–90 min or until there is significant pain relief. Recurrent pain may respond to repeated hot-water treatment. Cryotherapy is contraindicated, and no data support the use of antihistamines or steroids. Systemic opioids as well as local wound infiltration or regional nerve block with lidocaine or bupivacaine can help alleviate pain. After soaking, the wound must be explored and debrided. Radiography (in particular, ultrasound or MRI) may be helpful in identification of foreign bodies. After exploration and debridement, the wound should be irrigated vigorously with warm sterile water, saline, or 1% povidone-iodine in solution. Bleeding usually can be controlled by sustained local pressure. In general, wounds should be left open to heal by secondary intention or treated by delayed primary closure. Tetanus immunization should be provided as appropriate. Antibiotic treatment should be considered for serious wounds and for envenomations in immunocompromised hosts. The initial antibiotics should cover *Staphylococcus* and *Streptococcus* species. If the victim is immunocompromised, if a wound is primarily repaired, or if an infection develops, antibiotic coverage should be broadened to include *Vibrio* species. Infection with *Aeromonas* species is of similar concern for wounds sustained in freshwater.

APPROACH TO THE PATIENT

Marine Envenomations

It is useful to be familiar with the local marine fauna and to recognize patterns of injury.

A large puncture wound or jagged laceration (particularly on the lower extremity) that is more painful than one would expect from the size and configuration of the wound is likely to be a stingray envenomation. Smaller punctures, sometimes associated with purple or dark discoloration, represent the activity of a sea urchin **(Fig. 451-6)** or starfish. Stony corals cause rough abrasions and, in rare instances, lacerations or puncture wounds.

Coelenterate (marine invertebrate) stings sometimes create diagnostic skin patterns. A diffuse urticarial rash on exposed skin is often indicative of exposure to fragmented hydroids or larval anemones. A linear, whip-like print pattern appears where a jellyfish tentacle has contacted the skin. In the case of the box jellyfish, a cross-hatched appearance, followed by development of dark purple coloration within a few hours of the sting, heralds skin necrosis. An encounter with fire coral causes immediate pain and swollen red skin irritation in the pattern of contact, similar to but more severe than the imprint left by exposure to an intact feather hydroid. Seabather's eruption, caused by thimble jellyfish and larval anemones, is a diffuse, intensely pruritic rash consisting of clusters of erythematous macules or raised papules that follow the pattern of bathing attire **(Fig. 451-7)**. Toxic sponges create a burning and painful red rash

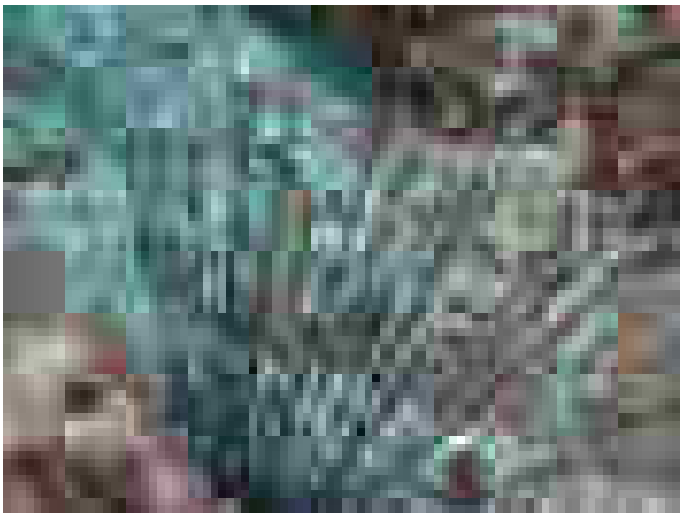


FIGURE 451-6 Spiny sea urchins. *(Courtesy of Paul Auerbach, with permission.)*

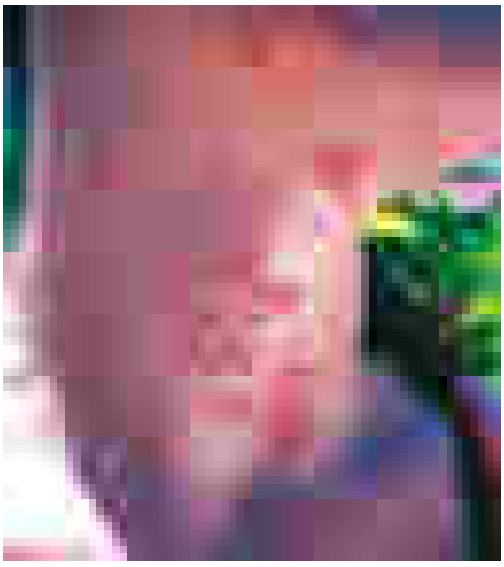


FIGURE 451-7 Erythematous, papular rash typical of seabather's eruption caused by thimble jellyfish and larval anemones. *(Courtesy of Paul Auerbach, with permission.)*

on exposed skin, which may blister and later desquamate. Virtually all marine stingers invoke the sequelae of inflammation; thus local erythema, swelling, and adenopathy are fairly nonspecific.

SOURCES OF ANTIVENOMS AND OTHER ASSISTANCE

In the United States, assistance in locating a specific antivenom can be obtained from a regional poison-control center. Divers Alert Network, a nonprofit organization designed to assist in the care of injured divers, also may help with the treatment of marine injuries. The network can be reached on the Internet at *www.diversalertnetwork.org* or by telephone 24 h a day at (919) 684-9111. Antivenom for box jellyfish (*C. fleckeri*) and stonefish (and severe scorpionfish) envenomations is made in Australia by Commonwealth Serum Laboratories (CSL; 45 Poplar Road, Parkville, Victoria, Australia 3052; *www.csl.com.au*; 61-3-9389-1911). When administering the box jellyfish antivenom, time is of the essence. For cardiac or respiratory decompensation, a minimum of 1 ampoule and up to 6 ampoules consecutively should be given IV, preferably in a 1:10 dilution with normal saline. For stonefish (or severe scorpionfish) envenomation, 1 ampoule of specific antivenom should be administered IM for every one or two punctures, to a maximum of 3 ampoules.

MARINE POISONINGS

CIGUATERA

Epidemiology and Pathogenesis Ciguatera poisoning is the most common nonbacterial food poisoning associated with fish in the United States; most U.S. cases occur in Florida and Hawaii, although, with transportation of imported fish worldwide, all clinicians need to be aware of ciguatera. The poisoning almost exclusively involves tropical coral reef fish common in the Indian Ocean, the South Pacific, and the Caribbean Sea. Global estimates predict that 20,000–50,000 people may be affected by this poisoning each year, although it is suspected that 90% of cases go unreported. More than 400 different fish have been associated with ciguatera toxicity, but 75% of poisonings involve the reef-dwelling barracuda, snapper, jack, or grouper. Ciguatera toxin is created by warm-water marine dinoflagellates, primarily of the genus *Gambierdiscus toxicus*, whose consumption by grazing fish allows the toxin to bioaccumulate in the food chain. Ciguatoxins act on neuron voltage-gated sodium channels, most notably producing cardiac, gastrointestinal, and neurologic symptoms. These toxins are unaffected by freeze-drying, heat, cold, or gastric acid, and the fact that none of the toxins affects the odor, color, or taste of fish makes identification difficult. Toxins are found in the highest concentrations in the fish's skin, head, and viscera; therefore, consumption of these portions should be avoided.

TABLE 451-3 Representative Symptoms and Signs of Ciguatera Poisoning

SYSTEM	SYMPTOMS/SIGNS
Gastrointestinal	Abdominal pain, nausea, vomiting, diarrhea
Neurologic	Paresthesias, pruritus, tongue and throat numbness or burning, sensation of "carbonation" during swallowing, odontalgia or dental dysesthesias, dysphagia, tremor, fasciculations, athetosis, meningismus, aphonia, ataxia, vertigo, pain and weakness in the lower extremities, visual blurring, transient blindness, hyporeflexia, seizures, coma
Dermatologic	Conjunctivitis, maculopapular rash, skin vesiculations, dermographism
Cardiovascular	Bradycardia, heart block, hypotension, central respiratory failure[a]
Other	Chills, dysuria, dyspnea, dyspareunia, weakness, fatigue, nasal congestion and dryness, insomnia, hypersalivation, diaphoresis, headache, arthralgias, myalgias

[a]Tachycardia and hypertension may occur after potentially severe transient bradycardia and hypotension. Death is rare.

Clinical Manifestations Symptoms may develop within 15–30 min of ingestion but typically do so within 2–6 h and increase in severity over the ensuing 4–6 h. Most victims develop symptoms within 12 h of ingestion, and virtually all are afflicted within 24 h. The more than 150 symptoms and signs reported include those shown in Table 451-3. Diarrhea, vomiting, and abdominal pain usually occur first and develop 3–6 h after ingestion. Bradycardia, hypotension, dysesthesias, and paresthesias may occur. Symptoms may persist for 48 h and then generally resolve (even without treatment). A pathognomonic symptom is the reversal of hot and cold tactile perception, which develops in some persons after 3–5 days and may last for months. More severe reactions tend to occur on repeat exposure. Persons who have ingested parrotfish (scaritoxin) may develop classic ciguatera poisoning as well as a "second-phase" syndrome (after a delay of 5–10 days) of disequilibrium with locomotor ataxia, dysmetria, and resting or kinetic tremor. This syndrome may persist for 2–6 weeks.

Diagnosis The differential diagnosis of ciguatera includes paralytic shellfish poisoning, eosinophilic meningitis, type E botulism, organophosphate insecticide poisoning, tetrodotoxin poisoning, and psychogenic hyperventilation. At present, the diagnosis of ciguatera poisoning is made on clinical grounds. Liquid chromatography–mass spectrometry is available for ciguatoxins, and a ciguatoxin enzyme immunoassay or radioimmunoassay may be used to test suspected fish; however, these tests are of limited clinical value because most health care institutions do not have the equipment needed for their performance.

TREATMENT

Ciguatera Poisoning

Therapy is supportive and based on symptoms. Hypotension should be treated with IV crystalloid and, in rare cases, a vasopressor. Bradyarrhythmias that lead to cardiac insufficiency and hypotension generally respond well to atropine (0.5 mg IV, up to 2 mg). Nausea and vomiting may be controlled with an antiemetic such as ondansetron (4–8 mg IV). Cool showers or the administration of hydroxyzine (25 mg orally every 6–8 h) may relieve pruritus. Amitriptyline (25 mg orally twice a day) reportedly alleviates pruritus and dysesthesias and may decrease rates of subsequent development of chronic nerve symptoms. Gabapentin has shown some efficacy in the treatment of long-term sequelae of nerve pain. IV infusion of mannitol may be beneficial in moderate or severe cases in fluid-repleted patients, particularly for the relief of distressing neurologic or cardiovascular symptoms; however, the efficacy of this therapy has been challenged and has not been definitively proven. An initial IV dose of mannitol at 1 g/kg may be given over 45–60 min. If symptoms are alleviated,

a second dose may be given within 3–4 h and a third dose the next day. Care must be taken to avoid dehydration. The mechanism of the drug's benefit against ciguatera intoxication is perhaps hyperosmotic water-drawing action, which reverses ciguatoxin-induced Schwann cell edema. Mannitol may also act in some fashion as a "hydroxyl scavenger" or may competitively inhibit ciguatoxin at the cell membrane. Activated charcoal is not recommended for ciguatera poisoning.

During recovery from ciguatera poisoning, the victim should exclude the following from the diet for 6 months: fish (fresh or preserved), fish sauces, shellfish, shellfish sauces, alcoholic beverages, nuts, and nut oils. Consumption of fish in ciguatera-endemic regions should be avoided. All oversized fish of any predacious reef species should be suspected of harboring ciguatoxin. Neither moray eels nor the viscera, head, or skin of tropical marine fish should be eaten.

DIARRHETIC SHELLFISH POISONING

Diarrhetic shellfish poisoning occurs with consumption of shellfish producing diarrheal illness. The causative agents are the lipophilic compound okadaic acid and the dinophysistoxins, which inhibit serine and threonine protein phosphatases, with consequent protein accumulation and continued secretion of fluid by intestinal cells leading to diarrhea. Shellfish acquire these toxins by feeding on dinoflagellates, particularly of the genera *Dinophysis* and *Prorocentrum*.

Symptoms include diarrhea, nausea, vomiting, abdominal pain, and chills. Onset occurs within 30 min to 12 h. The illness is usually self-limited; most patients recover in 3–4 days and only a few require hospitalization. Treatment is supportive and focused on hydration. Toxins can be detected in food samples by a mouse bioassay, an immunoassay, and fluorometric high-performance liquid chromatography (HPLC).

PARALYTIC SHELLFISH POISONING

Paralytic shellfish poisoning is induced by ingestion of any of a variety of feral or aquacultured filter-feeding organisms, including clams, oysters, scallops, mussels, chitons, limpets, starfish, and sand crabs. The source of their toxicity is the chemical toxin they accumulate and concentrate by feeding on various planktonic dinoflagellates (e.g., *Alexandrium*, *Pyrodinium*, *Gymnodinium*). These unicellular phytoplanktonic organisms form the foundation of the food chain, and in warm summer months these organisms "bloom" in nutrient-rich coastal temperate and semitropical waters. In the United States, paralytic shellfish poisoning is acquired primarily from seafood harvested in the Northeast, the Pacific Northwest, and Alaska. These planktonic species can release massive amounts of toxic metabolites into the water and cause mortality in bird and marine populations. The paralytic shellfish toxins are water soluble as well as heat and acid stable; they cannot be destroyed by ordinary cooking or freezing. Contaminated seafood looks, smells, and tastes normal. The best-characterized, most potent, and most frequently identified paralytic shellfish toxin is saxitoxin, which appears to block sodium conductance, inhibiting neuromuscular transmission at the axonal and muscle membrane levels. A toxin concentration of >75 µg/100 g of foodstuff is considered hazardous to humans. During a "red tide" algal bloom, the concentration of saxitoxin in shellfish can exceed 9000 µg/100 g. A direct human serum assay to identify the toxin responsible for paralytic shellfish poisoning is not yet clinically available; a mouse bioassay that identifies saxitoxin in suspected shellfish is currently in use.

Intraoral and perioral paresthesias (notably of the lips, tongue, and gums) can occur within minutes to a few hours after ingestion of contaminated shellfish, and these paresthesias can progress rapidly to involve the neck and distal extremities. Other symptoms develop rapidly and can include lightheadedness, disequilibrium, incoordination, weakness, hyperreflexia, incoherence, dysarthria, hypersalivation, dysphagia, thirst, diarrhea, abdominal pain, nausea, vomiting, nystagmus, dysmetria, headache, diaphoresis, loss of vision, chest pain, and tachycardia. Flaccid paralysis and respiratory insufficiency may follow 2–12 h after ingestion. In the absence of hypoxia, the victim often remains alert but paralyzed. Up to 12% of patients may die.

TREATMENT

Paralytic Shellfish Poisoning

Treatment is supportive and based on symptoms. If the victim comes to medical attention within the first few hours after poison ingestion, the stomach should be emptied by gastric lavage and then irrigated with 2 L (in 200-mL aliquots) of a solution of 2% sodium bicarbonate; this intervention has not been proven to be of benefit but is based on the notion that gastric acidity may enhance the potency of saxitoxin. Because respiratory insufficiency can be rapid in onset, induction of emesis is not advised. The administration of activated charcoal (50–100 g) and a cathartic (e.g., sorbitol, 20–50 g) makes empirical sense because these shellfish toxins are believed to bind well to charcoal. Some authors advise against administration of magnesium-based solutions (e.g., certain cathartics), cautioning that hypermagnesemia may contribute to suppression of nerve conduction.

The most serious concern is respiratory paralysis. The victim should be closely observed for respiratory distress for at least 24 h in a hospital. With prompt recognition of ventilatory failure and establishment of endotracheal intubation and assisted ventilation, anoxic myocardial and brain injury may be prevented. If the patient survives for 18 h, the prognosis is good for a complete recovery.

DOMOIC ACID POISONING (AMNESIC SHELLFISH POISONING)

Domoic acid poisoning occurs when humans consume shellfish containing the marine toxin domoic acid. The toxin is produced by marine diatoms of the *Pseudonitzchia* species and during algal blooms can bioaccumulate in filter-feeding shellfish. Clams, mussels, oysters, anchovies, and Dungeness crabs have all been found to cause domoic acid poisoning. A water-soluble, heat-stable neuroexcitatory amino acid with biochemical analogues of kainic acid and glutamic acid, domoic acid binds to the kainate type of glutamate receptor with 3 times the affinity and 20 times the toxicity of kainic acid. The toxin is heat stable and is not affected by cooking or freezing. Shellfish can be tested for domoic acid by mouse bioassay and HPLC. The regulatory limit for domoic acid in shellfish is 20 parts per million.

The abnormalities noted within 24 h of ingesting contaminated shellfish include arousal, confusion, disorientation, and memory loss. The median time of onset is 5.5 h. Other prominent symptoms and signs include severe headache, nausea, vomiting, diarrhea, abdominal cramps, hiccups, arrhythmias, hypotension, seizures, ophthalmoplegia, pupillary dilation, piloerection, hemiparesis, mutism, grimacing, agitation, emotional lability, coma, copious bronchial secretions, and pulmonary edema. Histologic study of brain tissue taken at autopsy has shown neuronal necrosis or cell loss and astrocytosis, most prominently in the hippocampus and amygdaloid nucleus—findings similar to those in animals poisoned with kainic acid. Several months after the primary intoxication, victims may still display chronic residual memory deficits and motor neuropathy or axonopathy. Nonneurologic sequelae generally do not persist.

TREATMENT

Domoic Acid Intoxication

Therapy is supportive and based on symptoms. IV fluids and antiemetics may be used for severe nausea, vomiting, and diarrhea. Because kainic acid neuropathology seems to be nearly entirely seizure mediated, the emphasis should be on anticonvulsive therapy; benzodiazepines, propofol, and barbiturates may be used.

HISTAMINE (SCOMBROID) FISH POISONING

Histamine fish poisoning, most often referred to as scombroid or pseudoallergenic fish poisoning, may be the most common type of seafood poisoning worldwide. It follows consumption of both scombroid (mackerel-like) fish (including albacore, bluefin, and yellowfin

tuna; mackerel; saury; needlefish; wahoo; skipjack; and bonito) and nonscombroid fish (including dolphinfish, kahawai, sardine, black marlin, pilchard, anchovy, herring, amberjack, Australian ocean salmon, and bluefish).

Under conditions of inadequate preservation or refrigeration, the musculature of these dark- or red-fleshed fish undergoes decomposition by *Morganella morganii*, *Escherichia coli*, *Proteus* species, and *Klebsiella* species, with consequent decarboxylation of the amino acid L-histidine to histamine, histamine phosphate, and histamine hydrochloride. Histamine levels of 20–50 mg/100 g are noted in toxic fish, with levels >400 mg/100 g on occasion. Toxic levels can be reached with as few as 12 h of inadequate refrigeration. A second compound is thought to play a causative role in this intoxication because large doses of oral histamine do not reproduce the affliction. It is proposed that this unknown agent (possibly saurine, cadaverine, or putrescine) works by inhibiting the metabolism of histamine, promoting degranulation of mast cells to release endogenous histamine, or acting as a histamine receptor agonist. The toxin or toxins involved are heat stable and are not destroyed by cooking or freezing. Affected fish may have a sharply metallic or peppery taste, although more often they are normal in appearance, color, and flavor. Not all persons who eat a contaminated fish necessarily become ill, perhaps because of uneven distribution of decay within the fish.

Symptoms develop within 15–90 min of ingestion. Most cases are mild, with tingling of lips and mouth, mild abdominal discomfort, and nausea. The more severe and commonly described presentation includes flushing (sharply demarcated; exacerbated by ultraviolet exposure; particularly pronounced on the face, neck, and upper trunk), a sensation of warmth, conjunctival hyperemia, pruritus, urticaria, and angioedema. This syndrome may progress to bronchospasm, nausea, vomiting, diarrhea, epigastric pain, abdominal cramps, dysphagia, headache, palpitations, tachycardia, dizziness, hypotension and cardiogenic shock. Without treatment, the symptoms generally resolve within 8–12 h. Because of blockade of gastrointestinal tract histaminase, the reaction may be more severe in a person who is concurrently ingesting isoniazid.

TREATMENT

Scombroid Poisoning

Therapy is directed at reversing the histamine effect with either oral or IV antihistamines. If bronchospasm is severe, an inhaled bronchodilator—or in rare, severe circumstances, injected epinephrine—may be used. The use of activated charcoal is not recommended. Protracted nausea and vomiting may be controlled with a specific antiemetic, such as ondansetron or prochlorperazine. Hypotension should be treated with IV fluids. The persistent headache of scombroid poisoning may respond to cimetidine or a similar antihistamine if standard analgesics are not effective. It is important to inform the patient that the symptoms are related to eating improperly refrigerated fish and are not due to a fish allergy.

FURTHER READING

Auerbach PS et al: Envenomation by aquatic invertebrates, in *Auerbach's Wilderness Medicine*, 7th ed. PS Auerbach et al (eds). Philadelphia, Elsevier, 2017, pp 1679–1720.

Auerbach PS et al: Envenomations by aquatic vertebrates, in *Auerbach's Wilderness Medicine*, 7th ed. PS Auerbach et al (eds). Philadelphia, Elsevier, 2017, pp 1721–1742.

Currie BJ: Marine antivenoms. J Toxicol Clin Toxicol 41:301, 2003.

Fenner P: Marine envenomation: An update—a presentation on the current status of marine envenomation first aid and medical treatments. Emerg Med Australas 12:295, 2000.

Kanaan NC et al: Wilderness Medical Society practice guidelines for the treatment of pitviper envenomations in the United States and Canada. Wilderness Environ Med 26:472, 2015.

Minns AB et al: Seafood toxidromes, in *Auerbach's Wilderness Medicine*, 7th ed. PS Auerbach et al (eds). Philadelphia, Elsevier, 2017, pp 1766–1793.

Norris RL et al: Bites by venomous reptiles in Canada, the United States, and Mexico, in *Auerbach's Wilderness Medicine*, 7th ed. PS Auerbach et al (eds). Philadelphia, Elsevier, 2017, pp 729–759.

Warrell DA: Venomous bites, stings, and poisoning. Infect Dis Clin North Am 26:207, 2012.

Warrell DA: Bites by venomous and nonvenomous reptiles worldwide, in *Auerbach's Wilderness Medicine*, 7th ed. PS Auerbach et al (eds.). Philadelphia, Elsevier, 2017, pp 760–827.

452 Ectoparasite Infestations and Arthropod Injuries

Richard J. Pollack, Scott A. Norton

Ectoparasites include arthropods and creatures from other phyla that infest the skin or hair of animals; the host animals provide them with sustenance and shelter. The ectoparasites may penetrate within or beneath the surface of the host or may attach by mouthparts and specialized claws. These organisms may inflict direct mechanical injury, consume blood or nutrients, induce hypersensitivity reactions, inoculate toxins, transmit pathogens, create openings in the skin for secondary bacterial infection, and incite fear or disgust. Human beings are the sole or obligate hosts for only a few kinds of ectoparasites but serve as facultative, dead-end, or paratenic (accidental) hosts for many others.

Arthropods that are ectoparasitic or otherwise cause injury include insects (such as lice, fleas, bedbugs, wasps, ants, bees, and flies), arachnids (spiders, scorpions, mites, and ticks), millipedes, and centipedes. Certain nematodes (helminths), such as the hookworms (Chap. 226), are ectoparasitic in that they penetrate and migrate through the skin. Infrequently encountered ectoparasites in other phyla include the pentastomes (tongue worms) and leeches.

Arthropods may cause injury when they attempt to take a blood meal or as they defend themselves by biting, stinging, or exuding venoms. Papular urticaria and other lesions caused by arthropod bites and stings are so diverse and variable (depending upon the host's health status and prior exposure to the arthropod's saliva, venom, or other exudates) that it is difficult to identify the precise causative organism without a bona fide specimen and taxonomic expertise.

SCABIES

The human itch mite, *Sarcoptes scabiei* var. *hominis*, is an obligate human ectoparasite and a common cause of itchy dermatosis, infesting ~300 million persons worldwide. Gravid female mites (~0.3 mm in length) burrow superficially within the stratum corneum, depositing several eggs per day. Six-legged larvae mature to eight-legged nymphs and then to adults. Gravid adult females emerge to the surface of the skin about 8 days later and then (re)invade the skin of the same or another host. Newly fertilized female mites are transferred from person to person mainly by direct skin-to-skin contact; transfer is facilitated by crowding, poor hygiene, and sex with multiple partners. Generally, scabies mites die within a day or so in the absence of a suitable host. Transmission via sharing of contaminated bedding or clothing occurs less frequently than is often thought. In the United States, scabies may account for up to 5% of visits to dermatologists. Outbreaks are known to occur in preschools, hospitals, nursing homes, and other institutional residences.

The itching and rash associated with scabies derive from a sensitization reaction to the mites and their secretions/excretions. A person's initial infestation remains asymptomatic for up to 6 weeks before the onset of intense pruritus, but a re-infestation produces a hypersensitivity reaction without delay. Burrows become surrounded by inflammatory infiltrates composed of eosinophils, lymphocytes, and histiocytes, and a generalized hypersensitivity rash later develops in remote sites.

Immunity and associated scratching limit most infestations to <15 mites per person. Hyperinfestation with thousands of mites, a condition known as *crusted scabies* (formerly termed *Norwegian scabies*), may result from glucocorticoid use, immunodeficiency, and neurologic or psychiatric illnesses that limit the itch and/or the scratch response.

Pruritus typically intensifies at night and after hot showers. Classic burrows are often difficult to find because they are few in number and may be obscured by excoriations. Burrows appear as dark wavy lines in the upper epidermis and are 3–15 mm long. Scabetic lesions are most common on the volar wrists and along the digital web spaces. In males, the penis and scrotum become involved. Small papules and vesicles, often accompanied by eczematous plaques, pustules, or nodules, appear symmetrically at those sites and within intertriginous areas, around the navel and belt line, in the axillae, and on the buttocks and upper thighs. Except in infants, the face, scalp, neck, palms, and soles are usually spared. Crusted scabies often resembles psoriasis: both are characterized by widespread thick keratotic crusts, scaly plaques, and dystrophic nails. Characteristic burrows are not seen in crusted scabies, and patients usually do not itch, although their infestations are highly contagious and have been responsible for outbreaks of classic scabies in hospitals.

Scabies should be considered in patients with pruritus and symmetric superficial, excoriated, papulovesicular skin lesions in characteristic locations, particularly if there is a history of direct and prolonged contact with an infested person. Burrows should be sought and unroofed with a sterile needle or scalpel blade, and the scrapings should be examined microscopically for mites, eggs, and fecal pellets. Examination of skin biopsies (including superficial cyanoacrylate biopsy) or scrapings, dermatoscopic imaging of papulovesicular lesions, and microscopic inspection of clear cellophane tape lifted from lesions also may be diagnostic. In the absence of identifiable mites or eggs, the diagnosis is based on a history of pruritus, a clinical examination, and an epidemiologic link. Diverse kinds of dermatitis from other causes frequently are misdiagnosed as scabies, particularly in presumed "outbreak" situations. Scabies mites of other animals may cause transient irritation, but they do not reside or reproduce in human hosts. In some Aboriginal communities, household dogs may serve as reservoirs for human scabies mites.

TREATMENT

Scabies

Permethrin cream (5%) is less toxic than 1% lindane preparations and is effective against lindane-tolerant infestations. Scabicides are applied thinly but thoroughly from the jawline down after bathing—with careful application to interdigital spaces and the umbilicus and under the fingernails—and are removed 8–14 h later with soap and water. Successful treatment of crusted scabies requires preapplication of a keratolytic agent such as 6% salicylic acid and then of scabicides to the scalp, face, and ears. Repeated treatments or the sequential use of several agents may be necessary. Ivermectin has not been approved by the U.S. Food and Drug Administration (FDA) for treatment of any form of scabies, but a single oral dose (200 µg/kg) is effective in otherwise healthy persons; patients with crusted scabies may require two doses separated by an interval of 1–2 weeks. All FDA-approved scabicides are available solely by prescription.

Within 1 day of effective treatment, scabies infestations become non-communicable, but the pruritic hypersensitivity dermatitis induced by the dead mites and their remnant products frequently persists for weeks. Unnecessary re-treatment with topical agents may provoke contact dermatitis. Antihistamines, salicylates, and calamine lotion relieve itching during treatment, and topical glucocorticoids are useful for pruritus that lingers after effective treatment. To prevent reinfestations, bedding and clothing should be washed and dried on high heat or heat-pressed. Close contacts of confirmed cases, even if asymptomatic, should be treated simultaneously.

CHIGGERS AND OTHER BITING MITES

Chiggers are the larvae of trombiculid (harvest) mites that normally feed on mice in grassy or brush-covered sites in tropical, subtropical, and (less frequently) temperate areas during warm months. They reside on low vegetation and attach themselves to passing vertebrate hosts. While feeding, larvae secrete saliva with proteolytic enzymes to create a tube-like invagination in the host's skin; this *stylostome* allows the mite to imbibe tissue fluids. The stylostomal saliva is highly antigenic and causes exceptionally pruritic papular, papulovesicular, or papulourticarial lesions (≤2 cm in diameter). In persons previously sensitized to salivary antigens, the papules develop within hours of attachment. While attached, mites appear as tiny red vesicles on the skin. Generally, lesions vesicate and develop a hemorrhagic base. Scratching invariably destroys the body of a mite, but itching and burning often persist for weeks. The rash is common on the ankles and areas where clothing obstructs the further wanderings of the mites. Repellents are useful for preventing chigger bites. Chiggers serve as vectors for *Orientia tsutsugamushi*, the agent of scrub typhus in Palearctic, Indomalayan, and Australasian regions.

Many kinds of mites associated with peridomestic birds and rodents are particularly bothersome when they invade homes and bite people. In North America, the northern fowl mite, chicken mite, tropical rat mite, and house mouse mite normally feed on poultry and diverse other birds as well as small mammals; these mites are abundant in and around their hosts' nests. After their natural hosts die or leave the nest, the mites disperse and may invade homes. Although the mites are rarely seen because of their small size, their bites can be painful and pruritic. House mouse mites (*Liponyssoides sanguineus*) serve as vectors for the agent of rickettsialpox, *Rickettsia akari*.

Once confirmed as the cause of irritation, rodent- and bird-associated mites are best eliminated by excluding their hosts, removing the nests, and cleaning and treating the nesting area with appropriate acaricides. *Pyemotes* and other mites that infest grain, straw, cheese, hay, oak leaf galls, or other products occasionally produce similar episodes of rash and discomfort and may produce a unique dermatologic "comet sign" lesion—a paisley-shaped urticarial plaque.

Diagnosis of mite-induced dermatitides (including those caused by chiggers) relies on confirmation of the mite's identity or elicitation of a history of exposure to the mite's source. Treatment of the patient with acaricides is not necessary, but oral antihistamines or topical steroids may suppress mite-induced pruritus temporarily.

TICK BITES AND TICK PARALYSIS

Ticks attach superficially to skin and feed painlessly; blood is their only food. Their salivary secretions are biologically active and can produce local reactions, induce fevers, and cause paralysis in addition to transmitting diverse pathogens. The two main families of ticks are the soft (argasid) and hard (ixodid) ticks. Generally, soft ticks attach for <1 h, leaving red macules after they drop off. Some species in Africa, the western United States, and Mexico produce painful hemorrhagic lesions. Hard ticks are much more common and transmit most of the tick-borne infections that are familiar to physicians and patients. Hard ticks attach to the host and feed for several days or sometimes for >1 week (depending upon the tick's species and stage of development). At the site of hard-tick bites, small areas of induration, often purpuric, develop and may be surrounded by an erythematous rim. A necrotic eschar, called a *tâche noire* ("black spot"), occasionally develops. Chronic nodules (persistent tick-bite granulomas) can be several centimeters in diameter and may linger for months after the feeding tick has been removed. These granulomas can be treated with injected intralesional glucocorticoids or by surgical excision. Tick-induced fever, unassociated with transmission of any pathogen, is often accompanied by headache, nausea, and malaise, but usually resolves ≤36 h after the tick is removed. Salivary antigens of the lone star tick, *Amblyomma americanum*, may induce antibodies to galactose-α-1,3-galactose (alpha-gal) that result in mammalian meat allergy–alpha-gal syndrome.

Tick paralysis, an acute ascending flaccid paralysis that resembles Guillain-Barré syndrome, is believed to be caused by one or more toxins in tick saliva that block neuromuscular transmission and

decrease nerve conduction. This rare complication has followed the bites of more than 60 kinds of ticks, although in the United States dog and wood ticks (*Dermacentor* species) are most commonly involved. Weakness begins symmetrically in the lower extremities ≤6 days after the tick's attachment, ascends symmetrically during several days, and may culminate in complete paralysis of the extremities and cranial nerves. Deep tendon reflexes are diminished or absent, but sensory examination and findings on lumbar puncture are typically normal. Removal of the tick generally leads to rapid improvement within a few hours and complete recovery after several days, although the patient's condition may continue to deteriorate for a full day. Failure to remove the tick may lead to dysarthria, dysphagia, and ultimately death from aspiration or respiratory paralysis. Diagnosis depends on finding the tick, which is often hidden beneath scalp hair. An antiserum to the saliva of *Ixodes holocyclus*, the usual cause of tick paralysis in Australia, effectively reverses paralysis caused by these ticks.

Removal of hard ticks during the first 36 h of attachment generally prevents transmission of the agents of Lyme disease, babesiosis, anaplasmosis, and ehrlichiosis, although tick-borne viruses may be transmitted more quickly. Ticks should be removed by traction with fine-tipped forceps placed firmly around the tick's mouthparts. Careful handling (to avoid rupture of ticks) and use of gloves may avert accidental contamination with pathogens contained in tick fluids. Use of occlusive dressings, heat, or other substances (in an attempt to induce the tick to detach) merely delays tick removal. Afterward, the site of attachment should be disinfected. Tick mouthparts sometimes remain in the skin but generally are shed spontaneously within days without the need for surgical removal. Although somewhat controversial, current guidelines from the Centers for Disease Control and Prevention suggest that, rather than awaiting the onset of erythema migrans, the results of tick testing, or seroconversion to antigens diagnostic for Lyme disease, administration of prophylaxis with a single oral dose of doxycycline (200 mg) within 72 h of tick removal is appropriate in adult patients with bites thought to be associated with *Ixodes scapularis* (deer ticks) **(Fig. 452-1)** in Lyme disease–endemic areas.

LOUSE INFESTATION (PEDICULIASIS AND PTHIRIASIS)

Nymphs and adults of all three kinds of human lice feed at least once a day, ingesting human blood exclusively. Head lice (*Pediculus capitis*) infest mainly the hair of the scalp, body lice (*Pediculus humanus*) the clothing, and crab or pubic lice (*Pthirus pubis*) mainly the hair of the pubis. The saliva of lice produces a pruritic morbilliform or urticarial rash in some sensitized persons. Female head and pubic lice cement their eggs (nits) firmly to hair, whereas female body lice cement their eggs to clothing, particularly to threads along clothing seams. After ~10 days of development within the egg, a nymph hatches. Empty eggs may remain affixed for months or years thereafter.

In North America, the prevalence of head lice is ~1% among children 6–10 years old and considerably lower among persons of other ages. Infestations can be far more prevalent elsewhere. Head lice are transmitted mainly by direct head-to-head contact rather than by fomites such as shared headgear, bed linens, and grooming implements. Chronic infestations by head lice tend to be asymptomatic. Pruritus, due mainly to hypersensitivity to the louse's saliva, generally is transient and mild and is most evident around the posterior hairline. Head lice removed from a person succumb to desiccation and starvation within ~1 day. Head lice are not known to serve as a natural vector for any pathogens.

Body lice remain on clothing except when feeding and generally succumb in ≤2 days if separated from their host. In most Western countries, body lice are generally found on a small proportion of indigent persons but may become increasingly prevalent after societal upheaval from natural or human-caused disasters, when displaced persons are in close contact with infested individuals with whom they share accommodations and lack the wherewithal to wash or change their clothes. Body lice are acquired by direct contact or by sharing of infested clothing and bedding. These lice are vectors for the agents of louse-borne (epidemic) typhus **(Chap. 182)**, louse-borne relapsing fever **(Chap. 180)**, and trench fever **(Chap. 167)**. Chronic infestations result in a postinflammatory hyperpigmentation and thickening of the skin known as *vagabond's disease*.

The crab or pubic louse is transmitted mainly by sexual contact. These lice occur predominantly on pubic hair and less frequently on axillary or facial hair, including the eyelashes. Children and adults may acquire pubic lice by sexual or close nonsexual contact. Intensely pruritic, bluish macules ~3 mm in diameter (*maculae ceruleae*) develop at the site of bites. Blepharitis commonly accompanies infestations of the eyelashes.

Pediculiasis is often suspected upon the detection of nits firmly cemented to hairs or in clothing. Many bona fide nits, however, are dead or hatched relics of prior infestation, and pseudo-nits are frequently misconstrued to be signs of a louse infestation. Confirmation of a louse infestation, therefore, best relies on the discovery of a live louse.

TREATMENT

Louse Infestation

Generally, treatment is justified only if live lice are discovered. The presence of nits alone is evidence of a former—not necessarily current—infestation. Mechanical removal of head lice and their eggs with a fine-toothed louse or nit comb **(Fig. 452-2)** often fails to eliminate infestations. Treatment of newly identified active infestations traditionally relies on a 10-min topical application of ~1% permethrin or pyrethrins, with a second application ~10 days later. Lice persisting after this treatment may be resistant to pyrethroids. Chronic infestations may be treated for ≤12 h with 0.5% malathion. Lindane is applied for just 4 min but seems less effective and may pose a greater risk of adverse reactions, particularly when misused.

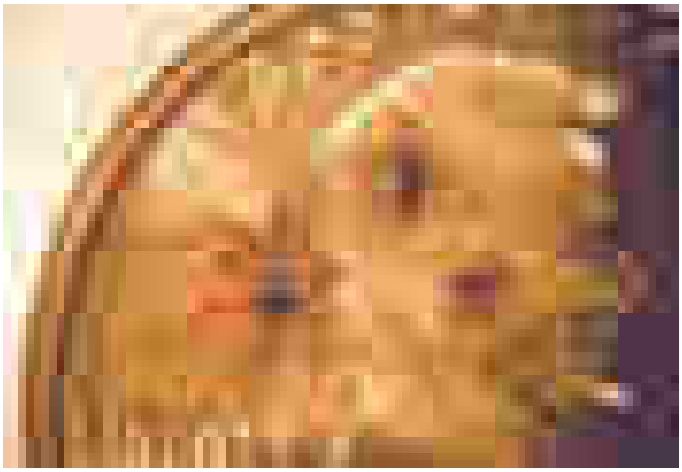


FIGURE 452-1 Deer ticks (*Ixodes scapularis*, black-legged ticks) on a U.S. penny: larva (*below ear*), nymph (*right*), adult male (*above*), and adult female (*left*).

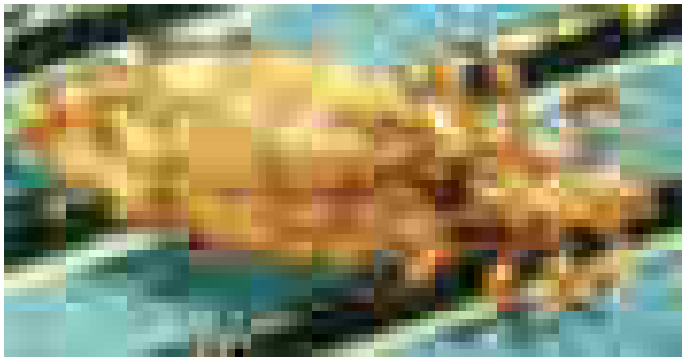


FIGURE 452-2 Adult female human head louse (*Pediculus capitis*) on a nit (louse-egg) comb.

Resistance of head lice to permethrin, malathion, and lindane has been reported. Newer FDA-approved topical pediculicides contain benzyl alcohol, dimethicone, spinosad, and ivermectin. Although children infested by head lice—or those who simply have remnant nits from a prior infestation—are frequently isolated or excluded from school, this practice increasingly is considered to be unjustified, ineffective, and counterproductive.

Body lice usually are eliminated by bathing and by changing to laundered clothes. Application of topical pediculicides from head to foot may be necessary for hirsute patients. Clothes and bedding are effectively deloused by heating in a clothes dryer at ≥55°C (≥131°F) for 30 min or by heat-pressing. Emergency mass delousing of persons and clothing may be warranted during periods of civil strife and after natural disasters to reduce the risk of pathogen transmission by body lice.

Pubic louse infestations are treated with topical pediculicides, except for eyelid infestations (*pthiriasis palpebrum*), which generally respond to a coating of petrolatum applied for 3–4 days.

MYIASIS (FLY INFESTATION)

Myiasis refers to infestations by fly larvae (maggots) that invade living or necrotic tissues or body cavities and produce different clinical syndromes, depending on the species of fly.

In forested parts of Central and South America, larvae of the human botfly (*Dermatobia hominis*) produce furuncular (boil-like) papules or subcutaneous nodules ≤3 cm in diameter. A gravid adult female botfly captures a mosquito or another bloodsucking insect and deposits her eggs on its abdomen. When the carrier insect attacks a human or bovine host several days later, the warmth and moisture of the host's skin stimulate the eggs to hatch. The emerging larvae promptly penetrate intact skin. After 6–12 weeks of development, mature larvae emerge from the skin and drop to the ground to pupate and then become adults.

The African tumbu fly (*Cordylobia anthropophaga*) deposits its eggs on damp sand or leaf litter or on drying laundry, particularly items contaminated by urine or sweat. Larvae hatch from eggs upon contact with a host's body and penetrate the skin, producing boil-like lesions from which mature larvae emerge ~9 days later. Furuncular myiasis is suggested by uncomfortable lesions with a central breathing pore that emit bubbles when submerged in water. A sensation of movement under the patient's skin may cause severe emotional distress.

Larvae that cause furuncular myiasis may be induced to emerge if the air pore is coated with petrolatum or another occlusive substance. Removal may be facilitated by injection of a local anesthetic into the surrounding tissue, but surgical excision is sometimes necessary because upward-pointing spines of some species hold the larvae firmly in place.

Other fly larvae cause nonfuruncular myiasis. Larvae of the horse botfly (*Gasterophilus intestinalis*) emerge from eggs deposited on the horse's flanks and may come into contact with and infest human beings. After penetrating human skin, these larvae rarely mature but instead may migrate for weeks in the dermis. The resulting pruritic and serpiginous eruption resembles cutaneous larva migrans caused by canine or feline hookworms **(Chap. 226)**. Larvae of rabbit and rodent botflies (*Cuterebra* species) occasionally cause dermal or tracheopulmonary myiasis.

Certain flies are attracted to blood and pus, laying their eggs on open or draining sores. Newly hatched larvae enter wounds or diseased skin. Larvae of several types of green bottle flies (*Lucilia*/*Phaenicia* species) usually remain superficial and confined to necrotic tissue. Specially raised, sterile "surgical maggots" are sometimes used intentionally for wound debridement. Larvae of screwworm flies (*Cochliomyia*) and the flesh fly invade viable tissues more deeply and produce large suppurating lesions. Larvae that infest wounds also may enter body cavities such as the mouth, nose, ears, sinuses, anus, vagina, and lower urinary tract, particularly in unconscious or otherwise debilitated patients. The consequences range from harmless colonization to destruction of the nose, meningitis, and deafness. Treatment involves removal of maggots and debridement of tissue.

Larvae of the sheep botfly, *Oestrus ovis*, and others responsible for furuncular and wound myiasis also may cause ophthalmomyiasis. Sequelae include nodules in the eyelid, retinal detachment, and destruction of the globe. Most instances in which maggots are found in human feces result from deposition of eggs or larvae by flies on recently passed stools, not from an intestinal maggot infestation.

PENTASTOMIASIS

Pentastomids (tongue worms) inhabit the respiratory passages of reptiles and carnivorous mammals. Human infestation by *Linguatula serrata* is common in the Middle East and results from the consumption of encysted larval stages in raw liver or lymph nodes of sheep and goats, which are true intermediate hosts for the tongue worms. Larvae migrate to the nasopharynx and produce an acute self-limiting syndrome—known as *halzoun* or *marrara*—characterized by pain and itching of the throat and ears, coughing, hoarseness, dysphagia, and dyspnea. Severe edema may cause obstruction that requires tracheostomy. In addition, ocular invasion has been described. Diagnostic larvae measuring ≤10 mm in length appear in copious nasal discharge or vomitus. Individuals become infected with another type of tongue worm, *Armillifer armillatus*, by consuming its eggs in contaminated food or drink or after handling the definitive host, the African python. Larvae encyst in various organs but rarely cause symptoms. Cysts may require surgical removal as they enlarge during molting, but they usually are encountered as an incidental finding at autopsy. Parasite-induced lesions may be misinterpreted as a malignancy, with the correct diagnosis confirmed histopathologically. Cutaneous larva migrans–type syndromes of other pentastomes have been reported from Southeast Asia and Central America.

LEECH INFESTATIONS

Medically important leeches are annelid worms that attach to their hosts with chitinous cutting jaws and draw blood through muscular suckers. The medicinal leech (*Hirudo medicinalis*) is still used occasionally for medical purposes to reduce venous congestion in surgical flaps or replanted body parts. This practice has been complicated by intractable bleeding, wound infections, myonecrosis, and sepsis due to *Aeromonas hydrophila*, which colonizes the gullets of commercially available leeches.

Ubiquitous aquatic leeches that parasitize fish, frogs, and turtles readily attach to the skin of human beings and avidly suck blood. More notorious are arboreal land leeches that live among moist vegetation of tropical rain forests. Attachment is usually painless, and the leeches will detach themselves when satiated with a blood meal. Hirudin, a powerful anticoagulant secreted by the leech, causes continued bleeding after the leech has detached. Healing of a leech-bite wound is slow, and bacterial infections are not uncommon. Several kinds of aquatic leeches in Africa, Asia, and southern Europe can enter the mouth, nose, and genitourinary tract and attach to mucosal surfaces at sites as deep as the esophagus and trachea. Externally attached leeches generally drop off after they have engorged, but removal is hastened by gentle scraping aside of the anterior and posterior suckers the leech uses for attachment and feeding. Some authorities dispute the wisdom of removing leeches with alcohol, salt, vinegar, insect repellent, a flame or heated instrument, or applications of other noxious substances. Internally attached leeches may detach on exposure to gargled saline or may be removed by forceps.

SPIDER BITES

Of the more than 30,000 recognized species of spiders, only ~100 defend themselves aggressively and have fangs sufficiently long to penetrate human skin. The venom that some spiders use to immobilize and digest their prey can cause necrosis of skin and systemic toxicity. Whereas the bites of most spiders are painful but not harmful, envenomations by recluse or fiddleback spiders (*Loxosceles* species) and widow spiders (*Latrodectus* species) may be life-threatening. Identification of the offending spider is important because specific treatments exist for bites of widow spiders and because injuries attributed to spiders are frequently due to other causes. Except in cases where the

patient actually observes a spider immediately associated with the bite or fleeing from the site, lesions reported as spider-bite reactions are most often due to other injuries or to infections with bacteria such as methicillin-resistant *Staphylococcus aureus* (MRSA) or, infrequently, *Franciscella tularensis.*

Recluse Spider Bites and Necrotic Arachnidism Brown recluse spiders (*Loxosceles reclusa*) live mainly in the south-central United States and have close relatives in Central and South America, Africa, and the Middle East. Bites by brown recluse spiders usually cause only minor injuries, with edema and erythema. Envenomation, however, occasionally causes severe necrosis of skin and subcutaneous tissue and more rarely causes systemic hemolysis. These spiders are not aggressive toward humans and bite only if threatened or pressed against the skin. They hide under rocks and logs or in caves and animal burrows. They invade homes and seek dark and undisturbed hiding spots in closets, in folds of clothing, or under furniture and rubbish in storage rooms, garages, and attics. Despite their impressive abundance in some homes, these spiders rarely bite humans. Bites tend to occur while the victim is dressing and are sustained primarily on the hands, arms, neck, and lower abdomen.

Initially, the bite is painless or may produce a stinging sensation. Within the next few hours, the site becomes painful and pruritic, with central induration surrounded by a pale ischemic zone that itself is encircled by a zone of erythema. In most cases, the lesion resolves without treatment in just a few days. In severe cases, the erythema spreads, and the center of the lesion becomes hemorrhagic or necrotic with an overlying bulla. A black eschar forms and sloughs several weeks later, leaving an ulcer that eventually may create a depressed scar. Healing usually takes place in ≤3 months. Local complications include injury to nerves and secondary bacterial infection. Fever, chills, weakness, headache, nausea, vomiting, myalgia, arthralgia, maculopapular rash, and leukocytosis may develop ≤72 h after the bite. Reports of deaths attributed to bites of North American brown recluse spiders have not been verified.

The Mediterranean recluse spider (*Loxosceles rufescens*) is a widely invasive species in urban areas of both the Old and New Worlds. The dorsal surface of *L. rufescens* and *L. reclusa* is adorned with a fiddle-shaped pattern. *L. rufescens* is warier than *L. reclusa*, is less likely to bite, and rarely causes necrosis. Misidentification of this spider may create spurious reports of *L. reclusa* activity outside the known range of that species.

TREATMENT

Recluse Spider Bites

Initial management includes rest, ice, compression, and elevation (RICE). Analgesics, antihistamines, antibiotics, and tetanus prophylaxis should be administered if indicated. Early debridement or surgical excision of the wound without closure delays healing. Routine use of antibiotics or dapsone lacks utility. Patients should be monitored closely for signs of hemolysis, renal failure, and other systemic complications.

Widow Spider Bites The black widow spider, common in the southeastern United States, measures ≤1 cm in body length and 5 cm in leg span and is shiny black with a red hourglass marking on the ventral abdomen. Other dangerous *Latrodectus* species occur elsewhere in temperate and subtropical parts of the world. The bites of the female widow spiders are notorious for their potent neurotoxins.

Widow spiders spin their webs under stones, logs, plants, or rock piles and in dark spaces in barns, garages, and outhouses. Bites are most common in the summer and early autumn and occur when a web is disturbed or a spider is trapped or provoked. The initial bite is perceived as a sharp pinprick or may go unnoticed. Fang-puncture marks are uncommon. The venom that is injected does not produce local necrosis, and some persons experience no other symptoms. α-Latrotoxin, the most active component of the venom, binds irreversibly to presynaptic nerve terminals and causes release and eventual depletion of acetylcholine, norepinephrine, and other neurotransmitters from those terminals. Painful cramps may spread within 60 min from the bite site to large muscles of the extremities and trunk. Extreme rigidity of the abdominal muscles and excruciating pain may suggest peritonitis, but the abdomen is not tender on palpation and surgery is not warranted. The pain begins to subside during the first 12 h but may recur during several days or weeks before resolving spontaneously. A wide range of other sequelae may include salivation, diaphoresis, vomiting, hypertension, tachycardia, labored breathing, anxiety, headache, weakness, fasciculations, paresthesia, hyperreflexia, urinary retention, uterine contractions, and premature labor. Rhabdomyolysis and renal failure have been reported, and respiratory arrest, cerebral hemorrhage, or cardiac failure may end fatally, especially in very young, elderly, or debilitated persons.

TREATMENT

Widow Spider Bites

Treatment consists of RICE and tetanus prophylaxis. Hypertension that does not respond to analgesics and antispasmodics (e.g., benzodiazepines or methocarbamol) requires specific antihypertensive medication. The efficacy and safety of antivenoms for black widow and redback spiders are controversial because of concerns about potential anaphylaxis or serum sickness.

Tarantulas and Other Spiders Tarantulas are hairy spiders of which 30 species are found in the United States, mainly in the Southwest. The tarantulas that have become popular household pets are usually imported from Central or South America. Tarantulas bite only when threatened and usually cause no more harm than a bee sting, but on occasion the venom causes deep pain and swelling. Several species of tarantulas are covered with urticating hairs that are brushed off in the thousands when a threatened spider rubs its hind legs across its dorsal abdomen. These hairs can penetrate human skin and produce pruritic papules that may persist for weeks. Failure to wear gloves or to wash the hands after handling the Chilean Rose tarantula, a popular pet spider, has resulted in transfer of hairs to the eye with subsequent devastating ocular inflammation. Treatment of bites includes local washing and elevation of the bitten area, tetanus prophylaxis, and analgesic administration. Antihistamines and topical or systemic glucocorticoids are given for exposure to urticating hairs.

Atrax robustus, a funnel-web spider of Australia, and *Phoneutria* species, the South American banana spiders, are among the most dangerous spiders in the world because of their aggressive behavior and potent neurotoxins. Envenomation by *A. robustus* causes a rapidly progressive neuromotor syndrome that can be fatal within 2 h. The bite of a banana spider causes severe local pain followed by profound systemic symptoms and respiratory paralysis that can lead to death within 2–6 h. Specific antivenoms for use after bites by each of these spiders are available. Yellow sac spiders (*Cheiracanthium* species) are common in homes worldwide. Their bites, though painful, generally lead to only minor erythema, edema, and pruritus.

SCORPION STINGS

Scorpions are arachnids that feed on ground-dwelling arthropods and small lizards. They paralyze their prey and defend themselves by injecting venom from a stinger on the tip of the tail. Painful but relatively harmless scorpion stings need to be distinguished from the potentially lethal envenomations that are produced by ~30 of the ~1000 known species and that cause more than 5000 deaths worldwide each year. Scorpions are nocturnal and remain hidden during the day in crevices or burrows or under wood, loose bark, or rocks. They occasionally enter houses and tents and may hide in shoes, clothing, or bedding. Scorpions sting humans only when threatened.

Of the 40 or so scorpion species in the United States, only bark scorpions (*Centruroides sculpturatus*/*C. exilicauda*) in the Southwest produce venom that is potentially lethal to humans. This venom

contains neurotoxins that cause sodium channels to remain open. Such envenomations usually are associated with little swelling, but prominent pain, paresthesia, and hyperesthesia can be accentuated by tapping on the affected area (the *tap test*). These symptoms soon spread to other locations; dysfunction of cranial nerves and hyperexcitability of skeletal muscles develop within hours. Patients present with restlessness, blurred vision, abnormal eye movements, profuse salivation, lacrimation, rhinorrhea, slurred speech, difficulty in handling secretions, diaphoresis, nausea, and vomiting. Muscle twitching, jerking, and shaking may be mistaken for a seizure. Complications include tachycardia, arrhythmias, hypertension, hyperthermia, rhabdomyolysis, and acidosis. Symptoms progress to maximal severity in ~5 h and subside within a day or two, although pain and paresthesia can last for weeks. Fatal respiratory arrest is most common among young children and the elderly.

Envenomations by *Leiurus quinquestriatus* in the Middle East and North Africa, by *Mesobuthus tamulus* in India, by *Androctonus* species along the Mediterranean littoral and in North Africa and the Middle East, and by *Tityus serrulatus* in Brazil cause massive release of endogenous catecholamines with hypertensive crises, arrhythmias, pulmonary edema, and myocardial damage. Acute pancreatitis occurs with stings of *Tityus trinitatis* in Trinidad, and central nervous toxicity complicates stings of *Parabuthus* and *Buthotus* scorpions of South Africa. Tissue necrosis and hemolysis may follow stings of the Iranian *Hemiscorpius lepturus*.

Stings of most other species cause immediate sharp local pain followed by edema, ecchymosis, and a burning sensation. Symptoms typically resolve within a few hours, and skin does not slough. Allergic reactions to the venom sometimes develop.

TREATMENT

Scorpion Stings

Identification of the offending scorpion helps to determine the course of treatment. Stings of nonlethal species require at most ice packs, analgesics, or antihistamines. Because most victims experience only local discomfort, they can be managed at home with instructions to return to the emergency department if signs of cranial-nerve or neuromuscular dysfunction develop. Aggressive supportive care and judicious use of antivenom can reduce or eliminate deaths from more severe envenomations. Keeping the patient calm and applying pressure dressings and cold packs to the sting site are measures that decrease the absorption of venom. A continuous IV infusion of midazolam controls the agitation, flailing, and involuntary muscle movements produced by scorpion stings. Close monitoring during treatment with this drug and other sedatives or narcotics is necessary for persons with neuromuscular symptoms because of the risk of respiratory arrest. Hypertension and pulmonary edema respond to nifedipine, nitroprusside, hydralazine, or prazosin. Dangerous bradydysrhythmia can be controlled with atropine.

Commercially prepared antivenoms are available in several countries for some of the most dangerous species. An FDA-approved *C. sculpturatus* antivenom in horse serum is now available. IV administration of antivenom rapidly reverses cranial-nerve dysfunction and muscular symptoms. Although effective, cost analyses suggest that antivenoms should be reserved for only the most severe envenomations.

HYMENOPTERA STINGS

Bees, wasps, hornets, yellow jackets, and ants (all of the insect order Hymenoptera) sting in defense or to subdue their prey. Their venoms contain a wide array of amines, peptides, and enzymes that cause local and systemic reactions. Although the toxic effect of multiple stings can be fatal to a human, nearly all of the ≥100 deaths due to hymenopteran stings in the United States each year result from allergic reactions.

Bee and Wasp Stings The stinger of the honeybee (*Apis mellifera*) is unique in being barbed. The stinging apparatus and attached venom sac tear loose from the honeybee's body, and muscular contraction of the venom sac continues to inject venom into the skin. Other kinds of bees, ants, and wasps have smooth stinging mechanisms and can sting numerous times in succession. Generally, a person sustains just one sting from a bee or social wasp unless a nest was disturbed. Africanized honeybees (now present in South and Central America and the southern and western United States) respond to minimal intrusions more aggressively. The sting of an Africanized bee contains less venom than that of its non-Africanized relatives, but victims tend to sustain far more stings and receive a far greater overall volume of venom. Most patients who report having sustained a "bee sting" are more likely to have encountered stinging wasps instead.

The venoms of different kinds of hymenopterans are biochemically and immunologically distinct. Direct toxic effects are mediated by mixtures of low-molecular-weight compounds such as serotonin, histamine, acetylcholine, and several kinins. Polypeptide toxins in honeybee venom include mellitin, which damages cell membranes; mast cell–degranulating protein, which causes histamine release; the neurotoxin apamin; and the anti-inflammatory compound adolapin. Enzymes in venom include hyaluronidase and phospholipases. There appears to be little cross-sensitization between the venoms of honeybees and wasps.

Uncomplicated hymenopteran stings cause immediate pain, a wheal-and-flare reaction, and local edema, all which usually subside in a few hours. Multiple stings can lead to vomiting, diarrhea, generalized edema, dyspnea, hypotension, and non-anaphylactic circulatory collapse. Rhabdomyolysis and intravascular hemolysis may cause renal failure. Death from the direct (nonallergic) effects of venom has followed stings of several hundred honeybees. Stings to the tongue or mouth may induce life-threatening edema of the upper airways.

Large local reactions accompanied by erythema, edema, warmth, and tenderness that spread ≥10 cm around the sting site over 1–2 days are not uncommon. These reactions may resemble bacterial cellulitis but are caused by hypersensitivity rather than by secondary infection. Such reactions tend to recur on subsequent exposure but are seldom accompanied by anaphylaxis and are not prevented by venom immunotherapy.

An estimated 0.4–4.0% of the U.S. population exhibits clinical immediate-type hypersensitivity to hymenopteran stings, and 15% may have asymptomatic sensitization manifested by positive skin tests. Persons who experience severe allergic reactions are likely to have similar or more severe reactions after subsequent stings by the same or closely related species. Mild anaphylactic reactions from insect stings, as from other causes, consist of nausea, abdominal cramping, generalized urticaria or angioedema, and flushing. Serious reactions, including upper airway edema, bronchospasm, hypotension, and shock, may be rapidly fatal. Severe reactions usually begin within 10 min of the sting and only rarely develop after 5 h.

TREATMENT

Bee and Wasp Stings

Honeybee stingers embedded in the skin should be removed as soon as possible to limit the quantity of venom delivered. The stinger and venom sac may be scraped off with a blade, a fingernail, or the edge of a credit card or may be removed with forceps. The site should be cleansed and disinfected and ice packs applied to slow the spread of venom. Elevation of the affected site and administration of oral analgesics, oral antihistamines, and topical calamine lotion help relieve symptoms.

Anaphylactic reactions to bee or wasp venom can be a life-threatening emergency that requires prompt life-saving actions. If the individual carries a bee-sting kit, then a subcutaneous injection of epinephrine hydrochloride (0.3 mL of a 1:1000 dilution) should be considered, with treatment repeated every 20–30 min as necessary. A tourniquet may slow the spread of venom. The patient should be transferred to a hospital emergency room where treatment for profound shock, if required, can be administered safely. Such treatment may entail the use of intravenous epinephrine and

other vasopressors, intubation or supplemental oxygen, fluid resuscitation, bronchodilators, and parenteral antihistamines. Patients should be observed for 24 h for recurrent anaphylaxis, renal failure, or coagulopathy.

Persons with a history of allergy to insect stings should carry an anaphylaxis kit with a preloaded syringe containing epinephrine for self-administration. These patients should seek medical attention immediately after using the kit.

Prophylactic immunotherapy may greatly reduce the risk of life-threatening reactions to bee and wasp stings. Repeated injections of purified venom produce a blocking IgG antibody response to venom and reduce the incidence of recurrent anaphylaxis. Honeybee, wasp, and yellow jacket venoms are commercially available for desensitization and for skin testing. Results of skin tests and venom-specific radioallergosorbent tests (RASTs) aid in the selection of patients for immunotherapy and guide the design of such treatment.

STINGING ANTS

Stinging ants are an important medical problem in the United States. Imported fire ants (*Solenopsis* species) infest southern states from Texas to North Carolina, with colonies now established in California, New Mexico, Arizona, and Virginia. Slight disturbances of their mound nests have provoked massive outpourings of ants and as many as 10,000 stings on a single person. Elderly and immobile persons are at high risk for attacks when fire ants invade dwellings.

Fire ants attach to skin with powerful mandibles and rotate their bodies while repeatedly injecting venom with posteriorly situated stingers. The alkaloid venom consists of cytotoxic and hemolytic piperidines and several proteins with enzymatic activity. The initial wheal-and-flare reaction, burning, and itching resolve in ~30 min, and a sterile pustule develops within 24 h. The pustule ulcerates over the next 48 h and then heals in ≥1 week. Large areas of erythema and edema lasting several days are not uncommon and in extreme cases may compress nerves and blood vessels. Anaphylaxis occurs in fewer than 2% of victims; seizures and mononeuritis have been reported. Stings are treated with ice packs, topical glucocorticoids, and oral antihistamines. Pustules should be cleansed and then covered with bandages and antibiotic ointment to prevent bacterial infection. Epinephrine administration and supportive measures are indicated for anaphylactic reactions. Fire ant whole-body extracts are available for skin testing and immunotherapy, which appear to lower the rate of anaphylactic reactions.

European fire (red) ants (*Myrmica rubra*) have recently become public health pests in the northeastern United States and southern Canada. The western United States is home to harvester ants (*Pogonomyrmex* species). The painful local reaction that follows harvester ant stings often extends to lymph nodes and may be accompanied by anaphylaxis. The bullet or conga ant (*Paraponera clavata*) of South America is known locally as *hormiga veinticuatro* ("24-hour ant"), a designation that refers to the 24 h of throbbing, excruciating pain following a sting that delivers the potent paralyzing neurotoxin poneratoxin.

DIPTERAN (FLY AND MOSQUITO) BITES

In the process of feeding on vertebrate blood and tissue fluids, adults of certain flies inflict painful bites, produce local allergic reactions, and may transmit pathogenic agents. Bites of mosquitoes, tiny "no-see-um" (ceratopogonid) midges, and phlebotomine sand flies typically produce a wheal and a pruritic papule. Small humpbacked black flies (simuliids) lacerate skin, resulting in a lesion with serosanguineous discharge that is often painful and pruritic. Regional lymphadenopathy, fever, or anaphylaxis occasionally ensues. The widely distributed deerflies and horseflies as well as the tsetse flies of Africa are stout flies measuring ≤25 mm in length that attack during the day and produce large and painful bleeding punctures. Houseflies (*Musca domestica*) do not consume blood but use rasping mouthparts to scarify skin and feed upon tissue fluids and salt. Beyond direct injury from bites of any kind of fly, risks include transmission of diverse pathogens and secondary infection of the lesion.

TREATMENT

Fly and Mosquito Bites

Treatment of fly bites is symptom based. Topical application of antipruritic agents, glucocorticoids, or antiseptic lotions may relieve itching and pain. Allergic reactions may require oral antihistamines. Antibiotics may be necessary for the treatment of large bite wounds that become secondarily infected.

FLEA BITES

Common human-biting fleas include the dog and cat fleas (*Ctenocephalides* species) and the rat flea (*Xenopsylla cheopis*), which infest their respective hosts and their nests and resting sites. Sensitized persons develop erythematous pruritic papules (*papular urticaria*) and occasionally vesicles and bacterial superinfection at the site of the bite. Symptom-based treatment consists of antihistamines, topical glucocorticoids, and topical antipruritic agents.

Flea infestations are eliminated by removal and treatment of animal nests, frequent cleaning of pet bedding, and application of contact and systemic insecticides to pets and the dwelling. Flea infestations in the home may be abated or prevented if pets are regularly treated with veterinary antiparasitic agents, insect growth regulators, or chitin inhibitors.

Tunga penetrans, like other fleas, is a wingless, laterally flattened insect that feeds on blood. Also known as the chigoe flea, sand flea, or jigger (not to be confused with the chigger), it occurs in tropical regions of Africa and the Americas. Adult female chigoes live in sandy soil and burrow under the skin, usually between toes, under nails, or on the soles of bare feet. Gravid chigoes engorge on the host's blood and grow from pinpoint to pea size during a 2-week interval. They produce lesions that resemble a white pustule with a central black depression and that may be pruritic or painful. Occasional complications include tetanus, bacterial infections, and autoamputation of toes (*ainhum*). Tungiasis is treated by removal of the intact flea with a sterile needle or scalpel, tetanus vaccination, and topical application of antibiotics.

HEMIPTERAN / HETEROPTERAN (TRUE BUG) BITES

Most true bugs feed on plants, but some are predaceous or feed on blood. They may inflict bites in order to feed or in defense that produce allergic reactions and are sometimes painful. Bites of the cone-nose or "kissing bugs" (family Reduviidae) tend to occur at night and are painless. Reactions to such bites depend on prior sensitization and include tender and pruritic papules, vesicular or bullous lesions, extensive urticaria, fever, lymphadenopathy, and (rarely) anaphylaxis. Bug bites are treated with topical antipruritics or oral antihistamines. Persons with anaphylactic reactions to reduviid bites should keep an epinephrine kit available. Some reduviids transmit *Trypanosoma cruzi*, the agent of New World trypanosomiasis (Chagas disease) **(Chap. 222)**.

The cosmopolitan and tropical bedbugs (*Cimex lectularius and C. hemipterus*) hide in crevices of mattresses, bed frames and other furniture, walls, and picture frames and under loose wallpaper, actively seeking blood meals at night. Bedbug populations have resurged, recently attaining levels and spreading to an extent not encountered since the mid-twentieth century. These bugs are now a common pest in homes, dormitories, and hotels; on cruise ships; and even in medical facilities. Their bite is painless, but minutes to days later, sensitized persons develop erythema, itching, and wheals around a central hemorrhagic punctum. Bedbugs are not known to transmit pathogens in nature.

CENTIPEDE BITES AND MILLIPEDE DERMATITIS

The fangs of centipedes of the genus *Scolopendra* can penetrate human skin and deliver a venom that produces intense burning pain, swelling, erythema, and sterile lymphangitis. Dizziness, nausea, and anxiety are described occasionally, and rhabdomyolysis and renal failure have been reported. Treatment includes washing of the site, application of cold dressings, oral analgesic administration or local lidocaine infiltration, and tetanus prophylaxis.

Millipedes are docile and do not bite, but some secrete defensive fluids that may burn and discolor human skin. Affected skin turns brown overnight and may blister and exfoliate. Secretions in the eye cause intense pain and inflammation that can result in corneal ulcers and even blindness. Management includes irrigation with copious amounts of water or saline, use of analgesics, and local care of denuded skin.

CATERPILLAR STINGS AND DERMATITIS

Caterpillars of several moth species are covered with hairs or spines that produce mechanical irritation and may contain or be coated with venom. Contact with these caterpillars or their hairs may lead to erucism (a pruritic urticarial or papular rash) or caterpillar envenomation. The response typically consists of an immediate burning sensation followed by local swelling and erythema and occasionally by regional lymphadenopathy, nausea, vomiting, and headache. A rare reaction to a South American caterpillar, *Lonomia obliqua*, can cause disseminated coagulopathy and fatal hemorrhagic shock. In the United States, dermatitis is most often associated with caterpillars of the io, puss, saddleback, and brown-tail moths. Even contact with detached hairs of other caterpillars, such as gypsy moth larvae, can later produce erucism. Spines may be deposited on tree trunks or drying laundry or may be airborne and cause irritation of the eyes and upper airways. Treatment of caterpillar stings consists of repeated application of adhesive or cellophane tape to remove the hairs, which can then be identified microscopically. Local ice packs, topical glucocorticoids, and oral antihistamines relieve symptoms.

BEETLE VESICATION AND DERMITITIS

Several families of beetles have independently developed the ability to produce chemically unrelated vesicating toxins. When disturbed, blister beetles (family Meloidae) exude cantharidin, a low-molecular-weight toxin that produces thin-walled blisters (≤5 cm in diameter) 2–5 h after contact. The blisters are not painful or pruritic unless broken and resolve without treatment in ≤10 days. Nephritis may follow unusually heavy cantharidin exposure. The hemolymph of certain rove beetles (Staphylinidae) contains pederin, a potent vesicant. When these beetles are crushed or brushed against the skin, the released fluid causes painful, red, flaccid bullae. These beetles occur worldwide but are most numerous and problematic in parts of Africa (where they are called "Nairobi fly") and southwestern Asia. Ocular lesions may develop after impact with flying beetles at night or unintentional transfer of the vesicant on the fingers. Treatment is rarely necessary, although ruptured blisters should be kept clean and bandaged.

Larvae of common carpet beetles are adorned with dense arrays of ornate hairs called *hastisetae*. Contact with these larvae or their setae results in delayed dermal reactions in sensitized individuals. The lesions are commonly mistaken as bites of bedbugs.

DELUSIONAL INFESTATIONS

The groundless conviction that one is infested with arthropods or other parasites (Ekbom's syndrome, delusory parasitosis, delusions of parasitosis, and perhaps Morgellons syndrome) is extremely difficult to treat and, unfortunately, is not uncommon. Patients describe uncomfortable sensations of something moving in or on their skin. Excoriations and self-induced ulcerations typically accompany the pruritus, dysesthesias, and imaginary insect bites. Patients often believe that some invisible or as yet undescribed creatures are infesting their skin, clothing, homes, or environment in general. Frequently, patients submit as evidence of infestation specimens that consist of plant-feeding and nonbiting peridomestic arthropods, pieces of skin, vegetable matter, lint, and other inanimate detritus. When evaluating a patient with possible delusional parasitosis, it is imperative to rule out true infestations and bites by arthropods, endocrinopathies, sensory disorders due to neuropathies, opiate and other drug use, environmental irritants (e.g., fiberglass threads), and other causes of tingling or prickling sensations. Frequently, such patients repeatedly seek medical consultations, resist alternative explanations for their symptoms, and exacerbate their discomfort by self-treatment. Long-term pharmacotherapy with pimozide or other psychotropic agents has been more helpful than psychotherapy in treating this disorder. Patients with delusory parasitosis often develop the unshakeable conviction that they are infested by a previously unknown pathogen, while their personal lives, family support, and employment collapse around them.

FURTHER READING

Arlia LG, Morgan MS: A review of *Sarcoptes scabiei*: Past, present and future. Parasit Vectors 10:297, 2017.

Goddard J: *Physician's Guide to Arthropods of Medical Importance*, 5th ed. Boca Raton, FL, CRC Press, 2007.

Hinkle N: Ekbom syndrome: The challenge of "invisible bug" infestations. Annu Rev Entomol 55:77, 2010.

McGraw TA, Turiansky MC: Cutaneous myiasis. J Am Acad Dermatol 58:907, 2008.

Mullen G, Durden L: *Medical and Veterinary Entomology*, 2nd ed. Amsterdam, Academic Press, 2009.

Pollack RJ, Marcus L: A travel medicine guide to arthropods of medical importance. Infect Dis Clin North Am 19:169, 2005.

Richards SL et al: Do tick attachment times vary between different tick–pathogen systems? Environments 4:37, 2017.

Saucier JR: Arachnid envenomation. Emerg Med Clin North Am 22:405, 2004.

Steen CJ et al: Insect sting reactions to bees, wasps, and ants. Int J Dermatol 44:91, 2005.

Vetter RS, Isbister GK: Medical aspects of spider bites. Annu Rev Entomol 53:409, 2008.

453 Altitude Illness

Buddha Basnyat, Geoffrey Tabin

EPIDEMIOLOGY

Mountains cover one-fifth of the earth's surface; 140 million people live permanently at altitudes ≥2500 m, and 100 million people travel to high-altitude locations each year. Skiers in the Alps or Aspen; tourists to La Paz, Ladakh, or Lahsa; religious pilgrims to Kailash-Manasarovar or Gosainkunda; trekkers and climbers to Kilimanjaro, Aconcagua, or Everest; miners working in high-altitude sites in South America; and military personnel deployed to high-altitude locations are all at risk of developing acute mountain sickness (AMS), high-altitude cerebral edema (HACE), high-altitude pulmonary edema (HAPE), and other altitude-related problems. AMS is the benign form of altitude illness, whereas HACE and HAPE are life-threatening. Altitude illness is likely to occur above 2500 m but has been documented even at 1500–2500 m. In the Mount Everest region of Nepal, ~50% of trekkers who walk to altitudes >4000 m over ≥5 days develop AMS, as do 84% of people who fly directly to 3860 m. The incidences of HACE and HAPE are much lower than that of AMS, with estimates in the range of 0.1–4%. Finally, reentry HAPE, which in the past was generally limited to highlanders (long-term residents of altitudes >2500 m) in the Americas, is now being seen in Himalayan and Tibetan highlanders—and often misdiagnosed as a viral illness—as a result of recent rapid air, train, and motorable-road access to high-altitude settlements.

PHYSIOLOGY

Ascent to a high altitude subjects the body to a decrease in barometric pressure that results in a decreased partial pressure of oxygen in the inspired gas in the lungs. This change leads in turn to less pressure driving oxygen diffusion from the alveoli and throughout the oxygen cascade. A normal initial "struggle response" to such an ascent includes increased ventilation—the cornerstone of acclimatization—mediated by the carotid bodies. Hyperventilation may cause respiratory alkalosis and dehydration. Respiratory alkalosis may be extreme, with an arterial blood pH of >7.7 (e.g., at the summit of Everest). Alkalosis may depress the ventilatory drive during sleep, with consequent periodic breathing and hypoxemia. During early acclimatization, renal suppression of carbonic anhydrase and excretion of dilute alkaline urine combat alkalosis and tend to bring the pH of the blood to normal. Other physiologic changes during normal acclimatization include increased sympathetic tone; increased erythropoietin levels, leading to increased hemoglobin levels and red blood cell mass; increased tissue capillary density and mitochondrial numbers; and higher levels of 2,3-bisphosphoglycerate, enhancing oxygen utilization. Even with normal acclimatization, however, ascent to a high altitude decreases maximal exercise capacity (by ~1% for every 100 m gained above 1500 m) and increases susceptibility to cold injury due to peripheral vasoconstriction. If the ascent is made faster than the body can adapt to the stress of hypobaric hypoxemia, altitude-related disease states can result.

GENETICS

Hypoxia-inducible factor, which acts as a master switch in high-altitude adaptation, controls transcriptional responses to hypoxia throughout the body and is involved in the release of vascular endothelial growth factor (VEGF) in the brain, erythropoiesis, and other pulmonary and cardiac functions at high altitudes. In particular, the gene *EPAS1*, which codes for transcriptional regulator hypoxia-inducible factor 2α, appears to play an important role in the adaptation of Tibetans living at high altitude, resulting in lower hemoglobin concentrations than are found in Han Chinese or South American highlanders. Other genes implicated include *EGLN1* and *PPARA*, which are also associated with hemoglobin concentration. Some evidence indicates that these genetic changes occurred within the past 3000 years, which is very fast in evolutionary terms. An intriguing question is whether the Sherpas' well-known mountain-climbing ability is partially attributable to their Tibetan ancestry, with overrepresentation of variants of *EPAS*. A striking recent finding is that some of these genetic characteristics may stem from those of Denisovan hominids who were contemporaries of the Neanderthals.

For acute altitude illness, a single gene variant is unlikely to be found, but differences in the susceptibility of individuals and populations, familial clustering of cases, and a positive association of some genetic variants all clearly support a role for genetics. Approximately 58 candidate genes have been tested, and at least one variant from 17 of these genes is associated with altitude illness.

ACUTE MOUNTAIN SICKNESS AND HIGH-ALTITUDE CEREBRAL EDEMA

AMS is a neurologic syndrome characterized by nonspecific symptoms (headache, nausea, fatigue, and dizziness), with a paucity of physical findings, developing 6–12 h after ascent to a high altitude. AMS is a clinical diagnosis. For uniformity in research studies, the Lake Louise Scoring System, created at the 1991 International Hypoxia Symposium, is generally used. AMS must be distinguished from exhaustion, dehydration, hypothermia, alcoholic hangover, and hyponatremia. AMS and HACE are thought to represent opposite ends of a continuum of altitude-related neurologic disorders. HACE (but not AMS) is an encephalopathy whose hallmarks are ataxia and altered consciousness with diffuse cerebral involvement but generally without focal neurologic deficits. Progression to these signal manifestations can be rapid. Papilledema and, more commonly, retinal hemorrhages may develop. In fact, retinal hemorrhages occur frequently at ≥5000 m, even in individuals without clinical symptoms of AMS or HACE. It is unclear whether retinal hemorrhage and cerebral hemorrhage at high altitude are caused by the same mechanism.

Risk Factors The most important risk factors for the development of altitude illness are the rate of ascent and a prior history of high-altitude illness. Exertion is a risk factor, but lack of physical fitness is not. An attractive but still speculative hypothesis proposes that AMS develops in people who have inadequate cerebrospinal capacity to buffer the brain swelling that occurs at high altitude. Children and adults seem to be equally affected, but people >50 years of age may be less likely to develop AMS than younger people. Most studies reveal no gender difference in AMS incidence. One study showed that, in women, adaptive responses to hypoxia with aging are blunted by menopause but can be maintained with endurance training. Sleep desaturation—a common phenomenon at high altitude—is associated with AMS. Debilitating fatigue consistent with severe AMS on descent from a summit is an important risk factor for death in mountaineers. A prospective study involving trekkers and climbers who ascended to altitudes between 4000 m and 8848 m showed that high oxygen desaturation and low ventilatory response to hypoxia during exercise are independent predictors of severe altitude illness. However, because there may be a large overlap between groups of susceptible and nonsusceptible individuals, accurate cutoff values are hard to define. Prediction is made more difficult because the pretest probabilities of HAPE and HACE are low. Neck irradiation or surgery damaging the carotid bodies, respiratory tract infections, and dehydration appear to be other potential risk factors for altitude illness. Unless guided by clinical signs and symptoms, pulse oximeter readings alone on a trek should not be used to predict AMS.

Pathophysiology Hypobaric hypoxia is the main trigger for altitude illness. In established AMS, raised intracranial pressure, increased sympathetic activity, relative hypoventilation, fluid retention and redistribution, and impaired gas exchange have all been well noted; these factors may play an important role in the pathophysiology of AMS. Severe hypoxemia can lead to a greater than normal increase in cerebral blood flow.

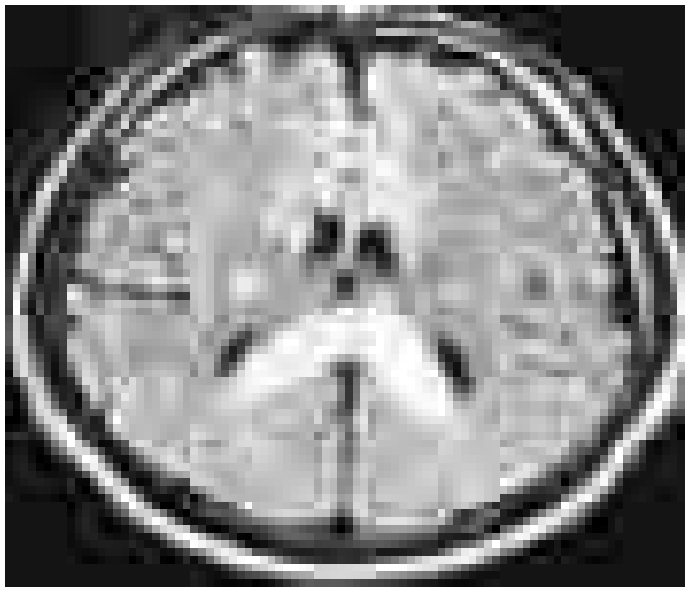


FIGURE 453-1 T2 magnetic resonance image of the brain of a patient with high-altitude cerebral edema (HACE) shows marked swelling and a hyperintense posterior body and splenium of the corpus callosum (area with dense opacity). The patient, a climber, went on to climb Mount Everest about 9 months after this episode of HACE. *(With permission from Wilderness Environ Med 15:53–55, 2004.)*

However, the exact mechanisms underlying AMS and HACE are unknown. Evidence points to a central nervous system process. MRI studies have suggested that vasogenic (interstitial) cerebral edema is a component of the pathophysiology of HACE. In the setting of high-altitude illness, the MRI findings shown in Fig. 453-1 are confirmatory of HACE, with increased signal in the white matter and particularly in the splenium of the corpus callosum. In addition, hemosiderin deposits in the corpus callosum have been characterized as long-lasting footprints of HACE. Quantitative analysis in a 3-tesla MRI study revealed that hypoxia is associated with mild vasogenic cerebral edema irrespective of AMS. This finding is in keeping with case reports of suddenly symptomatic brain tumors and of cranial nerve palsies without AMS at high altitudes. Vasogenic edema may become cytotoxic (intracellular) in severe HACE.

Impaired cerebral autoregulation in the presence of hypoxic cerebral vasodilation and altered permeability of the blood–brain barrier due to hypoxia-induced chemical mediators like histamine, arachidonic acid, and VEGF may all contribute to brain edema. In 1995, VEGF was first proposed as a potent promoter of capillary leakage in the brain at high altitude, and studies in mice have borne out this role. Although studies of VEGF in climbers have yielded inconsistent results regarding its association with altitude illness, indirect evidence of a role for this growth factor in AMS and HACE comes from the observation that dexamethasone, when used in the prevention and treatment of these conditions, blocks hypoxic upregulation of VEGF. Other factors in the development of cerebral edema may be the release of calcium-mediated nitric oxide and neuronally mediated adenosine, which may promote cerebral vasodilation. Venous outflow obstruction resulting in increased brain capillary pressure is also thought to play an important role in the development of HACE. Lesions in the globus pallidum (which is sensitive to hypoxia) leading to Parkinson's disease have been reported to be complications of HACE.

The pathophysiology of the most common and prominent symptom of AMS—headache—remains unclear because the brain itself is an insensate organ; only the meninges contain trigeminal sensory nerve fibers. The cause of high-altitude headache is multifactorial. Various chemicals and mechanical factors activate a final common pathway, the trigeminovascular system. In the genesis of high-altitude headache, the response to nonsteroidal anti-inflammatory drugs and glucocorticoids provides indirect evidence for involvement of the arachidonic acid pathway and inflammation. Although high altitude may be a trigger for migraine, it is unclear whether high-altitude headache and migraine share the same pathophysiology.

TABLE 453-1 Management of Altitude Illness

CONDITION	MANAGEMENT
Acute mountain sickness (AMS), mild[a]	Discontinuation of ascent Treatment with acetazolamide (250 mg q12h) Descent[b]
AMS, moderate[a]	Immediate descent for worsening symptoms Use of low-flow oxygen if available Treatment with acetazolamide (250 mg q12h) and/or dexamethasone (4 mg q6h)[c] Hyperbaric therapy[d]
High-altitude cerebral edema (HACE)	Immediate descent or evacuation Administration of oxygen (2–4 L/min) Treatment with dexamethasone (8 mg PO/IM/IV; then 4 mg q6h) Hyperbaric therapy if descent is not possible
High-altitude pulmonary edema (HAPE)	Immediate descent or evacuation Minimization of exertion while patient is kept warm Administration of oxygen (4–6 L/min) to bring O_2 saturation to >90% Adjunctive therapy with nifedipine[e] (30 mg, extended-release, q12h) Hyperbaric therapy if descent is not possible

[a]Categorization of cases as mild or moderate is a subjective judgment based on the severity of headache and the presence and severity of other manifestations (nausea, fatigue, dizziness, insomnia). [b]No fixed altitude is specified; the patient should descend to a point below that at which symptoms developed. [c]Acetazolamide treats and dexamethasone masks symptoms. For prevention (as opposed to treatment) of AMS, 125–250 mg of acetazolamide q12h or (when acetazolamide is contraindicated—e.g., in people with sulfa allergy) 4 mg of dexamethasone q12h may be used. [d]In hyperbaric therapy (Fig. 453-2), the patient is placed in a portable altitude chamber or bag to simulate descent. [e]Nifedipine at this dose is also effective for the prevention of HAPE, as is salmeterol (125 mg inhaled twice daily), tadalafil (10 mg twice daily), or dexamethasone (8 mg twice daily).

Prevention and Treatment (Table 453-1) Gradual ascent, with adequate time for acclimatization, is the best method for the prevention of altitude illness. Even though there may be individual variation in the rate of acclimatization, a conservative approach would be a graded ascent of ≤300 m from the previous day's sleeping altitude above 3000 m, and taking every third day of gain in sleeping altitude as an extra day for acclimatization is helpful. Spending one night at an intermediate altitude before proceeding to a higher altitude may enhance acclimatization and attenuate the risk of AMS. Another protective factor in AMS is high-altitude exposure during the preceding 2 months; for example, the incidence and severity of AMS at 4300 m are reduced by 50% with an ascent after 1 week at an altitude ≥2000 m rather than with an ascent from sea level. Studies have examined whether exposure to a normobaric hypoxic environment (in a room or a tent) before an ascent can provide protection against AMS. In double-blind placebo-controlled trials, repeated intermittent exposure (60–90 min) to normobaric hypoxia (up to 4500 m) or continuous exposure to 3000 m during 8 h of sleep for 7 consecutive days failed to reduce the incidence of AMS at altitudes of 4300–4559 m. However, a recent study with 14 consecutive nights in a similar setting demonstrated reduced symptoms and incidence of AMS. More uniform studies need to be done to clarify the utility, if any, of this approach to preacclimatization.

Clearly, a flexible itinerary that permits additional rest days will be helpful. Sojourners to high-altitude locations must be aware of the symptoms of altitude illness and should be encouraged not to ascend further if these symptoms develop. Any hint of HAPE (see below) or HACE mandates descent. Proper hydration (but not overhydration) in high-altitude trekking and climbing, aimed at countering fluid loss due to hyperventilation and sweating, may play a role in avoiding AMS. Pharmacologic prophylaxis at the time of travel to high altitudes is warranted for people with a history of AMS or when a graded ascent and acclimatization are not possible—e.g., when rapid ascent is necessary for rescue purposes or when flight to a high-altitude location is required. Acetazolamide is the drug of choice for AMS prevention. It inhibits renal carbonic anhydrase, causing prompt bicarbonate diuresis

that leads to metabolic acidosis and hyperventilation. Acetazolamide (125–250 mg twice a day), administered for 1 day before ascent and continued for 2 or 3 days, is effective. Treatment can be restarted if symptoms return after discontinuation of the drug. Higher doses are not required. A meta-analysis limited to randomized controlled trials revealed that 125 mg of acetazolamide twice daily was effective in the prevention of AMS, with a relative-risk reduction of ~48% from values obtained with placebo. Paresthesia and a tingling sensation are common side effects of acetazolamide. This drug is a nonantibiotic sulfonamide that has low-level cross-reactivity with sulfa antibiotics; as a result, severe reactions are rare. Dexamethasone (8 mg/d in divided doses) is also effective. A large-scale, randomized, double-blind, placebo-controlled trial in partially acclimatized trekkers clearly showed that *Ginkgo biloba* is ineffective in the prevention of AMS. In randomized studies, ibuprofen (600 mg three times daily) has been shown to be beneficial in the prevention of AMS. Recently, acetaminophen (1 g three times daily) was as effective as ibuprofen at the above dosage in a randomized, double-blind study. However, more definitive studies and (for ibuprofen) a proper gastrointestinal-bleeding risk assessment need to be conducted before these drugs can be routinely recommended for AMS prevention. Many drugs, including spironolactone, medroxyprogesterone, magnesium, calcium channel blockers, and antacids, confer no benefit in the prevention of AMS. Starkly conflicting results from a number of trials of inhaled budesonide for the prevention of AMS have recently been published, but, in all likelihood, the drug is ineffective. Similarly, no efficacy studies are available for coca leaves (a weak form of cocaine), which are offered to high-altitude travelers in the Andes, or for *soroche* pills, which contain aspirin, caffeine, and acetaminophen and are sold over the counter in Bolivia and Peru. Finally, a word of caution applies in the pharmacologic prevention of altitude illness. A fast-growing population of climbers in pursuit of a summit are using prophylactic drugs such as glucocorticoids in an attempt to improve their performance; the outcome can be tragic because of potentially severe side effects of these drugs.

For the treatment of mild AMS, rest alone with analgesic use may be adequate. Descent and the use of acetazolamide and (if available) oxygen are sufficient to treat most cases of moderate AMS. Even a minor descent (400–500 m) may be adequate for symptom relief. For moderate AMS or early HACE, dexamethasone (4 mg orally or parenterally) is highly effective. For HACE, immediate descent is mandatory. When descent is not possible because of poor weather conditions or darkness, a simulation of descent in a portable hyperbaric chamber **(Fig. 453-2)** can be very effective. Pressurization in the bag for 1–2 h often leads to spectacular improvement and, like dexamethasone administration, "buys time." Thus, in certain high-altitude locations (e.g., remote pilgrimage sites), the decision to bring along the lightweight hyperbaric chamber may prove lifesaving. Like nifedipine, phosphodiesterase-5 inhibitors have no role in the treatment of AMS or HACE.

FIGURE 453-2 A hyperbaric bag. The cylindrical, portable (<7 kg) nylon bag has a one-way valve to prevent carbon dioxide buildup. A patient with severe acute mountain sickness (AMS), high-altitude cerebral edema (HACE), or high-altitude pulmonary edema (HAPE) is zipped inside the bag, which is continuously inflated with a foot pedal. The increased barometric pressure (2 psi) inside the bag simulates descent; for example, at 4250 m, the equivalent "elevation" inside the bag is ~2100 m. No supplemental oxygen is required.

HIGH-ALTITUDE PULMONARY EDEMA

Risk Factors and Manifestations Unlike HACE (a neurologic disorder), HAPE is primarily a pulmonary problem and therefore is not necessarily preceded by AMS. HAPE develops within 2–4 days after arrival at high altitude; it rarely occurs after more than 4 or 5 days at the same altitude, probably because of remodeling and adaptation that render the pulmonary vasculature less susceptible to the effects of hypoxia. A rapid rate of ascent, a history of HAPE, respiratory tract infections, and cold environmental temperatures are risk factors. Men are more susceptible than women. People with abnormalities of the cardiopulmonary circulation leading to pulmonary hypertension—e.g., a large patent foramen ovale, mitral stenosis, primary pulmonary hypertension, and unilateral absence of the pulmonary artery—may be at increased risk of HAPE, even at moderate altitudes. For example, patent foramen ovale is four times more common among HAPE-susceptible individuals than in the general population. Echocardiography is recommended when HAPE develops at relatively low altitudes (<3000 m) and whenever cardiopulmonary abnormalities predisposing to HAPE are suspected.

The initial manifestation of HAPE may be a reduction in exercise tolerance greater than that expected at the given altitude. Although a dry, persistent cough may presage HAPE and may be followed by the production of blood-tinged sputum, cough in the mountains is almost universal and the mechanism is poorly understood. Tachypnea and tachycardia, even at rest, are important markers as illness progresses. Crackles may be heard on auscultation but are not diagnostic. HAPE may be accompanied by signs of HACE. Patchy or localized opacities **(Fig. 453-3)** or streaky interstitial edema may be noted on chest radiography. In the past, HAPE was mistaken for pneumonia due to the cold or for heart failure due to hypoxia and exertion. Kerley B lines or a bat-wing appearance are not seen on radiography. Electrocardiography may reveal right ventricular strain or even hypertrophy. Hypoxemia and respiratory alkalosis are consistently present unless the patient is taking acetazolamide, in which case metabolic acidosis may supervene. Assessment of arterial blood gases is not necessary in the evaluation of HAPE; an oxygen saturation reading with a pulse oximeter is generally adequate. The existence of a subclinical form of HAPE has been suggested by an increased alveolar–arterial oxygen gradient in Everest climbers near the summit, but hard evidence correlating this abnormality with the development of clinically relevant HAPE is lacking. Comet-tail scoring—an ultrasound technique—has been used

FIGURE 453-3 Chest radiograph of a patient with high-altitude pulmonary edema shows opacity in the right middle and lower zones simulating pneumonic consolidation. The opacity cleared almost completely in 2 days with descent and supplemental oxygen.

for evaluation of extravascular lung water at high altitude; the results have been encouraging, and this technique may potentially prove useful in detecting HAPE (clinical or subclinical) and even in ascertaining whether the presence of extravascular lung water is a harbinger of HAPE in patients with AMS.

Pathophysiology HAPE is a noncardiogenic pulmonary edema with normal pulmonary artery wedge pressure. It is characterized by patchy pulmonary hypoxic vasoconstriction that leads to overperfusion in some areas. This abnormality leads in turn to increased pulmonary capillary pressure (>18 mmHg) and capillary "stress" failure. The exact mechanism for this hypoxic vasoconstriction is unknown. Endothelial dysfunction due to hypoxia may play a role by impairing the release of nitric oxide, an endothelium-derived vasodilator. At high altitude, HAPE-prone persons have reduced levels of exhaled nitric oxide. The effectiveness of phosphodiesterase-5 inhibitors in alleviating altitude-induced pulmonary hypertension, decreased exercise tolerance, and hypoxemia supports the role of nitric oxide in the pathogenesis of HAPE. One study demonstrated that prophylactic use of tadalafil, a phosphodiesterase-5 inhibitor, decreases the risk of HAPE by 65%. In contrast, the endothelium also synthesizes endothelin-1, a potent vasoconstrictor whose concentrations are higher than average in HAPE-prone mountaineers.

Exercise and cold lead to increased pulmonary intravascular pressure and may predispose to HAPE. In addition, hypoxia-triggered increases in sympathetic drive may lead to pulmonary venoconstriction and extravasation into the alveoli from the pulmonary capillaries. Consistent with this concept, phentolamine, which elicits α-adrenergic blockade, improves hemodynamics and oxygenation in HAPE more than do other vasodilators. The study of tadalafil cited above also investigated dexamethasone in the prevention of HAPE. Surprisingly, dexamethasone reduced the incidence of HAPE by 78%—a greater decrease than with tadalafil. Besides possibly increasing the availability of endothelial nitric oxide, dexamethasone may have altered the excessive sympathetic activity associated with HAPE: the heart rate of participants in the dexamethasone arm of the study was significantly lowered. Finally, people susceptible to HAPE also display enhanced sympathetic activity during short-term hypoxic breathing at low altitudes.

Because many patients with HAPE have fever, peripheral leukocytosis, and an increased erythrocyte sedimentation rate, inflammation has been considered an etiologic factor in HAPE. However, strong evidence suggests that inflammation in HAPE is an epiphenomenon rather than the primary cause. Nevertheless, inflammatory processes (e.g., those elicited by viral respiratory tract infections) do predispose persons to HAPE—even those who are constitutionally resistant to its development.

Another proposed mechanism for HAPE is impaired transepithelial clearance of sodium and water from the alveoli. β-Adrenergic agonists upregulate the clearance of alveolar fluid in animal models. In a double-blind, randomized, placebo-controlled study of HAPE-susceptible mountaineers, prophylactic inhalation of the β-adrenergic agonist salmeterol reduced the incidence of HAPE by 50%. Other effects of β agonists may also contribute to the prevention of HAPE, but these findings are in keeping with the concept that alveolar fluid clearance may play a pathogenic role in this illness.

Prevention and Treatment (Table 453-1) Allowing sufficient time for acclimatization by ascending gradually (as discussed above for AMS and HACE) is the best way to prevent HAPE. Sustained-release nifedipine (30 mg), given once or twice daily, prevents HAPE in people who must ascend rapidly or who have a history of HAPE. Other drugs for the prevention of HAPE are listed in Table 453-1 (footnote e). Although dexamethasone is listed for prevention, its adverse-effect profile requires close monitoring. Acetazolamide has been shown to blunt hypoxic pulmonary vasoconstriction in animal models, and this observation warrants further study in HAPE prevention. However, one large study failed to show a decrease in pulmonary vasoconstriction in partially acclimatized individuals given acetazolamide.

Early recognition is paramount in the treatment of HAPE, especially when it is not preceded by the AMS symptoms of headache and nausea. Fatigue and dyspnea at rest may be the only initial manifestations. Descent and the use of supplementary oxygen (aimed at bringing oxygen saturation to >90%) are the most effective therapeutic interventions. Exertion should be kept to a minimum, and the patient should be kept warm. Hyperbaric therapy (Fig. 453-2) in a portable altitude chamber may be lifesaving, especially if descent is not possible and oxygen is not available. Oral sustained-release nifedipine (30 mg once or twice daily) can be used as adjunctive therapy. Inhaled β agonists, which are safe and convenient to carry, are useful in the prevention of HAPE and may be effective in its treatment, although no trials have yet been carried out. Inhaled nitric oxide and expiratory positive airway pressure may also be useful therapeutic measures but may not be available in high-altitude settings. No studies have investigated phosphodiesterase-5 inhibitors in the treatment of HAPE, but reports have described their use in clinical practice. The mainstays of treatment remain descent and (if available) oxygen.

In AMS, if symptoms abate (with or without acetazolamide), the patient may reascend gradually to a higher altitude. Unlike that in acute respiratory distress syndrome (another noncardiogenic pulmonary edema), the architecture of the lung in HAPE is usually well-preserved, with rapid reversibility of abnormalities (Fig. 453-3). This fact has allowed some people with HAPE to reascend slowly after a few days of descent and rest. In HACE, reascent after a few days may not be advisable during the same trip.

OTHER HIGH-ALTITUDE PROBLEMS

Sleep Impairment The mechanisms underlying sleep problems, which are among the most common adverse reactions to high altitude, include increased periodic breathing; changes in sleep architecture, with increased time in lighter sleep stages; and changes in rapid eye movement sleep. Sojourners should be reassured that sleep quality improves with acclimatization. In cases where drugs do need to be used, acetazolamide (125 mg before bedtime) is especially useful because this agent decreases hypoxemic episodes and alleviates sleeping disruptions caused by excessive periodic breathing. Whether combining acetazolamide with temazepam or zolpidem is more effective than administering acetazolamide alone is unknown. In combinations, the doses of temazepam and zolpidem should not be increased by >10 mg at high altitudes. Limited evidence suggests that diazepam causes hypoventilation at high altitudes and therefore is contraindicated. For trekkers with obstructive sleep apnea who are using a continuous positive airway pressure (CPAP) machine, the addition of acetazolamide, which will decrease centrally mediated sleep apnea, may be helpful. There is evidence to show that obstructive sleep apnea at high altitude may decrease and "convert" to central sleep apnea.

Gastrointestinal Issues High-altitude exposure may be associated with increased gastric and duodenal bleeding, but further studies are required to determine whether there is a causal effect. Because of decreased atmospheric pressure and consequent intestinal gas expansion at high altitudes, many sojourners experience abdominal bloating and distension as well as excessive flatus expulsion. In the absence of diarrhea, these phenomena are normal, if sometimes uncomfortable. Accompanying diarrhea, however, may indicate the involvement of bacteria or *Giardia* parasites, which are common at many high-altitude locations in the developing world. Prompt treatment with fluids and empirical antibiotics may be required to combat dehydration in the mountains. Hemorrhoids are common on high-altitude treks; treatment includes hot soaks, application of hydrocortisone ointment, and measures to avoid constipation.

High-Altitude Cough High-altitude cough can be debilitating and is sometimes severe enough to cause rib fracture, especially at >5000 m. The etiology of this common problem is probably multifactorial. Although high-altitude cough has been attributed to inspiration of cold dry air, this explanation appears not to be sufficient by itself; in long-duration studies in hypobaric chambers, cough has occurred

despite controlled temperature and humidity. The implication is that hypoxia also plays a role. Exercise can precipitate cough at high altitudes, possibly because of water loss from the respiratory tract. Long-acting β agonists and glucocorticoids prevent bronchoconstriction that otherwise may be brought on by cold and exercise. In general, infection does not seem to be a common etiology. Anecdotal reports have described the efficacy of an inhaled combination of fluticasone and salmeterol in the treatment of high-altitude cough. Many trekkers find it useful to wear a balaclava to trap some moisture and heat. In most situations, cough resolves upon descent.

High-Altitude Neurologic Events Unrelated to "Altitude Illness" Transient ischemic attacks (TIAs) and strokes have been well described in high-altitude sojourners outside the setting of altitude sickness. However, these descriptions are not based on cause (hypoxia) and effect. In general, symptoms of AMS present gradually, whereas many of these neurologic events happen suddenly. The population that suffers strokes and TIAs at sea level is generally an older age group with other risk factors, whereas those so afflicted at high altitudes are generally younger and probably have fewer risk factors for atherosclerotic vascular disease. Other mechanisms (e.g., migraine, vasospasm, focal edema, hypocapneic vasoconstriction, hypoxia in the watershed zones of minimal cerebral blood flow, or cardiac right-to-left shunt) may be operative in TIAs and strokes at high altitude.

Subarachnoid hemorrhage, transient global amnesia, delirium, and cranial nerve palsies (e.g., lateral rectus palsy) occurring at high altitudes but outside the setting of altitude sickness have been well described. Syncope is common at moderately high altitudes, generally occurs shortly after ascent, usually resolves without descent, and appears to be a vasovagal event related to hypoxemia. Seizures occur rarely with HACE, but hypoxemia and hypocapnia, which are prevalent at high altitudes, are well-known triggers that may contribute to new or breakthrough seizures in predisposed individuals. Nevertheless, the consensus among experts is that sojourners with well-controlled seizure disorders can ascend to high altitudes.

Finally, persons with hypercoagulable conditions (e.g., antiphospholipid syndrome, protein C deficiency) who are asymptomatic at sea level may experience cerebral venous thrombosis (possibly due to enhanced blood viscosity triggered by polycythemia and dehydration) at high altitudes. Proper history taking, examination, and prompt investigations where possible will help define these conditions as entities separate from altitude sickness. Administration of oxygen (where available) and prompt descent are the cornerstones of treatment of most of these neurologic conditions.

Ocular Problems Ocular issues are common in sojourners to high altitudes. Hypoxemia induced by altitude leads to increased retinal blood flow, which can be visible as engorged retinal veins on ophthalmoscopic examination. Both high flow and hypoxemic vascular damage causing permeability have been implicated in a breakdown of the blood–retina barrier and the formation of retinal hemorrhages. Blot, dot, flame, and white-centered hemorrhages can be observed. These hemorrhages usually resolve spontaneously with descent, with only mild symptoms and no lasting visual damage in most healthy eyes. The exception is hemorrhage in the macular area. Macular hemorrhages can cause devastating initial visual loss, particularly if bilateral, and have been reported to cause permanently decreased vision in a few cases.

Stroke syndromes such as retinal vein occlusion, retinal artery occlusion, ischemic optic neuropathy, and cortical visual loss have all been reported. With unilateral vision loss, it is always important to check for a relative afferent pupillary defect. Increased hematocrit combined with dehydration may contribute to these maladies. Glaucomatous optic nerve damage may progress with hypoxemia of altitude. Acetazolamide is helpful both in combating the respiratory alkalosis that comes with increased ventilation at high altitude and in lowering the interocular pressure; its use should be considered in patients with stable controlled glaucoma. Macular degeneration and diabetic eye disease are not directly exacerbated by ascent to high altitude. Dry eye and solar damage to the cornea, known as "snow blindness," are common. Wearing of high-quality UV blocking sunglasses, even on cloudy days, and attention to protecting and supplementing the tear film with artificial tear drops can greatly improve comfort and vision. Although modern refractive surgeries, such as photorefractive keratectomy (PRK) and laser in situ keratomileusis (LASIK), are stable at high altitude, patients who have undergone radial keratotomy should be cautioned that hypoxemia to the cornea can lead to swelling that shifts the refraction during ascent.

Psychological/Psychiatric Problems Delirium characterized by a sudden change in mental status, a short attention span, disorganized thinking, and an agitated state during the period of confusion has been well described in mountain climbers and trekkers without a prior history. In addition, anxiety attacks, often triggered at night by excessive periodic breathing, are well documented. The contribution of hypoxia to these conditions is unknown. Expedition medical kits need to include antipsychotic injectable drugs to control psychosis in patients in remote high-altitude locations.

PREEXISTING MEDICAL ISSUES

Because travel to high altitudes is increasingly popular, common conditions such as hypertension, coronary artery disease, and diabetes are more frequently encountered among high-altitude sojourners. This situation is of particular concern for the millions of elderly pilgrims with medical problems who visit high-altitude sacred areas (e.g., in the Himalayas) each year. In recent years, high-altitude travel has attracted intrepid trekkers who are taking immunosuppressive medications (e.g., kidney transplant recipients or patients undergoing chemotherapy). Recommended vaccinations and other precautions (e.g., hand washing) may be especially important for this group. Although most of these medical conditions do not appear to influence susceptibility to altitude illness, they may be exacerbated by ascent to altitude, exertion in cold conditions, and hypoxemia. Advice regarding the advisability of high-altitude travel and the impact of high-altitude hypoxia on these preexisting conditions is becoming increasingly relevant, but there are no evidence-based guidelines. In addition, recommendations made for relatively low altitudes (~3000 m) may not hold true for higher altitudes (>4000 m), where hypoxic stress is greater. Personal risks and benefits must be clearly thought through before ascent.

Hypertension At high altitudes, enhanced sympathetic activity may lead to a transient rise in blood pressure. Occasionally, nonhypertensive, healthy, asymptomatic trekkers have pathologically high blood pressure at high altitude that rapidly normalizes without medicines on descent. Sojourners should continue to take their antihypertensive medications at high altitudes. Hypertensive patients are not more likely than others to develop altitude illness. Because the probable mechanism of high-altitude hypertension is α-adrenergic activity, anti-α-adrenergic drugs like prazosin have been suggested for symptomatic patients and those with labile hypertension. It is best to start taking the drug several weeks before the trip and to carry a sphygmomanometer if a trekker has labile hypertension. Sustained-release nifedipine may also be useful. A recent observational cohort study of 672 hypertensive and nonhypertensive trekkers in the Himalayas showed that most travelers, including those with well-controlled hypertension, can be reassured that their blood pressure will remain relatively stable at high altitude. Although blood pressure may be extremely elevated at high altitude in normotensive and hypertensive people, it is unlikely to cause symptoms.

Coronary Artery Disease Myocardial oxygen demand and maximal heart rate are reduced at high altitudes because the VO_2 max (maximal oxygen consumption) decreases with increasing altitude. This effect may explain why signs of cardiac ischemia or dysfunction usually are not seen in healthy persons at high altitudes. Asymptomatic, fit individuals with no risk factors need not undergo any tests for coronary artery disease before ascent. For persons with ischemic heart disease, previous myocardial infarction, angioplasty, and/or bypass surgery, an exercise treadmill test is indicated. A strongly positive treadmill test is a contraindication for high-altitude trips. Patients with poorly controlled arrhythmias should avoid high-altitude travel, but

patients with arrhythmias that are well controlled with antiarrhythmic medications do not seem to be at increased risk. Sudden cardiac deaths are not noted with a greater frequency in the Alps than at lower altitudes; although sudden cardiac deaths are encountered every trekking season in the higher Himalayan range, accurate documentation is lacking.

Asthma Although cold air and exercise may provoke acute bronchoconstriction, asthmatic patients usually have fewer problems at high than at low altitudes, possibly because of decreased allergen levels and increased circulating catecholamine levels. Nevertheless, asthmatic individuals should carry all their medications, including oral glucocorticoids, with proper instructions for use in case of an exacerbation. Severely asthmatic persons should be cautioned against ascending to high altitudes.

Pregnancy In general, low-risk pregnant women ascending to 3000 m are not at special risk except for the relative unavailability of medical care in many high-altitude locations, especially in developing countries. Despite the lack of firm data on this point, venturing higher than 3000 m to altitudes at which oxygen saturation drops steeply seems unadvisable for pregnant women.

Obesity Although living at a high altitude has been suggested as a means of controlling obesity, obesity has also been reported to be a risk factor for AMS, probably because nocturnal hypoxemia is more pronounced in obese individuals. Hypoxemia may also lead to greater pulmonary hypertension, thus possibly predisposing the trekker to HAPE.

Sickle Cell Disease High altitude is one of the rare environmental exposures that occasionally provokes a crisis in persons with the sickle cell trait. Even when traversing mountain passes as low as 2500 m, people with sickle cell disease have been known to have a vaso-occlusive crisis. Sickle cell disease needs to be considered when persons traveling to high altitudes become unwell and develop left-upper-quadrant pain. Patients with known sickle cell disease who need to travel to high altitudes should use supplemental oxygen and travel with caution. Thalassemia has not been known to cause problems at high altitude.

Diabetes Mellitus Trekking at high altitudes may enhance sugar uptake. Thus, high-altitude travel may not pose problems for persons with diabetes that is well controlled with oral hypoglycemic agents. An eye examination before travel may be useful. Patients taking insulin may require lower doses on trekking/climbing days than on rest days. Because of these variations, diabetic patients need to carry a reliable glucometer and use fast-acting insulin. Ready access to sweets is also essential. It is important for companions of diabetic trekkers to be fully aware of potential problems like hypoglycemia.

Chronic Lung Disease Depending on disease severity and access to medical care, preexisting lung disease may not always preclude high-altitude travel. A proper pretravel evaluation must be conducted. Supplemental oxygen may be required if the predicted PaO_2 for the altitude is <50–55 mmHg. Preexisting pulmonary hypertension may also need to be assessed in these patients. If the result is positive, patients should be discouraged from ascending to high altitudes; if such travel is necessary, treatment with sustained-release nifedipine (20 mg twice a day) should be considered. Small-scale studies have revealed that when patients with bullous disease reach ~5000 m, bullous expansion and pneumothorax are not noted. Compared with information on chronic obstructive pulmonary disease, fewer data exist about the safety of travel to high altitude for people with pulmonary fibrosis, but acute exacerbation of pulmonary fibrosis has been seen at high altitude. A handheld pulse oximeter can be useful to check for oxygen saturation.

Chronic Kidney Disease Patients with chronic kidney disease can tolerate short-term stays at high altitudes, but theoretical concern persists about progression to end-stage renal disease. Acetazolamide, the drug most commonly used for altitude sickness, should be avoided by anyone with preexisting metabolic acidosis, which can be exacerbated by this drug. In addition, the acetazolamide dosage should be adjusted when the glomerular filtration rate falls to <50 mL/min, and the drug should not be used at all if this value falls to <10 mL/min.

Cirrhosis Of patients with cirrhosis, 16% may have portopulmonary arterial hypertension and 32% may have hepatopulmonary syndrome; these conditions may be detrimental at high altitude as they may cause exaggerated hypoxemia. Thus, screening for these problems is important in cirrhotic patients planning a high-altitude trip. In addition, acetazolamide may be inadvisable in these patients as the drug may increase the risk of hepatic encephalopathy.

CHRONIC MOUNTAIN SICKNESS AND HIGH-ALTITUDE PULMONARY HYPERTENSION IN HIGHLANDERS

The largest populations of highlanders live in the South American Andes, the Tibetan Plateau, and parts of Ethiopia. Chronic mountain sickness (*Monge's disease*) is a disease in highlanders that is characterized by excessive erythrocytosis with moderate to severe pulmonary hypertension leading to cor pulmonale. This condition was originally described in South America and has also been documented in Colorado and in the Han Chinese population in Tibet; it is much less common in Tibetans or in Ethiopian highlanders. Migration to a low altitude results in the resolution of chronic mountain illness. Venesection and acetazolamide are helpful.

High-altitude pulmonary hypertension is also a subacute disease of long-term high-altitude residents. Unlike Monge's disease, this syndrome is characterized primarily by pulmonary hypertension (not erythrocytosis) leading to heart failure. Indian soldiers living at extreme altitudes for prolonged periods and Han Chinese infants born in Tibet have presented with the adult and infantile forms, respectively. High-altitude pulmonary hypertension bears a striking pathophysiologic resemblance to brisket disease in cattle. Descent to a lower altitude is curative.

FURTHER READING

Basnyat B: High altitude pilgrimage medicine. High Alt Med Biol 15:434, 2014.

Basnyat B, Murdoch D: High altitude illness. Lancet 361:1967, 2003.

Basnyat B et al: Acute medical problems in the Himalayas outside the setting of altitude sickness. High Alt Med Biol 1:167, 2000.

Kanaan NC et al: Prophylactic acetaminophen or ibuprofen results in equivalent acute mountain sickness incidence at high altitude: A prospective randomized trial. Wilderness Environ Med 28:72, 2017.

Roach RC et al: Mountain medicine, in *Wilderness Medicine*, 7th ed. PS Auerbach et al (eds). Philadelphia, Elsevier, 2016, pp 2–39.

454 Hypothermia and Peripheral Cold Injuries

Daniel F. Danzl

HYPOTHERMIA

Accidental hypothermia occurs when there is an unintentional drop in the body's core temperature below 35°C (95°F). At this temperature, many of the compensatory physiologic mechanisms that conserve heat begin to fail. *Primary accidental hypothermia* is a result of the direct exposure of a previously healthy individual to the cold. The mortality rate is much higher for patients who develop *secondary hypothermia* as a complication of a serious systemic disorder or injury.

CAUSES

Primary accidental hypothermia is geographically and seasonally pervasive. Although most cases occur in the winter months and in colder

TABLE 454-1 Risk Factors for Hypothermia

Age extremes	Endocrine-related
Elderly	Diabetes mellitus
Neonates	Hypoglycemia
Environmental exposure	Hypothyroidism
Occupational	Adrenal insufficiency
Sports-related	Hypopituitarism
Inadequate clothing	Neurologic
Immersion	Cerebrovascular accident
Toxicologic and pharmacologic	Hypothalamic disorders
Ethanol	Parkinson's disease
Phenothiazines	Spinal cord injury
Barbiturates	Multisystemic
Anesthetics	Trauma
Neuromuscular blockers	Sepsis
Antidepressants	Shock
Insufficient fuel	Hepatic or renal failure
Malnutrition	Carcinomatosis
Marasmus	Burns and exfoliative dermatologic disorders
Kwashiorkor	Immobility or debilitation

climates, this condition is surprisingly common in warmer regions as well. Multiple variables render individuals at the extremes of age—both the elderly and neonates—particularly vulnerable to hypothermia **(Table 454-1)**. The elderly have diminished thermal perception and are more susceptible to immobility, malnutrition, and systemic illnesses that interfere with heat generation or conservation. Dementia, psychiatric illness, and socioeconomic factors often compound these problems by impeding adequate measures to prevent hypothermia. Neonates have high rates of heat loss because of their increased surface-to-mass ratio and their lack of effective shivering and adaptive behavioral responses. At all ages, malnutrition can contribute to heat loss because of diminished subcutaneous fat and as a result of depleted energy stores used for thermogenesis.

Individuals whose occupations or hobbies entail extensive exposure to cold weather are at increased risk for hypothermia. Military history is replete with hypothermic tragedies. Hunters, sailors, skiers, and climbers also are at great risk of exposure, whether it involves injury, changes in weather, or lack of preparedness.

Ethanol causes vasodilation (which increases heat loss), reduces thermogenesis and gluconeogenesis, and may impair judgment or lead to obtundation. Phenothiazines, barbiturates, benzodiazepines, heterocyclic antidepressants, and other medications reduce centrally mediated vasoconstriction. Many hypothermic patients are admitted to intensive care because of drug overdose. Anesthetics can block shivering responses; their effects are compounded when patients are not insulated adequately in the operating or recovery units.

Several types of endocrine dysfunction can lead to hypothermia. Hypothyroidism—particularly when extreme, as in myxedema coma—reduces the metabolic rate and impairs thermogenesis and behavioral responses. Adrenal insufficiency and hypopituitarism also increase susceptibility to hypothermia. Hypoglycemia, most commonly caused by insulin or oral hypoglycemic drugs, is associated with hypothermia, in part because of neuroglycopenic effects on hypothalamic function. Increased osmolality and metabolic derangements associated with uremia, diabetic ketoacidosis, and lactic acidosis can lead to altered hypothalamic thermoregulation.

Neurologic injury from trauma, cerebrovascular accident, subarachnoid hemorrhage, and hypothalamic lesion increases susceptibility to hypothermia. Agenesis of the corpus callosum (*Shapiro's syndrome*) is one cause of episodic hypothermia; in this syndrome, profuse perspiration is followed by a rapid fall in temperature. Acute spinal cord injury disrupts the autonomic pathways that lead to shivering and prevents cold-induced reflex vasoconstrictive responses.

Hypothermia associated with sepsis is a poor prognostic sign. Hepatic failure causes decreased glycogen storage and gluconeogenesis as well as a diminished shivering response. In acute myocardial infarction associated with low cardiac output, hypothermia may be reversed after adequate resuscitation. With extensive burns, psoriasis, erythrodermas, and other skin diseases, increased peripheral-blood flow leads to excessive heat loss.

THERMOREGULATION

Heat loss occurs through five mechanisms: radiation (55–65% of heat loss), conduction (10–15% of heat loss, much increased in cold water), convection (increased in the wind), respiration, and evaporation; both of the latter two mechanisms are affected by the ambient temperature and the relative humidity.

The preoptic anterior hypothalamus normally orchestrates thermoregulation **(Chap. 15)**. The immediate defense of thermoneutrality is via the autonomic nervous system, whereas delayed control is mediated by the endocrine system. Autonomic nervous system responses include the release of norepinephrine, increased muscle tone, and shivering, leading to thermogenesis and an increase in the basal metabolic rate. Cutaneous cold thermoreception causes direct reflex vasoconstriction to conserve heat. Prolonged exposure to cold also stimulates the thyroid axis, leading to an increased metabolic rate.

CLINICAL PRESENTATION

In most cases of hypothermia, the history of exposure to environmental factors (e.g., prolonged exposure to the outdoors without adequate clothing) makes the diagnosis straightforward. In urban settings, however, the presentation is often more subtle, and other disease processes, toxin exposures, or psychiatric diagnoses should be considered. Predicting the core temperature based on the clinical presentation is very difficult.

After initial stimulation by hypothermia, there is progressive depression of all organ systems. The timing of the appearance of these clinical manifestations varies widely **(Table 454-2)**. Without knowing the core temperature, it can be difficult to interpret other vital signs. For example, tachycardia disproportionate to the core temperature suggests secondary hypothermia resulting from hypoglycemia, hypovolemia, or a toxin overdose. Because carbon dioxide production declines progressively, the respiratory rate should be low; persistent hyperventilation suggests a central nervous system (CNS) lesion or one of the organic acidoses. A markedly depressed level of consciousness in a patient with mild hypothermia raises suspicion of an overdose or CNS dysfunction due to infection or trauma.

Physical examination findings can also be altered by hypothermia. For instance, the assumption that areflexia is solely attributable to hypothermia can obscure and delay the diagnosis of a spinal cord lesion. Patients with hypothermia may be confused or combative; these symptoms abate more rapidly with rewarming than with chemical or physical restraint. A classic example of maladaptive behavior in patients with hypothermia is paradoxical undressing, which involves the inappropriate removal of clothing in response to a cold stress. The cold-induced ileus and abdominal rectus spasm can mimic or mask the presentation of an acute abdomen **(Chap. 12)**.

When a patient in hypothermic cardiac arrest is first discovered, cardiopulmonary resuscitation (CPR) is indicated unless (1) a do-not-resuscitate status is verified, (2) obviously lethal injuries are identified, or (3) the depression of a frozen chest wall is not possible. Continuous CPR is normally recommended and interruptions should be avoided. In the field, when the core temperature is below 28°C, intermittent CPR may be effective.

As the resuscitation proceeds, the prognosis is grave if there is evidence of widespread cell lysis, as reflected by potassium levels >10–12 mmol/L (10–12 meq/L). Other findings that may preclude continuing resuscitation include a core temperature <10–12°C (<50–54°F), a pH <6.5, and evidence of intravascular thrombosis with a fibrinogen value <0.5 g/L (<50 mg/dL). The decision to terminate resuscitation before rewarming the patient past 33°C (91°F) should be predicated on the type and severity of the precipitants of hypothermia. Survival

TABLE 454-2 Physiologic Changes Associated with Accidental Hypothermia

SEVERITY	BODY TEMPERATURE	CENTRAL NERVOUS SYSTEM	CARDIOVASCULAR	RESPIRATORY	RENAL AND ENDOCRINE	NEUROMUSCULAR
Mild	35°C (95°F)– 32.2°C (90°F)	Linear depression of cerebral metabolism; amnesia; apathy; dysarthria; impaired judgment; maladaptive behavior	Tachycardia, then progressive bradycardia; cardiac cycle prolongation; vasoconstriction; increase in cardiac output and blood pressure	Tachypnea, then progressive decrease in respiratory minute volume; declining oxygen consumption; bronchorrhea; bronchospasm	Diuresis; increase in catecholamines, adrenal steroids, triiodothyronine, and thyroxine; increase in metabolism with shivering	Increased preshivering muscle tone, then fatiguing
Moderate	<32.2°C (90°F)–28°C (82.4°F)	EEG abnormalities; progressive depression of level of consciousness; pupillary dilation; paradoxical undressing; hallucinations	Progressive decrease in pulse and cardiac output; increased atrial and ventricular arrhythmias; suggestive (J-wave) ECG changes	Hypoventilation; 50% decrease in carbon dioxide production per 8°C (17.6°F) drop in temperature; absence of protective airway reflexes	50% increase in renal blood flow; renal autoregulation intact; impaired insulin action	Hyporeflexia; diminishing shivering-induced thermogenesis; rigidity
Severe	<28°C (<82.4°F)	Loss of cerebrovascular autoregulation; decline in cerebral blood flow; coma; loss of ocular reflexes; progressive decrease in EEG abnormalities	Progressive decrease in blood pressure, heart rate, and cardiac output; reentrant dysrhythmias; maximal risk of ventricular fibrillation; asystole	Pulmonic congestion and edema; 75% decrease in oxygen consumption; apnea	Decrease in renal blood flow that parallels decrease in cardiac output; extreme oliguria; poikilothermia; 80% decrease in basal metabolism	No motion; decreased nerve-conduction velocity; peripheral areflexia; no corneal or oculocephalic reflexes

Abbreviations: ECG, electrocardiogram; EEG, electroencephalogram.

Source: Modified from DF Danzl, RS Pozos: N Engl J Med 331:1756, 1994.

has occurred with a cardiac arrest time over 7 h. There are no validated prognostic indicators for recovery from hypothermia. A history of asphyxia with secondary cooling is the most important negative predictor of survival.

DIAGNOSIS AND STABILIZATION

Hypothermia is confirmed by measurement of the core temperature, preferably at two sites. Rectal probes should be placed to a depth of 15 cm and not adjacent to cold feces. A simultaneous esophageal probe should be placed 24 cm below the larynx; it may read falsely high during heated inhalation therapy. Relying solely on infrared tympanic thermography is not advisable.

After a diagnosis of hypothermia is established, cardiac monitoring should be instituted, along with attempts to limit further heat loss. If the patient is in ventricular fibrillation, it is unclear at what core temperature ventricular defibrillation (2 J/kg) should first be attempted. One attempt below 30°C is warranted. Further defibrillation attempts should be deferred until some rewarming (1°–2°C) is achieved and ventricular fibrillation is coarser. Although cardiac pacing for hypothermic bradydysrhythmias is rarely indicated, the transthoracic technique is preferable.

Supplemental oxygenation is always warranted, since tissue oxygenation is affected adversely by the leftward shift of the oxyhemoglobin dissociation curve. Pulse oximetry may be unreliable in patients with vasoconstriction. If protective airway reflexes are absent, gentle endotracheal intubation should be performed. Adequate preoxygenation will prevent ventricular arrhythmias.

Insertion of a gastric tube prevents dilation secondary to decreased bowel motility. Indwelling bladder catheters facilitate monitoring of cold-induced diuresis. Dehydration is encountered commonly with chronic hypothermia, and most patients benefit from an intravenous or intraosseous bolus of crystalloid. Normal saline is preferable to lactated Ringer's solution, as the liver in hypothermic patients inefficiently metabolizes lactate. The placement of a pulmonary artery catheter can cause perforation of the less compliant pulmonary artery. Insertion of a central venous catheter deeply into the cold right atrium should be avoided since this procedure, similar to transvenous pacing, can precipitate arrhythmias.

Arterial blood gases should not be corrected for temperature (Chap. 51). An uncorrected pH of 7.42 and a Pco_2 of 40 mmHg reflect appropriate alveolar ventilation and acid-base balance at any core temperature. Acid-base imbalances should be corrected gradually, since the bicarbonate buffering system is inefficient. A common error is overzealous hyperventilation in the setting of depressed CO_2 production. When the Pco_2 decreases by 10 mmHg at 28°C (82°F), it doubles the pH increase of 0.08 that occurs at 37°C (99°F).

The severity of anemia may be underestimated because the hematocrit increases 2% for each 1°C drop in temperature. White blood cell sequestration and bone marrow suppression are common, potentially masking an infection. Although hypokalemia is more common in chronic hypothermia, hyperkalemia also occurs; the expected electrocardiographic changes can be obscured by hypothermia. Patients with renal insufficiency, metabolic acidoses, or rhabdomyolysis are at greatest risk for electrolyte disturbances.

Coagulopathies are common because cold inhibits the enzymatic reactions required for activation of the intrinsic cascade. In addition, thromboxane B_2 production by platelets is temperature dependent, and platelet function is impaired. The administration of platelets and fresh-frozen plasma is therefore not effective. The prothrombin or partial thromboplastin times or the international normalized ratio can be deceptively normal and contrast with the observed in vivo coagulopathy. This contradiction occurs because all coagulation tests are routinely performed at 37°C (99°F), and the enzymes are thus rewarmed.

REWARMING STRATEGIES

The key initial decision is whether to rewarm the patient passively or actively. *Passive external rewarming* simply involves covering and insulating the patient in a warm environment. With the head also covered, the rate of rewarming is usually 0.5°–2°C (1.10°–4.4°F) per hour. This technique is ideal for previously healthy patients who develop acute, mild primary accidental hypothermia. The patient must have sufficient glycogen to support endogenous thermogenesis.

The application of heat directly to the extremities of patients with chronic severe hypothermia should be avoided because it can induce peripheral vasodilation and precipitate core temperature "afterdrop," a response characterized by a continual decline in the core temperature after removal of the patient from the cold. Truncal heat application reduces the risk of afterdrop.

Active rewarming is necessary under the following circumstances: core temperature <32°C (<90°F) (*poikilothermia*), cardiovascular instability, age extremes, CNS dysfunction, hormone insufficiency, and suspicion of secondary hypothermia. *Active external rewarming* is best accomplished with forced-air heating blankets. Other options include devices that circulate water through external heat exchange pads,

radiant heat sources, and hot packs. Monitoring a patient with hypothermia in a heated tub is extremely difficult. Electric blankets should be avoided because vasoconstricted skin is easily burned.

There are numerous widely available options for *active core rewarming*. Airway rewarming with heated humidified oxygen (40°–45°C [104°–113°F]) via mask or endotracheal tube is a convenient option. Although airway rewarming provides less heat than do some other forms of active core rewarming, it eliminates respiratory heat loss and adds 1°–2°C (2.2°–4.4°F) to the overall rewarming rate. Crystalloids should be heated to 40°–42°C (104°–108°F), but the quantity of heat provided is significant only during massive volume resuscitation. The most efficient method for heating and delivering fluid or blood is with a countercurrent in-line heat exchanger. Heated irrigation of the gastrointestinal tract or bladder transfers minimal heat because of the limited available surface area. These methods should be reserved for patients in cardiac arrest and then used in combination with all available active rewarming techniques.

Closed thoracic lavage is far more efficient in severely hypothermic patients with cardiac arrest. The hemithoraxes are irrigated through two inserted large-bore thoracostomy tubes. Thoracostomy tubes should not be placed in the left chest of a spontaneously perfusing patient for purposes of rewarming. Peritoneal lavage with the dialysate at 40°–45°C (104°–113°F) efficiently transfers heat when delivered through two catheters with outflow suction. Like peritoneal dialysis, standard hemodialysis is especially useful for patients with electrolyte abnormalities, rhabdomyolysis, or toxin ingestion. Another option involves the central venous insertion of a rapid endovascular warming device.

Extracorporeal blood rewarming options **(Table 454-3)** should be considered in severely hypothermic patients, especially those with *primary accidental hypothermia*. Cardiopulmonary bypass should be considered in nonperfusing patients without documented contraindications to resuscitation. Circulatory support may be the only effective option in patients with completely frozen extremities or those with significant tissue destruction coupled with rhabdomyolysis. There is no evidence that extremely rapid rewarming improves survival in perfusing patients. The best strategy is usually a combination of passive, truncal active, and active core rewarming techniques.

TABLE 454-3 Options for Extracorporeal Blood Rewarming

EXTRACORPOREAL REWARMING TECHNIQUE	CONSIDERATIONS
Continuous venovenous (CVV) rewarming	Circuit: CV catheter to CV, dual-lumen CV, or peripheral catheter No oxygenator/circulatory support Flow rates 150–400 mL/min ROR 2°–3°C (4.4°–6.6°F)/h
Hemodialysis	Circuit: single- or dual-vessel cannulation Stabilizes electrolyte or toxicologic abnormalities Exchange cycle volumes 200–500 mL/min ROR 2°–3°C (4.4°– 6.6°F)/h
Continuous arteriovenous rewarming (CAVR)	Circuit: percutaneous 8.5-Fr femoral catheters Requires systolic blood pressure of 60 mmHg No perfusionist/pump/anticoagulation Flow rates 225–375 mL/min ROR 3°–4°C (6.6°–8.8°F)/h
Cardiopulmonary bypass (CPB)	Circuit: full circulatory support with pump and oxygenator Perfusate-temperature gradient (5°–10°C [41°–50°F]) Flow rates 2–7 L/min (average 3–4 L/min) ROR up to 9.5°C (20.9°F)/h
Extracorporeal membrane oxygenation (ECMO)	Decreased risk of post-rewarming cardiorespiratory failure

Abbreviations: CV, central venous; ROR, rate of rewarming.

TREATMENT

Hypothermia

When a patient is hypothermic, target organs and the cardiovascular system respond minimally to most medications. Generally, IV medications are withheld below 30°C (86°F). In contrast to antiarrhythmics, low-dose vasopressor medications may improve the intra-arrest rates of return of spontaneous circulation. Because of increased binding of drugs to proteins as well as impaired metabolism and excretion, either a lower dose or a longer interval between doses should be used to avoid toxicity. As an example, the administration of repeated doses of digoxin or insulin would be ineffective while the patient is hypothermic, and the residual drugs would be potentially toxic during rewarming.

Achieving a mean arterial pressure of at least 60 mmHg should be an early objective. If the hypotension does not respond to crystalloid/colloid infusion and rewarming, low-dose dopamine support (2–5 μg/kg per min) should be considered. Perfusion of the vasoconstricted cardiovascular system also may be improved with low-dose IV nitroglycerin.

Atrial arrhythmias should be monitored initially without intervention, as the ventricular response should be slow and, unless preexistent, most will convert spontaneously during rewarming. The role of prophylaxis and treatment of ventricular arrhythmias is problematic. Preexisting ventricular ectopy may be suppressed by hypothermia and reappear during rewarming. None of the class I agents has proved to be safe and efficacious. Initiating empirical therapy for adrenal insufficiency usually is not warranted unless the history suggests steroid dependence or hypoadrenalism or efforts to rewarm with standard therapy fail. The administration of parenteral levothyroxine to euthyroid patients with hypothermia, however, is potentially hazardous. Because laboratory results can be delayed and confounded by the presence of the sick euthyroid syndrome **(Chap. 375)**, historic clues or physical findings suggestive of hypothyroidism should be sought. When myxedema is the cause of hypothermia, the relaxation phase of the Achilles reflex is prolonged more than is the contraction phase.

Hypothermia obscures most of the symptoms and signs of infection, notably fever and leukocytosis. Shaking rigors from infection may be mistaken for shivering. Except in mild cases, extensive cultures and repeated physical examinations are essential. Unless an infectious source is identified, empirical antibiotic prophylaxis is most warranted in the elderly, neonates, and immunocompromised patients.

Preventive measures should be discussed with high-risk individuals, such as the elderly and people whose work frequently exposes them to extreme cold. The importance of layered clothing and headgear, adequate shelter, increased caloric intake, and the avoidance of ethanol should be emphasized, along with access to rescue services.

FROSTBITE

Peripheral cold injuries include both freezing and nonfreezing injuries to tissue. Tissue freezes quickly when in contact with thermal conductors such as metal and volatile solutions. Other predisposing factors include constrictive clothing or boots, immobility, and vasoconstrictive medications. Frostbite occurs when the tissue temperature drops below 0°C (32°F). Ice-crystal formation subsequently distorts and destroys the cellular architecture. Once the vascular endothelium is damaged, stasis progresses rapidly to microvascular thrombosis. After the tissue thaws, there is progressive dermal ischemia. The microvasculature begins to collapse, arteriovenous shunting increases tissue pressures, and edema forms. Finally, thrombosis, ischemia, and superficial necrosis appear. The development of mummification and demarcation may take weeks to months.

CLINICAL PRESENTATION

The initial presentation of frostbite can be deceptively benign. The symptoms always include a sensory deficiency affecting light touch, pain, or temperature perception. The acral areas and distal extremities

are the most common insensate areas. Some patients describe a clumsy or "chunk of wood" sensation in the extremity.

Deep frostbitten tissue can appear waxy, mottled, yellow, or violaceous-white. Favorable presenting signs include some warmth or sensation with normal color. The injury is often superficial if the subcutaneous tissue is pliable or if the dermis can be rolled over bony prominences.

Frostnip may precede frostbite. Frostnip is a nonfreezing cold injury resulting from intense vasoconstriction of exposed acral skin.

Clinically, frostbite is superficial or deep. Superficial frostbite does not entail tissue loss but rather causes only anesthesia and erythema. The appearance of vesiculation surrounded by edema and erythema implies deeper involvement **(Fig. 454-1)**. Hemorrhagic vesicles reflect a serious injury to the microvasculature and indicate severe frostbite. Damages in subcuticular, muscular, or osseous tissues may result in amputation. An alternative classification establishes grades based on the location of presenting cyanosis; that is Grade 1, absence of cyanosis; Grade 2, cyanosis on the distal phalanx; Grade 3, cyanosis up to the MP joint; and Grade 4 cyanosis proximal to the MP joint.

The two most common nonfreezing peripheral cold injuries are *chilblain* (*pernio*) and *immersion* (*trench*) *foot*. Chilblain results from neuronal and endothelial damage induced by repetitive exposure to damp cold above the freezing point. Young females, particularly those with a history of Raynaud's phenomenon, are at greatest risk. Persistent vasospasticity and vasculitis can cause erythema, mild edema, and pruritus. Eventually plaques, blue nodules, and ulcerations develop. These lesions typically involve the dorsa of the hands and feet. In contrast, immersion foot results from repetitive exposure to wet cold above the freezing point. The feet initially appear cyanotic, cold, and edematous.

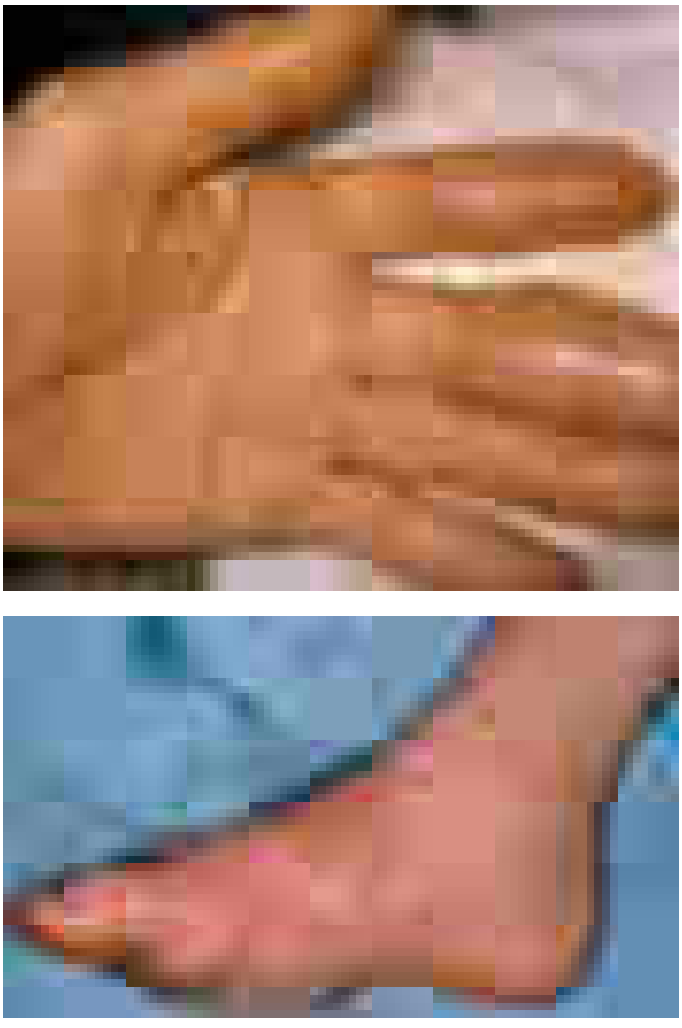


FIGURE 454-1 Frostbite with vesiculation, surrounded by edema and erythema.

The subsequent development of bullae is often indistinguishable from frostbite. This vesiculation rapidly progresses to ulceration and liquefaction gangrene. Patients with milder cases report hyperhidrosis, cold sensitivity, and painful ambulation for many years.

TREATMENT

Peripheral Cold Injuries

When frostbite accompanies hypothermia, hydration may improve vascular stasis. Frozen tissue should be thawed rapidly and completely by immersion in circulating water at 37°–40°C (99°–104°F) for 30–60 min. Rapid rewarming often produces an initial hyperemia. The early formation of large clear distal blebs is more favorable than that of smaller proximal dark hemorrhagic blebs. A common error is the premature termination of thawing, since the reestablishment of perfusion is intensely painful. Parenteral narcotics will be necessary with deep frostbite. If cyanosis persists after rewarming, the tissue compartment pressures should be monitored carefully.

Many antithrombotic and vasodilatory primary and adjunctive treatment regimens have been evaluated. The prostacyclin analogue iloprost given within 48 h after rewarming may prove useful. There is no conclusive evidence that sympathectomy, steroids, calcium channel blockers, or hyperbaric oxygen salvages tissue.

Patients who have deep frostbite injuries with the potential for significant morbidity should be considered for intravenous or intraarterial thrombolytic therapy. Angiography or pyrophosphate scanning may help evaluate the injury and monitor the progress of tissue plasminogen activator therapy (rt-PA). Heparin is recommended as adjunctive therapy. Intraarterial thrombolysis may reduce the need for digital and more proximal amputations when administered within 24 h of severe injuries. A treatment protocol for frostbite is summarized in **Table 454-4**.

Unless infection develops, any decision regarding debridement or amputation should generally be deferred. Angiography or

TABLE 454-4 Treatment for Frostbite

BEFORE THAWING	DURING THAWING	AFTER THAWING
Remove from environment.	Consider parenteral analgesia and ketorolac.	Gently dry and protect part; elevate; place pledgets between toes, if macerated.
Prevent partial thawing and refreezing.	Administer ibuprofen (400 mg PO).	If clear vesicles are intact, aspirate sterilely; if broken, debride and dress with antibiotic or sterile aloe vera ointment.
Stabilize core temperature and treat hypothermia.	Immerse part in 37°–40°C (99°–104°F) (thermometer-monitored) circulating water containing an antiseptic soap until distal flush (10–45 min).	Leave hemorrhagic vesicles intact to prevent desiccation and infection.
Protect frozen part—no friction or massage.	Encourage patient to gently move part.	Continue ibuprofen (400– 600 mg PO [12 mg/kg per day] q8 to 12h).
Address medical or surgical conditions.	If pain is refractory, reduce water temperature to 35°–37°C (95°–99°F) and administer parenteral narcotics.	Consider tetanus and streptococcal prophylaxis; elevate part. Administer hydrotherapy at 37°C (99°F). Consider dextran or phenoxybenzamine or, in severe cases, thrombolysis rt-PA (IV or intraarterial)

technetium-99 bone scan may assist in the determination of surgical margins. Magnetic resonance angiography may also demonstrate the line of demarcation earlier than does clinical demarcation.

The most common symptomatic sequelae reflect neuronal injury and persistently abnormal sympathetic tone, including paresthesias, thermal misperception, and hyperhidrosis. Delayed findings include nail deformities, cutaneous carcinomas, and epiphyseal damage in children.

Management of the chilblain syndrome is usually supportive. With refractory perniosis, alternatives include nifedipine, steroids, and limaprost, a prostaglandin E_1 analogue.

FURTHER READING

Brandstrom H et al: Accidental cold-related injury leading to hospitalization in northern Sweden: An eight-year retrospective analysis. Scan J Trauma Resusc Emerg Med 22:6, 2014.

Cauchy E et al: A controlled trial of a prostacyclin and rt-PA in the treatment of severe frostbite. N Engl J Med 364:189, 2011.

McIntosh SE et al: Wilderness Medical Society practice guidelines for the prevention and treatment of frostbite. Wilderness Environ Med 25(4 suppl):S43, 2014.

Petrone P et al: Management of accidental hypothermia and cold injury. [Review]. Curr Prob Surg 51:417, 2014.

Zafren K et al: Wilderness Medical Society practice guidelines for the out-of-hospital evaluation and treatment of accidental hypothermia. Wilderness Environ Med 25:425, 2014.

455 Heat-Related Illnesses

Daniel F. Danzl

Heat-related illnesses include a spectrum of disorders ranging from heat syncope, muscle cramps, and heat exhaustion to medical emergencies such as heatstroke. The core body temperature is normally maintained within a very narrow range. Although significant levels of hypothermia are tolerated **(Chap. 454)**, multiorgan dysfunction occurs rapidly at temperatures >41°–43°C. In contrast to heatstroke, the far more common sign of fever reflects intact thermoregulation.

THERMOREGULATION

Humans are capable of significant heat generation. Strenuous exercise can increase heat generation twentyfold. The heat load from metabolic heat production and environmental heat absorption is balanced by a variety of heat dissipation mechanisms. These central integrative dissipation pathways are orchestrated by the central thermostat, which is located in the preoptic nucleus of the anterior hypothalamus. Efferent signals sent via the autonomic nervous system trigger cutaneous vasodilation and diaphoresis to facilitate heat loss.

Normally, the body dissipates heat into the environment via four mechanisms. The *evaporation* of skin moisture is the single most efficient mechanism of heat loss but becomes progressively ineffective as the relative humidity rises to >70%. The *radiation* of infrared electromagnetic energy directly into the surrounding environment occurs continuously. (Conversely, radiation is a major source of heat gain in hot climates.) *Conduction*—the direct transfer of heat to a cooler object—and *convection*—the loss of heat to air currents—become ineffective when the environmental temperature exceeds the skin temperature.

Factors that interfere with the evaporation of diaphoresis significantly increase the risk of heat illness. Examples include dripping of sweat off the skin, constrictive or occlusive clothing, dehydration, and excessive humidity. While air is an effective insulator, the thermal conductivity of water is 25 times greater than that of air at the same temperature. The *wet-bulb globe temperature* is a commonly used index to assess the environmental heat load. This calculation considers the ambient air temperature, the relative humidity, and the degree of radiant heat.

The regulation of this heat load is complex and involves the central nervous system (CNS), thermosensors, and thermoregulatory effectors. The central thermostat activates the effectors that produce peripheral vasodilation and sweating. The skin surface is in effect the radiator and the principal location of heat loss, since skin blood flow can increase 25–30 times over the basal rate. This dramatic increase in skin blood flow, coupled with the maintenance of peripheral vasodilation, efficiently radiates heat. At the same time, there is a compensatory vasoconstriction of the splanchnic and renal beds.

Acclimatization to heat reflects a constellation of physiologic adaptations that permit the body to lose heat more efficiently. This process often requires one to several weeks of exposure and work in a hot environment. During acclimatization, the thermoregulatory set point is altered, and this alteration affects the onset, volume, and content of diaphoresis. The threshold for the initiation of sweating is lowered, and the amount of sweat increases, with a lowered salt concentration. Sweating rates can be 1–2 L/h in acclimated individuals during heat stress. Plasma volume expansion also occurs and improves cutaneous vascular flow. The heart rate lowers, with a higher stroke volume. After the individual leaves the hot environment, improved tolerance to heat stress dissipates rapidly, the plasma volume decreases, and de-acclimatization occurs within weeks.

PREDISPOSING FACTORS AND DIFFERENTIAL DIAGNOSIS

When there is an excessive heat load, unacclimated individuals can develop a variety of heat-related illnesses. Heat waves exacerbate the mortality rate, particularly among the elderly and among persons lacking adequate nutrition and access to air-conditioned environments. Secondary vascular events, including cerebrovascular accidents and myocardial infarctions, occur at least 10 times more often in conditions of extreme heat.

Exertional heat illness continues to occur when laborers, military personnel, or athletes exercise strenuously in the heat. A variety of common factors predispose to heat illness. In addition to the very young and very old, preadolescents and teenagers are at risk since they may use poor judgment when vigorously exercising in high humidity and heat. Other risk factors include obesity, poor conditioning and lack of acclimatization, and mild dehydration.

Cardiovascular inefficiency is a common feature of heat illness. Any physiologic or pharmacologic impediment to cutaneous perfusion will impair heat loss. Many patients are unaware of the heat risk associated with their medications. Anticholinergic agents impair sweating and blunt the normal cardiovascular response to heat. Phenothiazines and heterocyclic antidepressants also have anticholinergic properties that interfere with the function of the preoptic nucleus of the anterior hypothalamus due to central depletion of dopamine.

Calcium channel blockers, beta blockers, and various stimulants also inhibit sweating by reducing peripheral blood flow. To maintain the mean arterial blood pressure, increased cardiac output must be capable of compensating for progressive dehydration. A variety of stimulants and substances of abuse also increase muscle activity and heat production. Careful consideration of the differential diagnosis is important in the evaluation of a patient for a potential heat-related illness. The clinical setting may suggest other etiologies, such as malignant hyperthermia after general anesthesia. Neuroleptic malignant syndrome can be triggered by certain antipsychotic medications, including selective serotonin reuptake inhibitors. A variety of infectious and endocrine disorders as well as conditions with toxicologic or CNS etiologies may initially mimic heatstroke **(Table 455-1)**.

MINOR HEAT-EMERGENCY SYNDROMES

Heat edema is characterized by mild swelling of the hands, feet, and ankles during the first few days of significant heat exposure. The principal mechanism involves cutaneous vasodilation and pooling of interstitial fluid in response to heat stress. Heat also increases the secretion

TABLE 455-1 Heat-Related Illness: Predisposing Factors and Differential Diagnosis

ILLNESS	PREDISPOSING FACTORS
Cardiovascular inefficiency	Age extremes
	Beta/calcium channel blockade
	Congestive heart failure
	Dehydration
	Diuresis
	Obesity
	Poor physical fitness
Central nervous system illness	Cerebral hemorrhage
	Hypothalamic cerebrovascular accident
	Psychiatric disorders
	Status epilepticus
Impaired heat loss	Antihistamines
	Heterocyclic antidepressants
	Occlusive clothing
	Skin abnormalities
Endocrine-related illness	Diabetic ketoacidosis
	Pheochromocytoma
	Thyroid storm
Excessive heat load	Environmental conditions
	Exertion
	Fever
	Hypermetabolic state
	Lack of acclimatization
Infectious illness	Cerebral abscess
	Encephalitis
	Malaria
	Meningitis
	Sepsis syndrome
	Tetanus
	Typhoid
Toxicologic illness	Amphetamines
	Anticholinergic toxidrome
	Cocaine
	Dietary supplements
	Hallucinogens
	Malignant hyperthermia
	Neuroleptic malignant syndrome
	Salicylates
	Serotonin syndrome
	Strychnine
	Sympathomimetics
	Withdrawal syndromes (ethanol, hypnotics)

of antidiuretic hormone and aldosterone. Systemic causes of edema, including cirrhosis, nephrotic syndrome, and congestive heart failure, can usually be excluded by the history and physical examination. Heat edema generally resolves without treatment in several days. Simple leg elevation or thigh-high support hose will usually suffice. Diuretics are *not* effective and, in fact, predispose to volume depletion and the development of more serious heat-related illnesses.

Prickly heat (*miliaria rubra*, *lichen tropicus*) is a maculopapular, pruritic, erythematous rash that commonly occurs in clothed areas. Blockage of the sweat pores by debris from macerated stratum corneum causes inflammation in the sweat ducts. As the ducts dilate, they rupture and produce superficial vesicles. The predominant symptom is pruritus. In addition to antihistamines, chlorhexidine in a light cream or lotion provides some relief. In adults, localized areas may benefit from 1% salicylic acid TID, with caution taken to avoid salicylate intoxication. Clothing with breathable fabric should be clean and loose fitting, and activities or environments that induce diaphoresis should be avoided.

Heat syncope (exercise-associated collapse) can follow endurance exercise or occur in the elderly. Other common clinical scenarios include prolonged standing while stationary in the heat and sudden standing after prolonged exposure to heat. Heat stress routinely causes relative volume depletion, decreased vasomotor tone, and peripheral vasodilation. The cumulative effect of this decrease in venous return is postural hypotension, especially in nonacclimated elderly individuals. Many of those affected also have comorbidities. Therefore, other cardiovascular, neurologic, and metabolic causes of syncope should be considered. After removal from the heat source, most patients will recover promptly with cooling and rehydration.

Hyperventilation tetany occurs in some individuals when exposure to heat stimulates hyperventilation, producing respiratory alkalosis, paresthesias, and carpopedal spasm. Unlike heat cramps, heat tetany causes very little muscle-compartment pain. Treatment includes providing reassurance, moving the patient out of the heat, and addressing the hyperventilation.

HEAT CRAMPS

Heat cramps (exercise-associated muscle cramps) are intermittent, painful, and involuntary spasmodic contractions of skeletal muscles. They typically occur in an unacclimated individual who is at rest after vigorous exertion in a humid, hot environment. In contrast, cramps that occur in athletes during exercise last longer, are relieved by stretching and massage, and resolve spontaneously.

Of note, not all muscle cramps are related to exercise, and the differential diagnosis includes many other disorders. A variety of medications, myopathies, endocrine disorders, and sickle cell trait are other possible causes.

The typical patient with heat cramps is usually profusely diaphoretic and has been replacing fluid losses with copious water or other hypotonic fluids. Roofers, firefighters, military personnel, athletes, steel workers, and field workers are commonly affected. Other predisposing factors include insufficient sodium intake before intense activity in the heat and lack of heat acclimatization, resulting in sweat with a high salt concentration.

The precise pathogenesis of heat cramps appears to involve a relative deficiency of sodium, potassium, and fluid at the intracellular level. Coupled with copious hypotonic fluid ingestion, large amounts of sodium in the diaphoresis cause hyponatremia and hypochloremia, resulting in muscle cramps due to calcium-dependent muscle relaxation. Total-body depletion of potassium may be observed during the period of heat acclimatization. Rhabdomyolysis is very rare with routine exercise-associated muscle cramps.

Heat cramps that are not accompanied by significant dehydration can be treated with commercially available electrolyte solutions. Although the flavored electrolyte solutions are far more palatable, two 650-mg salt tablets dissolved in 1 quart of water produce a 0.1% saline solution. Individuals should avoid the ingestion of undissolved salt tablets, which are a gastric irritant and may induce vomiting.

HEAT EXHAUSTION

The physiologic hallmarks of heat exhaustion—in contrast to heatstroke—are the maintenance of thermoregulatory control and CNS function. The core temperature is usually elevated but is generally <40.5°C (<105°F). The two physiologic precipitants are water depletion and sodium depletion, which often occur in combination. Laborers, athletes, and elderly individuals exerting themselves in hot environments, without adequate fluid intake, tend to develop *water-depletion heat exhaustion*. Persons working in the heat frequently consume only two-thirds of their net water loss and are voluntarily dehydrated. In contrast, *salt-depletion heat exhaustion* occurs more slowly in unacclimated persons who have been consuming large quantities of hypotonic solutions.

Heat exhaustion is usually a diagnosis of exclusion because of the multitude of nonspecific symptoms. If any signs of heatstroke are present, rapid cooling and crystalloid resuscitation should be initiated immediately during stabilization and evaluation. Mild neurologic and gastrointestinal influenza-like symptoms are common. These symptoms may include headache, vertigo, ataxia, impaired judgment, malaise, dizziness,

nausea, and muscle cramps. Orthostatic hypotension and sinus tachycardia develop frequently. More significant CNS impairment suggests heatstroke or other infectious, neurologic, or toxicologic diagnoses.

Hemoconcentration does not always develop, and rapid infusion of isotonic IV fluids should be guided by frequent electrolyte determinations and perfusion requirements. Most cases of heat exhaustion reflect mixed sodium and water depletion. Sodium-depletion heat exhaustion is characterized by hyponatremia and hypochloremia. Hepatic aminotransferases are mildly elevated in both types of heat exhaustion. Urinary sodium and chloride concentrations are usually low.

Some patients with heat exhaustion develop heatstroke after removal from the heat-stress environment. Aggressive cooling of nonresponders is indicated until their core temperature is 39°C (102.2°F). Except in mild cases, free water deficits should be replaced slowly over 24–48 h to avoid a decrease of serum osmolality by >2 mOsm/h.

The disposition of younger, previously healthy heat-exhaustion patients who have no major laboratory abnormalities may include hospital observation and discharge after IV rehydration. Older patients with comorbidities (including cardiovascular disease) or predisposing factors often require inpatient fluid and electrolyte replacement, monitoring, and reassessment.

HEATSTROKE

The clinical manifestations of heatstroke reflect a total loss of thermoregulatory function. Typical vital-sign abnormalities include tachypnea, various tachycardias, hypotension, and a widened pulse pressure. Although there is no single specific diagnostic test, the historical and physical triad of exposure to a heat stress, CNS dysfunction, and a core temperature >40.5°C helps establish the preliminary diagnosis. Some patients with impending heat stroke will initially appear lucid. The definitive diagnosis should be reserved until the other potential causes of hyperthermia are excluded. Many of the usual laboratory abnormalities seen with heatstroke overlap with other conditions. If the patient's mental status does not improve with cooling, toxicologic screening may be indicated, and cranial CT and spinal fluid analysis can be considered.

The premonitory clinical characteristics may be nonspecific and include weakness, dizziness, disorientation, ataxia, and gastrointestinal or psychiatric symptoms. These prodromal symptoms often resemble heat exhaustion. The sudden onset of heatstroke occurs when the maintenance of adequate perfusion requires peripheral vasoconstriction to stabilize the mean arterial blood pressure. As a result, the cutaneous radiation of heat ceases. At this juncture, the core temperature rises dramatically. Since many patients with heatstroke also meet the criteria for systemic inflammatory response syndrome and have a broad differential diagnosis, rapid cooling is essential during the extensive diagnostic evaluation (Table 455-1).

There are two forms of heatstroke with significantly different manifestations **(Table 455-2)**. Classic (epidemic) heatstroke (CHS) usually occurs during long periods of high ambient temperature and humidity, as during summer heat waves. Patients with CHS commonly have chronic diseases that predispose to heat-related illness, and they may have limited access to oral fluids. Heat dissipation mechanisms are overwhelmed by both endogenous heat production and exogenous heat stress. Patients with CHS are often compliant with prescribed medications that can impair tolerance to a heat stress. In many of these dehydrated CHS patients, sweating has ceased and the skin is hot and dry.

TABLE 455-2 Typical Manifestations of Heatstroke

CLASSIC	EXERTIONAL
Older patient	Younger patient
Predisposing health factors/medications	Healthy condition
Epidemiology (heat waves)	Sporadic cases
Sedentary	Exercising
Anhidrosis (possible)	Diaphoresis (common)
Central nervous system dysfunction	Myocardial/hepatic injury
Oliguria	Acute renal failure
Coagulopathy (mild)	Disseminated intravascular coagulation
Mild lactic acidosis	Marked lactic acidosis
Mild creatine kinase elevation	Rhabdomyolysis
Normoglycemia/calcemia	Hypoglycemia/calcemia
Normokalemia	Hyperkalemia
Normonatremia	Hyponatremia

If cooling is delayed, severe hepatic dysfunction, renal failure, disseminated intravascular coagulation, and fulminant multisystem organ failure may occur. Hepatocytes are very heat sensitive. On presentation, the serum level of aspartate aminotransferase (AST) is routinely elevated. Eventually, levels of both AST and alanine aminotransferase (ALT) often increase to >100 times the normal values. Coagulation studies commonly demonstrate decreased platelets, fibrinogen, and prothrombin. Most patients with CHS require cautious crystalloid resuscitation, electrolyte monitoring, and—in certain refractory cases—consideration of central venous pressure (CVP) measurements. Hypernatremia is secondary to dehydration in CHS. Many patients exhibit significant stress leukocytosis, even in the absence of infection.

Patients with exertional heatstroke (EHS), in contrast to those with CHS, are often young and previously healthy, and their diagnosis is usually more obvious from the history. Athletes, laborers, and military recruits are common victims. Unlike those with CHS, many EHS patients present profusely diaphoretic despite significant dehydration. As a result of muscular exertion, rhabdomyolysis and acute renal failure are more common in EHS. Studies to detect rhabdomyolysis and its complications, including hypocalcemia and hyperphosphatemia, should be considered. Hyponatremia, hypoglycemia, and coagulopathies are frequent findings. Elevated creatine kinase and lactate dehydrogenase levels also suggest EHS. Oliguria is a common finding. Renal failure can result from direct thermal injury, untreated rhabdomyolysis, or volume depletion. Common urinalysis findings include microscopic hematuria, myoglobinuria, and granular or red cell casts.

With both CHS and EHS, heat-related reversible increases in cardiac biomarker levels are often present. Heatstroke often causes thermal cardiomyopathy. As a result, the CVP may be elevated despite significant dehydration. In addition, the patient often presents with potentially deceptive noncardiogenic pulmonary edema and basilar rales despite being significantly hypovolemic. The electrocardiogram commonly displays a variety of tachyarrhythmias, nonspecific ST-T wave changes, and heat-related ischemia or infarction. Rapid cooling—not the administration of antiarrhythmic medications—is essential.

Above 42°C (107.6°F), heat can rapidly produce direct cellular injury. Thermosensitive enzymes become nonfunctional, and eventually there is irreversible uncoupling of oxidative phosphorylation. The production of heat-shock proteins increases, and cytokines mediate a systemic inflammatory response. The vascular endothelium is also damaged, and this injury activates the coagulation cascade. Significant shunting away from the splanchnic circulation produces gastrointestinal ischemia. Endotoxins further impair normal thermoregulation. As a result, if cooling is delayed, severe hepatic dysfunction, permanent renal failure, disseminated intravascular coagulation, and fulminant multisystem organ failure may occur.

COOLING STRATEGIES

Before cooling is initiated, endotracheal intubation and continuous core-temperature monitoring should be considered. Peripheral methods to measure temperature are *not* reliable. Hypoglycemia is a frequent finding and can be addressed by glucose infusion. Since peripheral vasoconstriction delays heat dissipation, repeated administration of discrete boluses of isotonic crystalloid for hypotension is preferable to the administration of α-adrenergic agonists.

Evaporative cooling is frequently the most practical and effective technique. Rapid cooling is essential in both CHS and EHS, and an immediate improvement in vital signs and mental status may prove valuable for diagnostic purposes. Cool water (15°C [60°F]) is sprayed on the exposed skin while fans direct continuous airflow over the moistened skin. Cold packs applied to the neck, axillae, and groin are

useful cooling adjuncts. If cardiac electrodes will not adhere, they can be applied to the patient's back.

Immersion cooling in ice-cold water is an alternative option in EHS but can induce peripheral vasoconstriction and shivering. The initial increase in temperature from peripheral vasoconstriction will rapidly be overcome by the large conductive thermal transfer into cold water. This technique presents significant monitoring and resuscitation challenges in many clinical settings. The safety of immersion cooling is best established for young, previously healthy patients with EHS (but not for those with CHS). To avoid hypothermic afterdrop (continued cooling after immersion), active cooling should be terminated ~ 38°–39°C (100.4°F–102.2°F).

Cooling with commercially available cooling blankets should not be the sole technique used, since the rate of cooling is far too slow. Other methods are less efficacious and rarely indicated, such as IV infusion of cold fluids and cold irrigation of the bladder or gastrointestinal tract. Cold thoracic and peritoneal lavage are efficient maneuvers but are invasive and rarely necessary. Cardiopulmonary bypass provides effective cooling but is labor intensive and is rarely necessary.

RESUSCITATION

Aspiration commonly occurs in heatstroke, and endotracheal intubation is usually necessary. The metabolic demands are high, and supplemental oxygenation is essential due to hypoxemia induced by thermal stress and pulmonary dysfunction. The oxyhemoglobin dissociation curve is shifted to the right. Pneumonitis, pulmonary infarction, hemorrhage, edema, and acute respiratory distress syndrome occur frequently in heatstroke patients. Seizures are common, and can occur during therapeutic cooling. Cold induced tonic-clonic muscular rigidity mimics seizure activity.

The circulatory fluid requirements, particularly in CHS, may be deceptively modest. Aggressive cooling and modest volume repletion usually elevate the CVP to 12–14 mmHg. The reading, however, may be deceptive. Many patients present with a thermally induced hyperdynamic circulation accompanied by a high cardiac index, low peripheral vascular resistance, and an elevated CVP caused by right-sided heart failure. In contrast, most patients with EHS require far more zealous isotonic crystalloid resuscitation.

The hypotension that is initially common among patients with heatstroke results from both dehydration and high-output cardiac failure caused by peripheral vasodilation. Inotropes causing α-adrenergic stimulation (e.g., norepinephrine) can impede cooling by causing significant vasoconstriction. Vasoactive catecholamines such as dopamine or dobutamine may be necessary if the cardiac output remains depressed despite an elevated CVP, particularly in patients with a hyperdynamic circulation.

A wide variety of tachyarrhythmias are routinely observed on presentation and usually resolve spontaneously during cooling. The administration of atrial or ventricular antiarrhythmic medications is rarely indicated during cooling. Anticholinergic medications (including atropine) inhibit sweating and should be avoided. With a cardiac rhythm that sustains perfusion, electrical cardioversion of the hyperthermic myocardium should be deferred until the myocardium is cooled. Significant shivering, discomfort, or extreme agitation is preferably mitigated with short-acting benzodiazepines, which are ideal due to their renal clearance. On the other hand, chlorpromazine may lower the seizure threshold, has anticholinergic properties, and can exacerbate the hypotension or cause neuroleptic malignant syndrome. With hepatic dysfunction, barbiturates should be avoided and seizures treated with benzodiazepines.

Coagulopathies more commonly occur after the first day of illness. After cooling, the patient should be monitored for disseminated intravascular coagulation, and replacement therapy with fresh-frozen plasma and platelets should be considered.

There is no therapeutic role for antipyretics in the control of environmentally induced hyperthermia; these drugs block the actions of pyrogens at hypothalamic receptor sites. Salicylates can further uncouple oxidative phosphorylation in heatstroke and exacerbate coagulopathies. Acetaminophen may further stress hepatic function. The safety and efficacy of dantrolene is not established. Although aminocaproic acid impedes fibrinolysis, it may cause rhabdomyolysis and is not recommended in heatstroke.

DISPOSITION

Most patients with minor heat-emergency syndromes (including heat edema, heat syncope, and heat cramps) require only stabilization and treatment with outpatient follow-up. Although there are no decision rules to guide disposition choices in heat exhaustion, many of these patients have multiple predisposing factors and comorbidities that will require prolonged observation or hospital admission.

Essentially all patients with actual heatstroke require admission to a monitored setting, and most require intensive care. Many of these patients also require prolonged tracheal intubation, invasive hemodynamic monitoring, and support for various degrees of multiorgan dysfunction syndrome. The prognosis worsens if the initial core temperature exceeds 42°C (107.6°F) or if there was a prolonged period during which the core temperature exceeded this level. Other features of a negative prognosis include acute renal failure, massively elevated liver enzymes, and significant hyperkalemia. As expected, the number of dysfunctional organ systems also correlates directly with mortality risk.

FURTHER READING

Atha WF: Heat-related illness. Emerg Med Clin North Am 31:1097, 2013.

Casa DJ et al: National Athletic Trainers' Association Position Statement: Exertional Heat Illnesses. J Athl Train 50:986, 2015.

Lipman GS et al: Wilderness Medical Society practice guidelines for the prevention and treatment of heat-related illness. Wilderness Environ Med 24:351, 2013.

Maron BJ et al: Sudden deaths in young competitive athletes: Analysis of 1866 deaths in the United States, 1980–2006. Circulation 119:1085, 2009.

O'Connor FG et al: American College of Sports Medicine Roundtable on exertional heat stroke—Return to duty/return to play: Conference proceedings. Curr Sports Med Rep 9:314, 2010.

Platt M et al: Heat illnesses, in *Rosen's Emergency Medicine: Concepts and Clinical Practices*, 9th ed, Marx JA et al (eds). Philadelphia, Elsevier, 2018, pp 1755–1764.

Racinais S et al: Consensus recommendations on training and competing in the heat. Br J Sports Med 49:1164, 2015.

Zeller L et al: Exertional heatstroke: Clinical characteristics, diagnostic and therapeutic considerations. Eur J Intern Med 22:296, 2011.

456 Principles of Human Genetics

J. Larry Jameson, Peter Kopp

IMPACT OF GENETICS AND GENOMICS ON MEDICAL PRACTICE

Human genetics refers to the study of individual genes, their role and function in disease, and their mode of inheritance. Genomics refers to an organism's entire genetic information, the *genome*, and the function and interaction of DNA within the genome, as well as with environmental or nongenetic factors, such as a person's lifestyle. With the characterization of the human genome, genomics not only complements traditional genetics in our efforts to elucidate the etiology and pathogenesis of disease, but it plays an increasingly prominent role in diagnostics, prevention, and therapy **(Chap. 457)**. These transformative developments, emerging from the Human Genome Project, have been variably designated *genomic medicine*, *personalized medicine*, or *precision medicine*. Precision medicine aims at customizing medical decisions to an individual patient. For example, a patient's genetic characteristics (genotype) can be used to optimize drug therapy and predict efficacy, adverse events, and drug dosing of selected medications (*pharmacogenomics*) **(Chap. 64)**. The characterization of the mutational profile of a malignancy allows to identify driver mutations or overexpressed signaling molecules, which then facilitates the selection of targeted therapies. Genomic risk prediction models for common diseases are also beginning to emerge.

Genetics has traditionally been viewed through the window of relatively rare single-gene diseases. These disorders account for ~10% of pediatric admissions and childhood mortality. Historically, genetics has focused predominantly on chromosomal and metabolic disorders, reflecting the long-standing availability of techniques to diagnose these conditions. For example, conditions such as trisomy 21 (Down's syndrome) or monosomy X (Turner's syndrome) can be diagnosed using cytogenetics. Likewise, many metabolic disorders (e.g., phenylketonuria, familial hypercholesterolemia) are diagnosed using biochemical analyses. The advances in DNA diagnostics have extended the field of genetics to include virtually all medical specialties and have led to the elucidation of the pathogenesis of numerous monogenic disorders. In addition, it is apparent that virtually every medical condition has a genetic component. As is often evident from a patient's family history, many common disorders such as hypertension, heart disease, asthma, diabetes mellitus, and mental illnesses are significantly influenced by the genetic background. These polygenic or multifactorial (complex) disorders involve the contributions of many different genes, as well as environmental factors that can modify disease risk. Genome-wide association studies (GWAS) have elucidated numerous disease-associated loci and are providing novel insights into the allelic architecture of complex traits. These studies have been facilitated by the availability of comprehensive catalogues of human single-nucleotide polymorphism (SNP) haplotypes (HapMap, International Genome Sample Resource/1000 genomes project). Next-generation DNA sequencing (NGS) technologies have evolved rapidly and the cost of sequencing whole exomes (the exons within the genome, WES) or genomes (WGS) has plummeted. Comprehensive unbiased sequence analyses are now frequently used to characterize individuals with complex undiagnosed conditions or to determine the mutational profile of advanced malignancies in order to select better targeted therapies.

Cancer has a genetic basis because it results from acquired somatic mutations in genes controlling growth, apoptosis, and cellular differentiation **(Chap. 67)**. In addition, the development of many cancers is associated with a hereditary predisposition. Characterization of the genome (and epigenome) in various malignancies has led to fundamental new insights into cancer biology and reveals that the genomic profile of mutations is in many cases more important in determining the appropriate therapy than the organ in which the tumor originates. The Cancer Genome Atlas (TCGA) initiative of the National Cancer Institute and the National Human Genome Research Institute has already characterized the genomic landscape of >30 malignancies and several others will be completed in the near future. TCGA consists of comprehensive analyses of genomic and proteomic alterations and is providing fundamental new insights into the molecular pathogenesis of cancer. This knowledge has direct clinical ramifications as it impacts cancer taxonomy and the development of targeted therapies.

Genetic and genomic approaches have proven invaluable for the detection of infectious pathogens and are used clinically to identify agents that are difficult to culture such as mycobacteria, viruses, and parasites, or to track infectious agents locally or globally. In many cases, molecular genetics has improved the feasibility and accuracy of diagnostic testing and is beginning to open new avenues for therapy, including gene and cellular therapies **(Chap. 458)**. Molecular genetics has also provided the opportunity to characterize the *microbiome*, a new field that characterizes the population dynamics of bacteria, viruses, and parasites that coexist with humans and other animals **(Chap. 459)**. Emerging data indicate that the microbiome has significant effects on normal physiology as well as various disease states and the field is now focusing on defining the mechanisms underlying these interactions.

Molecular biology has significantly changed the treatment of human disease. Peptide hormones, growth factors, cytokines, and vaccines can now be produced in large amounts using recombinant DNA technology. Targeted modifications of these peptides provide the practitioner with improved therapeutic tools, as illustrated by genetically modified insulin analogues with more favorable kinetics. Lastly, there is reason to believe that a better understanding of the genetic basis of human disease will also have an increasing impact on disease prevention.

The astounding rate at which new genetic and genomic information is being generated creates a major challenge for physicians, health care providers, and basic investigators. Although many functional aspects of the genome remain unknown, there are many clinical situations where sufficient evidence exists for the use of genetic and genomic information to optimize patient care and treatment. Much genetic information resides in databases that provide easy access to the expanding information about the human genome, genetic disease, and genetic testing **(Table 456-1)**. For example, several thousand monogenic disorders are summarized in a large, continuously evolving compendium, referred to as the *Online Mendelian Inheritance in Man* (OMIM) catalogue (Table 456-1). The constant refinement of bioinformatics and new developments in *big data analytics*, together with the widespread adoption of electronic health records (EHRs), is simplifying the access, analysis and integration of this daunting amount of new information. Importantly, genomic data can be integrated readily into EHRs, and will accelerate the impact on clinical practice.

THE HUMAN GENOME

Structure of the Human Genome The Human Genome Project was initiated in the mid-1980s as an ambitious effort to characterize the entire human genome and culminated in the completion of the DNA sequence for the last of the human chromosomes in 2006. The scope of a whole genome sequence analysis can be illustrated by the following analogy. Human DNA consists of ~3 billion base pairs (bp) of DNA per haploid genome, which is nearly 1000-fold greater than that of the *Escherichia coli* genome. If the human DNA sequence were printed out, it would correspond to about 120 volumes of *Harrison's Principles of Internal Medicine*.

In addition to the human genome, the genomes of numerous organisms have been sequenced completely (~4000) or partially (~10,000) (Genomes Online Database [GOLD]; Table 456-1). They include, among

TABLE 456-1 Selected Databases Relevant for Genomics and Genetic Disorders

SITE	URL	COMMENT
National Center for Biotechnology Information (NCBI)	*http://www.ncbi.nlm.nih.gov/*	Broad access to biomedical and genomic information, literature (PubMed), sequence databases, software for analyses of nucleotides and proteins Extensive links to other databases, genome resources, and tutorials
National Human Genome Research Institute	*http://www.genome.gov/*	An institute of the National Institutes of Health focused on genomic and genetic research; links providing information about the human genome sequence, genomes of other organisms, and genomic research
Catalog of Published Genome-Wide Association Studies	*http://www.genome.gov/GWAStudies/*	Published high-resolution genome-wide association studies (GWAS)
Ensembl Genome browser	*http://www.ensembl.org*	Maps and sequence information of eukaryotic genomes
Online Mendelian Inheritance in Man	*http://www.ncbi.nlm.nih.gov/omim*	Online compendium of Mendelian disorders and human genes causing genetic disorders
Office of Biotechnology Activities, National Institutes of Health	*http://oba.od.nih.gov/oba*	Information about recombinant DNA and gene transfer; medical, ethical, legal, and social issues raised by genetic testing; medical, ethical, legal, and social issues raised by xenotransplantation
American College of Medical Genetics and Genomics	*http://www.acmg.net/*	Extensive links to other databases relevant for the diagnosis, treatment, and prevention of genetic disease
American Society of Human Genetics	*http://www.ashg.org*	Information about advances in genetic research, professional and public education, social and scientific policies
The Cancer Genome Atlas	*https://cancergenome.nih.gov/*	Comprehensive, multi-dimensional characterization of the genomic and proteomic landscape of malignancies with high public health impact.
Cancer Genome Anatomy Project (CGAP)	*http://cgap.nci.nih.gov/*	Information about gene expression profiles of normal, precancer, and cancer cells
GeneTests	*http://www.genetests.org/*	International directory of genetic testing laboratories and prenatal diagnosis clinics; reviews and educational materials
Genomes Online Database (GOLD)	*http://www.genomesonline.org/*	Information on published and unpublished genomes
HUGO Gene Nomenclature	*http://www.genenames.org/*	Gene names and symbols
GENECODE	*https://www.gencodegenes.org/*	High quality reference gene annotation and experimental validation for human and mouse genomes
MITOMAP, a human mitochondrial genome database	*http://www.mitomap.org/*	A compendium of polymorphisms and mutations of the human mitochondrial DNA
The International Genome Sample Resource (IGSR)	*http://www.internationalgenome.org*	Public catalogue of human variation and genotype data from numerous ethnic groups
ENCODE	*http://www.genome.gov/10005107*	*Encyclopedia of DNA Elements*; catalogue of all functional elements in the human genome
Dolan DNA Learning Center, Cold Spring Harbor Laboratories	*http://www.dnalc.org/*	Educational material about selected genetic disorders, DNA, eugenics, and genetic origin
The Online Metabolic and Molecular Bases of Inherited Disease (OMMBID)	*http://ommbid.mhmedical.com*	Online version of the comprehensive text on the metabolic and molecular bases of inherited disease
Online Mendelian Inheritance in Animals (OMIA)	*http://omia.angis.org.au/*	Online compendium of Mendelian disorders in animals
The Jackson Laboratory	*http://www.jax.org/*	Information about murine models and the mouse genome
Mouse genome informatics	*http://www.informatics.jax.org*	Mouse genome informatics

Note: Databases are evolving constantly. Pertinent information may be found by using links listed in the few selected databases.

others, eukaryotes such as the mouse (*Mus musculus*), *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Drosophila melanogaster*; bacteria (e.g., *E. coli*); and archaea, viruses, organelles (mitochondria, chloroplasts), and plants (e.g., *Arabidopsis thaliana*). Genomic information of infectious agents has significant impact for the characterization of infectious outbreaks and epidemics. Other ramifications arising from the availability of genomic data include, among others, (1) the comparison of entire genomes (*comparative genomics*), (2) the study of large-scale expression of RNAs (*functional genomics*) and proteins (*proteomics*) to detect differences between various tissues in health and disease, (3) the characterization of the variation among individuals by establishing catalogues of sequence variations and SNPs, and (4) the identification of genes that play critical roles in the development of polygenic and multifactorial disorders.

CHROMOSOMES The human genome is divided into 23 different chromosomes, including 22 autosomes (numbered 1–22) and the X and Y sex chromosomes **(Fig. 456-1)**. Adult cells are *diploid*, meaning they contain two homologous sets of 22 autosomes and a pair of sex chromosomes. Females have two X chromosomes (XX), whereas males have one X and one Y chromosome (XY). As a consequence of meiosis, germ cells (sperm or oocytes) are haploid and contain one set of 22 autosomes and one of the sex chromosomes. At the time of fertilization, the diploid genome is reconstituted by pairing of the homologous chromosomes from the mother and father. With each cell division (mitosis), chromosomes are replicated, paired, segregated, and divided into two daughter cells.

STRUCTURE OF DNA DNA is a double-stranded helix composed of four different bases: adenine (A), thymidine (T), guanine (G), and cytosine (C). Adenine is paired to thymidine, and guanine is paired to cytosine, by hydrogen bond interactions that span the double helix (Fig. 456-1). DNA has several remarkable features that make it ideal for the transmission of genetic information. It is relatively stable, and the double-stranded nature of DNA and its feature of strict base-pair complementarity permit faithful replication during cell division. Complementarity also allows the transmission of genetic information from DNA → RNA → protein **(Fig. 456-2)**. mRNA is encoded by the so-called sense or coding strand of the DNA double helix and is translated into proteins by ribosomes.

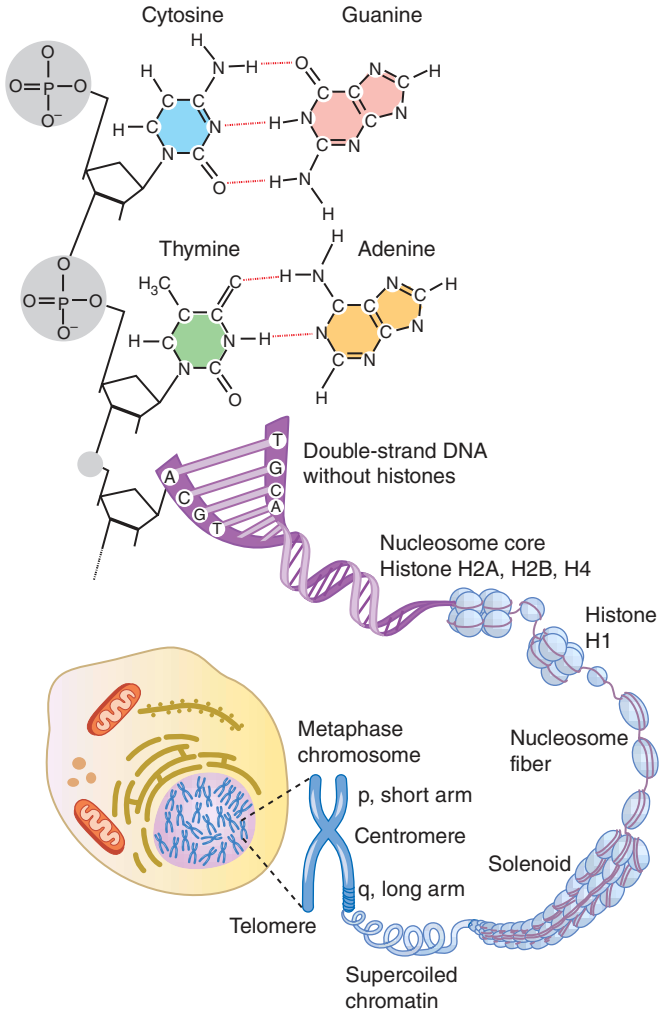


FIGURE 456-1 Structure of chromatin and chromosomes. Chromatin is composed of double-strand DNA that is wrapped around histone and nonhistone proteins forming nucleosomes. The nucleosomes are further organized into solenoid structures. Chromosomes assume their characteristic structure, with short (p) and long (q) arms at the metaphase stage of the cell cycle.

The presence of four different bases provides surprising genetic diversity. In the protein-coding regions of genes, the DNA bases are arranged into codons, a triplet of bases that specifies a particular amino acid. It is possible to arrange the four bases into 64 different triplet codons (4^3). Each codon specifies 1 of the 20 different amino acids, or a regulatory signal such as initiation and stop of translation. Because there are more codons than amino acids, the genetic code is degenerate; that is, most amino acids can be specified by several different codons. By arranging the codons in different combinations and in various lengths, it is possible to generate the tremendous diversity of primary protein structure.

DNA length is normally measured in units of 1000 bp (kilobases, kb) or 1,000,000 bp (megabases, Mb). Not all DNA encodes genes. In fact, genes account for only ~10–15% of DNA. Much of the remaining DNA consists of sequences, often of highly repetitive nature, the function of which remains incompletely understood. These repetitive DNA regions, along with nonrepetitive sequences that do not encode genes, serve, in part, a structural role in the packaging of DNA into chromatin (i.e., DNA bound to histone proteins, and chromosomes) and exert regulatory functions (Fig. 456-1).

GENES A *gene* is a functional unit that is regulated by transcription (see below) and encodes an RNA product, which is most commonly, but not always, translated into a protein that exerts activity within or outside the cell **(Fig. 456-3)**. Historically, genes were identified because they conferred specific traits that are transmitted from one generation to the next. Increasingly, they are characterized based on expression in various tissues (*transcriptome*). The size of genes is quite broad; some genes are only a few hundred base pairs, whereas others are extraordinarily large (2 Mb). The number of genes greatly underestimates the complexity of genetic expression, because single genes can generate multiple spliced messenger RNA (mRNA) products (*isoforms*), which are translated into proteins that are subject to complex posttranslational modification such as phosphorylation. *Exons* refer to the portion of genes that are eventually spliced together to form mRNA. *Introns* refer to the spacing regions between the exons that are spliced out of precursor RNAs during RNA processing. The gene locus also includes regions that are necessary to control its expression (Fig. 456-2). Current estimates predict roughly 20,000 protein-coding genes in the human genome with an average of about four different coding transcripts per gene. Remarkably, the exome only constitutes 1.14% of the genome. Of note, the number of transcripts is close to 200,000 and includes thousands of noncoding transcripts (RNAs of various length such as microRNAs [miRNA] and long noncoding RNAs [lncRNA]). These non-coding RNAs are involved in numerous cellular processes such as transcriptional and posttranscriptional regulation of gene expression, chromatin remodeling, and protein trafficking, among others. Not surprisingly, aberrant expression and/or mutations in these RNAs play a pathogenic role in numerous diseases.

SINGLE-NUCLEOTIDE POLYMORPHISMS Each individual has roughly 5 million sequence variants that differentiate one person from another. Some of these variants have no impact on health, whereas others may increase or lower the risk for developing a specific disease. Remarkably, however, the primary DNA sequence of humans has ~99.9% similarity compared to that of any other human. A SNP is a variation of a single base pair in the DNA. The identification of the ~10 million SNPs estimated to occur in the human genome has generated a catalogue of common genetic variants that occur in human beings from distinct ethnic backgrounds **(**Fig. 456-3). SNPs are the most common type of sequence variation and account for ~90% of all sequence variation. They occur on average every 100–300 bases and are the major source of genetic heterogeneity. SNPs that are in close proximity are inherited together (e.g., they are linked) and are referred to as *haplotypes* **(Fig. 456-4)**. Haplotype maps describe the nature and location of these SNP haplotypes and how they are distributed among individuals within and among populations, information that is facilitating GWAS designed to elucidate the complex interactions among multiple genes and lifestyle factors in multifactorial disorders (see below). Moreover, haplotype analyses are useful to assess variations in responses to medications (*pharmacogenomics*) and environmental factors, as well as the prediction of disease predisposition.

COPY NUMBER VARIATIONS Copy number variations (CNVs) are relatively large genomic regions (1 kb to several Mb) that have been duplicated or deleted on certain chromosomes and hence alter the diploid status of the DNA **(Fig. 456-5)**. It has been estimated that 5–10% of the genome can display CNVs. When comparing the genomes of two individuals, ~0.4–0.8% of their genomes differ in terms of CNVs scattered throughout the genome. Of note, de novo CNVs have been observed between monozygotic twins, who otherwise have identical genomes. Some CNVs have no functional consequences, whereas others have been associated with susceptibility or resistance to disease, and CNVs also occur in cancer cells.

Replication of DNA and Mitosis

Genetic information in DNA is transmitted to daughter cells under two different circumstances: (1) somatic cells divide by *mitosis*, allowing the diploid ($2n$) genome to replicate itself completely in conjunction with cell division; and (2) germ cells (sperm and ova) undergo *meiosis*, a process that enables the reduction of the diploid ($2n$) set of chromosomes to the haploid state ($1n$).

Prior to mitosis, cells exit the resting, or G_0 state, and enter the cell cycle. After traversing a critical checkpoint in G_1, cells undergo DNA synthesis (S phase), during which the DNA in each chromosome is replicated, yielding two pairs of sister chromatids ($2n \rightarrow 4n$). The process

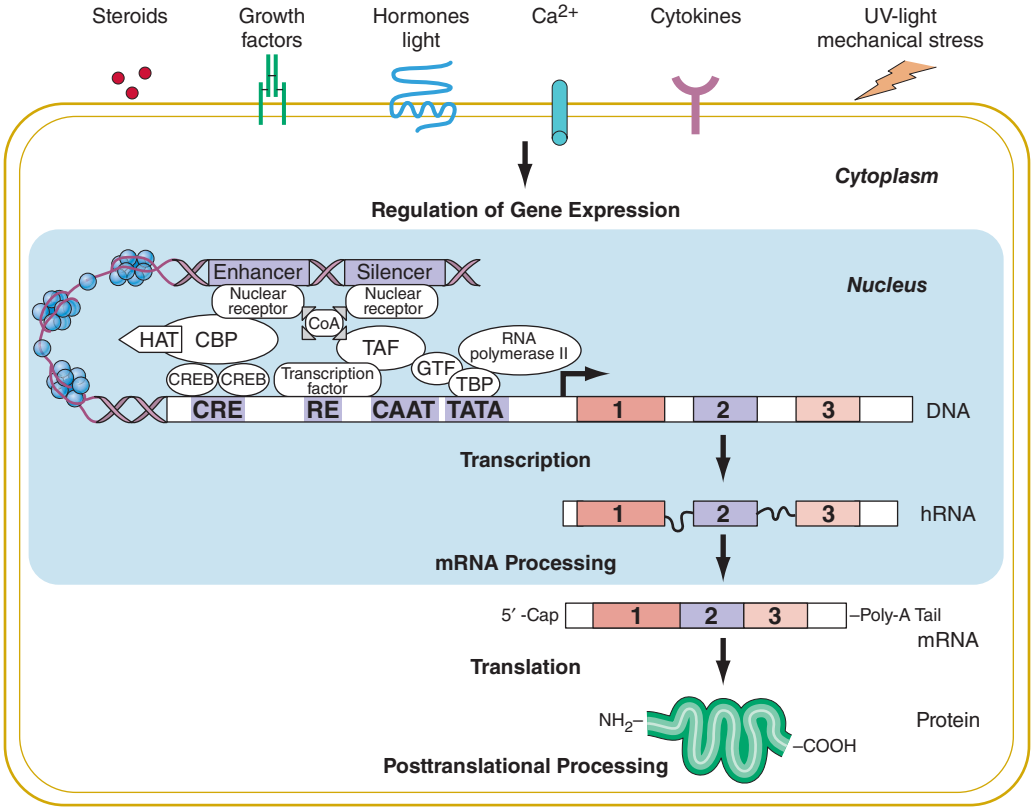


FIGURE 456-2 Flow of genetic information. Multiple extracellular signals activate intracellular signal cascades that result in altered regulation of gene expression through the interaction of transcription factors with regulatory regions of genes. RNA polymerase transcribes DNA into RNA that is processed to mRNA by excision of intronic sequences. The mRNA is translated into a polypeptide chain to form the mature protein after undergoing posttranslational processing. CBP, CREB-binding protein; CoA, co-activator; COOH, carboxyterminus; CRE, cyclic AMP responsive element; CREB, cyclic AMP response element–binding protein; GTF, general transcription factors; HAT, histone acetyl transferase; NH2, aminoterminus; RE, response element; TAF, TBP-associated factors; TATA, TATA box; TBP, TATA-binding protein.

of DNA synthesis requires stringent fidelity in order to avoid transmitting errors to subsequent generations of cells. Genetic abnormalities of DNA mismatch/repair include xeroderma pigmentosum, Bloom's syndrome, ataxia telangiectasia, and hereditary nonpolyposis colon cancer (HNPCC), among others. Many of these disorders strongly predispose to neoplasia because of the rapid acquisition of additional mutations **(Chap. 67)**. After completion of DNA synthesis, cells enter G_2 and progress through a second checkpoint before entering mitosis. At this stage, the chromosomes condense and are aligned along the equatorial plate at metaphase. The two identical sister chromatids, held together at the centromere, divide and migrate to opposite poles of the cell. After formation of a nuclear membrane around the two separated sets of chromatids, the cell divides and two daughter cells are formed, thus restoring the diploid ($2n$) state.

Assortment and Segregation of Genes During Meiosis

Meiosis occurs only in germ cells of the gonads. It shares certain features with mitosis but involves two distinct steps of cell division that reduce the chromosome number to the haploid state. In addition, there is active recombination that generates genetic diversity. During the first cell division, two sister chromatids ($2n \rightarrow 4n$) are formed for each chromosome pair and there is an exchange of DNA between homologous paternal and maternal chromosomes. This process involves the formation of *chiasmata*, structures that correspond to the DNA segments that cross over between the maternal and paternal homologues **(Fig. 456-6)**. Usually there is at least one crossover on each chromosomal arm; recombination occurs more frequently in female meiosis than in male meiosis. Subsequently, the chromosomes segregate randomly. Because there are 23 chromosomes, there exist 2^{23} (>8 million) possible combinations of chromosomes. Together with the genetic exchanges that occur during recombination, chromosomal segregation generates tremendous diversity, and each gamete is genetically unique. The process of recombination and the independent segregation of chromosomes provide the foundation for performing linkage analyses, whereby one attempts to correlate the inheritance of certain chromosomal regions (or linked genes) with the presence of a disease or genetic trait (see below).

After the first meiotic division, which results in two daughter cells ($2n$), the two chromatids of each chromosome separate during a second meiotic division to yield four gametes with a haploid state ($1n$). When the egg is fertilized by sperm, the two haploid sets are combined, thereby restoring the diploid state ($2n$) in the zygote.

REGULATION OF GENE EXPRESSION

Regulation by Transcription Factors The expression of genes is regulated by DNA-binding proteins that activate or repress transcription. The number of DNA sequences and transcription factors that regulate transcription is much greater than originally anticipated. Most genes contain at least 15–20 discrete regulatory elements within 300 bp of the transcription start site. This densely packed promoter region often contains binding sites for ubiquitous transcription factors. However, factors involved in cell-specific expression may also bind to these sequences. Key regulatory elements may also reside at a large distance from the proximal promoter. The globin and the immunoglobulin genes, for example, contain *locus control regions* that are several kilobases away from the structural sequences of the gene. Specific groups of transcription factors that bind to these promoter and enhancer sequences provide a combinatorial code for regulating transcription. In this manner, relatively ubiquitous factors interact with more restricted factors to allow each gene to be expressed and regulated in a unique manner that is dependent on developmental state, cell type, and numerous extracellular stimuli. Regulatory factors also bind within the gene itself, particularly in the intronic regions. The transcription factors that bind to DNA actually represent only the first level of regulatory control. Other proteins—*co-activators* and *co-repressors*—interact with the DNA-binding transcription factors to

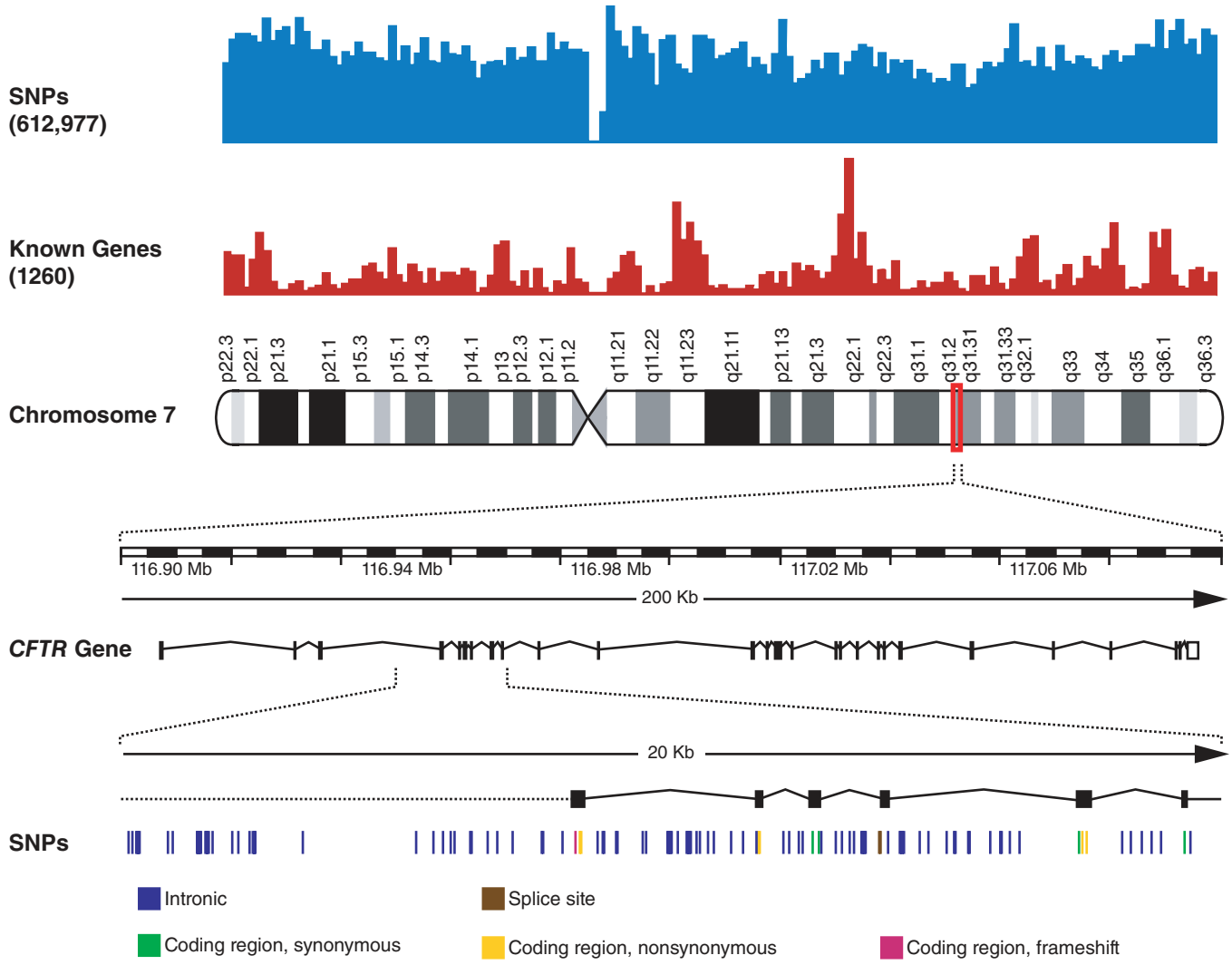


FIGURE 456-3 Chromosome 7 is shown with the density of single-nucleotide polymorphisms (SNPs) and genes above. A 200-kb region in 7q31.2 containing the *CFTR* gene is shown below. The *CFTR* gene contains 27 exons. Close to 2000 mutations in this gene have been found in patients with cystic fibrosis. A 20-kb region encompassing exons 4–9 is shown further amplified to illustrate the SNPs in this region.

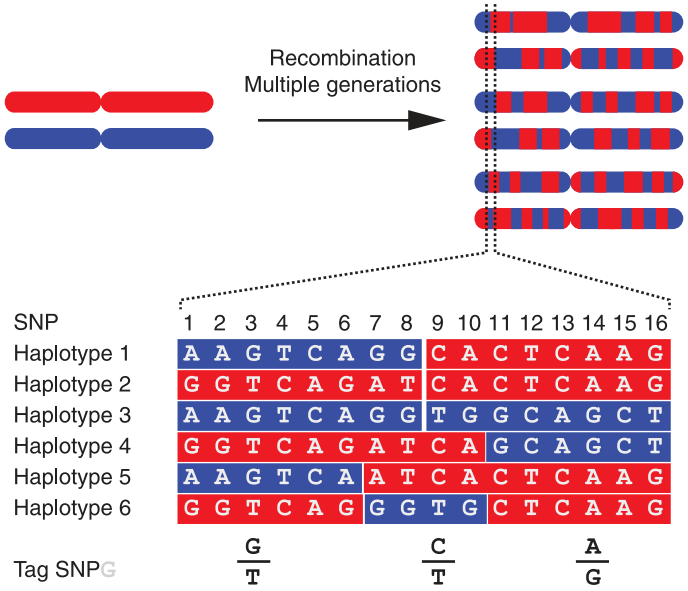


FIGURE 456-4 The origin of haplotypes is due to repeated recombination events occurring in multiple generations. Over time, this leads to distinct haplotypes. These haplotype blocks can often be characterized by genotyping selected Tag single-nucleotide polymorphisms (SNPs), an approach that facilitates performing genome-wide association studies (GWAS).

generate large regulatory complexes. These complexes are subject to control by numerous cell-signaling pathways and enzymes, leading to phosphorylation, acetylation, sumoylation, and ubiquitination. Ultimately, the recruited transcription factors interact with, and stabilize, components of the basal transcription complex that assembles at the site of the TATA box and initiator region. This basal transcription factor complex consists of >30 different proteins. Gene transcription occurs when RNA polymerase begins to synthesize RNA from the DNA template. A large number of identified genetic diseases involve transcription factors (Table 456-2).

The field of *functional genomics* is based on the concept that understanding alterations of gene expression under various physiologic and pathologic conditions provides insight into the underlying functional role of the gene. The ENCODE (ENCyclopedia Of DNA Elements) project aims to compile and annotate all functional sequences in the human genome. By revealing specific gene expression profiles, this knowledge may be of diagnostic and therapeutic relevance. The large-scale study of expression profiles, which takes advantage of micro and bead array technologies, is also referred to as *transcriptomics* because the complement of mRNAs transcribed by the cellular genome is called the *transcriptome*.

Most studies of gene expression have focused on the regulatory DNA elements of genes that control transcription. However, it should be emphasized that gene expression requires a series of steps, including mRNA processing, protein translation, and posttranslational modifications, all of which are actively regulated (Fig. 456-2).

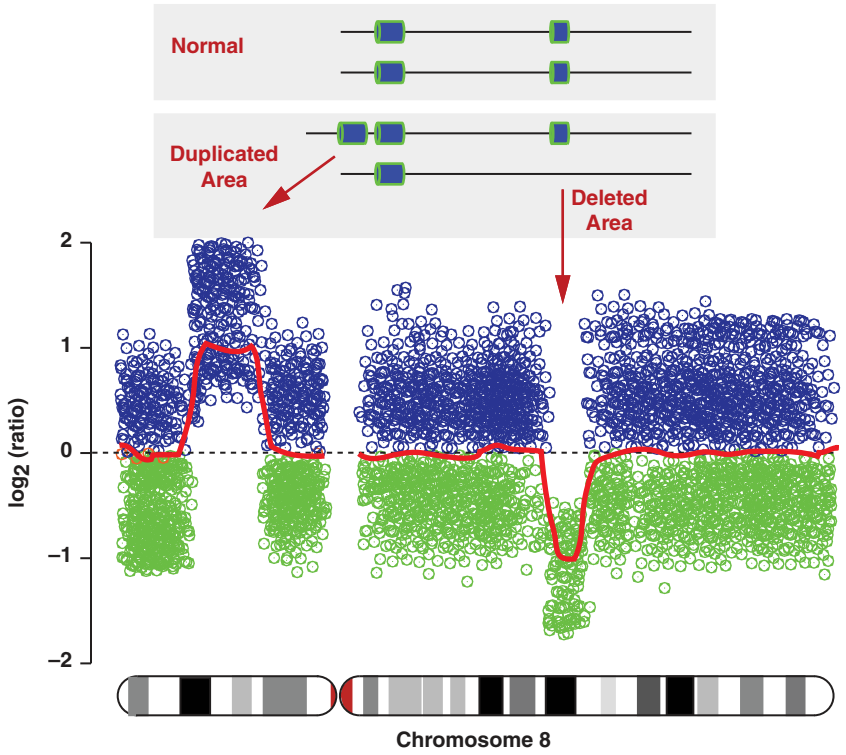


FIGURE 456-5 Copy number variations (CNV) encompass relatively large regions of the genome that have been duplicated or deleted. Chromosome 8 is shown with CNV detected by genomic hybridization. An increase in the signal strength indicates a duplication, a decrease reflects a deletion of the covered chromosomal regions.

Epigenetic Regulation of Gene Expression (see Chap. 471)

Epigenetics describes mechanisms and phenotypic changes that are not a result of variation in the primary DNA nucleotide sequence, but are caused by secondary modifications of DNA or histones. These modifications include heritable changes such as X-inactivation and imprinting, but they can also result from dynamic posttranslational protein modifications in response to environmental influences such as diet, age, or drugs. The epigenetic modifications result in altered expression of individual genes or chromosomal loci encompassing multiple genes. The term *epigenome* describes the constellation of covalent modifications of DNA and histones that impact chromatin structure, as well as noncoding transcripts that modulate the transcriptional activity of DNA. Although the primary DNA sequence is usually identical in all cells of an organism, tissue-specific changes in the epigenome contribute to determining the transcriptional signature of a cell (transcriptome) and hence the protein expression profile (proteome).

Mechanistically, DNA and histone modifications can result in the activation or silencing of gene expression (Fig. 456-7). DNA methylation involves the addition of a methyl group to cytosine residues. This is usually restricted to cytosines of CpG dinucleotides, which are abundant throughout the genome. Methylation of these dinucleotides is thought to represent a defense mechanism that minimizes the expression of sequences that have been incorporated into the genome such as retroviral sequences. CpG dinucleotides also exist in so-called *CpG islands*, stretches of DNA characterized by a high CG content, which are found in the majority of human gene promoters. CpG islands in promoter regions are typically unmethylated, and the lack of methylation facilitates transcription.

Histone methylation involves the addition of a methyl group to lysine residues in histone proteins (Fig. 456-7). Depending on the specific lysine residue being methylated, this alters chromatin configuration, either making it more open or tightly packed. Acetylation of histone proteins is another well-characterized mechanism that results in an open chromatin configuration, which favors active transcription. Acetylation is generally more dynamic than methylation, and many transcriptional activation complexes have histone acetylase activity, whereas repressor complexes often contain deacetylases and remove acetyl groups from histones. Other histone modifications, whose effects are incompletely characterized, include phosphorylation and sumoylation. Lastly, noncoding RNAs that bind to DNA can have a significant impact on transcriptional activity.

Physiologically, epigenetic mechanisms play an important role in several instances. For example, X-inactivation refers to the relative silencing of one of the two X chromosome copies present in females. The inactivation process is a form of dosage compensation such that females (XX) do not generally express twice as many X-chromosomal gene products as males (XY). In a given cell, the choice of which chromosome is inactivated occurs randomly in humans. But once the maternal or paternal X chromosome is inactivated, it will remain inactive, and this information is transmitted with each cell division. The *X-inactive specific transcript* (*Xist*) gene encodes a large noncoding RNA that mediates the silencing of the X chromosome from which it is transcribed by coating it with Xist RNA. The inactive X chromosome is highly methylated and has low levels of histone acetylation. While the majority of X-chromosomal genes are silenced by X-inactivation, about 15% escape inactivation and are expressed.

Epigenetic gene inactivation also occurs on selected chromosomal regions of autosomes, a phenomenon referred to as *genomic imprinting*. Through this mechanism, a small subset of genes is only expressed in a monoallelic fashion. Imprinting is heritable and leads to the preferential expression of one of the parental alleles, which deviates from the usual biallelic expression seen for the majority of genes. Remarkably, imprinting can

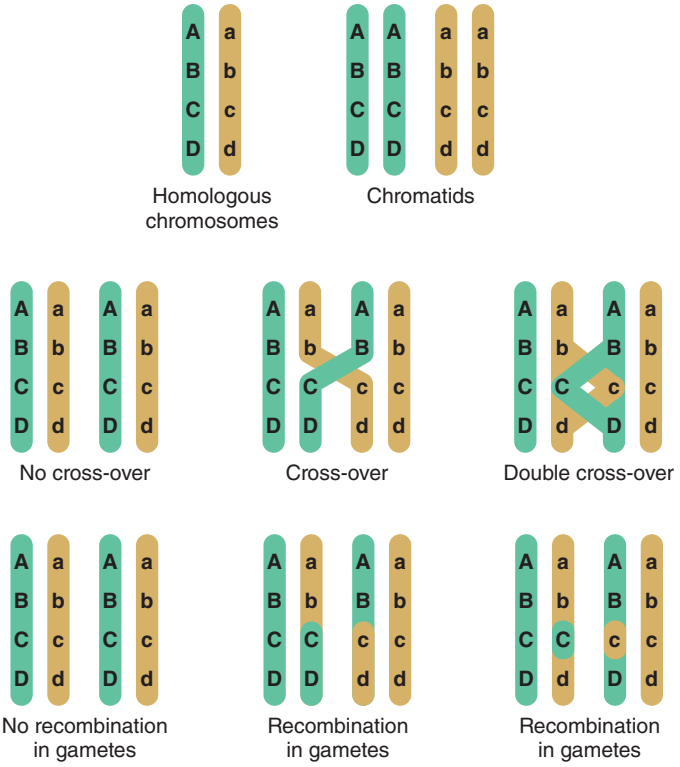


FIGURE 456-6 Crossing-over and genetic recombination. During chiasma formation, either of the two sister chromatids on one chromosome pairs with one of the chromatids of the homologous chromosome. Genetic recombination occurs through crossing-over and results in recombinant and nonrecombinant chromosome segments in the gametes. Together with the random segregation of the maternal and paternal chromosomes, recombination contributes to genetic diversity and forms the basis of the concept of linkage.

TABLE 456-2 Selected Examples of Diseases Caused by Mutations and Rearrangements in Transcription Factors

TRANSCRIPTION FACTOR CLASS	EXAMPLE	ASSOCIATED DISORDER
Nuclear receptors	Androgen receptor	Complete or partial androgen insensitivity (recessive missense mutations) Spinobulbar muscular atrophy (CAG repeat expansion)
Zinc finger proteins	WT1	WAGR syndrome: *W*ilms' tumor, *a*niridia, *g*enitourinary malformations, mental *r*etardation
Basic helix-loop-helix	MITF	Waardenburg's syndrome type 2A
Homeobox	IPF1	Maturity onset of diabetes mellitus type 4 (heterozygous mutation/haploinsufficiency) Pancreatic agenesis (homozygous mutation)
Leucine zipper	Retina leucine zipper (NRL)	Autosomal dominant retinitis pigmentosa
High mobility group (HMG) proteins	SRY	Sex reversal
Forkhead	HNF4α, HNF1α, HNF1β	Maturity onset of diabetes mellitus types 1, 3, 5
Paired box	PAX3	Waardenburg's syndrome types 1 and 3
T-box	TBX5	Holt-Oram syndrome (thumb anomalies, atrial or ventricular septum defects, phocomelia)
Cell cycle control proteins	P53	Li-Fraumeni syndrome, other cancers
Co-activators	CREB binding protein (CBP)	Rubinstein-Taybi syndrome
General transcription factors	TATA-binding protein (TBP)	Spinocerebellar ataxia 17 (CAG expansion)
Transcription elongation factor	VHL	von Hippel–Lindau syndrome (renal cell carcinoma, pheochromocytoma, pancreatic tumors, hemangioblastomas) Autosomal dominant inheritance, somatic inactivation of second allele (Knudson two-hit model)
Runt	CBFA2	Familial thrombocytopenia with propensity to acute myelogenous leukemia
Chimeric proteins due to translocations	PML-RAR	Acute promyelocytic leukemia t(15;17)(q22;q11.2-q12) translocation

Abbreviations: CREB, cAMP responsive element–binding protein; HNF, hepatocyte nuclear factor; PML, promyelocytic leukemia; RAR, retinoic acid receptor; SRY, sex-determining region Y; VHL, von Hippel–Lindau.

be limited to a subset of tissues. Imprinting is mediated through DNA methylation of one of the alleles. The epigenetic marks on imprinted genes are maintained throughout life, but during zygote formation, they are activated or inactivated in a sex-specific manner (imprint reset) **(Fig. 456-8)**, which allows a differential expression pattern in the fertilized egg and the subsequent mitotic divisions. Appropriate expression of imprinted genes is important for normal development and cellular functions. Imprinting defects and uniparental disomy, which is the inheritance of two chromosomes or chromosomal regions from the same parent, are the cause of several developmental disorders such as Beckwith-Wiedemann syndrome, Silver-Russell syndrome, Angelman's syndrome, and Prader-Willi syndrome (see below). Monoallelic loss-of-function mutations in the *GNAS1* gene lead to Albright's hereditary osteodystrophy (AHO). Paternal transmission of *GNAS1* mutations leads to an isolated AHO phenotype (pseudopseudohypoparathyroidism), whereas maternal transmission leads to AHO in combination with hormone resistance to parathyroid hormone, thyrotropin, and gonadotropins (pseudohypoparathyroidism type IA). These phenotypic differences are explained by tissue-specific imprinting of the *GNAS1* gene, which is expressed primarily from the maternal allele in the thyroid, gonadotropes, and the proximal renal tubule. In most other tissues, the *GNAS1* gene is expressed biallelically. In patients with isolated renal resistance to parathyroid hormone (pseudohypoparathyroidism type IB), defective imprinting of the *GNAS1* gene results in decreased Gsα expression in the proximal renal tubules. Rett's syndrome is an X-linked dominant disorder resulting in developmental regression and stereotypic hand movements in affected girls. It is caused by mutations in the *MECP2* gene, which encodes a methyl-binding protein. The ensuing aberrant methylation results in abnormal gene expression in neurons, which are otherwise normally developed.

Remarkably, epigenetic differences also occur among monozygotic twins. Although twins are epigenetically indistinguishable during the early years of life, older monozygotic twins exhibit differences in the overall content and genomic distribution of DNA methylation and histone acetylation, which would be expected to alter gene expression in various tissues.

In cancer, the epigenome is characterized by simultaneous losses and gains of DNA methylation in different genomic regions, as well as repressive histone modifications. Hyper- and hypomethylation are associated with mutations in genes that control DNA methylation. Hypomethylation is thought to remove normal control mechanisms that prevent expression of repressed DNA regions. It is also associated with genomic instability. Hypermethylation, in contrast, results in the silencing of CpG islands in promoter regions of genes, including tumor-suppressor genes. Epigenetic alterations are considered to be more easily reversible compared to genetic changes and modification of the epigenome with demethylating agents and histone deacetylases is being used in the treatment of various malignancies.

TRANSMISSION OF GENETIC DISEASE

Origins and Types of Mutations The term *mutation* is used to designate the process of generating genetic variations as well as the outcome of these alterations. A *mutation* can be defined as any change in the primary nucleotide sequence of DNA regardless of its functional consequences, although it often has a negative connotation. The more neutral term *variation* is now increasingly used to describe sequence changes and is recommended by several professional organizations and guidelines instead of *mutation*. Some variations may be lethal, others are less deleterious, and some may confer an evolutionary advantage. Variations can occur in the germline (sperm or oocytes); these can be transmitted to progeny. Alternatively, variations can occur during embryogenesis or in somatic tissues. Variations that occur during development lead to *mosaicism*, a situation in which tissues are composed of cells with different genetic constitutions. If the germline is mosaic, a mutation can be transmitted to some progeny but not others, which sometimes leads to confusion in assessing the pattern of inheritance. Somatic mutations that do not affect cell survival can sometimes be detected because of variable phenotypic effects in tissues (e.g., pigmented lesions in McCune-Albright syndrome). Other somatic mutations are associated with neoplasia because they confer a growth advantage to cells. Epigenetic events may also influence gene expression or facilitate genetic damage. With the exception of triplet nucleotide repeats, which can expand (see below), variations are usually stable.

Mutations are structurally diverse—they can involve the entire genome, as in triploidy (one extra set of chromosomes), or gross numerical or structural alterations in chromosomes or individual genes. Large deletions may affect a portion of a gene or an entire gene, or, if several genes are involved, they may lead to a *contiguous gene syndrome*. Unequal crossing-over between homologous genes can result in fusion gene mutations, as illustrated by color blindness. Variations involving single nucleotides are referred to as *point mutations*. Substitutions are called *transitions* if a purine is replaced by another purine base (A ↔ G) or if a pyrimidine is replaced by another pyrimidine (C ↔ T).

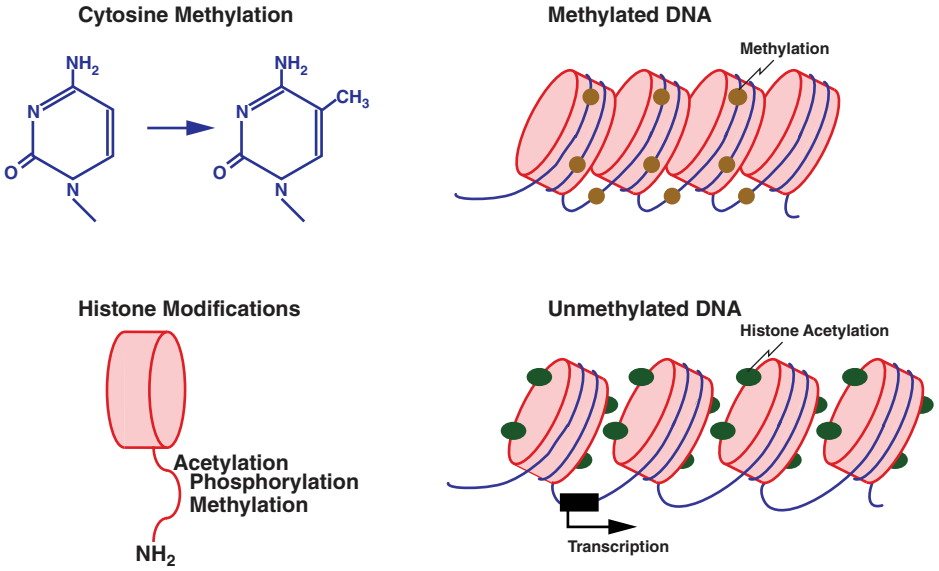


FIGURE 456-7 Epigenetic modifications of DNA and histones. Methylation of cytosine residues is associated with gene silencing. Methylation of certain genomic regions is inherited (imprinting), and it is involved in the silencing of one of the two X chromosomes in females (X-inactivation). Alterations in methylation can also be acquired, e.g., in cancer cells. Covalent posttranslational modifications of histones play an important role in altering DNA accessibility and chromatin structure and hence in regulating transcription. Histones can be reversibly modified in their amino-terminal tails, which protrude from the nucleosome core particle, by acetylation of lysine, phosphorylation of serine, methylation of lysine and arginine residues, and sumoylation. Acetylation of histones by histone acetylases (HATs), e.g., leads to unwinding of chromatin and accessibility to transcription factors. Conversely, deacetylation by histone deacetylases (HDACs) results in a compact chromatin structure and silencing of transcription.

Changes from a purine to a pyrimidine, or vice versa, are referred to as *transversions*. If the DNA sequence change occurs in a coding region and alters an amino acid, it is called a *missense mutation*. Depending on the functional consequences of such a missense mutation, amino acid substitutions in different regions of the protein can lead to distinct phenotypes.

Variations can occur in all domains of a gene **(Fig. 456-9)**. A point mutation occurring within the coding region leads to an amino acid substitution if the codon is altered **(Fig. 456-10)**. Point mutations that introduce a premature stop codon result in a truncated or missing protein. Large deletions may affect a portion of a gene or an entire gene, whereas small deletions and insertions alter the reading frame if they do not represent a multiple of three bases. These "frameshift" mutations, now also designated as *amphigoric* amino acid changes, lead to an entirely altered carboxy terminus. Mutations in intronic sequences or in exon junctions may destroy or create splice donor or splice acceptor sites. Variations may also be found in the regulatory sequences of genes, resulting in reduced or enhanced gene transcription.

Certain DNA sequences are particularly susceptible to mutagenesis. Successive pyrimidine residues (e.g., T-T or C-C) are subject to the formation of ultraviolet light–induced photoadducts. If these pyrimidine dimers are not repaired by the nucleotide excision repair pathway, mutations will be introduced after DNA synthesis. The dinucleotide C-G, or CpG, is also a hot spot for a specific type of mutation. In this case, methylation of the cytosine is associated with an enhanced rate of deamination to uracil, which is then replaced with thymine. This C → T transition (or G → A on the opposite strand) accounts for at least one-third of point mutations associated with polymorphisms and mutations. In addition to the fact that certain types of mutations (C → T or G → A) are relatively common, the nature of the genetic code also results in overrepresentation of certain amino acid substitutions.

Polymorphisms are sequence variations that have a frequency of at least 1%. Usually, they do not result in a perceptible phenotype but because allele frequency and functional consequences are often not known, the term variation is now increasingly recommended for the description of these sequence changes. Often they consist of single base-pair substitutions that do not alter the protein coding sequence because of the degenerate nature of the genetic code (synonymous polymorphism), although it is possible that some might alter mRNA stability, translation, or the amino acid sequence (nonsynonymous polymorphism) (Fig. 456-10). The detection of sequence variants poses a practical problem because it is often unclear whether it creates a change with functional consequences or a benign variation. In this situation, the sequence alteration is also described as *variant of unknown significance (VUS)*.

MUTATION RATES Mutations represent an important cause of genetic diversity as well as disease. Mutation rates are difficult to determine in humans because many mutations are silent and because testing is often not adequate to detect the phenotypic consequences. Mutation rates vary in different genes but are estimated to occur at a rate of ~10^{-10}/bp per cell division. Germline mutation rates (as opposed to somatic mutations) are relevant in the transmission of genetic disease. Because the population of oocytes is established very early in development, only ~20 cell divisions are required for completed oogenesis, whereas spermatogenesis involves ~30 divisions by the time of puberty and 20 cell divisions each year thereafter. Consequently, the probability of acquiring new point mutations is much greater in the male germline than the female germline, in which rates of aneuploidy are increased. Thus, the incidence of new point mutations in spermatogonia increases with paternal age (e.g., achondrodysplasia, Marfan's syndrome, neurofibromatosis). It is estimated that about 1 in 10 sperm carries a new deleterious mutation. The rates for new mutations are calculated most readily for autosomal dominant and X-linked disorders and are ~10^{-5}–10^{-6}/locus per generation. Because most monogenic diseases are relatively rare, new mutations account for a significant fraction of cases. This is important in the context of genetic counseling, because a new mutation can be transmitted to the affected individual but does not necessarily imply that the parents are at risk to transmit the disease to other children. An exception to this is when the new mutation occurs early in germline development, leading to *gonadal mosaicism*.

UNEQUAL CROSSING-OVER Normally, DNA recombination in germ cells occurs with remarkable fidelity to maintain the precise junction sites for the exchanged DNA sequences (Fig. 456-6). However, mispairing of homologous sequences leads to unequal crossover, with gene duplication on one of the chromosomes and gene deletion on the other chromosome. A significant fraction of growth hormone (*GH*) gene deletions, for example, involve unequal crossing-over **(Chap. 372)**. The *GH* gene is a member of a large gene cluster that includes a GH variant gene as well as several structurally related chorionic somatomammotropin genes and pseudogenes (highly homologous but functionally inactive relatives of a normal gene). Because such gene clusters contain multiple homologous DNA sequences arranged in tandem, they are particularly prone to undergo recombination and, consequently, gene duplication or deletion. On the other hand, duplication of the *PMP22* gene because of unequal crossing-over results in increased gene dosage and type IA Charcot-Marie-Tooth disease. Unequal crossing-over resulting in deletion of *PMP22* causes a distinct neuropathy called *hereditary liability to pressure palsy* **(Chap. 438)**.

Glucocorticoid-remediable aldosteronism (GRA) is caused by a gene fusion or rearrangement involving the genes that encode aldosterone synthase (*CYP11B2*) and steroid 11β-hydroxylase (*CYP11B1*), normally arranged in tandem on chromosome 8q. These two genes are 95%

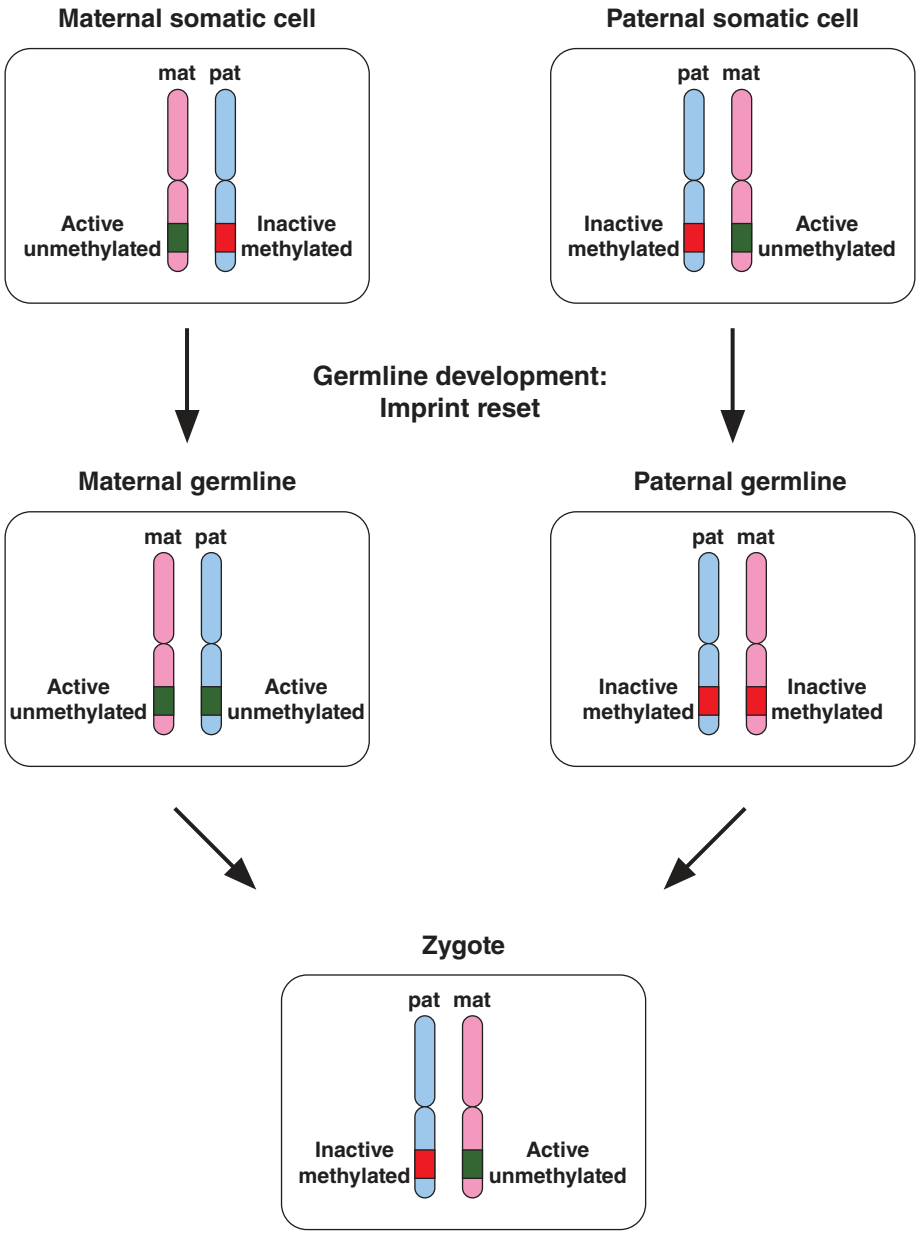


FIGURE 456-8 A few genomic regions are imprinted in a parent-specific fashion. The unmethylated chromosomal regions are actively expressed, whereas the methylated regions are silenced. In the germline, the imprint is reset in a parent-specific fashion: both chromosomes are unmethylated in the maternal (mat) germline and methylated in the paternal (pat) germline. In the zygote, the resulting imprinting pattern is identical with the pattern in the somatic cells of the parents.

identical, predisposing to gene duplication and deletion by unequal crossing-over. The rearranged gene product contains the regulatory regions of 11β-hydroxylase fused to the coding sequence of aldosterone synthetase. Consequently, the latter enzyme is expressed in the adrenocorticotropic hormone (ACTH)–dependent zona fasciculata of the adrenal gland, resulting in overproduction of mineralocorticoids and hypertension **(Chap. 379)**.

Gene conversion refers to a nonreciprocal exchange of homologous genetic information. It has been used to explain how an internal portion of a gene is replaced by a homologous segment copied from another allele or locus; these genetic alterations may range from a few nucleotides to a few thousand nucleotides. As a result of gene conversion, it is possible for short DNA segments of two chromosomes to be identical, even though these sequences are distinct in the parents. A practical consequence of this phenomenon is that nucleotide substitutions can occur during gene conversion between related genes, often altering the function of the gene. In disease states, gene conversion often involves intergenic exchange of DNA between a gene and a related pseudogene. For example, the 21-hydroxylase gene (*CYP21A2*) is adjacent to a nonfunctional pseudogene (*CYP21A1P*). Many of the nucleotide substitutions that are found in the *CYP21A2* gene in patients with congenital adrenal hyperplasia correspond to sequences that are present in the *CYP21A1P* pseudogene, suggesting gene conversion as one cause of mutagenesis. In addition, mitotic gene conversion has been suggested as a mechanism to explain revertant mosaicism in which an inherited mutation is "corrected" in certain cells. For example, patients with autosomal recessive generalized atrophic benign epidermolysis bullosa have acquired reverse mutations in one of the two mutated *COL17A1* alleles, leading to clinically unaffected patches of skin.

INSERTIONS AND DELETIONS Although many instances of insertions and deletions occur as a consequence of unequal crossing-over, there is also evidence for internal duplication, inversion, or deletion of DNA sequences. The fact that certain deletions or insertions appear to occur repeatedly as independent events indicates that specific regions within the DNA sequence predispose to these errors. For example, certain regions of the *DMD* gene, which encodes dystrophin, appear to be hot spots for deletions and result in muscular dystrophy **(Chap. 441)**. Some regions within the human genome are rearrangement hot spots and lead to CNVs.

ERRORS IN DNA REPAIR Because mutations caused by defects in DNA repair accumulate as somatic cells divide, these types of mutations are particularly important in the context of neoplastic disorders. Several genetic disorders involving DNA repair enzymes underscore their importance. Patients with xeroderma pigmentosum have defects in DNA damage recognition or in the nucleotide excision and repair pathway **(Chap. 72)**. Exposed skin is dry and pigmented and is extraordinarily sensitive to the mutagenic effects of ultraviolet irradiation. More than 10 different genes have been shown to cause the different forms of xeroderma pigmentosum. This finding is consistent with the earlier classification of this disease into different complementation groups in which normal function is rescued by the fusion of cells derived from two different forms of xeroderma pigmentosum.

Ataxia telangiectasia causes large telangiectatic lesions of the face, cerebellar ataxia, immunologic defects, and hypersensitivity to ionizing radiation **(Chap. 431)**. The discovery of the ataxia telangiectasia mutated (*ATM*) gene reveals that it is homologous to genes involved in DNA repair and control of cell cycle checkpoints. Mutations in the *ATM* gene give rise to defects in meiosis as well as increasing susceptibility to damage from ionizing radiation. Fanconi's anemia is also associated with an increased risk of multiple acquired genetic abnormalities. It is characterized by diverse congenital anomalies and a strong predisposition to develop aplastic anemia and acute myelogenous leukemia **(Chap. 100)**. Cells from these patients are susceptible to chromosomal breaks caused by a defect in genetic recombination. Currently, at least 16 different complementation groups have been identified, and the genes associated with Fanconi's anemia have been cloned. HNPCC (Lynch's syndrome) is characterized by autosomal dominant transmission of colon cancer, young age (<50 years) of presentation, predisposition

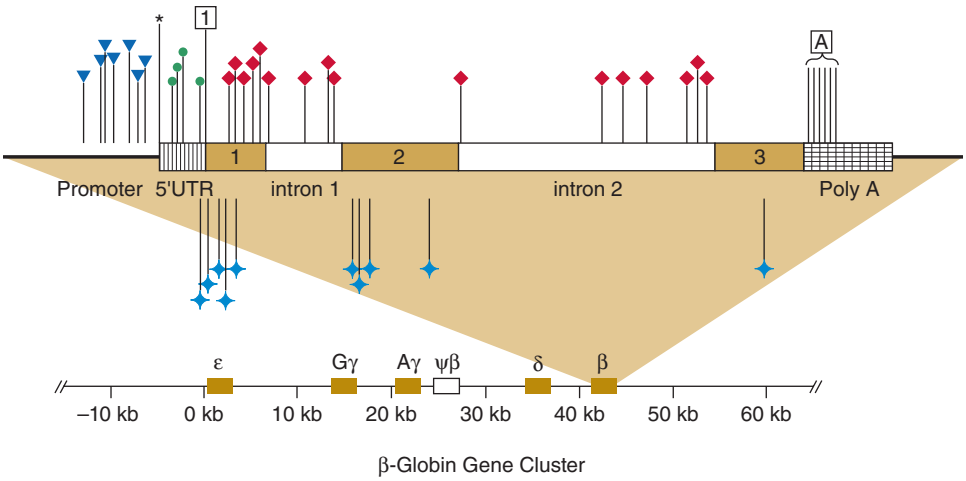


FIGURE 456-9 Point mutations causing β thalassemia as example of allelic heterogeneity. The β-globin gene is located in the globin gene cluster. Point mutations can be located in the promoter, the CAP site, the 5'-untranslated region, the initiation codon, each of the three exons, the introns, or the polyadenylation signal. Many mutations introduce missense or nonsense mutations, whereas others cause defective RNA splicing. Not shown here are deletion mutations of the β-globin gene or larger deletions of the globin locus that can also result in thalassemia. ▼, promoter mutations; *, CAP site; ●, 5'UTR; 1, initiation codon; ◆, defective RNA processing; ✦, missense and nonsense mutations; A, Poly A signal.

to lesions in the proximal large bowel, and associated malignancies such as uterine cancer and ovarian cancer. HNPCC is predominantly caused by mutations in one of several different mismatch repair (MMR) genes including MutS homologue 2 (*MSH2*), MutL homologue 1 and 6 (*MLH1, MLH6*), *MSH6*, *PMS1*, and *PMS2* **(Chap. 77)**. These proteins are involved in the detection of nucleotide mismatches and in the recognition of slipped-strand trinucleotide repeats. Germline mutations in these genes lead to microsatellite instability and a high mutation rate in colon cancer. Genetic screening tests for this disorder are now being used for families considered to be at risk. Recognition of HNPCC allows early screening with colonoscopy and the implementation of prevention strategies using nonsteroidal anti-inflammatory drugs.

UNSTABLE DNA SEQUENCES *Trinucleotide repeats* may be unstable and expand beyond a critical number. Mechanistically, the expansion is thought to be caused by unequal recombination and slipped mispairing. A premutation represents a small increase in trinucleotide copy number. In subsequent generations, the expanded repeat may increase further in length and result in an increasingly severe phenotype, a process called *dynamic mutation* (see below for discussion of anticipation). Trinucleotide expansion was first recognized as a cause of the fragile X syndrome, one of the most common causes of intellectual disability. Other disorders arising from a similar mechanism include Huntington's disease, X-linked spinobulbar muscular atrophy, and myotonic dystrophy. Malignant cells are also characterized by genetic instability, indicating a breakdown in mechanisms that regulate DNA repair and the cell cycle.

Functional Consequences of Mutations Functionally, mutations can be broadly classified as gain-of-function and loss-of-function mutations. Gain-of-function mutations are typically dominant (e.g., they result in phenotypic alterations when a single allele is affected). Inactivating mutations are usually recessive, and an affected individual is homozygous or compound heterozygous (e.g., carrying two different mutant alleles of the same gene) for the disease-causing mutations. Alternatively, mutation in a single allele can result in *haploinsufficiency*, a situation in which one normal allele is not sufficient to maintain a normal phenotype. Haploinsufficiency is a commonly observed mechanism in diseases associated with mutations in transcription factors (Table 456-2). Remarkably, the clinical features among patients with an identical mutation often vary significantly. One mechanism underlying this variability consists in the influence of modifying genes. Haploinsufficiency can also affect the expression of rate-limiting enzymes. For example, haploinsufficiency in enzymes involved in heme synthesis can cause porphyrias **(Chap. 409)**.

An increase in dosage of a gene product may also result in disease, as illustrated by the duplication of the *DAX1* gene in dosage-sensitive

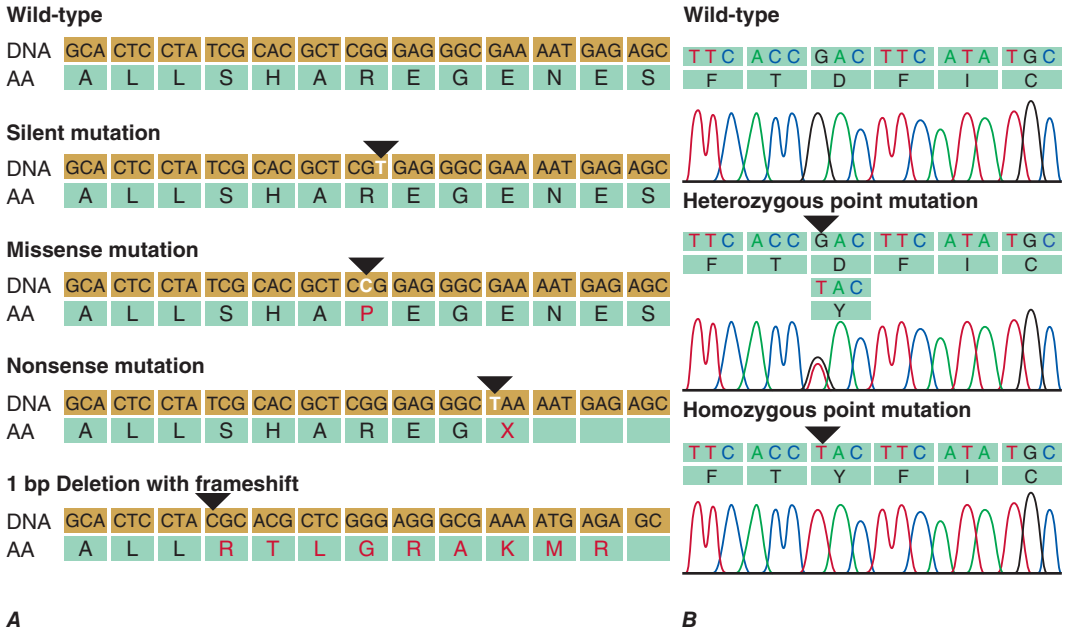


FIGURE 456-10 ***A.*** Examples of mutations (now commonly referred to as *variations*). The coding strand is shown with the encoded amino acid sequence. ***B.*** Chromatograms of sequence analyses after amplification of genomic DNA by polymerase chain reaction.

sex reversal **(Chap. 383)**. Mutation in a single allele can also result in loss of function due to a dominant-negative effect. In this case, the mutated allele interferes with the function of the normal gene product by one of several different mechanisms: (1) a mutant protein may interfere with the function of a multimeric protein complex, as illustrated by mutations in type 1 collagen (*COL1A1*, *COL1A2*) genes in osteogenesis imperfecta **(Chap. 406)**; (2) a mutant protein may occupy binding sites on proteins or promoter response elements, as illustrated by thyroid hormone resistance β, a disorder in which inactivated thyroid hormone receptor β binds to target genes and functions as an antagonist of normal receptors **(Chap. 375)**; or (3) a mutant protein can be cytotoxic as in α_1 antitrypsin deficiency **(Chap. 286)** or autosomal dominant neurohypophyseal diabetes insipidus **(Chap. 374)**, in which the abnormally folded proteins are trapped within the endoplasmic reticulum and ultimately cause cellular damage.

Genotype and Phenotype

• ALLELES, GENOTYPES, AND HAPLOTYPES An observed trait is referred to as a *phenotype*; the genetic information defining the phenotype is called the *genotype*. Alternative forms of a gene or a genetic marker are referred to as *alleles*. Alleles may be polymorphic variants of nucleic acids that have no apparent effect on gene expression or function. In other instances, these variants may have subtle effects on gene expression, thereby conferring adaptive advantages associated with genetic diversity. On the other hand, allelic variants may reflect mutations that clearly alter the function of a gene product. The common Glu6Val (E6V) sickle cell mutation in the β-*globin* gene and the ΔF508 deletion of phenylalanine (F) in the *CFTR* gene are examples of allelic variants of these genes that result in disease. Because each individual has two copies of each chromosome (one inherited from the mother and one inherited from the father), an individual can have only two alleles at a given locus. However, there can be many different alleles in the population. The normal or common allele is usually referred to as *wild type*. When alleles at a given locus are identical, the individual is *homozygous*. Inheriting identical copies of a mutant allele occurs in many autosomal recessive disorders, particularly in circumstances of consanguinity or isolated populations. If the alleles are different on the maternal and the paternal copy of the gene, the individual is *heterozygous* at this locus (Fig. 456-10). If two different mutant alleles are inherited at a given locus, the individual is said to be a *compound heterozygote*. *Hemizygous* is used to describe males with a mutation in an X chromosomal gene or a female with a loss of one X chromosomal locus.

Genotypes describe the specific alleles at a particular locus. For example, there are three common alleles (E2, E3, E4) of the apolipoprotein E (*APOE*) gene. The genotype of an individual can therefore be described as *APOE3/4* or *APOE4/4* or any other variant. These designations indicate which alleles are present on the two chromosomes in the *APOE* gene at locus 19q13.2. In other cases, the genotype might be assigned arbitrary numbers (e.g., 1/2) or letters (e.g., B/b) to distinguish different alleles.

A *haplotype* refers to a group of alleles that are closely linked together at a genomic locus (Fig. 456-4). Haplotypes are useful for tracking the transmission of genomic segments within families and for detecting evidence of genetic recombination, if the crossover event occurs between the alleles (Fig. 456-6). As an example, various alleles at the histocompatibility locus antigen (HLA) on chromosome 6p are used to establish haplotypes associated with certain disease states. For example, 21-hydroxylase deficiency, complement deficiency, and hemochromatosis are each associated with specific HLA haplotypes. It is now recognized that these genes lie in close proximity to the HLA locus, which explains why HLA associations were identified even before the disease genes were cloned and localized. In other cases, specific HLA associations with diseases such as ankylosing spondylitis (HLA-B27) or type 1 diabetes mellitus (HLA-DR4) reflect the role of specific HLA allelic variants in susceptibility to these autoimmune diseases. The characterization of common SNP haplotypes in numerous populations from different parts of the world has provided the necessary tools for association studies designed to detect genes involved in the pathogenesis of complex disorders (Table 456-1). The presence or absence of certain haplotypes can also be relevant for the customized choice of medical therapies (pharmacogenomics) or may have value for preventive strategies.

Genotype-phenotype correlation describes the association of a specific mutation and the resulting phenotype. The phenotype may differ depending on the location or type of the mutation in some genes. For example, in von Hippel–Lindau disease, an autosomal dominant multisystem disease that can include renal cell carcinoma, hemangioblastomas, and pheochromocytomas, among others, the phenotype varies greatly and the identification of the specific mutation can be clinically useful in order to predict the phenotypic spectrum.

ALLELIC HETEROGENEITY *Allelic heterogeneity* refers to the fact that different mutations in the same genetic locus can cause an identical or similar phenotype. For example, many different mutations of the β-globin locus can cause β thalassemia **(Table 456-3)** (Fig. 456-9). In essence, allelic heterogeneity reflects the fact that many different mutations are capable of altering protein structure and function. For this reason, maps of inactivating mutations in genes usually show a near-random distribution. Exceptions include (1) a founder effect, in which a particular mutation that does not affect reproductive capacity can be traced to a single individual; (2) "hot spots" for mutations, in which the nature of the DNA sequence predisposes to a recurring mutation; and (3) localization of mutations to certain domains that are particularly critical for protein function. Allelic heterogeneity creates a practical problem for genetic testing because one must often examine the entire genetic locus for mutations, because these can differ in each patient. For example, about 2000 variants have been identified in the *CFTR* gene to date, although some of them are very rare and some may not be disease-causing (Fig. 456-3). Mutational analysis may initially focus on a panel of mutations that are particularly frequent (often taking the ethnic background of the patient into account), but a negative result does not exclude the presence of a mutation elsewhere in the gene. One should also be aware that mutational analyses tend to focus on the coding region of a gene without considering regulatory and intronic regions. Because disease-causing mutations may be located outside the coding regions, negative results need to be interpreted with caution. The advent of more comprehensive sequencing technologies now greatly facilitates concomitant mutational analyses of several genes after targeted enrichment, or even mutational analysis of the whole exome or genome. However, comprehensive sequencing can result in significant diagnostic challenges because the detection of a sequence alteration alone is not always sufficient to establish that it has a causal role (variants of unknown significance, VUS).

PHENOTYPIC HETEROGENEITY *Phenotypic heterogeneity* occurs when more than one phenotype is caused by allelic mutations (e.g., different mutations in the same gene) (Table 456-3). For example, laminopathies are monogenic multisystem disorders that result from mutations in the *LMNA* gene, which encodes the nuclear lamins A and C. Twelve autosomal dominant and five autosomal recessive disorders are caused by mutations in the *LMNA* gene. They include several forms of lipodystrophies, Emery-Dreifuss muscular dystrophy, progeria syndromes, a form of neuronal Charcot-Marie-Tooth disease (type 2B1), and a group of overlapping syndromes. Remarkably, hierarchical cluster analysis has revealed that the phenotypes vary depending on the position of the mutation (*genotype-phenotype correlation*). Similarly, identical mutations in the *FGFR2* gene can result in very distinct phenotypes: Crouzon's syndrome (craniofacial synostosis) or Pfeiffer's syndrome (acrocephalopolysyndactyly).

LOCUS OR NONALLELIC HETEROGENEITY AND PHENOCOPIES *Nonallelic or locus heterogeneity* refers to the situation in which a similar disease phenotype results from mutations at different genetic loci (Table 456-3). This often occurs when more than one gene product produces different subunits of an interacting complex or when different genes are involved in the same genetic cascade or physiologic pathway. For example, osteogenesis imperfecta can arise from mutations in two different procollagen genes (*COL1A1* or *COL1A2*) that are located on different chromosomes, and can involve multiple other genes **(Chap. 406)**. The effects of inactivating mutations in these two genes are similar because

TABLE 456-3 Selected Examples of Phenotypic Heterogeneity and Locus Heterogeneity

Phenotypic Heterogeneity

GENE, PROTEIN	PHENOTYPE	INHERITANCE	OMIM
LMNA, Lamin A/C	Emery-Dreifuss muscular dystrophy (AD)	AD	181350
	Familial partial lipodystrophy Dunnigan	AD	151660
	Hutchinson-Gilford progeria	AD	176670
	Atypical Werner's syndrome	AD	150330
	Dilated cardiomyopathy	AD	115200
	Early-onset atrial fibrillation	AD	607554
	Emery-Dreifuss muscular dystrophy (AR)	AR	604929
	Limb-girdle muscular dystrophy type 1B	AR	159001
	Charcot-Marie-Tooth type 2B1	AR	605588
KRAS	Noonan's syndrome	AD	163950
	Cardio-facio-cutaneous syndrome	AD	115150

Locus Heterogeneity

PHENOTYPE	GENE	CHROMOSOMAL LOCATION	PROTEIN
Familial hypertrophic cardiomyopathy	*MYH7*	14q12	Myosin heavy chain beta
Genes encoding sarcomeric proteins	*TNNT2*	1q2	Troponin-T2
	TPM1	15q22.1	Tropomyosin alpha
	MYBPC3	11p11q	Myosin-binding protein C
	TNNI3	19q13.4	Troponin 1
	MYL2	12q23-24.3	Myosin light chain 2
	MYL3	3p	Myosin light chain 3
	TTN	2q24.3	Cardiac titin
	ACTC	15q11	Cardiac alpha actin
	MYH6	14q1	Myosin heavy chain alpha
	MYLK2	20q13.3	Myosin light-peptide kinase
	CAV3	3p25	Caveolin 3
Genes encoding nonsarcomeric proteins	*MTT1*	Mitochondrial	tRNA isoleucine
	MTTG	Mitochondrial	tRNA glycine
	PRKAG2	7q35-q36	AMP-activated protein kinase γ2 subunit
	DMPK	19q13.2-13.3	Myotonin protein kinase (myotonic dystrophy)
	FRDA	9q13	Frataxin (Friedreich's ataxia)
Polycystic kidney disease	*PKD1*	16p13.3-13.12	Polycystin 1 (AD)
	PKD2	4q21.-23	Polycystin 2 (AD)
	PKHD1	6p21.1-p12	Fibrocystin (AR)
Noonan's syndrome	*PTPN11*	12q24.1	Protein-tyrosine phosphatase 2c
	KRAS	12p12.1	KRAS

the protein products comprise different subunits of the helical collagen fiber. Similarly, muscular dystrophy syndromes can be caused by mutations in various genes, consistent with the fact that it can be transmitted in an X-linked (Duchenne or Becker), autosomal dominant (limb-girdle muscular dystrophy type 1), or autosomal recessive (limb-girdle muscular dystrophy type 2) manner **(Chap. 441)**. Mutations in the X-linked *DMD* gene, which encodes dystrophin, are the most common cause of muscular dystrophy. This feature reflects the large size of the gene as well as the fact that the phenotype is expressed in hemizygous males because they have only a single copy of the X chromosome. Dystrophin is associated with a large protein complex linked to the membrane-associated cytoskeleton in muscle. Mutations in several different components of this protein complex can also cause muscular dystrophy syndromes. Although the phenotypic features of some of these disorders are distinct, the phenotypic spectrum caused by mutations in different genes overlaps, thereby leading to nonallelic heterogeneity. It should be noted that mutations in dystrophin are also associated with allelic heterogeneity. For example, mutations in the *DMD* gene can cause either Duchenne's or the less severe Becker's muscular dystrophy, depending on the severity of the protein defect.

Recognition of nonallelic heterogeneity is important for several reasons: (1) the ability to identify disease loci in linkage studies is reduced by including patients with similar phenotypes but different genetic disorders; (2) genetic testing is more complex because several different genes need to be considered along with the possibility of different mutations in each of the candidate genes; and (3) novel information is gained about how genes or proteins interact, providing unique insights into molecular physiology.

Phenocopies refer to circumstances in which nongenetic conditions mimic a genetic disorder. For example, features of toxin- or drug-induced neurologic syndromes can resemble those seen in Huntington's disease, and vascular causes of dementia share phenotypic features with familial forms of Alzheimer's dementia **(Chap. 423)**. As in nonallelic heterogeneity, the presence of phenocopies has the potential to confound linkage studies and genetic testing. Patient history and subtle differences in phenotype can often provide clues that distinguish these disorders from related genetic conditions.

VARIABLE EXPRESSIVITY AND INCOMPLETE PENETRANCE The same genetic mutation may be associated with a phenotypic spectrum in

different affected individuals, thereby illustrating the phenomenon of *variable expressivity*. This may include different manifestations of a disorder variably involving different organs (e.g., multiple endocrine neoplasia [MEN]), the severity of the disorder (e.g., cystic fibrosis), or the age of disease onset (e.g., Alzheimer's dementia). MEN 1 illustrates several of these features. In this autosomal dominant tumor syndrome, affected individuals carry an inactivating germline mutation that is inherited in an autosomal dominant fashion. After somatic inactivation of the alternate allele (loss of heterozygosity; Knudson two-hit model), they can develop tumors of the parathyroid gland, endocrine pancreas, and the pituitary gland **(Chap. 381)**. However, the pattern of tumors in the different glands, the age at which tumors develop, and the types of hormones produced vary among affected individuals, even within a given family. In this example, the phenotypic variability arises, in part, because of the requirement for a second somatic mutation in the normal copy of the *MEN1* gene, as well as the large array of different cell types that are susceptible to the effects of *MEN1* gene mutations. In part, variable expression reflects the influence of modifier genes, or genetic background, on the effects of a particular mutation. Even in identical twins, in whom the genetic constitution is essentially the same, one can occasionally see variable expression of a genetic disease.

Interactions with the environment can also influence the course of a disease. For example, the manifestations and severity of hemochromatosis can be influenced by iron intake **(Chap. 407)**, and the course of phenylketonuria is affected by exposure to phenylalanine in the diet **(Chap. 413)**. Other metabolic disorders, such as hyperlipidemias and porphyria, also fall into this category. Many mechanisms, including genetic effects and environmental influences, can therefore lead to variable expressivity. In genetic counseling, it is particularly important to recognize this variability, because one cannot always predict the course of disease, even when the mutation is known.

Penetrance refers to the proportion of individuals with a mutant genotype that express the phenotype. If all carriers of a mutant express the phenotype, penetrance is complete, whereas it is said to be *incomplete* or *reduced* if some individuals do not exhibit features of the phenotype. Dominant conditions with incomplete penetrance are characterized by skipping of generations with unaffected carriers transmitting the mutant gene. For example, hypertrophic obstructive cardiomyopathy (HCM) caused by mutations in the *myosin-binding protein C* gene is a dominant disorder with clinical features in only a subset of patients who carry the mutation **(Chap. 254)**. Patients who have the mutation but no evidence of the disease can still transmit the disorder to subsequent generations. In many conditions with postnatal onset, the proportion of gene carriers who are affected varies with age. Thus, when describing penetrance, one has to specify age. For example, for disorders such as Huntington's disease or familial amyotrophic lateral sclerosis, which present later in life, the rate of penetrance is influenced by the age at which the clinical assessment is performed. *Imprinting* can also modify the penetrance of a disease. For example, in patients with AHO, mutations in the Gsα subunit (*GNAS1* gene) are expressed clinically only in individuals who inherit the mutation from their mother **(Chap. 403)**.

SEX-INFLUENCED PHENOTYPES Certain mutations affect males and females quite differently. In some instances, this is because the gene resides on the X or Y sex chromosomes (X-linked disorders and Y-linked disorders). As a result, the phenotype of mutated X-linked genes will be expressed fully in males but variably in heterozygous females, depending on the degree of X-inactivation and the function of the gene. For example, most heterozygous female carriers of factor VIII deficiency (hemophilia A) are asymptomatic because sufficient factor VIII is produced to prevent a defect in coagulation **(Chap. 112)**. On the other hand, some females heterozygous for the X-linked lipid storage defect caused by α-galactosidase A deficiency (Fabry's disease) experience mild manifestations of painful neuropathy, as well as other features of the disease **(Chap. 411)**. Because only males have a Y chromosome, mutations in genes such as *SRY*, which causes male-to-female sex reversal, or *DAZ* (deleted in azoospermia), which causes abnormalities of spermatogenesis, are unique to males **(Chap. 383)**.

Other diseases are expressed in a sex-limited manner because of the differential function of the gene product in males and females. Activating mutations in the luteinizing hormone receptor cause dominant male-limited precocious puberty in boys **(Chap. 384)**. The phenotype is unique to males because activation of the receptor induces testosterone production in the testis, whereas it is functionally silent in the immature ovary. Biallelic inactivating mutations of the follicle-stimulating hormone (FSH) receptor cause primary ovarian failure in females because the follicles do not develop in the absence of FSH action. In contrast, affected males have a more subtle phenotype, because testosterone production is preserved (allowing sexual maturation) and spermatogenesis is only partially impaired **(Chap. 384)**. In congenital adrenal hyperplasia, most commonly caused by 21-hydroxylase deficiency, cortisol production is impaired and ACTH stimulation of the adrenal gland leads to increased production of androgenic precursors **(Chap. 379)**. In females, the increased androgen level causes ambiguous genitalia, which can be recognized at the time of birth. In males, the diagnosis may be made on the basis of adrenal insufficiency at birth, because the increased adrenal androgen level does not alter sexual differentiation, or later in childhood, because of the development of precocious puberty. Hemochromatosis is more common in males than in females, presumably because of differences in dietary iron intake and losses associated with menstruation and pregnancy in females **(Chap. 407)**.

Chromosomal Disorders *Chromosomal disorders* and the techniques used for their characterization have been discussed in detail in a chapter in previous editions of this textbook. Chromosomal or cytogenetic disorders are caused by numerical (aneuploidy) or structural aberrations (deletions, duplications, translocations, inversions, dicentric and ring chromosomes, Robertsonian translocations) in chromosomes. They occur in about 1% of the general population, in 8% of stillbirths, and in close to 50% of spontaneously aborted fetuses. Indications for cytogenetic and cytogenomic chromosome analyses are summarized in **Table 456-4**. *Contiguous gene syndromes* (e.g., large deletions affecting several genes) have been useful for identifying the location of new disease-causing genes. Because of the variable size of gene deletions in different patients, a systematic comparison of phenotypes and locations of deletion breakpoints allows positions of particular genes to be mapped within the critical genomic region.

Monogenic Mendelian Disorders Monogenic human diseases are frequently referred to as *Mendelian disorders* because they obey

TABLE 456-4 Indications for Cytogenetic and Cytogenomic Analysis across the Lifespan

TIMING OF TESTING	INDICATIONS FOR TESTING
Prenatal	Advanced maternal age
	Abnormalities on ultrasound
	Increased risk for genetic disorder on maternal serum screen
Neonatal and Childhood	Multiple congenital anomalies
	Intellectual disability
	Autism
	Developmental delay
	Failure to thrive
	Short stature
	Disorders of sexual development
	History of familial chromosomal alteration
	Cancer
Adult	Infertility
	Recurrent miscarriage
	Familial cancer

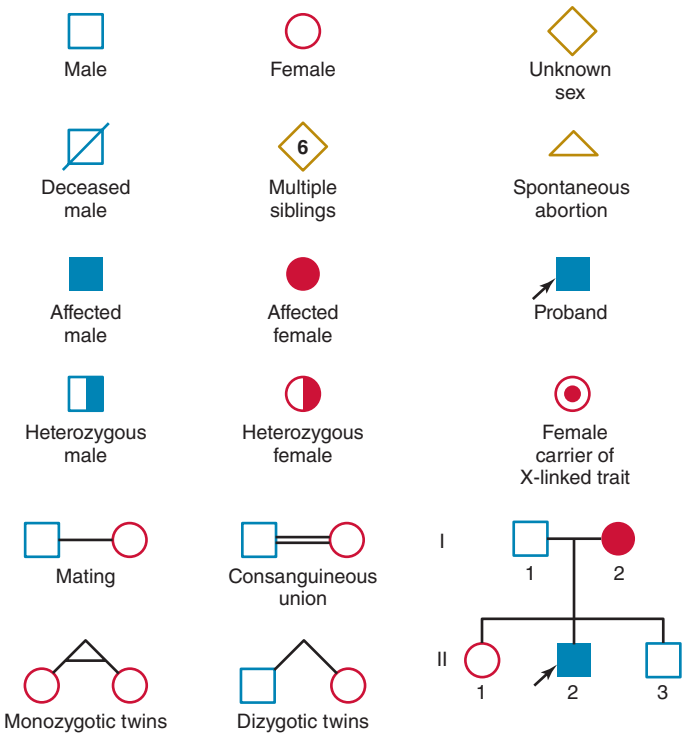


FIGURE 456-11 Standard pedigree symbols.

the principles of genetic transmission originally set forth in Gregor Mendel's classic work. The continuously updated OMIM catalogue lists several thousand of these disorders and provides information about the clinical phenotype, molecular basis, allelic variants, and pertinent animal models (Table 456-1). The mode of inheritance for a given phenotypic trait or disease is determined by pedigree analysis. All affected and unaffected individuals in the family are recorded in a pedigree using standard symbols **(Fig. 456-11)**. The principles of allelic segregation, and the transmission of alleles from parents to children, are illustrated in **Fig. 456-12**. One dominant (A) allele and one recessive (a) allele can display three Mendelian modes of inheritance: autosomal dominant, autosomal recessive, and X-linked. About 65% of human monogenic disorders are autosomal dominant, 25% are autosomal recessive, and 5% are X-linked. Genetic testing is now available for many of these disorders and plays an important role in clinical medicine **(Chap. 457)**.

AUTOSOMAL DOMINANT DISORDERS In autosomal dominant disorders, mutations in a single allele are sufficient to cause the disease. In contrast to recessive disorders, in which disease pathogenesis is relatively straightforward because there is a biallelic loss of gene function, dominant disorders can be caused by various disease mechanisms, many of which are unique to the function of the genetic pathway involved. Mechanistically, the mutation may confer constitutive activation (gain-of-function), exert a dominant negative effect, or result in loss-of-function and haploinsufficiency.

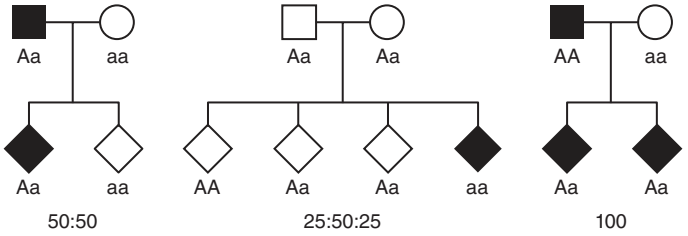


FIGURE 456-12 Segregation of alleles. Segregation of genotypes in the offspring of parents with one dominant (A) and one recessive (a) allele. The distribution of the parental alleles to their offspring depends on the combination present in the parents. Filled symbols = affected individuals.

In autosomal dominant disorders, individuals are affected in successive generations; the disease does not occur in the offspring of unaffected individuals. Males and females are affected with equal frequency because the defective gene resides on one of the 22 autosomes **(Fig. 456-13*A*)**. Autosomal dominant mutations alter one of the two alleles at a given locus. Because the alleles segregate randomly at meiosis, the probability that an offspring will be affected is 50%. Unless there is a new germline mutation, an affected individual has an affected parent. Children with a normal genotype do not transmit the disorder. Due to differences in penetrance or expressivity (see above), the clinical manifestations of autosomal dominant disorders may be variable. Because of these variations, it is sometimes challenging to determine the pattern of inheritance.

It should be recognized, however, that some individuals acquire a mutated gene from an unaffected parent. De novo germline mutations occur more frequently during later cell divisions in gametogenesis, which explains why siblings are rarely affected. As noted before, new germline mutations occur more frequently in fathers of advanced age. For example, the average age of fathers with new germline mutations that cause Marfan's syndrome is ~37 years, whereas fathers who transmit the disease by inheritance have an average age of ~30 years.

AUTOSOMAL RECESSIVE DISORDERS In recessive disorders, the mutated alleles result in a complete or partial loss of function. They frequently involve enzymes in metabolic pathways, receptors, or proteins in signaling cascades. In an autosomal recessive disease, the affected

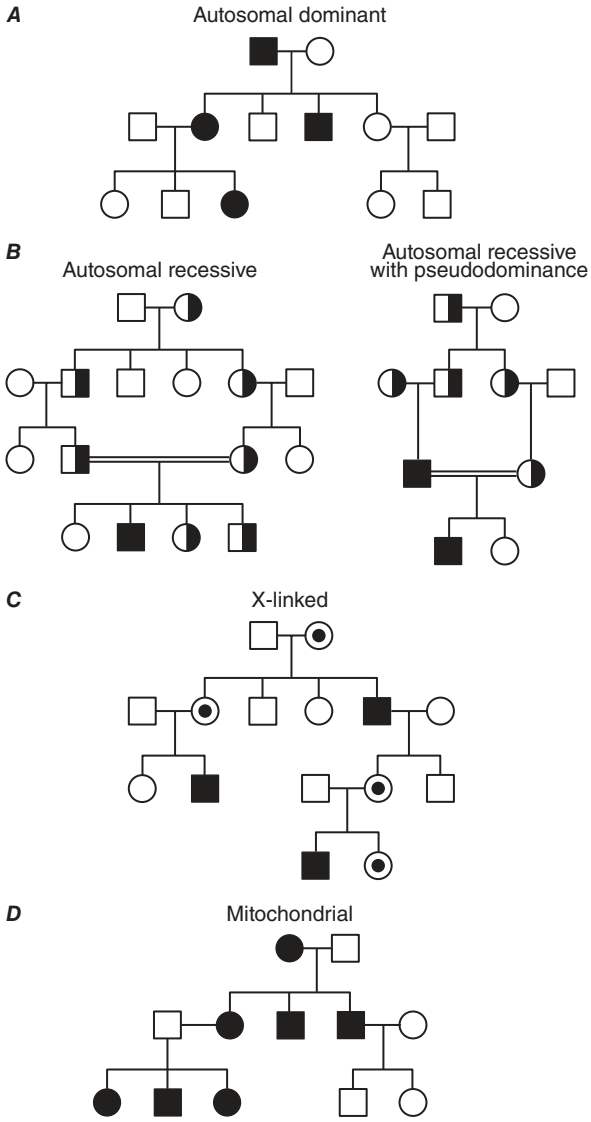


FIGURE 456-13 ***A.*** Dominant, ***B.*** recessive, ***C.*** X-linked, and ***D.*** mitochondrial (matrilinear) inheritance.

individual, who can be of either sex, is a homozygote or compound heterozygote for a single-gene defect. With a few important exceptions, autosomal recessive diseases are rare and often occur in the context of parental consanguinity. The relatively high frequency of certain recessive disorders such as sickle cell anemia, cystic fibrosis, and thalassemia, is partially explained by a selective biologic advantage for the heterozygous state (see below). Although heterozygous carriers of a defective allele are usually clinically normal, they may display subtle differences in phenotype that only become apparent with more precise testing or in the context of certain environmental influences. In sickle cell anemia, for example, heterozygotes are normally asymptomatic. However, in situations of dehydration or diminished oxygen pressure, sickle cell crises can also occur in heterozygotes **(Chap. 94)**.

In most instances, an affected individual is the offspring of heterozygous parents. In this situation, there is a 25% chance that the offspring will have a normal genotype, a 50% probability of a heterozygous state, and a 25% risk of homozygosity for the recessive alleles (**Figs.** 456-10, **456-13*B***). In the case of one unaffected heterozygous and one affected homozygous parent, the probability of disease increases to 50% for each child. In this instance, the pedigree analysis mimics an autosomal dominant mode of inheritance (*pseudodominance*). In contrast to autosomal dominant disorders, new mutations in recessive alleles are rarely manifest because they usually result in an asymptomatic carrier state.

X-LINKED DISORDERS Males have only one X chromosome; consequently, a daughter always inherits her father's X chromosome in addition to one of her mother's two X chromosomes. A son inherits the Y chromosome from his father and one maternal X chromosome. Thus, the characteristic features of X-linked inheritance are (1) the absence of father-to-son transmission, and (2) the fact that all daughters of an affected male are obligate carriers of the mutant allele **(Fig. 456-13*C*)**. The risk of developing disease due to a mutant X-chromosomal gene differs in the two sexes. Because males have only one X chromosome, they are hemizygous for the mutant allele; thus, they are more likely to develop the mutant phenotype, regardless of whether the mutation is dominant or recessive. A female may be either heterozygous or homozygous for the mutant allele, which may be dominant or recessive. The terms *X-linked dominant* or *X-linked recessive* are therefore only applicable to expression of the mutant phenotype in women. In addition, the expression of X-chromosomal genes is influenced by X chromosome inactivation.

Y-LINKED DISORDERS The Y chromosome has a relatively small number of genes. One such gene, the sex-region determining Y factor (*SRY*), which encodes the testis-determining factor (*TDF*), is crucial for normal male development. Normally there is infrequent exchange of sequences on the Y chromosome with the X chromosome. The *SRY* region is adjacent to the pseudoautosomal region, a chromosomal segment on the X and Y chromosomes with a high degree of homology. A crossing-over event occasionally involves the *SRY* region with the distal tip of the X chromosome during meiosis in the male. Translocations can result in XY females with the Y chromosome lacking the *SRY* gene or XX males harboring the *SRY* gene on one of the X chromosomes **(Chap. 383)**. Point mutations in the *SRY* gene may also result in individuals with an XY genotype and an incomplete female phenotype. Most of these mutations occur de novo. Men with oligospermia/azoospermia frequently have microdeletions on the long arm of the Y chromosome that involve one or more of the azoospermia factor (*AZF*) genes.

Exceptions to Simple Mendelian Inheritance Patterns •

MITOCHONDRIAL DISORDERS Mendelian inheritance refers to the transmission of genes encoded by DNA contained in the nuclear chromosomes. In addition, each mitochondrion contains several copies of a small circular chromosome **(Chap. 472)**. The mitochondrial DNA (mtDNA) is ~16.5 kb and encodes transfer and ribosomal RNAs and 13 core proteins that are components of the respiratory chain involved in oxidative phosphorylation and ATP generation. The mitochondrial genome does not recombine and is inherited through the maternal line because sperm does not contribute significant cytoplasmic components to the zygote. A noncoding region of the mitochondrial chromosome, referred to as D-loop, is highly polymorphic. This property, together with the absence of mtDNA recombination, makes it a valuable tool for studies tracing human migration and evolution, and it is also used for specific forensic applications.

Inherited mitochondrial disorders are transmitted in a matrilineal fashion; all children from an affected mother will inherit the disease, but it will not be transmitted from an affected father to his children **(Fig. 456-13*D*)**. Alterations in the mtDNA that involves enzymes required for oxidative phosphorylation lead to reduction of ATP supply, generation of free radicals, and induction of apoptosis. Several syndromic disorders arising from mutations in the mitochondrial genome are known in humans and they affect both protein-coding and tRNA genes. The broad clinical spectrum often involves (cardio) myopathies and encephalopathies because of the high dependence of these tissues on oxidative phosphorylation. The age of onset and the clinical course are highly variable because of the unusual mechanisms of mtDNA transmission, which replicates independently from nuclear DNA. During cell replication, the proportion of wild-type and mutant mitochondria can drift among different cells and tissues. The resulting heterogeneity in the proportion of mitochondria with and without a mutation is referred to as *heteroplasmia* and underlies the phenotypic variability that is characteristic of mitochondrial diseases.

Acquired somatic mutations in mitochondria are thought to be involved in several age-dependent degenerative disorders affecting predominantly muscle and the peripheral and central nervous system (e.g., Alzheimer's and Parkinson's diseases). Establishing that an mtDNA alteration is causal for a clinical phenotype is challenging because of the high degree of polymorphism in mtDNA and the phenotypic variability characteristic of these disorders. Certain pharmacologic treatments may have an impact on mitochondria and/or their function. For example, treatment with the antiretroviral compound azidothymidine (AZT) causes an acquired mitochondrial myopathy through depletion of muscular mtDNA.

MOSAICISM Mosaicism refers to the presence of two or more genetically distinct cell lines in the tissues of an individual. It results from a mutation that occurs during embryonic, fetal, or extrauterine development. The developmental stage at which the mutation arises will determine whether germ cells and/or somatic cells are involved. Chromosomal mosaicism results from nondisjunction at an early embryonic mitotic division, leading to the persistence of more than one cell line, as exemplified by some patients with Turner's syndrome **(Chap. 383)**. Somatic mosaicism is characterized by a patchy distribution of genetically altered somatic cells. The McCune-Albright syndrome, for example, is caused by activating mutations in the stimulatory G protein α (Gsα) that occur early in development **(Chap. 403)**. The clinical phenotype varies depending on the tissue distribution of the mutation; manifestations include ovarian cysts that secrete sex steroids and cause precocious puberty, polyostotic fibrous dysplasia, café-au-lait skin pigmentation, GH–secreting pituitary adenomas, and hypersecreting autonomous thyroid nodules.

X-INACTIVATION, IMPRINTING, AND UNIPARENTAL DISOMY According to traditional Mendelian principles, the parental origin of a mutant gene is irrelevant for the expression of the phenotype. There are, however, important exceptions to this rule. *X-inactivation* prevents the expression of most genes on one of the two X chromosomes in every cell of a female. Gene inactivation through genomic imprinting occurs on selected chromosomal regions of autosomes and leads to inheritable preferential expression of one of the parental alleles. It is of pathophysiologic importance in disorders where the transmission of disease is dependent on the sex of the transmitting parent and, thus, plays an important role in the expression of certain genetic disorders. Two classic examples are the Prader-Willi syndrome and Angelman's syndrome. Prader-Willi syndrome is characterized by diminished fetal activity, obesity, hypotonia, mental retardation, short stature, and hypogonadotropic hypogonadism. Deletions of the paternal copy of the Prader-Willi locus located on the short arm of chromosome 15 result in a contiguous gene syndrome involving missing paternal copies of the *necdin* and *SNRPN* genes, among others. In contrast, patients with

Angelman's syndrome, characterized by mental retardation, seizures, ataxia, and hypotonia, have deletions involving the maternal copy of this region on chromosome 15. These two syndromes may also result from *uniparental disomy*. In this case, the syndromes are not caused by deletions on chromosome 15 but by the inheritance of either two maternal chromosomes (Prader-Willi syndrome) or two paternal chromosomes (Angelman's syndrome). Lastly, the two distinct phenotypes can also be caused by an imprinting defect that impairs the resetting of the imprint during zygote development (defect in the father leads to Prader-Willi syndrome; defect in the mother leads to Angelman's syndrome).

Imprinting and the related phenomenon of allelic exclusion may be more common than currently documented, because it is difficult to examine levels of mRNA expression from the maternal and paternal alleles in specific tissues or in individual cells. Genomic imprinting, or uniparental disomy, is involved in the pathogenesis of several other disorders and malignancies. For example, hydatidiform moles contain a normal number of diploid chromosomes, but they are all of paternal origin. The opposite situation occurs in ovarian teratomata, with 46 chromosomes of maternal origin. Expression of the imprinted gene for insulin-like growth factor II (IGF-II) is involved in the pathogenesis of the cancer-predisposing Beckwith-Wiedemann syndrome (BWS). These children show somatic overgrowth with organomegalies and hemihypertrophy, and they have an increased risk of embryonal malignancies such as Wilms' tumor. Normally, only the paternally derived copy of the *IGF-II* gene is active and the maternal copy is inactive. Imprinting of the *IGF-II* gene is regulated by *H19*, which encodes an RNA transcript that is not translated into protein. Disruption or lack of *H19* methylation leads to a relaxation of *IGF-II* imprinting and expression of both alleles. Alterations of the epigenome through gain and loss of DNA methylation, as well as altered histone modifications, play an important role in the pathogenesis of malignancies.

SOMATIC MUTATIONS Cancer can be considered a genetic disease at the cellular level **(Chap. 67)**. Cancers are monoclonal in origin, indicating that they have arisen from a single precursor cell with one or several mutations in genes controlling growth (proliferation or apoptosis) and/or differentiation. These acquired somatic mutations are restricted to the tumor and its metastases and are not found in the surrounding normal tissue. The molecular alterations include dominant gain-of-function mutations in oncogenes, recessive loss-of-function mutations in tumor-suppressor genes and DNA repair genes, gene amplification, and chromosome rearrangements. Rarely, a single mutation in certain genes may be sufficient to transform a normal cell into a malignant cell. In most cancers, however, the development of a malignant phenotype requires several genetic alterations for the gradual progression from a normal cell to a cancerous cell, a phenomenon termed *multistep carcinogenesis*. Genome-wide analyses of cancers using deep sequencing often reveal somatic rearrangements resulting in fusion genes and mutations in multiple genes **(Fig. 456-14)**. Comprehensive sequence analyses provide further insight into genetic heterogeneity within malignancies; these include intratumoral heterogeneity among the cells of the primary tumor, intermetastatic and intrametastatic heterogeneity, and interpatient differences. These analyses further support the notion of cancer as an ongoing process of clonal evolution, in which successive rounds of clonal selection within the primary tumor and metastatic lesions result in diverse genetic and epigenetic alterations that require targeted (personalized) therapies (precision medicine). The heterogeneity of mutations within a tumor can also lead to resistance to target therapies because cells with mutations that are resistant to the therapy, even if they are a minor part of the tumor population, will be selected as the more sensitive cells are killed. Most human tumors express telomerase, an enzyme formed of a protein and an RNA component, which adds telomere repeats at the ends of chromosomes during replication. This mechanism impedes shortening of the telomeres, which is associated with senescence in normal cells and is associated with enhanced replicative capacity in cancer cells. Telomerase inhibitors provide a strategy for treating advanced human cancers.

In many cancer syndromes, there is frequently an inherited *predisposition* to tumor formation. In these instances, a germline mutation is inherited in an autosomal dominant fashion inactivating one allele of an autosomal tumor-suppressor gene. If the second allele is inactivated by a somatic mutation or by epigenetic silencing in a given cell, this will lead to neoplastic growth (Knudson two-hit model). Thus, the defective allele in the germline is transmitted in a dominant mode, although tumorigenesis results from a biallelic loss of the tumor-suppressor gene in an affected tissue. The classic example to illustrate this phenomenon is retinoblastoma, which can occur as a sporadic or hereditary tumor. In sporadic retinoblastoma, both copies of the retinoblastoma (*RB*) gene are inactivated through two somatic events. In hereditary retinoblastoma, one mutated or deleted *RB* allele is inherited in an autosomal

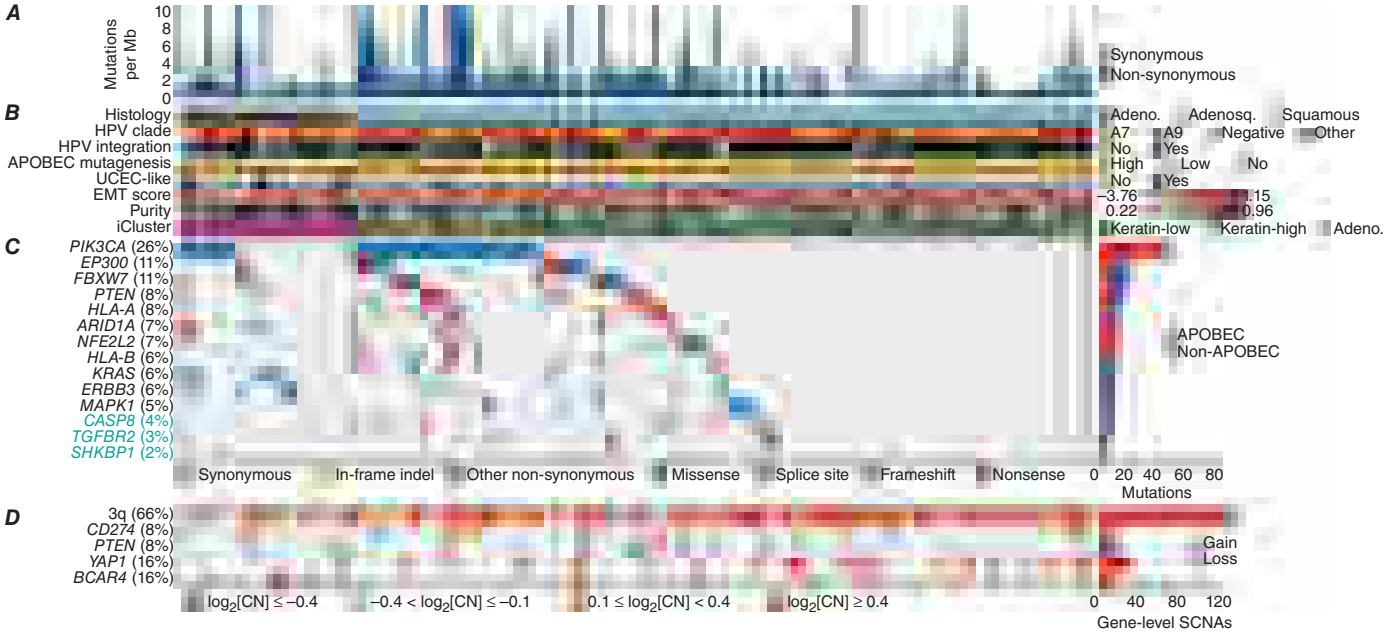


FIGURE 456-14 Somatic alterations in cervical cancer. ***A.*** Cervical carcinoma samples ordered by histology and mutation frequency; ***B.*** clinical and molecular platform features; ***C.*** Significantly Mutated Genes (SMGs); and ***D.*** select somatic copy number alterations. SMGs are ordered by the overall mutation frequency and color-coded by mutation type. Adeno, adenocarcinomas; Adenosq, adenosquamous cancers; CN, copy number; SCNAs, somatic copy number alterations; Squamous, squamous cell carcinomas. *(From The Cancer Genome Atlas Research Network. Integrated genomic and molecular characterization of cervical cancer. Nature 543:378–384, 2017. Permission for reproduction granted through a Creative Commons CC-BY (CC BY 4.0) license.)*

TABLE 456-5 Selected Trinucleotide Repeat Disorders

DISEASE	LOCUS	REPEAT	TRIPLET LENGTH (NORMAL/DISEASE)	INHERITANCE	GENE PRODUCT
X-chromosomal spinobulbar muscular atrophy (SBMA)	Xq11-q12	CAG	11–34/40–62	XR	Androgen receptor
Fragile X syndrome (FRAXA)	Xq27.3	CGG	6–50/200–300	XR	FMR-1 protein
Fragile X syndrome (FRAXE)	Xq28	GCC	6–25/>200	XR	FMR-2 protein
Dystrophia myotonica (DM)	19q13.2-q13.3	CTG	5–30/200–1000	AD, variable penetrance	Myotonin protein kinase
Huntington's disease (HD)	4p16.3	CAG	6–34/37–180	AD	Huntingtin
Spinocerebellar ataxia type 1 (SCA1)	6p21.3-21.2	CAG	6–39/40–88	AD	Ataxin 1
Spinocerebellar ataxia type 2 (SCA2)	12q24.1	CAG	15–31/34–400	AD	Ataxin 2
Spinocerebellar ataxia type 3 (SCA3); Machado-Joseph disease (MD)	14q21	CAG	13–36/55–86	AD	Ataxin 3
Spinocerebellar ataxia type 6 (SCA6, CACNAIA)	19p13.1-13.2	CAG	4–16/20–33	AD	Alpha 1A voltage-dependent L-type calcium channel
Spinocerebellar ataxia type 7 (SCA7)	3p21.1-p12	CAG	4–19/37 to >300	AD	Ataxin 7
Spinocerebellar ataxia type 12 (SCA12)	5q31	CAG	6–26/66–78	AD	Protein phosphatase 2A
Dentatorubral pallidoluysian atrophy (DRPLA)	12p	CAG	7–23/49–75	AD	Atrophin 1
Friedreich's ataxia (FRDA1)	9q13-21	GAA	7–22/200–900	AR	Frataxin

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; XR, X-linked recessive.

dominant manner and the second allele is inactivated by a subsequent somatic mutation. This two-hit model applies to other inherited cancer syndromes such as MEN 1 **(Chap. 381)** and neurofibromatosis type 2 **(Chap. 86)**. In contrast, in the autosomal dominant MEN2 syndrome, the predisposition for tumor formation in various organs is caused by a gain-of-function mutation in a single allele of the *RET* gene **(Chap. 381)**.

NUCLEOTIDE REPEAT EXPANSION DISORDERS Several diseases are associated with an increase in the number of nucleotide repeats above a certain threshold **(Table 456-5)**. The repeats are sometimes located within the coding region of the genes, as in Huntington's disease or the X-linked form of spinal and bulbar muscular atrophy (SBMA; Kennedy's syndrome). In other instances, the repeats probably alter gene regulatory sequences. If an expansion is present, the DNA fragment is unstable and tends to expand further during cell division. The length of the nucleotide repeat often correlates with the severity of the disease. When repeat length increases from one generation to the next, disease manifestations may worsen or be observed at an earlier age; this phenomenon is referred to as *anticipation*. In Huntington's disease, for example, there is a correlation between age of onset and length of the triplet codon expansion **(Chap. 417)**. Anticipation has also been documented in other diseases caused by dynamic mutations in trinucleotide repeats (Table 456-5). The repeat number may also vary in a tissue-specific manner. In myotonic dystrophy, the CTG repeat may be tenfold greater in muscle tissue than in lymphocytes **(Chap. 441)**.

Complex Genetic Disorders

The expression of many common diseases such as cardiovascular disease, hypertension, diabetes, asthma, psychiatric disorders, and certain cancers is determined by a combination of genetic background, environmental factors, and lifestyle. A trait is called *polygenic* if multiple genes contribute to the phenotype or *multifactorial* if multiple genes are assumed to interact with environmental factors. Genetic models for these complex traits need to account for genetic heterogeneity and interactions with other genes and the environment. Complex genetic traits may be influenced by modifier genes that are not linked to the main gene involved in the pathogenesis of the trait. This type of gene-gene interaction, or *epistasis*, plays an important role in polygenic traits that require the simultaneous presence of variations in multiple genes to result in a pathologic phenotype.

Type 2 diabetes mellitus provides a paradigm for considering a multifactorial disorder, because genetic, nutritional, and lifestyle factors are intimately interrelated in disease pathogenesis **(Table 456-6) (Chap. 396)**. The identification of genetic variations and environmental factors that either predispose to or protect against disease is essential for predicting disease risk, designing preventive strategies, and developing novel therapeutic approaches. The study of rare monogenic diseases may provide insight into some of the genetic and molecular mechanisms important in the pathogenesis of complex diseases. For example, the identification of the genes causing monogenic forms of permanent neonatal diabetes mellitus or maturity-onset diabetes defined them as *candidate genes* in the pathogenesis of diabetes mellitus type 2 (Tables 456-2 and 456-6) **(Fig. 456-15)**. Genome scans have identified numerous genes and loci that may be associated with susceptibility to development of diabetes mellitus in certain populations **(Fig. 456-16)**. Efforts to identify susceptibility genes require very large sample sizes, and positive results may depend on ethnicity, ascertainment criteria, and statistical analysis. Association studies analyzing the potential influence of (biologically functional) SNPs and SNP haplotypes on a particular phenotype have revealed new insights into the genes involved in the pathogenesis of these common disorders. Large variants ([micro]deletions, duplications, and inversions) present in the human population also contribute to the pathogenesis of complex disorders, but their contributions remain poorly understood.

Linkage and Association Studies

There are two primary strategies for mapping genes that cause or increase susceptibility to human disease: (1) classic linkage can be performed based on a known genetic model or, when the model is unknown, by studying pairs of affected relatives; or (2) disease genes can be mapped using allelic association studies **(Table 456-7)**.

GENETIC LINKAGE *Genetic linkage* refers to the fact that genes are physically connected, or linked, to one another along the chromosomes. Two fundamental principles are essential for understanding the concept of linkage: (1) when two genes are close together on a chromosome, they are usually transmitted together, unless a recombination event separates them (Figs. 456-6); and (2) the odds of a crossover, or recombination event, between two linked genes is proportional to the distance that separates them. Thus, genes that are farther apart are more likely to undergo a recombination event than genes that are very close together. The detection of chromosomal loci that segregate with a disease by linkage can be used to identify the gene responsible for the disease (*positional cloning*) and to predict the odds of disease gene transmission in genetic counseling.

Polymorphic variants are essential for linkage studies because they provide a means to distinguish the maternal and paternal chromosomes in an individual. On average, 1 out of every 1000 bp varies from one person to the next. Although this degree of variation seems low (99.9% identical), it means that >3 million sequence differences exist

TABLE 456-6 Genes and Loci Involved in Mono- and Polygenic Forms of Diabetes

DISORDER	GENES OR SUSCEPTIBILITY LOCUS	CHROMOSOMAL LOCATION	OTHER FACTORS
Monogenic permanent neonatal diabetes mellitus	*KCNJ11* (inwardly rectifying potassium channelKir6.2)	11p15.1	AD
	GCK (glucokinase)	7p15-p13	AR
	INS (insulin)	11p15.5	AR, hyperproinsulinemia
	ABCC8 (ATP-binding cassette, subfamily c, member 8; sulfonylurea receptor)	11p15.1	AD or AR
	GLIS3 (GLIS family zinc finger protein 3)	9p24.2	AR, diabetes, congenital hypothyroidism
Maturity-onset diabetes of the young (MODY): Monogenic forms of diabetes mellitus			
MODY 1	*HNF4α* (hepatocyte nuclear factor 4α)	20q12-q13.1	AD inheritance
MODY 2	*GCK* (glucokinase)	7p15-p13	
MODY 3	*HNF1α* (hepatocyte nuclear factor 1α)	12q24.2	
MODY 4	*IPF1* (insulin receptor substrate)	13q12.1	
MODY 5 (renal cysts, diabetes)	*HNF1β* (hepatocyte nuclear factor 1β)	17cen-q21.3	
MODY 6	*NeuroD1* (neurogenic differentiation factor 1)	2q32	
MODY 7	*KLF1* (Kruppel-like factor 1)	19p13.13-p13.12	
MODY 8	*CEL* (carboxyl ester lipase)	9q34.3	
MODY 9	*PAX4* (paired box transcription factor 4)	7q32	
MODY 10	*INS* (insulin)	11p15.5	
MODY 11	*BLK* (B-lymphocyte-specific tyrosine kinase)	8p23-p22	
Diabetes mellitus type 2; loci and genes linked and/or associated with susceptibility for diabetes mellitus type 2	Genes and loci identified by linkage/association studies		Heavily influenced by diet, energy expenditure, obesity
	PPARG, KCNJ11/ABCC8, TCF7L2, HNF1B, WFS1, SLC30A8, FTO, HHEX, IGF2BP2, CDKN2A/B, CDKAL1, TSPAN8, ADAMTs9, CDC123/CAMK1D, JAZF1, NOTCH2, THADA, KCNQ1, DUSP8, MTNR1B, IRS1, SPRY2, SRR, ZFAND6, GCK, KLF14, TP53INP1, PROX1, PRC1, BCL11A, ZBED3, RBMS1, HNF1A, DGKB/TMEM195, CCND2, C2CD4A/C2CD4B, PTPRD, ARAP1/CENTD2, HMGA2, TLE4/CHCHD9, ADCY5, UBE2E2, DUSP9, GCKR, COBLL1/GRB14, HMG20A, VPS26A, ST6GAL1, AP3S2, HNF4A, BCL2, LAMA1, GIPR, MC4R, TLE1, KCNK16, ANK1, KLHDC5, ZMIZ1, PSMD6 , FITM2/R3HDML/HNF4A, CILP2, ANKRD55, GLIS3, PEPD, GCC1/PAX4, ZFAND3, MAEA, BCAR1, RBM43/RND3 , MACF1, RASGRP1, GRK5, TMEM163, SGCG, LPP, FAF1, TMEM154, MPHOSPH9, ARL15, POU5F1/TCF19, SSR1/RREB1, HLA-B, INS-IGF2, GPSM1, LEP, SLC16A13, PAM/PPIP5K2, SLC16A11, CCDC63, C12orf51, CCND2, HNF1A, TBC1D4, CCDC85A, INAFM2, ASB3, FAM60A, ATP8B2, MIR4686, MTMR3, DMRTA1, SLC35D3, GLP2R, GIP, MAP3K11, PLEKHA1, HSD17B12, NRXN3, CMIP, ZZEF1, MNX1, ABO, ACSL1, HLA-DQA1		

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; MODY, maturity onset diabetes of the young.

between any two unrelated individuals and the probability that the sequence at such loci will differ on the two homologous chromosomes is high (often >70–90%). These sequence variations include variable number of tandem repeats (VNTRs), short tandem repeats (STRs), and SNPs. Most STRs, also called *polymorphic microsatellite markers*, consist of di-, tri-, or tetranucleotide repeats that can be characterized readily using the polymerase chain reaction (PCR). Characterization of SNPs, using DNA chips or beads, permits comprehensive analyses of genetic variation, linkage, and association studies. Although these sequence variations often have no apparent functional consequences, they provide much of the basis for variation in genetic traits.

In order to identify a chromosomal locus that segregates with a disease, it is necessary to characterize polymorphic DNA markers from affected and unaffected individuals of one or several pedigrees. One can then assess whether certain marker alleles cosegregate with the disease. Markers that are closest to the disease gene are less likely to undergo recombination events and therefore receive a higher linkage score. Linkage is expressed as a lod (logarithm of odds) score—the ratio of the probability that the disease and marker loci are linked rather than unlinked. Lod scores of +3 (1000:1) are generally accepted as supporting linkage, whereas a score of –2 is consistent with the absence of linkage.

ALLELIC ASSOCIATION, LINKAGE DISEQUILIBRIUM, AND HAPLOTYPES

Allelic association refers to a situation in which the frequency of an allele is significantly increased or decreased in individuals affected by a particular disease in comparison to controls. Linkage and association differ in several aspects. Genetic linkage is demonstrable in families or sibships. Association studies, on the other hand, compare a population of affected individuals with a control population. Association studies can be performed as case-control studies that include unrelated affected individuals and matched controls or as family-based studies that compare the frequencies of alleles transmitted or not transmitted to affected children.

Allelic association studies are particularly useful for identifying susceptibility genes in complex diseases. When alleles at two loci occur more frequently *in combination* than would be predicted (based on known allele frequencies and recombination fractions), they are said to be in *linkage disequilibrium*. Evidence for linkage disequilibrium can be helpful in mapping disease genes because it suggests that the two loci are tightly linked.

Detecting the genetic factors contributing to the pathogenesis of common complex disorders is challenging. In many instances, these are low-penetrance alleles (e.g., variations that individually have a subtle effect on disease development, and they can only be identified by unbiased GWAS) (Catalog of published Genome-Wide Association

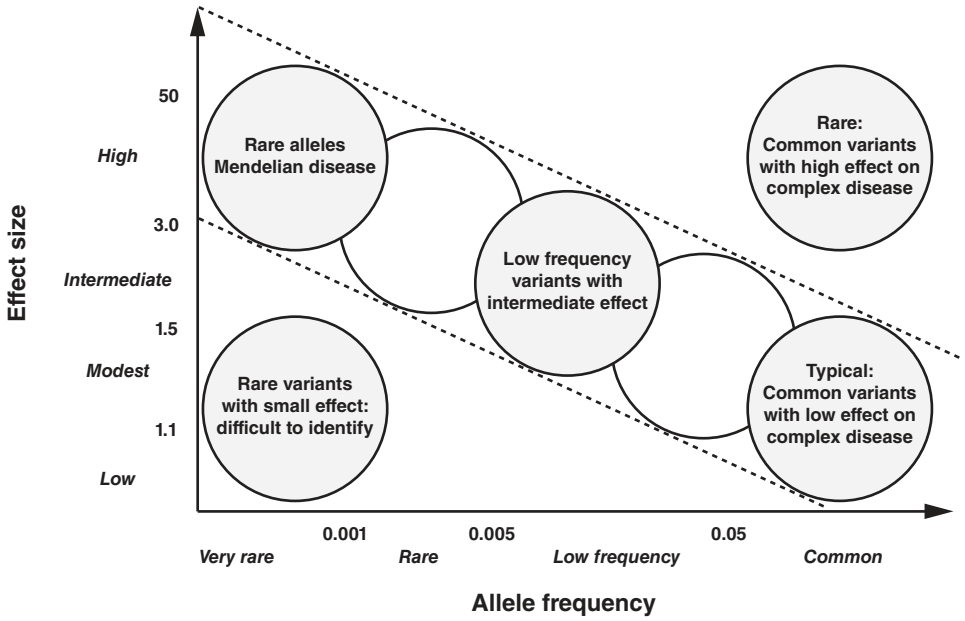


FIGURE 456-15 Relationship between allele frequency and effect size in monogenic and polygenic disorders. In classic Mendelian disorders, the allele frequency is typically low but has a high impact (single gene disorder). This contrasts with polygenic disorders that require the combination of multiple low impact alleles that are frequently quite common in the general population.

Studies; Table 456-1) (Fig. 456-16). Most variants occur in noncoding or regulatory sequences but do not alter protein structure. The analysis of complex disorders is further complicated by ethnic differences in disease prevalence, differences in allele frequencies in known susceptibility genes among different populations, locus and allelic heterogeneity, gene-gene and gene-environment interactions, and the possibility of phenocopies. Catalogues of human variation and genotype data (HapMap, International Genome Sample Resource) are greatly facilitating GWAS for the characterization of complex disorders. Adjacent SNPs are inherited together as blocks, and these blocks can be identified by genotyping selected marker SNPs, so-called *Tag SNPs*, thereby reducing cost and workload (Fig. 456-4). The availability of this information permits the characterization of a limited number of SNPs to identify the set of haplotypes present in an individual (e.g., in cases and controls). This, in turn, permits performing GWAS by searching for associations of certain haplotypes with a disease phenotype of interest, an essential step for unraveling the genetic factors contributing to complex disorders.

POPULATION GENETICS In population genetics, the focus changes from alterations in an individual's genome to the distribution pattern of different genotypes in the population. In a case where there are only two alleles, A and a, the frequency of the genotypes will be $p^2 + 2pq + q^2 = 1$, with p^2 corresponding to the frequency of AA, $2pq$ to the frequency of Aa, and q^2 to aa. When the frequency of an allele is known, the frequency of the genotype can be calculated. Alternatively, one can determine an allele frequency if the genotype frequency has been determined.

Allele frequencies vary among ethnic groups and geographic regions. For example, heterozygous mutations in the *CFTR* gene are relatively common in populations of European origin but are rare in the African population. Allele frequencies may vary because certain allelic variants confer a selective advantage. For example, heterozygotes for the sickle cell mutation, which is particularly common in West Africa, are more resistant to malarial infection because the erythrocytes of heterozygotes provide a less favorable environment for *Plasmodium* parasites. Although homozygosity for the sickle cell mutation is associated with severe anemia and sickle crises, heterozygotes have a higher probability of survival because of the reduced morbidity and mortality from malaria; this phenomenon has led to an increased frequency of the mutant allele. Recessive conditions are more prevalent in geographically isolated populations because of the more restricted gene pool.

APPROACH TO THE PATIENT

Inherited Disorders

For the practicing clinician, the family history remains an essential step in recognizing the possibility of a hereditary predisposition to disease. When taking the history, it is useful to draw a detailed pedigree of the first-degree relatives (e.g., parents, siblings, and children), because they share 50% of genes with the patient. Standard symbols for pedigrees are depicted in Fig. 456-11. The family history should include information about ethnic background, age, health status, and deaths, including infants. Next, the physician should explore whether there is a family history of the same or related illnesses to the current problem. An inquiry focused on commonly occurring disorders such as cancers, heart disease, and diabetes mellitus should follow. Because of the possibility of age-dependent expressivity and penetrance, the family history will need intermittent updating. If the findings suggest a genetic disorder, the clinician should assess whether some of the patient's relatives may be at risk of carrying or transmitting the disease. In this circumstance, it is useful to confirm and extend the pedigree based on input from several family members. This information may form the basis for genetic counseling, carrier detection, early intervention, and disease prevention in relatives of the index patient (Chap. 457).

In instances where a diagnosis at the molecular level may be relevant, it is important to identify an appropriate laboratory that can perform the appropriate test. Genetic testing is available for a large number of monogenic disorders through commercial laboratories. For uncommon disorders, the test may only be performed in a specialized research laboratory. Approved laboratories offering testing for inherited disorders can be identified in continuously updated online resources (e.g., GeneTests; Table 456-1). If genetic testing is considered, the patient and the family should be counseled about the potential implications of positive results, including psychological distress and the possibility of discrimination. The patient or caretakers should be informed about the meaning of a negative result, technical limitations, and the possibility of false-negative and inconclusive results. For these reasons, genetic testing should only be performed after obtaining *informed consent*. Published ethical guidelines address the specific aspects that should be considered when testing children and adolescents. Genetic testing should usually be limited

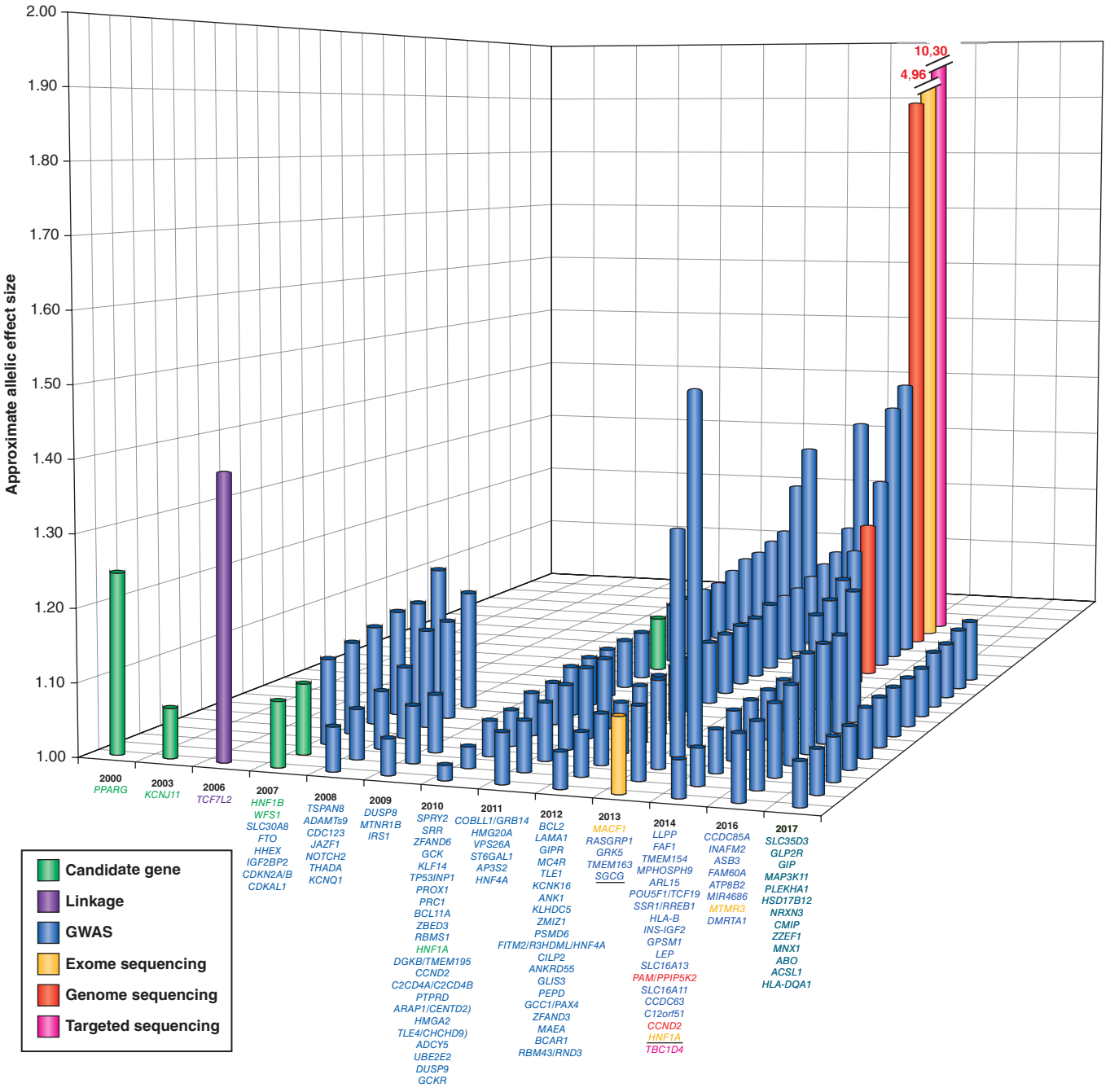


FIGURE 456-16 Loci and genes associated with diabetes mellitus type 2. Loci are listed by year of identification, and the color indicates discovery method. Gene names indicate the locus and does not necessarily imply that the gene itself is causally involved. Approximate allelic effect sizes were either derived from the discovery cohort or, if available, from the DIAGRAM (Diabetes Genetics Replication and Meta-analysis consortium) European ancestry meta-analysis and the Asian ancestry meta-analysis. Gene names that are underlined denote identification in population isolates. *(The data have been graciously provided by Dr. Miriam Udler and Dr. Jose Florez, Harvard Medical School, Boston.)*

to situations in which the results may have an impact on medical management.

IDENTIFYING THE DISEASE-CAUSING GENE

Precision medicine aims to enhance the quality of medical care through the use of genotypic analysis (DNA testing) to identify genetic predisposition to disease, to select more specific pharmacotherapy, and to design individualized medical care based on genotype. Genotype can be deduced by analysis of protein (e.g., hemoglobin, apoprotein E), mRNA, or DNA. However, technologic advances have made DNA analysis particularly useful because it can be readily applied.

DNA testing is performed by mutational analysis or linkage studies in individuals at risk for a genetic disorder known to be present in a family. Mass screening programs require tests of high sensitivity and specificity to be cost-effective. Prerequisites for the success of genetic screening programs include the following: that the disorder is potentially serious; that it can be influenced at a presymptomatic stage by changes in behavior, diet, and/or pharmaceutical manipulations; and that the screening does not result in any harm or discrimination. Screening in Jewish populations for the autosomal recessive neurodegenerative storage disease Tay-Sachs has reduced the number of affected individuals. In contrast, screening for sickle cell trait/disease in African Americans has led to unanticipated problems of discrimination by health insurers and employers. Mass screening programs harbor additional potential problems. For example, screening for the most common genetic alteration in cystic fibrosis, the ΔF508 mutation with a frequency of ~70% in northern Europe, is feasible and seems to be effective.

TABLE 456-7 Genetic Approaches for Identifying Disease Genes

METHOD	INDICATIONS AND ADVANTAGES	LIMITATIONS
Linkage Studies		
Classical linkage analysis (parametric methods)	Analysis of monogenic traits Suitable for genome scan Control population not required Useful for multifactorial disorders in isolated populations	Difficult to collect large informative pedigrees Difficult to obtain sufficient statistical power for complex traits
Allele-sharing methods (nonparametric methods)	Suitable for identification of susceptibility genes in polygenic and multifactorial disorders	Difficult to collect sufficient number of subjects
Affected sib and relative pair analyses	Suitable for genome scan	Difficult to obtain sufficient statistical power for complex traits
Sib pair analysis	Control population not required if allele frequencies are known Statistical power can be increased by including parents and relatives	Reduced power compared to classical linkage, but not sensitive to specification of genetic mode
Association Studies		
Case-control studies	Suitable for identification of susceptibility genes in polygenic and multifactorial disorders	Requires large sample size and matched control population
Linkage disequilibrium	Suitable for testing specific allelic variants of known candidate loci	False-positive results in the absence of suitable control population
Transmission disequilibrium test (TDT)	Facilitated by comprehensive catalogs of genotypes and variants	Candidate gene approach does not permit detection of novel genes and pathways
Whole-genome association studies	Does not necessarily need relatives	Susceptibility genes can vary among different populations
Next-Generations Sequencing Technologies		
Whole exome or genome sequencing	Unbiased approach, analysis can be performed without reference sequences from parents or siblings	Requires appropriate bioinformatics, may have low sensitivity if CNV analysis is not included, detects numerous VUS, can lead to the detection of unrelated deleterious alleles
Targeted sequencing of gene panels	Captures multiple candidate genes and loci with hybridization techniques followed by deep sequencing	Permits analyses of multiple candidate genes in parallel. Facilitates molecular characterization of disorders with locus heterogeneity.

Abbreviations: CNV, copy number variation; GWAS, genome-wide association study; VUS, variants of unknown significance.

One has to keep in mind, however, that there is pronounced allelic heterogeneity and that the disease can be caused by about 2000 other mutations. The search for these less common mutations would substantially increase costs, but not the effectiveness of the screening program as a whole. Next-generation genome sequencing permits comprehensive and cost-effective mutational analyses after selective enrichment of candidate genes. For example, tests that sequence all the common genes causing hereditary deafness or hereditary pheochromocytomas are commercially available. Occupational screening programs aim to detect individuals with increased risk for certain professional activities (e.g., α_1 antitrypsin deficiency and smoke or dust exposure). Integrating genomic data into electronic medical records is evolving and can provide significant decision support at the point of care, for example, by providing the clinician with genomic data and decision algorithms for the prescription of drugs that are subject to pharmacogenetic influences.

Mutational Analyses DNA sequence analysis is now widely used as a diagnostic tool and has significantly enhanced diagnostic accuracy. It is used for determining carrier status and for prenatal testing in monogenic disorders. Numerous techniques, discussed in previous versions of this chapter, are available for the detection of mutations. In a very broad sense, one can distinguish between techniques that allow for screening of known mutations (screening mode) or techniques that definitively characterize mutations. Analyses of large alterations in the genome are possible using classic methods such as karyotype analysis, cytogenetics, fluorescent in situ hybridization (FISH), as well as more sensitive array- or bead-based techniques that search for multiple single exon deletions or duplications.

More discrete sequence alterations rely heavily on the use of PCR, which allows rapid gene amplification and analysis. Moreover, PCR makes it possible to perform genetic testing and mutational analysis with small amounts of DNA extracted from leukocytes or even from single cells, buccal cells, or hair roots. DNA sequencing can be performed directly on PCR products or on fragments cloned into plasmid vectors amplified in bacterial host cells. Sequencing of the whole genome, exome, selected chromosomes, or sequencing of numerous candidate genes in a single run, is now possible with next-generation sequencing platforms and has entered the clinical realm. Genomic tests are also widely used for the detection of pathogens and for the identification of viral or bacterial sequence variations. The integration of genomic tests into clinical medicine is, however, associated with a number of ongoing challenges related to costs, variable sensitivities of the tests, bioinformatics analyses, storage and sharing of data, and the difficulty of interpreting all genetic variants identified with comprehensive testing. The discovery of incidental (or secondary) findings that are unrelated to the indication for the sequencing analysis but indicators of other disorders of potential relevance for patient care can pose a difficult ethical dilemma. It can lead to the detection of undiagnosed medically actionable genetic conditions, but can also reveal deleterious mutations that cannot be influenced, as numerous sequence variants are of unknown significance.

A general algorithm for the approach to mutational analysis is outlined in Fig. 456-17. The importance of a detailed clinical phenotype cannot be overemphasized. This is the step where one should also consider the possibility of genetic heterogeneity and phenocopies. If obvious candidate genes are suggested by the phenotype, they can be analyzed directly. After identification of a mutation, it is essential to demonstrate that it segregates with the phenotype. The functional characterization of novel mutations is labor intensive and may require analyses in vitro or in transgenic models in order to document the relevance of the genetic alteration.

Prenatal diagnosis of numerous genetic diseases in instances with a high risk for certain disorders is now possible by direct DNA analysis. *Amniocentesis* involves the removal of a small amount of amniotic fluid, usually at 16 weeks of gestation. Cells can be collected and submitted for karyotype analyses, FISH, and mutational analysis of selected genes Table 456-4). The main indications for amniocentesis include advanced maternal age (>35 years), presence of an abnormality of the fetus on ultrasound examination, an abnormal serum "quad" test (α-fetoprotein, β human chorionic gonadotropin, inhibin-A, and unconjugated estriol), a family history of chromosomal abnormalities, or a Mendelian disorder amenable to genetic testing.

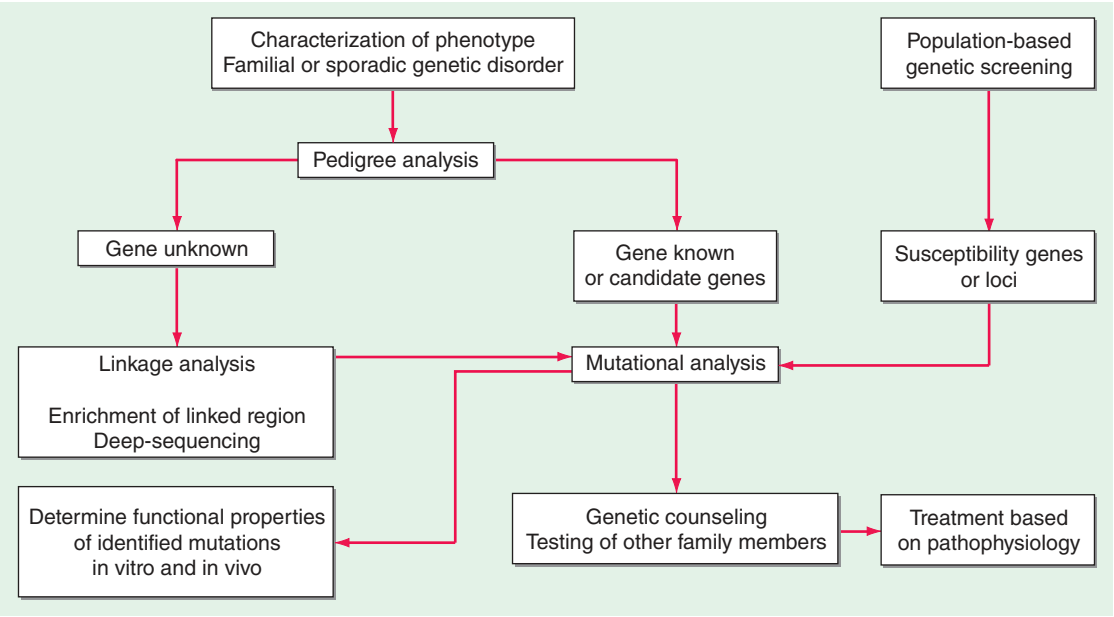


FIGURE 456-17 Approach to genetic disease.

Prenatal diagnosis can also be performed by *chorionic villus sampling* (CVS), in which a small amount of the chorion is removed by a transcervical or transabdominal biopsy. Chromosomes and DNA obtained from these cells can be submitted for cytogenetic and mutational analyses. CVS can be performed earlier in gestation (weeks 9–12) than amniocentesis, an aspect that may be of relevance when termination of pregnancy is a consideration. Later in pregnancy, beginning at about 18 weeks of gestation, percutaneous umbilical blood sampling (PUBS) permits collection of fetal blood for lymphocyte culture and analysis. These approaches enable screening for clinically relevant and deleterious alleles inherited from the parents, as well as for de novo germline mutations, and they may have the potential to change the diagnosis of genetic disorders in the prenatal setting. Although genomic sequencing of fetal DNA circulating in the bloodstream of the mother has been achieved, it is not commonly used for non-invasive prenatal testing.

In combination with in vitro fertilization (IVF) techniques, it is possible to perform genetic diagnoses in a single cell removed from the four- to eight-cell embryo or to analyze the first polar body from an oocyte. Preconceptual diagnosis thereby avoids therapeutic abortions but is costly and labor intensive. It should be emphasized that excluding a specific disorder by any of these approaches is never equivalent to the assurance of having a normal child. Postnatal indications for cytogenetic analyses in infants or children include multiple congenital anomalies, suspicion of a known cytogenetic syndrome, developmental delay, dysmorphic features, autism, short stature and disorders of sexual development, among others (Table 456-4).

Mutations in certain cancer susceptibility genes such as *BRCA1* and *BRCA2* may identify individuals with an increased risk for the development of malignancies and result in risk-reducing interventions. The detection of cytogenetic alterations and mutations is an important diagnostic and prognostic tool in leukemias and somatic mutational analysis is transforming oncology by providing diagnostic and prognostic information, and it informs the choice of targeted therapies. However, the cancer may recur due to treatment resistance of subclones due to the continuously evolving genomic landscape. The demonstration of the presence or absence of mutations and polymorphisms is also relevant for the rapidly evolving field of pharmacogenomics, including the identification of differences in drug treatment response or metabolism as a function of genetic background. For example, the thiopurine drugs 6-mercaptopurine and azathioprine are commonly used cytotoxic and immunosuppressive agents. They are metabolized by thiopurine methyltransferase (TPMT), an enzyme with variable activity associated with genetic polymorphisms in 10% of whites and complete deficiency in about 1 in 300 individuals. Patients with intermediate or deficient TPMT activity are at risk for excessive toxicity, including fatal myelosuppression. Characterization of these polymorphisms allows mercaptopurine doses to be modified based on TPMT genotype. Pharmacogenomics may increasingly permit individualized drug therapy, improve drug effectiveness, reduce adverse side effects, and provide cost-effective pharmaceutical care **(Chap. 64)**.

ETHICAL ISSUES

Determination of the association of genetic defects with disease, comprehensive data of an individual's genome, and studies of genetic variation raise many ethical and legal issues. Genetic information is generally regarded as sensitive information that should not be readily accessible without explicit consent (*genetic privacy*). The disclosure of genetic information may risk possible discrimination by insurers or employers. The scientific components of the Human Genome Project have been paralleled by efforts to examine ethical, social, and legal implications. An important milestone emerging from these endeavors consists in the Genetic Information Nondiscrimination Act (GINA), signed into law in 2008, which aims to protect asymptomatic individuals against the misuse of genetic information for health insurance and employment. It does not, however, protect the symptomatic individual. Provisions of the U.S. Patient Protection and Affordable Care Act, effective in 2014, have, in part, filled this gap and prohibit exclusion from, or termination of, health insurance based on personal health status. Potential threats to the maintenance of genetic privacy consist in the emerging integration of genomic data into electronic medical records, compelled disclosures of health records, and direct-to-consumer genetic testing.

It is widely accepted that identifying disease-causing genes can lead to improvements in diagnosis, treatment, and prevention. However, the information gleaned from genotypic results can have quite different impacts, depending on the availability of strategies to modify the course of disease. For example, the identification of mutations that cause MEN 2 or hemochromatosis allows specific interventions for affected family members. On the other hand, at present, the identification of an Alzheimer's or Huntington's disease gene does not currently alter therapy and outcomes. Most genetic disorders are likely to fall into an intermediate category where the

opportunity for prevention or treatment is significant but limited. However, the progress in this area is unpredictable, as underscored by the finding that angiotensin II receptor blockers may slow disease progression in Marfan's syndrome. Genetic test results can generate anxiety in affected individuals and family members. Comprehensive sequence analyses are particularly challenging because most individuals can be expected to harbor several serious recessive gene mutations. Moreover, the sensitivity of comprehensive sequence analyses are not always greater, for example, if CNV analysis is not integrated, but can be associated with higher costs.

The impact of genetic testing on health care costs remain unclear. It does vary among disorders and depends on the availability of effective therapeutic modalities. A significant problem arises from the marketing of genetic testing directly to consumers by commercial companies. The validity of these tests is, in part, not well defined, and there are persisting concerns about the lack of appropriate regulatory oversight, the accuracy and confidentiality of genetic information, the availability of counseling, and the handling of these results.

Many issues raised by the genome project are familiar, in principle, to medical practitioners. For example, an asymptomatic patient with increased low-density lipoprotein (LDL) cholesterol, high blood pressure, or a strong family history of early myocardial infarction is known to be at increased risk of coronary heart disease. In such cases, it is clear that the identification of risk factors and an appropriate intervention are beneficial. Likewise, patients with phenylketonuria, cystic fibrosis, or sickle cell anemia are often identified as having a genetic disease early in life. These precedents can be helpful for adapting policies that relate to genetic information. We can anticipate similar efforts, whether based on genotypes or other markers of genetic predisposition, to be applied to many disorders. One confounding aspect of the rapid expansion of information is that our ability to make clinical decisions often lags behind initial insights into genetic mechanisms of disease. For example, when genes that predispose to breast cancer such as *BRCA1* are described, they generate tremendous public interest in the potential to predict disease, but many years of clinical research are still required to rigorously establish genotype and phenotype correlations.

Genomics may contribute to improvements in global health by providing a better understanding of pathogens and diagnostics, and through contributions to drug development. There is, however, ongoing concern about the development of a "genomics divide" because of the costs associated with these developments and uncertainty as to whether these advances will be accessible to the populations of developing countries faced with pressing health needs associated with poverty, infectious diseases, and the relative lack of essential infrastructure. The World Health Organization has summarized these issues and inequities surrounding genomic medicine in a detailed report titled "Genomics and World Health."

Whether related to informed consent, participation in research, or the management of a genetic disorder that affects an individual or his or her family, there is a great need for more information about fundamental principles of genetics. The pervasive nature of the role of genetics in medicine makes it important for physicians and other health care professionals to become more informed about genetics and to provide advice and counseling in conjunction with trained genetic counselors **(Chap. 457)**. The application of screening and prevention strategies does therefore require continuing patient and physician education, changes in health care financing, and legislation to protect patient's rights.

Acknowledgment

Selected sections and Table 456-4 have been integrated from the chapter on Chromosome Disorders by Dr. Nancy B. Spinner and Dr. Laura K. Conlin, published in the 19th edition of Harrison's Principles in Internal Medicine. The data and concept for Figure 456-16 have been graciously provided by Dr. Miriam Udler and Dr. Jose Florez, Massachusetts General Hospital and Harvard Medical School, Boston.

FURTHER READING

Ashley EA: Towards precision medicine. Nat Rev Genet 17:507, 2016.

Jameson JL, Longo DL: Precision medicine—personalized, problematic, and promising. N Engl J Med 372:2229, 2015.

Jarvik GP, Evans JP: Mastering genomic terminology. Genet Med 19: 491, 2017.

Joly Y et al: Comparative approaches to genetic discrimination: Chasing shadows? Trends Genet 33: 299, 2017.

Manolio TA: Bringing genome-wide association findings into clinical use. Nat Rev Genet 14:549, 2013.

Manolio TA et al: Bedside back to bench: Building bridges between basic and clinical genomic research. Cell 169:6, 2017.

Rehm HL: Evolving health care through personal genomics. Nat Rev Genet 18:259, 2017.

Richards S et al: ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17:405, 2015.

Sidow A, Spies N: Concepts in solid tumor evolution. Trends Genet 31:208, 2015.

Wauters A, Van Hoyweghen I: Global trends on fears and concerns of genetic discrimination: A systematic literature review. J Hum Genet 61:275, 2016.

Wolf SM et al: Mapping the ethics of translational genomics: Situating return of results and navigating the research-clinical divide. J Law Med Ethics 43:486, 2015.

World Health Organization: Genomics and World Health: Report of the Advisory Committee on Health Research. 1-254, 2002.

Zheng S et al: Comprehensive pan-genomic characterization of adrenocortical carcinoma. Cancer Cell 29:723, 2016.

457 The Practice of Genetics in Clinical Medicine

Susan M. Domchek, J. Larry Jameson, Susan Miesfeldt

APPLICATIONS OF MOLECULAR GENETICS IN CLINICAL MEDICINE

Genetic testing for inherited abnormalities associated with disease risk is increasingly used in the practice of clinical medicine. Germline alterations include chromosomal abnormalities specific gene mutations with autosomal dominant or recessive patterns of transmission **(Chap. 456)**, and single nucleotide polymorphisms (SNPs) with small relative risks associated with disease. Germline alterations are responsible for disorders beyond classic Mendelian conditions with genetic susceptibility to common adult-onset diseases such as asthma, hypertension, diabetes mellitus, macular degeneration, and a number of types of cancer. For many of these diseases, there is a complex interplay of genes (often multiple) and environmental factors that effect lifetime risk, age of onset, disease severity, and treatment options.

The expansion of human genetic knowledge is changing our understanding of pathophysiology and influencing our classification of diseases. Awareness of genetic etiology can have an impact on clinical management, including prevention and screening for or treatment of a range of diseases. Primary care physicians are relied upon to help patients navigate testing and treatment options. Consequently, they must understand the genetic basis for a large number of genetically influenced diseases, incorporate personal and family history to determine the risk for a specific mutation, and be positioned to provide counseling. Even if patients are seen by genetic specialists who assess

genetic risk and coordinate testing, primary care providers should offer information to their patients regarding the indications, limitations, risks, and benefits of genetic counseling and testing. They must also be prepared to offer risk-based management following genetic risk assessment. Given the pace of genetics, this is an increasingly difficult task. The field of clinical genetics has rapidly transitioned from single gene testing to multigene panel testing, with techniques such as whole-exome and genome sequencing on the horizon, increasing the complexity of test selection and interpretation, as well as patient education and medical decision-making.

COMMON ADULT-ONSET GENETIC DISORDERS

INHERITANCE PATTERNS

Adult-onset hereditary diseases follow multiple patterns of inheritance. Some are autosomal dominant conditions. These include many common cancer susceptibility syndromes such as hereditary breast and ovarian cancer (due to germline *BRCA1* and *BRCA2* mutations) and Lynch syndrome (caused by germline mutations in the mismatch repair genes *MLH1*, *MSH2*, *MSH6*, and *PMS2*). In both of these examples, inherited mutations are associated with a high penetrance (lifetime risk) of cancer, although penetrance is incomplete (risk is not 100%). In other conditions, although there is autosomal dominant transmission, penetrance is lower, thereby making the disorders more difficult to recognize. For example, germline mutations in *CHEK2* increase the risk of breast cancer, but with a moderate lifetime risk in the range of 20–30%, as opposed to 50–70% for mutations in *BRCA1* or *BRCA2*. Other adult-onset hereditary diseases are transmitted in an autosomal recessive fashion where two mutant alleles are necessary to cause full expression of disease. Examples include hemochromatosis and *MUTYH*-associated polyposis. There are more pediatric-onset autosomal recessive disorders, such as lysosomal storage diseases and cystic fibrosis.

The genetic risk for many adult-onset disorders is multifactorial. Risk can be conferred by genetic factors at a number of loci (polygenic), which individually have very small effects (usually with relative risks of <1.5). These risk loci (generally SNPs) combine with other genes and environmental factors in ways that are not well understood. Despite our incomplete understanding of gene-environment interactions, recent data suggest that a healthy lifestyle can mitigate risk associated with elevated polygenic risk for cardiovascular disease. SNP panels are available to assess risk of disease, but the optimal way of using this information in the clinical setting to improve patient outcomes remains uncertain.

Many diseases have multiple patterns of inheritance, adding to the complexity of evaluating patients and families for these conditions. For example, colon cancer can be associated with a single germline mutation in a mismatch repair gene (Lynch syndrome; autosomal dominant), biallelic mutations in *MUTYH* (autosomal recessive), or multiple SNPs (polygenic). Many more individuals will have SNP risk alleles than germline mutations in high-penetrance genes, but cumulative lifetime risk of colon cancer related to the former is modest, whereas the risk related to the latter is significant. Personal and family histories provide important insights into the possible mode of inheritance.

FAMILY HISTORY

When two or more first-degree relatives are affected with asthma, cardiovascular disease, type 2 diabetes, breast cancer, colon cancer, or melanoma, the relative risk for disease among close relatives ranges from two- to fivefold, underscoring the importance of family history for these prevalent disorders. In most situations, the key to assessing the inherited risk for common adult-onset diseases is the collection and interpretation of a detailed personal and family medical history in conjunction with a directed physical examination.

Family history should be recorded in the form of a pedigree, conveying health-related data on first- and second-degree relatives. When such pedigrees suggest inherited disease, they should be expanded to include additional family members. The determination of risk for an asymptomatic individual will vary depending on the size of the family, the number of unaffected relatives, the types of diagnoses, and the ages of disease onset. For example, a woman with two first-degree relatives with breast cancer is at greater risk for a specific Mendelian disorder if she has a total of 3 female first-degree relatives (with only 1 unaffected) than if she has a total of 10 female first-degree relatives (with 8 unaffected). Factors such as adoption and limited family structure (few women in a family or multiple early deaths unrelated to the target disease) should to be taken into consideration in the interpretation of a pedigree. Additional considerations include young age of disease onset (e.g., a 30-year nonsmoking woman with a myocardial infarction), unusual diseases (e.g., male breast cancer or medullary thyroid cancer), and the finding of multiple potentially related diseases in an individual (e.g., a woman with a history of both colon and endometrial cancer). Some adult-onset diseases are more prevalent in certain ethnic groups. For instance, 2.5% of individuals of Ashkenazi Jewish ancestry carry one of three founder mutation in *BRCA1* and *BRCA2*. Factor V Leiden mutations are much more common in Caucasians than in Africans or Asians.

Additional variables that should be documented are nonhereditary risk factors among those with disease (such as cigarette smoking and myocardial infarction; asbestos exposure and lung disease; and mantle radiation and breast cancer). Significant associated environmental exposures or lifestyle factors decrease the likelihood of a specific genetic disorder. In contrast, the absence of nonhereditary risk factors typically associated with a disease raises concern about a genetic association. A personal or family history of deep-vein thrombosis in the absence of known environmental or medical risk factors suggests a hereditary thrombotic disorder. The physical examination may also provide important clues about the risk for a specific inherited disorder. A patient presenting with xanthomas at a young age should prompt consideration of familial hypercholesterolemia. The presence of trichilemmomas in a woman with breast cancer raises concern for Cowden syndrome, associated with *PTEN* mutations.

Recall of family history is often inaccurate. This is especially so when the history is remote and families lose contact or separate geographically. It can be helpful to ask patients to fill out family history forms before or after their visits, because this provides them with an opportunity to contact relatives. Ideally, this information should be embedded in electronic health records and updated intermittently. Attempts should be made to confirm the illnesses reported in the family history before making important and, in certain circumstances, irreversible management decisions. This process is often labor intensive and ideally involves interviews of additional family members or review of medical records (including pathology reports), and death certificates.

Although many inherited disorders will be suggested by the clustering of relatives with the same or related conditions, it is important to note that disease penetrance is incomplete for most genetic disorders. As a result, the pedigree obtained in such families may not exhibit a clear Mendelian inheritance pattern because not all family members carrying the disease-associated alleles will manifest clinical evidence of the condition. Furthermore, genes associated with some of these disorders often exhibit variable disease expression. For example, the breast cancer–associated gene *BRCA2* can predispose to several different malignancies in the same family, including cancers of the breast, ovary, pancreas, skin (melanoma), and prostate. For common diseases such as breast cancer, some family members without the susceptibility allele (or genotype) may develop breast cancer (or phenotype) sporadically. Such *phenocopies* represent another confounding variable in the pedigree analysis.

Some of the aforementioned features of the family history are illustrated in **Fig. 457-1**. In this example, the proband (the individual serving as the starting point for genetic assessment in a family), a 36-year-old woman (IV-1), has a strong history of breast and ovarian cancer on the paternal side of her family. The early age of onset and the co-occurrence of breast and ovarian cancer in this family suggest the possibility of an inherited mutation in *BRCA1* or *BRCA2*, associated with the hereditary breast and ovarian cancer syndrome. It is unclear however, without genetic testing, whether her father harbors such a mutation and transmitted it to her. After appropriate genetic counseling of the proband

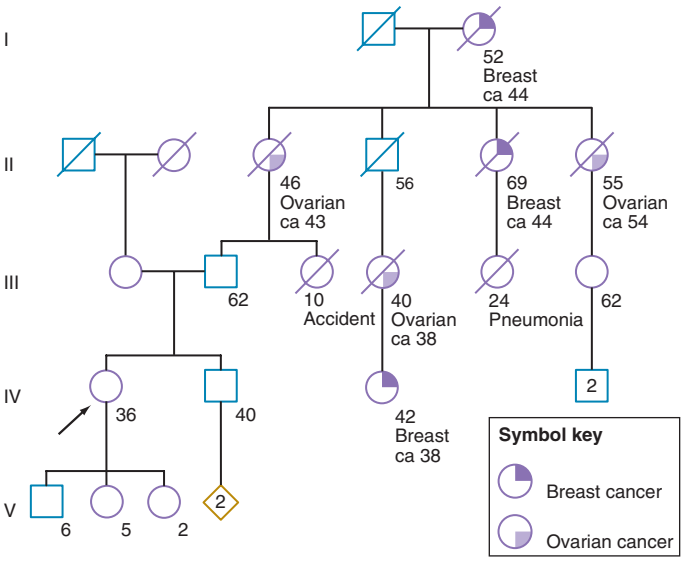


FIGURE 457-1 36-year-old woman (*arrow*) seeks consultation because of her family history of cancer. The patient expresses concern that the multiple cancers in her relatives imply an inherited predisposition to develop cancer. The family history is recorded, and records of the patient's relatives confirm the reported diagnoses.

and her family, the most informative and cost-effective approach to DNA analysis in this family is to test the cancer-affected 42-year-old living cousin for the presence of a *BRCA1* or *BRCA2* mutation. If a mutation is found, then it is possible to test for this particular alteration in other family members, if they so desire. In the example shown, if the proband's cousin has a *BRCA1* mutation, testing for this alteration (single site testing) would be offered to the proband's father. If he tests positive, there is a 50:50 probability that the mutation was transmitted to her, and genetic testing can be used to establish the absence or presence of this alteration. In contrast, if he tests negative for the known familial *BRCA1* mutation (a true negative result), the proband and her brother are not at risk for having inherited this variant.

GENETIC TESTING FOR ADULT-ONSET DISORDERS

A critical first step before initiating genetic testing is to ensure that the correct clinical diagnosis has been made, whether it is based on family history, characteristic physical findings, pathology, or biochemical testing. Such careful clinical assessment can define the phenotype. In the traditional model of genetic testing, testing is directed initially toward the most probable genes (determined by the phenotype), which prevents unnecessary testing. Many disorders exhibit the feature of locus heterogeneity, which refers to the fact that mutations in different genes can cause phenotypically similar disorders. For example, osteogenesis imperfecta **(Chap. 406)**, long QT syndrome **(Chap. 247)**, muscular dystrophy **(Chap. 441)**, and hereditary predisposition to breast **(Chap. 75)** or colon **(Chap. 77)** cancer can each be caused by mutations in a number of distinct genes. The patterns of disease transmission, disease risk, clinical course, and treatment may differ significantly depending on the specific gene affected. Historically, the choice of which gene to test has been determined by unique clinical and family history features and the relative prevalence of candidate genetic disorders. However, rapid changes in genetic testing techniques, as discussed below, are impacting this paradigm. It is now technically and financially feasible to sequence many genes (or even the whole exome) at one time. The incorporation of expanded testing for germline mutations is rapidly evolving both within the clinic as well as through direct-to-consumer marketing of genetic and genomic tests.

METHODOLOGIC APPROACHES TO GENETIC TESTING

Genetic testing is regulated and performed in much the same way as other specialized laboratory tests. In the United States, genetic testing laboratories are Clinical Laboratory Improvement Amendments (CLIA) approved to ensure that they meet quality and proficiency standards. A useful information source for various genetic tests is *www.genetests.org*. It should be noted that many tests need to be ordered through specialized laboratories.

Genetic testing is performed largely by DNA sequence analysis for mutations, although genotype can also be deduced through the study of RNA or protein (e.g., apolipoprotein E, hemoglobin S, and immunohistochemistry). For example, universal Lynch syndrome screening of colorectal and uterine cancers via immunohistochemical analysis for absence of expression of mismatch repair proteins is recommended by the National Comprehensive Cancer Center Network. The determination of DNA sequence alterations relies heavily on the use of polymerase chain reaction (PCR), which allows rapid amplification and analysis of the gene of interest. In addition, PCR enables genetic testing on minimal amounts of DNA extracted from a wide range of tissue sources including leukocytes, mucosal epithelial cells (obtained via saliva or buccal swabs), and archival tissues. Amplified DNA can be analyzed directly by DNA sequencing, or it can be hybridized to DNA chips or blots to detect the presence of normal and altered DNA sequences. Direct DNA sequencing is frequently used for determination of hereditary disease susceptibility and prenatal diagnosis. Analyses of large alterations (e.g., deletions, duplications, rearrangements, translocations) of the genome are possible using cytogenetics, fluorescent in situ hybridization (FISH), Southern blotting, or multiplex ligation-dependent probe amplification (MLPA).

Massively parallel sequencing (also called *next-generation sequencing*) has significantly altered the approach to genetic testing for adult-onset hereditary susceptibility disorder. This technology encompasses high-throughput approaches to DNA analysis which can reliably examine many genes at one time. Technically, this involves sequencing of millions of small fragments of DNA in parallel. Through bioinformatics, these fragments are pieced together by mapping the individual sequence reads to the human reference genome, a very different process than traditional Sanger sequencing which is time-consuming and expensive.

Multiplex panels for inherited susceptibility are commercially available and include testing of a number of genes that have been associated with the condition of interest. For example, panels are available for Brugada syndrome, hypertrophic cardiomyopathy, and Charcot-Marie-Tooth neuropathy. For many syndromes, this type of panel testing may make sense. However, in other situations, the clinical utility of panel testing is evolving and may be dependent on the particular composition of the panel. Currently available breast cancer susceptibility panels contain close to 30 genes with larger multi-cancer panels available. Some of the genes included in the larger multi-cancer panels have no known association with breast cancer or have only a modest associated risk and the clinical utility is uncertain. An additional problem of sequencing many genes, rather than focusing on leading candidate genes, is the identification of one or more variants of uncertain significance, discussed below, or an unexpected, yet clinically relevant result.

Whole-exome sequencing (WES) is also now commercially available, although largely used in individuals with syndromes unexplained by traditional genetic testing. As cost declines, WES may be more widely used. Whole-genome sequencing is also commercially available. Although it may be quite feasible to sequence the entire genome, there are many issues in doing so, including the daunting task of analyzing the vast amount of data generated. Other issues include: (1) the optimal way in which to obtain informed consent, (2) interpretation of frequent sequence variation of uncertain significance, (3) interpretation of alterations in genes with unclear relevance to specific human pathology, and (4) management of unexpected but clinically significant genetic findings.

Testing strategies are evolving as a result of these new genetic testing platforms. As the costs of multiplex gene panels and WES continue to fall, and as interpretation and understanding of the clinical relevance of such test results improve, there has been a shift to more extensive panel-based genetic testing in the clinic. For example, in the past, a 30-year-old woman with breast cancer but no family

history of cancer and no syndromic features would undergo *BRCA1/2* testing and would be offered *TP53* testing in light of her early onset disease (notably, a reasonable number of individuals offered *TP53* testing for Li-Fraumeni syndrome in the past declined because mutations are associated with extremely high cancer risks—including childhood cancers—in multiple organs and means to mitigate risk are uncertain).Without features consistent with other high-risk, breast cancer-related conditions like Cowden syndrome, the woman would not have been routinely offered *PTEN* analysis (associated with Cowden syndrome) or testing for other breast cancer-associated genes including *CHEK2* and *ATM*. It is now possible to synchronously analyze all of the aforementioned genes, along with genes such as BRIP and RAD51D (which are associated with moderate ovarian cancer risk but unclear risk for other cancers, including breast cancer) for a nominally higher cost than *BRCA1/2* testing alone. Concerns about such panels include appropriate consent strategies related to unclear findings, including one or more variants of uncertain significance, unanticipated results, and the uncertain clinical utility of some of the genes included on the panel **(Fig. 457-2)**.

Limitations to the accuracy and interpretation of genetic testing exist. In addition to technical errors, genetics tests are sometimes designed to detect only the most common mutations. In addition, genetic testing has evolved over time. For example, it was not possible to obtain commercially available comprehensive large genomic rearrangement testing for *BRCA1* and *BRCA2* until 2006. Therefore, a negative result must be qualified by the possibility that the individual may have a variant that was not detectable in the test. In addition, the individual may have a mutation in another untested cancer-associated gene or in a gene not yet reported to be associated with elevated disease risk. As such, unless there is known baseline mutation in the family, a negative result in an individual with a suggestive personal or family history is typically classified as an uninformative negative. In this circumstance, medical management decisions should be based on personal and family history. For example, a woman with a strong family history of breast cancer who receives an uninformative negative panel-based genetic test result may still be eligible for high-risk care, including consideration of MRI-based breast cancer screening in addition to close clinical surveillance and mammograms.

The finding of a VUS is another limitation to genetic testing. A VUS (also termed *unclassified variant*) is a sequence variation in a gene where the effect of the alteration on the function of the protein is not known. Many of these variants are single nucleotide substitutions (also called missense mutations) that result in a single amino acid change. Although many VUSs will ultimately be reclassified as benign variants, some will prove to be functionally important. As more genes are sequenced (for example, in a multiplex panel or through WES), the percentage of individuals found to have one or more VUSs increases significantly. The finding of a VUS is difficult for patients and providers alike and complicates decisions regarding medical management. In this setting, until there is further reclassification of the variant, ongoing screening, surveillance and care is typically determined based on personal and family history.

Clinical utility is an important consideration because genetic testing for susceptibility to chronic diseases is increasingly integrated into the practice of medicine. In some situations, there is proven clinical utility to genetic testing with significant evidence-based changes in medical management options and recommendations based on results. For example, there is clear evidence that risk-reducing bilateral salpingo-oophorectomy benefits women with a documented *BRCA1/2* mutation, relative to both ovarian and breast cancer-related risk. However, in many cases, the discovery of disease-associated genes has outpaced studies that assess how such information should be used in the clinical management of the patient and family. This is particularly true for moderate- and low-penetrance gene mutations. Therefore, predictive genetic testing should be approached with caution and offered to individuals who have been adequately counseled and have provided informed consent.

Predictive genetic testing falls into two distinct categories. Presymptomatic testing applies to diseases where a specific genetic alteration is associated with a near 100% likelihood of developing disease. In contrast, predisposition testing predicts a risk for disease that is <100%. For example, presymptomatic testing is available for those at risk for Huntington's disease; whereas, predisposition testing is considered for those at risk for hereditary colon cancer. It is important to note that for the majority of adult-onset disorders, testing is only predictive. Test results cannot reveal with confidence whether, when, or how the disease will manifest itself. For example, not everyone with the apolipoprotein ε4 allele will develop Alzheimer's disease, and individuals without this genetic marker can still develop the disorder.

The optimal testing strategy for a family is to initiate testing in a [target] disease-affected family member first. Identification of a mutation can direct the testing of other at-risk family members (whether symptomatic or not). In the absence of additional familial or environmental risk factors, individuals who test negative for the mutation found in the affected family member can be informed that they are at general population risk for that particular disease. Furthermore, they can be reassured that they are not at risk for passing the mutation on to their children. On the other hand, asymptomatic family members who test positive for the known mutation must be informed that they are at increased risk for disease development and for transmitting the alteration to their children.

Pretest counseling and education are important, as is an assessment of the patient's ability to understand and cope with test results. Genetic testing has implications for entire families, and thus individuals interested in pursuing genetic testing must consider how test results might

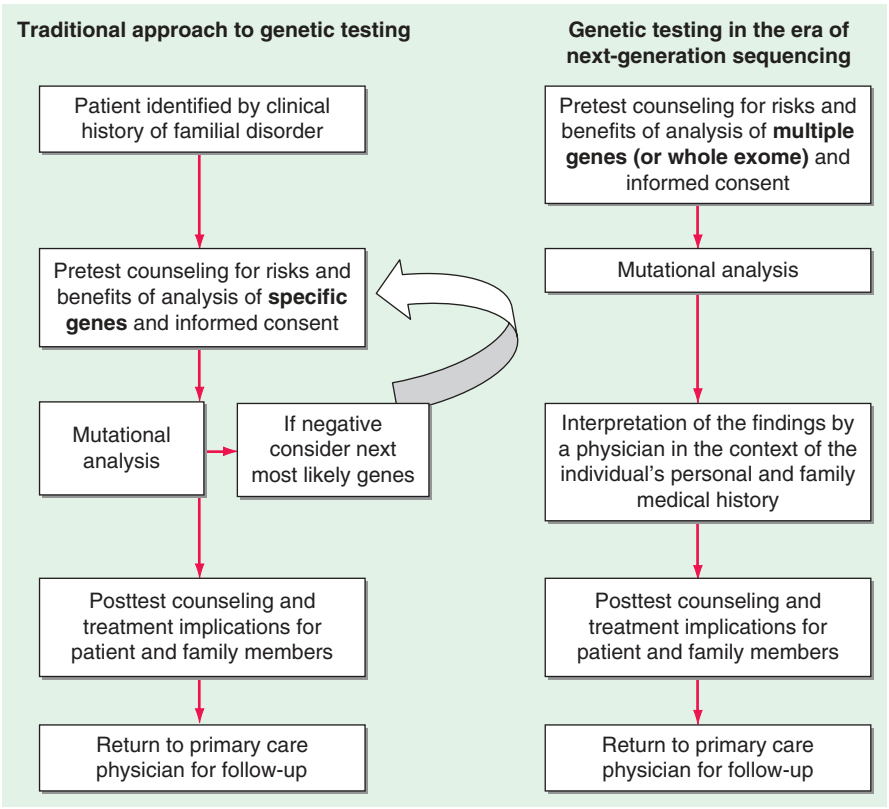


FIGURE 457-2 Approach to genetic testing.

impact their relationships with relatives, partners, spouses, and children. In families with a known genetic mutation, those who test positive must consider the impact of their carrier status on their present and future lifestyles; those who test negative may manifest survivor guilt. Parents who are found to have a disease-associated mutation often express considerable anxiety and despair as they address the issue of risk to their children. In addition, some individuals consider options such as preimplantation genetic diagnosis (PGD) in their reproductive decision-making.

When a condition does not manifest until adulthood, clinicians and parents are faced with the question of whether at-risk children should be offered genetic testing and, if so, at what age. Although the matter is debated, several professional organizations have cautioned that genetic testing for adult-onset disorders should not be offered to children. Many of these conditions have no known interventions in childhood to prevent disease; consequently, such information can pose significant psychosocial risk to the child. In addition, there is concern that testing during childhood violates a child's right to make an informed decision regarding testing upon reaching adulthood. On the other hand, testing should be offered in childhood for disorders that may manifest early in life, especially when management options are available. For example, children with multiple endocrine neoplasia 2 (MEN 2) may develop medullary thyroid cancer early in life and should be considered for prophylactic thyroidectomy **(Chap. 381)**. Similarly, children with familial adenomatous polyposis (FAP) due to a mutation in *APC* may develop polyps in their teens with progression to invasive cancer in the twenties, and therefore, colonoscopy screening is started between the ages of 10 and 15 years **(Chap. 77)**.

INFORMED CONSENT

Informed consent for genetic testing begins with education and counseling. The patient should understand the risks, benefits, and limitations of genetic testing, as well as the potential implications of test results. Informed consent should include a written document, drafted clearly and concisely in a language and format that is understandable to the patient. Because molecular genetic testing of an asymptomatic individual often allows prediction of future risk, the patient should understand all potential long-term medical, psychological, and social implications of testing. There have long been concerns about the potential for genetic discrimination. The Genetic Information Nondiscrimination Act (GINA) was passed in 2008 and provides some protections related to job and health insurance discrimination. It is important to explore with patients the potential impact of genetic test results on future health as well as disability and life insurance coverage. Patients should understand that alternatives remain available if they decide not to pursue genetic testing, including the option of delaying testing to a later date. The option of DNA banking should be presented so that samples are readily available for future use by family members, if needed.

FOLLOW-UP CARE AFTER TESTING

Depending on the nature of the genetic disorder, posttest interventions may include: (1) cautious surveillance and awareness; (2) specific medical interventions such as enhanced screening, chemoprevention, or risk-reducing surgery; (3) risk avoidance; and (4) referral to support services. For example, patients with known deleterious mutations in *BRCA1* or *BRCA2* are strongly encouraged to undergo risk-reducing salpingo-oophorectomy and are offered intensive breast cancer screening as well as the option of risk-reducing mastectomy. In addition, such women may wish to take chemoprevention with tamoxifen, raloxifene, or exemestane. Those with more limited medical management and prevention options, such as patients with Huntington's disease, should be offered continued follow-up and supportive services, including physical and occupational therapy and social services or support groups as indicated. Specific interventions will change as research continues to enhance our understanding of the medical management of these genetic conditions and more is learned about the functions of the gene products involved.

Individuals who test negative for a mutation in a disease-associated gene identified in an affected family member must be reminded that they may still be at risk for the disease. This is of particular importance for common diseases such as diabetes mellitus, cancer, and coronary artery disease. For example, a woman who finds that she does not carry the disease-associated mutation in *BRCA1* previously discovered in the family should be reminded that she still requires the same breast cancer screening recommended for the general population.

TABLE 457-1 Indications for Genetic Counseling

Advanced maternal age (>35 years)
Consanguinity
Previous history of a child with birth defects or a genetic disorder
Personal or family history suggestive of a genetic disorder
High-risk ethnic groups
Documented genetic alteration in a family member
Ultrasound or prenatal testing suggesting a genetic disorder

GENETIC COUNSELING AND EDUCATION

Genetic counseling is the process of helping people understand and adapt to the medical, psychological and familial implications of genetic contributions to disease. This process integrates the following: interpretation of family and medical histories to assess the chance of disease occurrence or recurrence; education about the natural history of the condition, inheritance pattern, testing, management, prevention, support resources and research; counseling to promote informed choices in view of risk assessment, family goals, ethical and religious values. Genetic counseling should be distinguished from genetic testing and risk-based medical screening and care, though genetic counselors are involved in these later issues.

Genetic risk assessment is complex and often involves elements of uncertainty. Genetic counseling can be useful in a wide range of situations **(Table 457-1)**. The role of the genetic counselor includes the following:

1. Gather and document a detailed family history.
2. Educate patients about general genetic principles related to disease risk, both for themselves and for others in the family.
3. Assess and enhance the patient's ability to cope with the genetic information offered.
4. Discuss how nongenetic factors may relate to the ultimate expression of disease.
5. Address medical management issues.
6. Assist in determining the role of genetic testing for the individual and the family.
7. Ensure the patient is aware of the indications, process, risks, benefits, and limitations of the various genetic testing options.
8. Assist the patient, family, and referring physician in the interpretation of the test results.
9. Ensure that the patient has the resources necessary to alert relatives to their risk, particularly in the face of a positive genetic test result.
10. Address the reproductive implications of a positive genetic test result, including the risk for a recessive disorder as well as discussion of reproductive options, including gamete donation or PGD.
11. Refer the patient and other at-risk family members for additional medical and support services, if necessary.

The principles of voluntary and informed decision-making, and protection of the individual's privacy and confidentiality are core principles in the practice of genetic counseling. Genetic counseling is generally offered in a *nondirective, noncoercive* manner, wherein patients learn to understand how their values factor into a particular medical decision. Nondirective counseling is particularly appropriate when there are no data demonstrating a clear benefit associated with a particular intervention or when an intervention is considered experimental. For example, nondirective genetic counseling is used when a person is deciding whether to undergo genetic testing for Huntington's disease. At this time, there is no clear benefit (in terms of medical outcome) to an at-risk individual undergoing genetic testing for this disease because its course cannot be altered by therapeutic interventions.

However, testing can have an important impact on the individual's perception of advanced care planning and his or her interpersonal relationships and plans for childbearing. Therefore, the decision to pursue testing rests on the individual's belief system and values. On the other hand, a more directive approach is appropriate when a condition can be treated. In a family with FAP, colon cancer screening and prophylactic colectomy should be recommended for known *APC* mutation carriers. The counselor and clinician following this family must ensure that the at-risk family members have access to the resources necessary to adhere to these recommendations.

Genetic education is central to an individual's ability to make an informed decision regarding testing options and treatment. An adequate knowledge of patterns of inheritance will allow patients to understand the probability of disease risk for themselves and other family members. It is also important to impart the concepts of disease penetrance and expression. For most complex adult-onset genetic disorders, asymptomatic patients should be advised that a positive test result does not always translate into future disease development. In addition, the role of nongenetic factors, such as environmental exposures and lifestyle, must be discussed in the context of multifactorial disease risk and disease prevention. Finally, patients should understand the natural history of the disease as well as the potential options for intervention, including screening, prevention, and in certain circumstances, pharmacologic treatment or prophylactic surgery.

THERAPEUTIC INTERVENTIONS BASED ON GENETIC RISK FOR DISEASE

Specific treatments are available for a number of genetic disorders. Strategies for the development of therapeutic interventions have a long history in childhood metabolic diseases; however, these principles have been applied in the diagnosis and management of adult-onset diseases as well **(Table 457-2)**. Hereditary hemochromatosis is usually caused by mutations in *HFE* (although other genes have been less commonly associated) and manifests as a syndrome of iron overload, which can lead to liver disease, skin pigmentation, diabetes mellitus, arthropathy, impotence in males, and cardiac issues **(Chap. 407)**. When identified

TABLE 457-2 Examples of Genetic Testing and Possible Interventions

GENETIC DISORDER	INHERITANCE	GENES	INTERVENTIONS
Oncologic			
Lynch syndrome (HNPCC)	AD	*MLH1, MSH2, MSH6, PMS2*	Early endoscopic screening; risk-reducing surgery
Familial adenomatous polyposis	AD	*APC*	Early and frequent endoscopy; prophylactic colectomy
Hereditary breast and ovarian cancer	AD	*BRCA1, BRCA2*	Risk reducing salpingo-oophorectomy; intensified breast surveillance including breast MRI; risk-reducing mastectomy
Hereditary diffuse gastric cancer	AD	*CDH1*	Prophylactic gastrectomy; enhanced breast cancer surveillance
Hematologic			
Factor V Leiden	AD	*F5*	Avoidance of thrombogenic risk factors
Hemophilia A	XL	*F8*	Factor VIII replacement
Hemophilia B	XL	*F9*	Factor IX replacement
Glucose 6-phosphate dehydrogenase deficiency	XL	*G6PD*	Avoidance of oxidant drugs and certain foods
Cardiovascular			
Hypertrophic cardiomyopathy	AD	>10 genes including *MYBPC3, MYH7, TNNT2, TPM1*	Echocardiographic screening; pharmacologic intervention; myomectomy
Long QT syndrome	AD, AR	>10 genes including *KCNQ1, SCN5A, KCNE1, KCNE2*	Electrocardiographic screening; pharmacologic intervention; implantable cardiac defibrillator devices
Marfan's syndrome	AD	*FBN1*	Echocardiographic screening; prophylactic beta blockers; aortic valve replacement as indicated
Gastrointestinal			
Familial Mediterranean fever	AR	*MEFV*	Colchicine
Hemochromatosis	AR	*HFE*	Phlebotomy
Pulmonary			
α_1 Antitrypsin deficiency	AR	*SERPINA1*	Avoidance of smoking and occupational and environmental toxins
Cystic fibrosis	AR	*CFTR*	Chest physiotherapy; agents to promote airway secretion clearance; *CFTR* modulators (G551D mutations); lung transplantation
Endocrine			
Neurohypophyseal diabetes insipidus	AD	*AVP*	Replace vasopressin
Familial hypocalciuric hypercalcemia	AD	*CASR*	Avoidance of parathyroidectomy
Multiple endocrine neoplasia type 2	AD	*RET*	Prophylactic thyroidectomy; screening for pheochromocytoma and hyperparathyroidism
Renal			
Polycystic kidney disease	AD, AR	*PKD1, PKD2, PKHD1*	Prevention of hypertension; prevention of urinary tract infections; kidney transplantation
Nephrogenic diabetes insipidus	XL, AR	*AVPR2, AQP2*	Fluid replacement; thiazides with or without amiloride
Neurologic			
Malignant hyperthermia	AD	*RYR1, CACNA1S*	Avoidance of precipitating anesthetics
Hyperkalemic periodic paralysis	AD	*SCN4A*	Diet rich in calcium and low in potassium; thiazides or acetazolamide
Duchenne's and Becker's muscular dystrophy	XL	*DMD*	Corticosteroids; physical therapy
Wilson's disease	AR	*ATP7B*	Zinc, trientene

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; HNPCC, hereditary nonpolyposis colorectal cancer; MRI, magnetic resonance imaging; XL, X-linked.

early, the disorder can be managed effectively with therapeutic phlebotomy. Therefore, when the diagnosis of hemochromatosis has been made in a proband, it is important to counsel other family members in order to minimize the impact of the disorder.

Preventative measures and therapeutic interventions are not restricted to metabolic disorders. Identification of familial forms of long QT syndrome, associated with ventricular arrhythmias, allows early electrocardiographic testing and the use of prophylactic antiarrhythmic therapy, overdrive pacemakers, or defibrillators. Individuals with familial hypertrophic cardiomyopathy can be screened by ultrasound, treated with beta blockers or other drugs, and counseled about the importance of avoiding strenuous exercise and dehydration. Those with Marfan's syndrome can be treated with beta blockers or angiotensin II receptor blockers and monitored for the development of aortic aneurysms.

The field of pharmacogenetics identifies genes that alter drug metabolism or confer susceptibility to toxic drug reactions. Pharmacogenetics seeks to individualize drug therapy in an attempt to improve treatment outcomes and reduce toxicity. Examples include thiopurine methyltransferase (TPMT) deficiency, dihydropyrimidine dehydrogenase deficiency, malignant hyperthermia, and glucose-6-phosphate deficiency. Despite successes in this area, it is not always clear how to incorporate pharmacogenetics into clinical care. For example, although there is an association with *CYP2C6* and *VKORC1* genotypes and warfarin dosing, there is no evidence that incorporating genotyping into clinical practice improves patient outcomes compared with clinical algorithms.

The identification of germline abnormalities that increase the risk of specific types of cancer is rapidly changing clinical management. Identifying family members with mutations that predispose to FAP or Lynch syndrome leads to recommendations of early cancer screening and prophylactic surgery, as well as consideration of chemoprevention and attention to healthy lifestyle habits. Similar principles apply to familial forms of melanoma as well as cancers of the breast, ovary, and thyroid. In addition to increased screening and prophylactic surgery, the identification of germline mutations associated with cancer may also lead to the development of targeted therapeutics, for example, the U.S. Food and Drug Administration (FDA) approval of the poly–ADP ribose polymerase (PARP) inhibitors olaparib and rucaparib for BRCA1/2-associated recurrent ovarian cancer.

Although the role of genetic testing in the clinical setting continues to evolve, such testing holds the promise of allowing early and more targeted interventions that can reduce morbidity and mortality. Rapid technologic advances are changing the ways in which genetic testing is performed. As genetic testing becomes less expensive and technically easier to perform, it is anticipated that there will be an expansion of its use. This will present challenges, but also opportunities. It is critical that physicians and other health care professionals keep current with advances in genetic medicine in order to facilitate appropriate referral for genetic counseling and judicious use of genetic testing, as well as to provide state-of-the-art, evidence-based care for affected or at-risk patients and their relatives.

FURTHER READING

Anya ER et al: Genetic information, non-discrimination, and privacy protections in genetic counseling practice. J Genet Couns 23:891, 2014.

Clayton EW: Ethical, legal, and social implications of genomic medicine. N Engl J Med 349:562, 2003.

Hampel H et al: A practice guideline from the American College of Medical Genetics and Genomics and the National Society of Genetic Counselors: Referral indications for cancer predisposition assessment. Genet Med 17:70, 2015.

Resta R et al: A new definition of genetic counseling: National Society of Genetic Counselors' Task Force report. J Genet Couns 15:77, 2006.

Roberts JS, Ostergren J: Direct-to-consumer genetic testing and personal genomics services: A review of recent empirical studies. Curr Genet Med Rep 1:182, 2013.

Robson ME et al: American Society of Clinical Oncology policy statement update: Genetic and genomic testing for cancer susceptibility. J Clin Oncol 33:3660, 2015.

458 Gene and Cell Based Therapy in Clinical Medicine

Katherine A. High, Malcolm K. Brenner

Gene transfer is a novel area of therapeutics in which the active agent is a nucleic acid sequence rather than a protein or small molecule. Because delivery of naked DNA or RNA to a cell is an inefficient process, most gene transfer is carried out using a vector, or gene delivery vehicle. These vehicles have generally been engineered from viruses by deleting some or all of the viral genome and replacing it with the therapeutic gene of interest under the control of a suitable promoter **(Table 458-1)**. Gene transfer strategies can thus be described in terms of three essential elements: (1) a vector; (2) a gene to be delivered, sometimes called the transgene; and (3) a physiologically relevant target cell to which the DNA or RNA is delivered. The series of steps in which the vector and donated DNA enter the target cell and express the transgene is referred to as *transduction*. Gene delivery can take place in vivo, in which the vector is directly injected into the patient, or, in the case of hematopoietic and some other target cells, ex vivo, with removal of the target cells from the patient, followed by return of the gene-modified autologous cells to the patient after manipulation in the laboratory. The latter approach effectively combines gene transfer techniques with cellular therapies **(Chap. 473)**.

Gene transfer is one of the most powerful concepts in modern molecular medicine and has the potential to address a host of diseases for which there are currently no available treatments. Clinical trials of gene therapy have been under way since 1990; the first gene therapy product to be licensed in the United States or Europe was licensed in 2012 (see below). Given that vector-mediated gene therapy is arguably one of the most complex therapeutics yet developed, consisting of both a nucleic acid and a protein component, this time course from first clinical trial to licensed product is noteworthy for being similar to those seen with other novel classes of therapeutics, i.e., monoclonal antibodies. Over 5000 subjects have been enrolled in gene transfer studies, and serious adverse events have been rare. Some of the initial trials were characterized by an overabundance of optimism and a failure to be appropriately critical of preclinical studies in animals; in addition, it was in some contexts not fully appreciated that animal studies are only a partial guide to safety profiles of products in humans (e.g., insertional mutagenesis, and human immune responses to the vector). Initial exuberance was driven by many factors, including an intense desire to develop therapies for hitherto untreatable diseases, lack of understanding of risks, and, in some cases, undisclosed financial conflicts of interest. After a teenager died of complications related to vector infusion, the field underwent a retrenchment; continued efforts led to a more nuanced understanding of the risks and benefits of these new therapies and more sophisticated selection of disease targets. Currently, gene therapies are being developed for a variety of disease entities **(Table 458-2)**.

GENE TRANSFER FOR GENETIC DISEASE

Gene transfer strategies for genetic disease generally involve gene addition therapy, an approach characterized by transfer of the missing gene to a physiologically relevant target cell. However, other strategies are possible, including supplying a gene that achieves a similar biologic effect through an alternative pathway (e.g., factor VIIa for hemophilia A); supplying an antisense oligonucleotide to splice out a mutant exon if the sequence is not critical to the function of the protein (as has been done with the dystrophin gene in Duchenne's muscular dystrophy); or downregulating a harmful effect through a small interfering RNA (siRNA). Two distinct strategies are used to achieve long-term gene expression: one is to transduce stem cells with an integrating vector, so that all progeny cells will carry the donated gene; and the other is to transduce long-lived cells, such as skeletal muscle or neurons. In the

TABLE 458-1 Characteristics of Gene Delivery Vehicles

	VIRAL VECTORS						
FEATURES	RETROVIRAL	LENTIVIRAL	ADENOVIRAL	AAV	HUMAN FOAMY VIRUS	HSV-1	ALPHA VIRUSES
Viral genome	RNA	RNA	DNA	DNA	RNA	DNA	RNA
Cell division requirement	Yes	G_1 phase	No	No	No	No	No
Packaging limitation	8 kb	8 kb	8–30 kb	5 kb	8.5 kb	40–150 kb	5 kb
Immune responses to vector	Few	Few	Extensive	Few	Few	Few in recombinant virus	Few
Genome integration	Yes	Yes	Poor	Poor	Yes	No	No
Long-term expression	Yes	Yes	No	Yes	Yes	No	No
Main advantages	Persistent gene transfer in dividing cells	Persistent gene transfer in transduced tissues	Highly effective in transducing various tissues	Elicits few inflammatory responses, nonpathogenic	Persistent gene expression in both dividing and nondividing cells	Large packaging capacity with persistent gene transfer	Limited immune responses against the vector
Main disadvantages	Theoretical risk of insertional mutagenesis (occurred in multiple cases)	Might induce oncogenesis in some cases (not yet observed)	Viral capsid elicits strong immune responses	Limited packaging capacity	In need of a stable packaging system	Residual cytotoxicity with neuron specificity	Transduced gene expression is transient

Abbreviations: AAV, adeno-associated virus; HSV, herpes simplex virus.

case of long-lived cells, integration into the target cell genome is unnecessary. Instead, because the cells are nondividing, the donated DNA, if stabilized in an episomal form, will give rise to expression for the life of the cell. This approach thus avoids problems related to integration and insertional mutagenesis.

IMMUNODEFICIENCY DISORDERS: PROOF OF PRINCIPLE

Early attempts to effect gene replacement into hematopoietic stem cells (HSCs) were stymied by the relatively low transduction efficiency of retroviral vectors, which require dividing target cells for integration. Because HSCs are normally quiescent, they are a formidable transduction target. However, identification of cytokines that induced cell division without promoting differentiation of stem cells, along with technical improvements in the isolation and transduction of HSCs, led to modest but real gains in transduction efficiency.

The first convincing therapeutic effect from gene transfer occurred with X-linked severe combined immunodeficiency disease (SCID), which results from mutations in the gene (*IL2RG*) encoding the γc subunit of cytokine receptors required for normal development of T and natural killer (NK) cells **(Chap. 344)**. Affected infants present in the first few months of life with overwhelming infections and/or failure to thrive. In this disorder, it was recognized that the transduced cells, even if few in number, would have a proliferative advantage compared to the nontransduced cells, which lack receptors for the cytokines required for lymphocyte development and maturation. Complete reconstitution of the immune system, including documented responses to standard childhood vaccinations, clearing of infections, and remarkable gains in growth occurred in most of the treated children. However, among 20 children treated in the initial trials, five eventually developed a syndrome similar to T cell acute lymphocytic leukemia, with splenomegaly, rising white counts, and the emergence of a single clone of T cells. Molecular studies revealed that, in most of these children, the retroviral vector had integrated within a gene, *LMO-2* (LIM only-2), which encodes a component of a transcription factor complex involved in hematopoietic development. The retroviral long terminal repeat increases the expression of *LMO-2*, resulting in T cell leukemia.

The X-linked SCID studies were a watershed event in the evolution of gene therapy. They demonstrated conclusively that gene therapy could cure disease; of the 20 children treated in these initial trials, 18 achieved correction of the immunodeficiency disorder. Unfortunately, 5 of the 20 patients later developed a leukemia-like disorder, and one died of this complication; the rest are alive and free of complications at time periods ranging up to 17 years after initial treatment. These studies demonstrated that insertional mutagenesis leading to cancer was more than a theoretical possibility **(Table 458-3)**. As a result of the experience in these trials, all protocols using integrating vectors in hematopoietic cells must include a plan for monitoring sites of insertion and clonal proliferation. Strategies to overcome this complication have included using a "suicide" gene cassette in the vector, so that errant clones can be quickly ablated, or using "insulator" elements in the cassette, which can limit the activation of genes surrounding the insertion site. Lentiviral vectors, which can efficiently transduce nondividing target cells, are also likely to be safer than retroviral vectors, based on patterns of integration; the field is thus gradually moving toward these to replace retroviral vectors.

More clear-cut success has been achieved in a gene therapy trial for another form of SCID, adenosine deaminase (ADA) deficiency **(Chap. 344)**. ADA-SCID is clinically similar to X-linked SCID, although it can be treated by enzyme replacement therapy with a pegylated form

TABLE 458-2 Most Common Indications in Gene Therapy Trials

INDICATION	NUMBER
Cancer	1554
Monogenic diseases	248
Infectious diseases	180

Source: Adapted from J Gene Med *http://www.abedia.com/wiley/indications.php, 2016.*

TABLE 458-3 Potential Complications of Gene Therapy

Gene silencing—repression of promoter
Genotoxicity—complications arising from insertional mutagenesis
Phenotoxicity—complications arising from overexpression or ectopic expression of the transgene
Immunotoxicity—harmful immune response to either the vector or transgene; or a harmful immune response of the vector (e.g. CAR T cells)
Risks of horizontal transmission—shedding of infectious vector into environment
Risks of vertical transmission—germline transmission of donated DNA

of the enzyme (PEG-ADA), which leads to immune reconstitution but not always to normal T cell counts. Enzyme replacement therapy is expensive (annual costs: $200,000–$300,000). The initial trials of gene therapy for ADA-SCID were unsuccessful, but modifications of this protocol to include the use of HSCs rather than T cells as the target for transduction; discontinuation of PEG-ADA at the time of vector infusion, so that the transduced cells have a proliferative advantage over the nontransduced; and the use of a mild conditioning regimen to facilitate engraftment of the transduced autologous cells have led to success without the complications seen in the X-linked SCID trials. There have been no complications in the 10 children treated on the Milan protocol, with a median follow-up of >11 years. This therapy was approved in 2016 by the European Medicines Agency. ADA-SCID, then, is an example where gene therapy has changed therapeutic options for patients. For those with a human leukocyte antigen (HLA)-identical sibling, bone marrow transplantation is still the best treatment option, but this includes only a minority of those affected. For those without an HLA-identical match, gene therapy has comparable efficacy to PEG-ADA, does not require repetitive injections, and does not present the risk of development of neutralizing antibodies to the bovine enzyme.

NEURODEGENERATIVE DISEASES: EXTENSION OF PRINCIPLE

The SCID trials gave support to the hypothesis that gene transfer into HSCs could be used to treat any disease for which allogeneic bone marrow transplantation was therapeutic. Moreover, the use of genetically modified autologous cells carried several advantages including no risk of graft-versus-host disease, guaranteed availability of a "donor" (unless the disease itself damages the stem cell population of the patient), and low likelihood of failure of engraftment. Cartier and Aubourg capitalized on this realization to conduct the first trial of lentiviral vector transduction of HSCs for a neurodegenerative disorder, X-linked adrenoleukodystrophy (ALD). X-linked ALD is a fatal demyelinating disease of the central nervous system caused by mutations in the gene encoding an adenosine triphosphate–binding cassette transporter. Deficiency of this protein leads to accumulation of very-long-chain fatty acids in oligodendrocytes and microglia, disrupting myelin maintenance by these cells. Affected boys present with clinical and neuroradiographic evidence of disease at age 6–8 and usually die before adolescence. Following lentiviral transduction of autologous HSCs in young boys with the disease, dramatic stabilization of disease occurred, demonstrating that stem cell transduction could work for neurodegenerative as well as immunologic disorders. Investigators in Milan carried this observation one step further to develop a treatment for another neurodegenerative disorder that has previously responded poorly to bone marrow transplantation. Metachromatic leukodystrophy is a lysosomal storage disorder caused by mutations in the gene encoding arylsulfatase A (ARSA). The late infantile form of the disease is characterized by progressive motor and cognitive impairment, and death within a few years of onset, due to accumulation of the ARSA substrate sulfatide in oligodendrocytes, microglia, and some neurons. Recognizing that endogenous levels of production of ARSA were too low to provide cross-correction by allogeneic transplant, Naldini and colleagues engineered a lentiviral vector that directed supraphysiologic levels of ARSA expression in transduced cells. Transduction of autologous HSCs from children born with the disease, at a point when they were still presymptomatic, led to preservation and continued acquisition of motor and cognitive milestones at time periods as long as 32 months after affected siblings had begun to lose milestones. These results illustrate that the ability to engineer levels of expression can allow gene therapy approaches to succeed where allogeneic bone marrow transplantation cannot. It is likely that a similar approach will be used in other neurodegenerative conditions.

Transduction of HSCs to treat the hemoglobinopathies is an obvious extension of studies already conducted but represents a higher hurdle in terms of the extent of transduction required to achieve a therapeutic effect. Trials are now under way for thalassemia and sickle cell disease, and for a number of other hematologic disorders, including Wiskott-Aldrich syndrome, and chronic granulomatous disease.

LONG-TERM EXPRESSION IN GENETIC DISEASE: IN VIVO GENE TRANSFER WITH RECOMBINANT ADENO-ASSOCIATED VIRAL VECTORS

Recombinant adeno-associated viral (AAV) vectors have emerged as attractive gene delivery vehicles for genetic disease. Engineered from a small replication-defective DNA virus, they are devoid of viral coding sequences and trigger very little immune response in experimental animals. They are capable of transducing nondividing target cells, and the donated DNA is stabilized primarily in an episomal form, thus minimizing risks arising from insertional mutagenesis. Because the vector has a tropism for certain long-lived cell types, such as skeletal muscle, the central nervous system (CNS), and hepatocytes, long-term expression can be achieved even in the absence of integration.

FIRST LICENSED PRODUCT

These features of AAV were used to develop the first licensed gene therapy product in Europe, an AAV vector for treatment of the autosomal recessive disorder lipoprotein lipase (LPL) deficiency. This rare disorder (1–2/million) is due to loss-of-function mutations in the gene encoding LPL, an enzyme normally produced in skeletal muscle and required for the catabolism of triglyceride-rich lipoproteins and chylomicrons. Affected individuals have lipemic serum and may have eruptive xanthomas, hepatosplenomegaly, and in some cases, recurrent bouts of acute pancreatitis. Clinical trials demonstrated the safety of intramuscular injection of AAV-LPL and its efficacy in reducing frequency of pancreatitis episodes in affected individuals, leading to drug approval in Europe. Additional clinical trials currently under way that use AAV vectors in the setting of genetic disease include those for muscular dystrophies, spinal muscular atrophy, Parkinson's disease, Batten's disease, ornithine transcarbamylase deficiency, hemophilia B and A, several forms of congenital blindness, and a variety of other inherited conditions.

HEMOPHILIA

Hemophilia **(Chap. 61)** has long been considered a promising disease model for gene transfer, because the gene product does not require precise regulation of expression and biologically active clotting factors can be synthesized in a variety of tissue types, permitting latitude in the choice of target tissue. Moreover, raising circulating factor levels from <1% (levels seen in those severely affected) into the range of 5% greatly improves the phenotype of the disease. Preclinical studies with recombinant AAV vectors infused into skeletal muscle or liver have resulted in long-term (>5 years) expression of factor VIII or factor IX in the hemophilic dog model. Administration to skeletal muscle of an AAV vector expressing factor IX in patients with hemophilia B was safe and resulted in long-term expression as measured on muscle biopsy, but circulating levels never rose to >1% for sustained periods, and a large number of IM injections (>80–100) was required to access a large muscle mass. Intravascular vector delivery has been used to access large areas of skeletal muscle in animal models of hemophilia and will likely be tested as a route of administration for muscular dystrophy disorders in upcoming trials.

The first trial of an AAV vector expressing factor IX delivered to the liver in humans with hemophilia B resulted in therapeutic circulating levels at the highest dose tested, but expression at these levels (>5%) lasted for only 6–10 weeks before declining to baseline (<1%). A memory T cell response to the viral capsid, present in humans but not in other animal species (which are not natural hosts for the virus), likely led to the loss of expression (Table 458-3). In response to these findings, a second trial included a short course of prednisolone, to be administered if factor IX levels began to decline. This approach resulted in long-term expression (>7 years, with observation ongoing) of factor IX, in the range of 2–7%, in men with severe hemophilia B. By using as the transgene a high specific activity variant of FIX, it has been possible to lower the vector dose required, reducing the risk of the immune response, and increasing the plateau levels of FIX into the range of 15–45%. Current efforts are focused on expanding these trials, and extending the approach to hemophilia A.

RETINAL GENE THERAPY

A logical conclusion from the early experience with AAV in liver in the hemophilia trial was that avoidance of immune responses was key to long-term expression. Thus immunoprivileged sites such as the retina began to attract substantial interest as therapeutic targets. This inference has been elegantly confirmed in the setting of the retinal degenerative disease Leber congenital amaurosis (LCA). Characterized by early-onset blindness, LCA was not previously treatable and is caused by mutations in several different genes; ~10% of cases of LCA are due to a mutation in a gene, *RPE65*, encoding a retinal pigment epithelial-associated 65-kDa protein. In dogs with a null mutation in *RPE65*, sight was restored after subretinal injection of an AAV vector expressing *RPE65*. Transgene expression appears to be stable, with the first animals treated >10 years ago continuing to manifest electroretinal and behavioral evidence of visual function. A Phase 3 trial, the first randomized controlled trial in human gene therapy, of an AAV vector encoding RPE65 was completed, and demonstrated improvement in multiple measures of retinal and visual function. This product has now been licensed by the U.S. Food and Drug Administration (FDA), and is the first licensed AAV gene therapy product in the United States. Trials for other inherited retinal degenerative disorders such as choroideremia are under way, as are studies for certain complex acquired disorders such as age-related macular degeneration, which affects several million people worldwide. The neovascularization that occurs in age-related macular degeneration can be inhibited by expression of vascular endothelial growth factor (VEGF) inhibitors such as angiostatin or through the use of RNA interference (RNAi)-mediated knockdown of VEGF. Early-phase trials of an AAV vector designed to achieve long-term inhibition of the biological effects of VEGF through a soluble VEGF receptor, however, failed to provide convincing evidence of efficacy, illustrating the challenges of developing genetic approaches for complex acquired disorders.

GENE THERAPY FOR CANCER

The majority of clinical gene transfer experience has been in subjects with cancer. The intent has been to increase the precision of cancer therapies and thereby make them less toxic and more effective. Most approaches have either modified the tumor directly, or altered the host's response to the malignancy to produce immune effector cells that are precisely targeted to the tumor phenotype.

MODIFYING THE CANCER

Since cancer is an (acquired) genetic disorder, initial efforts were directed at correcting the genetic deficits of the tumor or introducing lethal genes. Two major and persisting obstacles, however, are the poor biodistribution and transduction efficiency of all currently available vectors, and the heterogeneity and genetic instability of the tumor targets themselves, so that correction of single driver mutations does not preclude the evolution of a resistant population.

Tumor Correction One widely used direct intratumoral approach was adenoviral-mediated expression of the tumor suppressor p53, which is mutated in many different cancers. Initial studies showed some complete and partial responses in squamous cell carcinoma of the head and neck, esophageal cancer, and non-small cell lung cancer, but as yet there have been no successful product licensing studies for this approach except in China.

Pro-Drug Metabolizing Genes Efforts to overcome the above limitations have included the introduction of a prodrug or a suicide gene that would increase sensitivity of tumor cells to cytotoxic drugs. A frequently used strategy has been intratumoral injection of an adenoviral vector expressing the thymidine kinase (*TK*) gene. Cells that take up and express the *TK* gene can be killed after the administration of gancyclovir, which is phosphorylated to a toxic nucleoside by *TK*. The advantage of this approach is that the effects of transducing even a limited number of tumor cells are amplified by the spread of active drug to adjacent tumor cells. Although the approach continues to be examined in aggressive brain tumors and locally recurrent prostate, breast, and colon tumors, progress remains slow, and systemic benefits against metastatic disease have not been established.

MODIFYING THE HOST

Recruiting the Immune System The successful use of monoclonal antibodies that produce anti-tumor activity by activating the immune response has demonstrated the feasibility of manipulating the immune system to recognize the abnormal pattern of antigen expression on tumor cells. Immune cells are capable of almost unlimited expansion and persistence and can provide long-term tumor control. They can also traffic to tumor sites irrespective of location and, in principle, have the potential to evolve with the changing pattern of tumor cell phenotype and function.

Vaccination This strategy promotes more efficient recognition of tumor cells by the immune system, but the development of a therapeutic as opposed to the preventative vaccines required to combat infectious diseases has proved to be a considerable challenge. Approaches have included transduction of tumor cells with immune-enhancing genes encoding cytokines, chemokines, or co-stimulatory molecules, and the ex vivo manipulation of dendritic cells to enhance the presentation of tumor antigens. A dendritic cell vaccine for treatment of recurrent prostate cancer has received approval in the United States but its limited potency and high cost constrained commercial success.

Adoptive Cell Transfer Host immune cells such as T cells, NK cells, and others can be modified to express new transgenic receptors intended to recognize tumor cells and their microenvironment (Fig. 458-1). Retargeting may use a modification of the cells' own receptor or a synthetic chimeric antigen receptor (CAR) that is usually composed of the antigen recognition portion of an antibody and the signaling components of the cell's native antigen receptor. Both approaches have been successful, with significant responses reported with native receptors targeted to melanoma and synovial cell sarcoma and—most dramatically—with CARs targeted to CD19, an antigen expressed at high levels on normal and many malignant B cells. Infused CAR T cells can expand many thousand fold in vivo, persist long term, and have produced >90% complete response rates when targeting intractable B-acute lymphoblastic leukemia. Many responses are sustained long term and the approach has been licensed by the U.S. FDA. Broader application of adoptive T cell approaches is limited by: (1) The immune inhibitory microenvironment associated with most tumors; recent studies further modify the T cells with countermeasures to tumor inhibitory signals; (2) Acute and sometimes fatal systemic inflammatory and neurological toxicities during the phase of T-cell

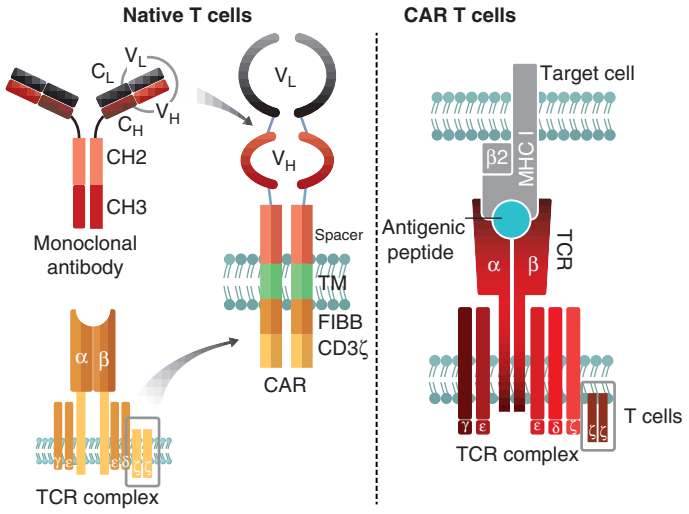


FIGURE 458-1 T cell receptors. A native T cell receptor (TCR) recognizes processed peptide antigens bound to major histocompatibility (MHC) molecules through its αβ chains. Signaling then occurs through a multichain intracellular CD3 complex. A chimeric antigen receptor (CAR) usually contains an extracellular receptor component derived from the antigen binding portion (V_H and V_L) of a monoclonal antibody. This produces a receptor that can recognize either protein or non-protein antigens independent of the MHC. A transmembrane (TM) domain then connects this receptor to the ζ chain of the CD3 complex derived from the native TCR.

expansion and tumor killing; (3) The off-target/or on-target but off-tumor effects that may damage normal host tissues (such as normal B cells following CD19 CAR therapy); (4) The cost, time, and complexity of manufacture; a particular problem when antigens unique to each tumor's individual mutations are targeted (neoantigens), rather than widely shared tumor associated antigens.

Non-Immunological Modifications to Host Gene transfer can be used to protect normal cells from the toxicities of chemotherapy and thereby increase the therapeutic index of these drugs. The most extensively studied approach has been to transduce hematopoietic cells with genes encoding resistance to chemotherapeutic agents, including the multidrug resistance gene *MDRI* or the gene encoding O^6-methylguanine DNA methyltransferase (*MGMT*). Although such approaches reduce hematologic toxicity, cytotoxic dose escalation quickly reveals dose-limiting toxicities to other organ systems.

Finally, gene transfer can be used to inhibit the host angiogenesis required for tumor support, for example by constitutive expression of inhibitors such as angiostatin and endostatin, or the transfer of T cells genetically modified to recognize antigens specific to newly forming vasculature. These studies are early-phase.

TABLE 458-4 Taking History from Subjects Enrolled in Gene Transfer Studies

Elements of History for Subjects Enrolled in Gene Transfer Trials
1. What vector was administered? Is it predominantly integrating (retroviral, lentiviral, herpesvirus [latency and reactivation]) or nonintegrating (plasmid, adenoviral, adeno-associated viral)?
2. What was the route of administration of the vector?
3. What was the target tissue?
4. What gene was transferred in? A disease-related gene? A marker?
5. Were there any adverse events noted after gene transfer?
Screening Questions for Long-Term Follow-Up in Gene Transfer Subjects[a]
1. Has a new malignancy been diagnosed?
2. Has a new neurologic/ophthalmologic disorder, or exacerbation of a preexisting disorder, been diagnosed?
3. Has a new autoimmune or rheumatologic disorder been diagnosed?
4. Has a new hematologic disorder been diagnosed?

[a]Factors influencing long-term risk include: integration of the vector into the genome, vector persistence without integration, and transgene-specific effects.

COMBINATION APPROACHES—MODIFICATION OF HOST AND TUMOR BY VIROTHERAPY

Immuno Oncolytic Viruses These viruses are genetically modified to replicate in malignant but not normal cells. The replicating vectors thus proliferate and spread within the tumor, facilitating eventual tumor clearance. However, physical limitations to viral spread, including fibrosis, intermixed normal cells, basement membranes, and necrotic areas within the tumor, may reduce clinical efficacy, and their activity against metastatic disease has proved limited. Recently, the FDA granted licensing approval to talimogene laherparepvec, an oncolytic herpes virus containing the human GM-CSF gene, for treatment of melanoma. This success has led to resurgent interest in combining the local tumor destruction and tumor antigen release mediated directly by oncolytic viruses with the recruitment of a systemic immune response mediated by immunostimulatory genes contained within the oncolytic virus. In principle such immune-oncolytic viruses should produce responses in both local and metastatic disease. Numerous novel viral agents are now entering early phase clinical testing.

OTHER APPROACHES

This chapter has focused on gene addition therapy, in which a normal gene is transferred to a target tissue to drive expression of a gene product with therapeutic effects. Another powerful technique under development is genome editing, in which a mutation is corrected in situ, generating a wild-type copy under the control of the endogenous regulatory signals. This approach makes use of novel reagents including zinc finger nucleases, TALENs and CRISPR, which introduce double-stranded breaks into the DNA near the site of the mutation and then rely on a donated repair sequence and cellular mechanisms for repair of double-strand breaks to reconstitute a functioning gene. These approaches have only recently entered the stage of clinical investigation. Another strategy recently introduced into clinical trials is the use of siRNAs or short hairpin RNAs as transgenes to knock down expression of deleterious genes (e.g., mutant huntingtin in Huntington's disease or genes of the hepatitis C genome in infected individuals).

SUMMARY

The power and versatility of gene transfer approaches are such that there are few serious disease entities for which gene transfer therapies are *not* under development. The development of new classes of therapeutics typically takes two to three decades; monoclonal antibodies and recombinant proteins are recent examples. Gene therapeutics, which entered clinical testing in the early 1990s, traversed the same time course. Examples of clinical success are now abundant, and gene therapy approaches are likely to become increasingly important as a therapeutic modality in the twenty-first century. A central question to be addressed is the long-term safety of gene transfer, and regulatory agencies have mandated a 15-year follow-up for subjects enrolled in gene therapy trials (Table 458-4). Realization of the therapeutic benefits of modern molecular medicine will depend on continued progress in gene transfer technology.

FURTHER READING

Aiuti A, Roncarolgo MG: Ten years of gene therapy for primary immune deficiencies. Hematology Am Soc Hematol Educ Program 2009:682, 2009.

Fesnak AD et al: Engineered T cells: The promise and challenges of cancer immunotherapy. Nat Rev Cancer. 16:566, 2016.

Hocquemiller M et al: Adeno-associated virus-based gene therapy for CNS diseases. Hum Gene Ther 27:478, 2016.

Klebanoff CA et al: Prospects for gene-engineered T cell immunotherapy for solid cancers. Nat Med 22:26, 2016.

Kumar SR et al: Clinical development of gene therapy: Results and lessons from recent successes. Mol Ther Methods Clin Dev 3:16034, 2016.

Rai P, Malik P: Gene Therapy for hemoglobin disorders—a mini-review. J Rare Dis Res Treat 1:25, 2016.

Ramos CA et al: CAR-T cell therapy for lymphoma. Annu Rev Med 67:165, 2016.

459 The Human Microbiome

Neeraj K. Surana, Dennis L. Kasper

"All disease begins in the gut."

–Hippocrates

Nearly two and a half millennia after Hippocrates made this statement, we are just coming to truly appreciate its profundity. Since the beginning of humankind, scholars have been investigating the underpinnings of disease with an almost singular focus on the human side of the equation. Microbes were not recognized as an important cause of disease until the inception of the "germ theory" in the late nineteenth century. During the first century of medical microbiology, research largely centered on the role of microbes as pathogens. Only recently has there been a resurgence of interest in understanding how commensal organisms—the bacteria, viruses, fungi, and Archaea that

make up the *microbiota*—impact human physiology. The idea that these microorganisms are vital to the well-being of humans has challenged our traditional notions of "self." Indeed, a human being can most accurately be described as a *holobiont*: a complex assemblage of human cells and microorganisms interacting in an elaborate *pas de deux* that drives normal physiologic processes.

Aimed at a better understanding of this relationship, myriad studies during the past decade have begun to catalogue the microbiota at various body sites and in a multitude of disease conditions. Diseases in virtually every organ system have been associated with changes in the microbiota. Indeed, the microbiota has been linked to intestinal disorders, disturbances in metabolic function, autoimmune diseases, and psychiatric conditions and has been shown to influence susceptibility to infection and the efficacy of pharmaceutical therapies. Knowledge of the specific mechanism(s) underlying most of these microbe–disease associations is lacking; it remains unclear whether the disease-associated alterations in the microbiota represent mere biomarkers of disease, a causal relationship, or a combination of the two. Although cause-and-effect relationships are still being elucidated for many diseases, it is clear that humans coexist in an intricate relationship with commensal organisms. This chapter explores in detail the nature of these host–commensal interactions, focusing on how this information might be translated into clinically meaningful interventions.

HISTORICAL PERSPECTIVE

Massive undertakings, such as the Human Microbiome Project (HMP) sponsored by the National Institutes of Health and MetaHIT sponsored by the European Commission, have catalogued all the bacteria present at multiple body sites in people with and without disease. Coupled with the confluence of advances in sequencing technologies **(Chap. 474)**, gnotobiotic animal availability, and microbial culture, significant progress has been made toward an understanding of the interplay between the microbiota and human health. However, recent findings were foreshadowed by work done centuries ago.

The human microbiota was first explored in 1683 when Antony van Leeuwenhoek described in a letter to the Royal Society of London the "very little living animalcules, very prettily a-moving" that he had observed in the plaque between his teeth. Leeuwenhoek went on to perform the first comparative "microbiota" studies by assessing how fecal and oral bacteria differ, how oral microbes change in the setting of disease (e.g., alcoholism and tobacco use), and how microbial composition changes across the age spectrum (e.g., in young children versus old men). He attempted—unsuccessfully—to eliminate these bacteria. Although Leeuwenhoek was not taken seriously when he first reported his findings, his studies laid the groundwork for what is now the field of microbiome research, and investigators are still trying to answer many of the same overarching questions that he raised more than three centuries ago.

Although Leeuwenhoek first reported the existence of bacteria and their association with humans at the end of the seventeenth century, the significance of commensal bacteria was not realized until late in the nineteenth century. In 1885, Pasteur suggested that animals could not survive if they were "artificially and completely deprived of the common microbes." Although Pasteur's preconceived ideas were proven incorrect in 1912 by the advent of germ-free animals (animals raised without exposure to any microorganisms), the underlying concept that commensal organisms are critical to health has held up. Élie Metchnikoff made another conceptual advance in this field by suggesting at the beginning of the twentieth century that clinical outcomes could be altered by the administration of specific beneficial organisms (*probiotics*). In particular, Metchnikoff believed that aging was caused by toxic bacteria in the gut and that lactic acid–producing bacteria (e.g., *Lactobacillus* species) present in sour milk and yogurt could mitigate against this process. The data behind this specific claim are still lacking, but recent discoveries offer continued hope that the microbiome can be effectively harnessed to protect against and treat a variety of diseases. Thus, although the field of microbiome research is sometimes considered to have emerged over the last one or two decades, the basic tenets—that the microbiota varies according to body site and clinical characteristics, that microbes are critical for human health, and that specific modulation of the microbiota may lead to improved clinical outcomes—are far from new.

A PRIMER ON TAXONOMY

Given that microbiome-based studies have identified and compared microbes at different levels of taxonomic resolution **(Fig. 459-1)**, some understanding of taxonomy is essential for better comprehension of the implications of these studies. Of the ~100 bacterial phyla that exist in nature, only five (Actinobacteria, Bacteroidetes, Firmicutes, Fusobacteria, and Proteobacteria) are dominant members of the human microbiome. Each of these phyla can be further categorized into multiple classes, orders, families, genera, and species. Early studies on the microbiota focused on changes in the relative abundance at the phylum level between different groups (e.g., obese versus normal-weight patients); however, these comparisons are at such a broad taxonomic level that they often provide little or no biological insight. As illustrated in Fig. 459-1, drawing comparisons between organisms in two different bacterial phyla is analogous to comparing humans to sea stars: the evolutionary distance between the two is tremendous. The limitations of current bioinformatic tools require lumping together of taxonomically related strains and thus cloud the richness of microbial ecology. Examining microbial profiles at the phylum, family, or even genus level—as is often done at present—ignores the great heterogeneity within different strains of the same bacterial species. The analytical pipelines are just now beginning to enable strain-level comparisons, and these improvements will likely facilitate our ongoing investigation of host–commensal interactions.

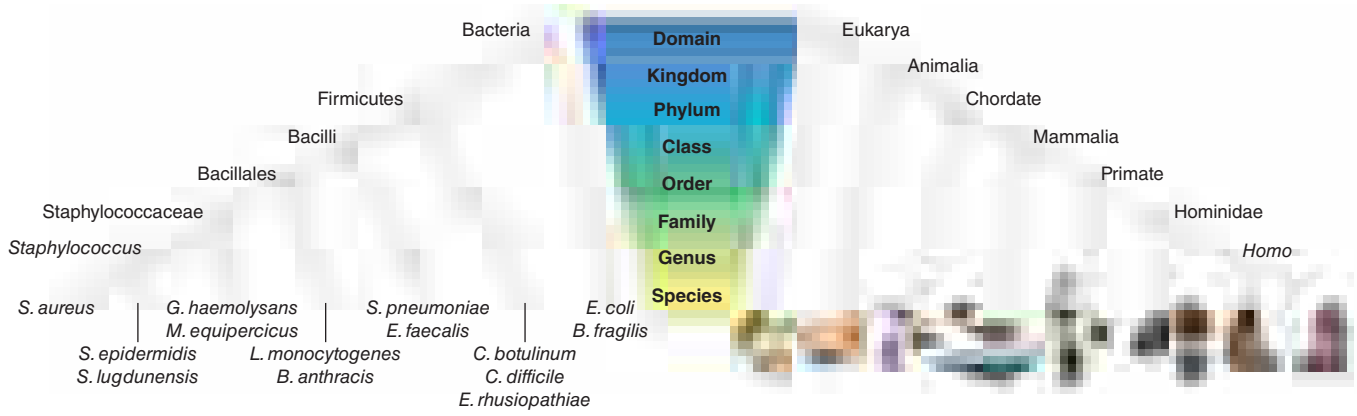


FIGURE 459-1 Juxtaposition of bacterial and human taxonomy highlights the evolutionary distance between different taxonomic levels. The listed species represent exemplars that are members of the taxon to which they are connected but that are not contained within the next-lower-level taxon listed. For example, *Clostridium botulinum*, *Clostridium difficile*, and *Erysipelothrix rhusiopathiae* are members of the phylum Firmicutes, but are in classes other than Bacilli. Similarly, starfish and humans are both members of the kingdom Animalia, but they are in different phyla.

THE MICROBIOTA AND HUMAN HEALTH

OVERVIEW OF THE HUMAN MICROBIOTA

The overwhelming majority of microbiota studies have focused on stool, given that this sample type represents the most ecologically rich anatomic site, is easy to obtain, and can readily be followed longitudinally in the same individual. A landmark study by the HMP sought to define the "normal" microbiota throughout the entire body in healthy Western adults. To this end, the microbial populations at 15–18 body sites were characterized in 242 people. One striking finding was that all samples from a given body region (e.g., skin) were more similar to each other than they were to samples from a different body region (e.g., stool), even in the same individual **(Fig. 459-2*A*)**. In essence, the effect of the anatomic site on microbial composition is far greater than the effect of heterogeneity between individuals. That said, there was a remarkable amount of inter-individual variation at any given body site **(Fig. 459-2*B*)**. In stool, for example, the abundance of the phylum Bacteroidetes ranged from ~10% in some individuals to >90% in others. Remarkably, even with person-to-person variability and differences among body sites, the functional capacity of the microbiota—assessed using metagenomic data to identify gene pathways—was quite similar across different people and different body sites **(Fig. 459-2*C*)**. This discrepancy between the substantial differences in microbial composition and the little or no resulting change in the functional properties of the microbiota reflects an important ecologic property of the microbiota: the microbial communities at different body sites and in different people assemble in such a way that all the core metabolic functions are maintained. This finding also hints at the likely possibility of significant functional redundancy within the microbiota, with different species executing the same biological functions at different anatomic sites.

While the HMP provided the first large-scale catalogue of the microbiome in multiple people and at many different body sites, the amount of data generated by what, at the time, was by far the largest microbiome study has been dwarfed by subsequent studies. These more recent studies have confirmed the HMP's major tenets: the composition of the microbiota differs by body site, there is tremendous inter-individual variation, and the microbial gene content is relatively conserved irrespective of the body site or individual. No microbial species are ubiquitous in all individuals and at all body sites, but some species are highly prevalent at a given body site: in the HMP study, *Staphylococcus epidermidis* was present in 93% of nares samples and *Escherichia coli* in 61% of stool samples. These findings highlight the remarkable personalization of the human microbiome. While the human genome is typically >99.5% identical in different people, the microbiotas of two individuals may not overlap at all. Although the "precision medicine" approach currently focuses on teasing out how differences in the human genome relate to different clinical end points, the human microbiome clearly represents a critical component for consideration.

THE MICROBIOTA BY THE NUMBERS

It has long been known that the human-associated microbiota is numerically dense. Leeuwenhoek estimated that there were more "animals living in the scrum on the teeth in man's mouth than there are men in a kingdom." Specific enumeration of the components of the microbiota has been challenging, in part because of its variability across time,

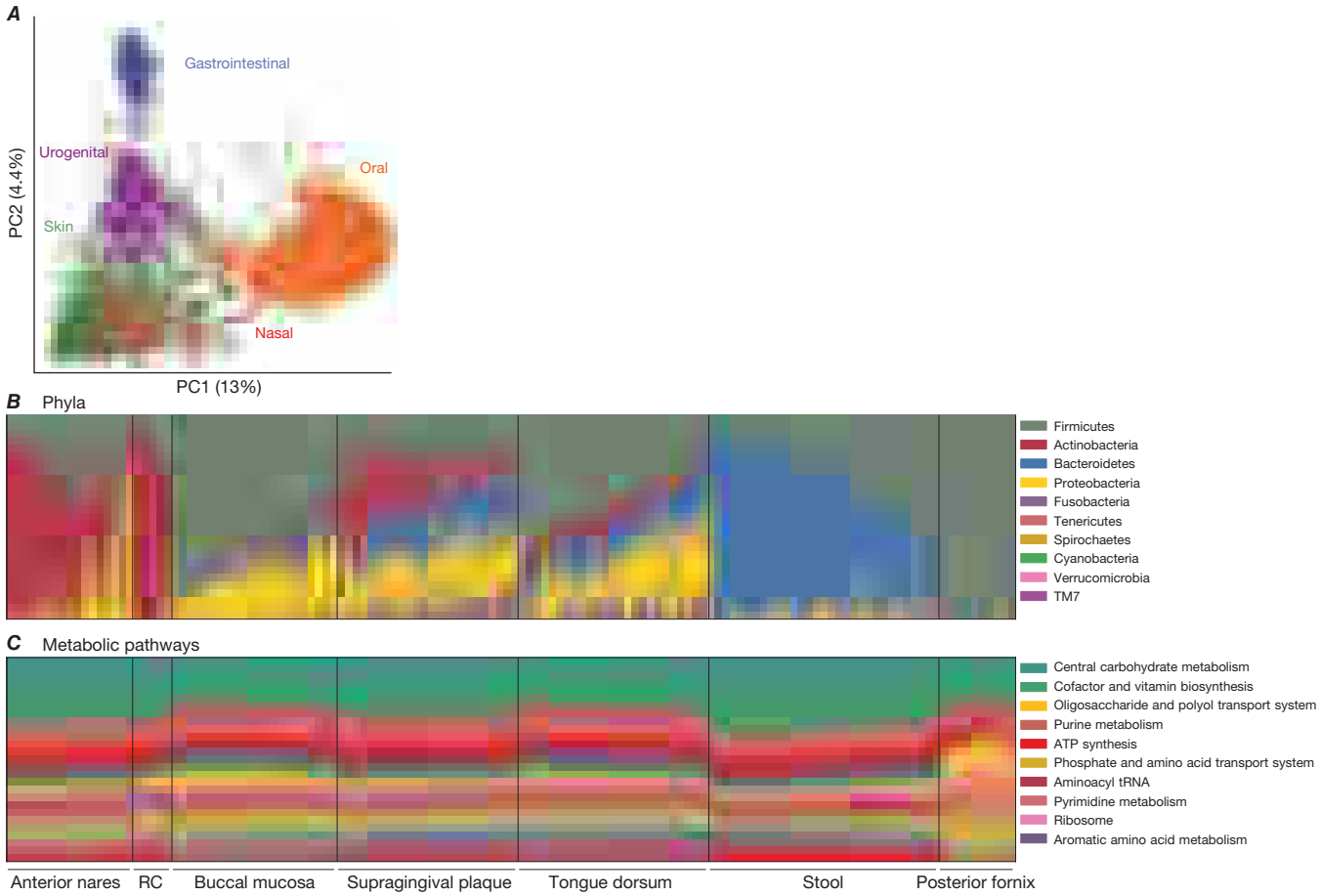


FIGURE 459-2 The human microbiome exhibits significant taxonomic variability among body sites and between individuals while maintaining core metabolic pathways. ***A.*** Principal coordinates (PC) plot showing variation among samples demonstrates that primary clustering is by body area, with the oral, gastrointestinal, skin, and urogenital habitats separate; the nares habitat bridges oral and skin habitats. Each circle represents an individual sample. ***B, C.*** Vertical bars represent microbiome samples by body habitat, with each bar within a given body site representing a different individual. Bars indicate relative abundances colored by microbial phyla (***B***) and metabolic pathways (***C***). The legend on the right indicates the most abundant phyla/pathways. RC, retroauricular crease. *(Reprinted by permission from Macmillan Publishers Ltd: Human Microbiome Project Consortium: Structure, function and diversity of the healthy human microbiome. Nature 486:207, 2012.)*

space (body region), and clinical conditions. Moreover, the majority of human-associated microbes are not readily cultivable—a situation that raises questions about the best methodology for such quantitation. Initial back-of-the-envelope calculations performed in the 1970s suggested that there were roughly tenfold more bacteria in the body than there were human cells. This rather astounding estimate suggested that humans are really only ~10% "human" and that by far the greatest part of the holobiont is represented by microbes. This stark numerical discrepancy has prompted some to question "who parasitizes whom." However, a more recent estimate has suggested that there are "only" ~1.3 times more bacteria in the body than there are human cells and thus that humans are ~56% "bacterial." Of note, this more recent study does not include the numbers of viruses (known to generally be approximately tenfold more abundant than other microbes), fungi, or Archaea. Given these additional microorganisms, the notion that microbes constitute >90% of the cells present in a human body is likely correct. These ratios are even starker when one considers the genetic potential of human cells versus that of commensal organisms. In contrast to the ~20,000 genes in the human genome, the estimated total number of genes in the microbiota (which together constitute the *microbiome*)—i.e., >2,000,000—indicates that the human genome contributes <1% to the total genetic potential of the overall holobiont. Most microbiome studies to date have focused almost exclusively on the bacterial component; much remains to be learned about the functional interplay of bacteria, viruses, fungi, and Archaea and how these other classes of microorganisms impact human health.

In terms of overall diversity, >10,000 different bacterial species are present in the human microbiota; the intestines alone contain >1000 species. At any given time, the body of any given individual harbors 500–1000 bacterial species, with 100–200 bacterial species in the gut alone. If one considers different strains of the same bacterial species, which may be functionally different from one another, the diversity of the microbiota is probably at least a magnitude greater. Although marked diversity exists at the strain and species level, only limited bacterial phyla are generally found in the human microbiota at any given body site **(Fig. 459-3)**.

INFLUENCES ON THE MICROBIOTA

An individual's specific microbial configuration is dynamic and is quickly altered in response to subtle changes in the microenvironments in which the bacteria reside. On a day-to-day basis, these changes usually reflect alterations in the relative abundance of the various microbes. However, some exposures have a greater effect on the microbiota and can shift the microbial population to a new equilibrium via the loss of specific species and/or the acquisition of others; this new microbial equilibrium can be associated with either health or a disease state **(Fig. 459-4)**. Identification of the factors that influence the microbiota's composition is critical to an understanding of what leads to and controls intra- and inter-individual variation. Moreover, an understanding of the influences on the microbiota will facilitate the design and proper interpretation of microbiota studies. While it is clear that the microbiota can be altered through these various mechanisms, it is not yet clear whether these changes are biologically significant.

Genetics Studies of monozygotic and dizygotic twins have revealed that host genetics have a small but statistically significant effect on the microbiota's composition. Notably, some taxa, such as *Christensenella* species, are more heritable than others. Thus, some inter-individual variation may be due to underlying differences in host genetics; however, it is likely that other factors explain more of the observed variability. That said, the host's genetic contribution to the microbiota, albeit small, may be meaningful. Studies in mice have demonstrated that genetic variation in the major histocompatibility complex, a specific set of immune-related genes, leads to changes in the microbiota that alter susceptibility to an autoimmune disease. These studies offer a proof of concept for the notion that the genetic predisposition observed for certain diseases may actually be mediated by indirect alterations in the microbiota.

Age Burgeoning evidence now indicates that microbial exposure begins in utero: bacterial DNA has been identified in otherwise healthy placentas, in amniotic fluid obtained at early stages of gestation, and in meconium of term newborns. Although some controversy persists about whether these results reflect contamination and/or the presence of nonviable bacteria, they raise the possibility that human exposure to the microbial world begins before birth. The delivery mode (vaginal versus cesarean section) and the method of feeding (breast milk versus formula, timing of solid food introduction) are major determinants of an infant's early microbiota. After birth, the infant's microbiota goes through a stereotyped succession process; with increases in bacterial diversity and functional capacity, the child's microbiota resembles that of an adult by the age of 2–3 years. Cross-sectional studies that have examined the microbiota across the entire age spectrum have revealed a general stability of the fecal microbiota after 2–3 years of age; however, the microbiota of the elderly (persons >80 years of age) demonstrates notable differences from those of their younger counterparts, with increases in *Bacteroides* and *Eubacterium* species and decreases in the bacterial family Lachnospiraceae.

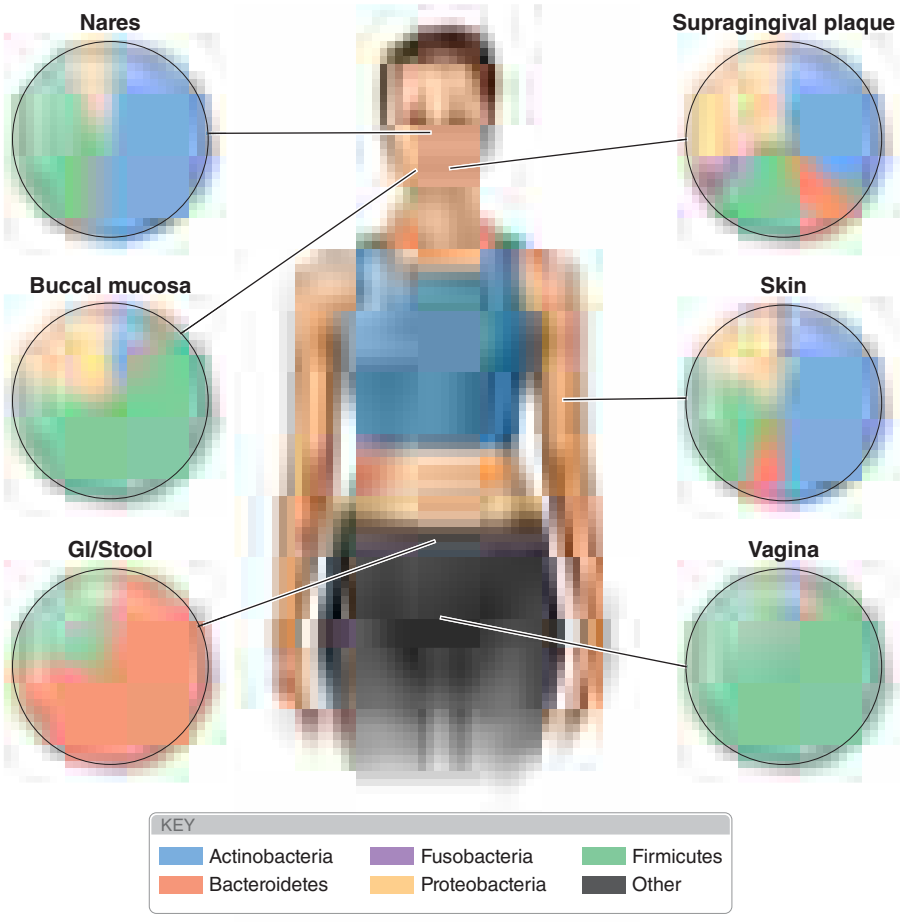


FIGURE 459-3 Different anatomic sites harbor very different microbiomes. The figure indicates the relative proportion of sequences determined at the taxonomic phylum level at six anatomic sites. *(Data for stool, vagina, nares, buccal mucosa, and supragingival plaque are from the Human Microbiome Project; data for the skin is from EA Grice et al: Topographical and temporal diversity of the human skin microbiome. Science 324:1190, 2009.)*

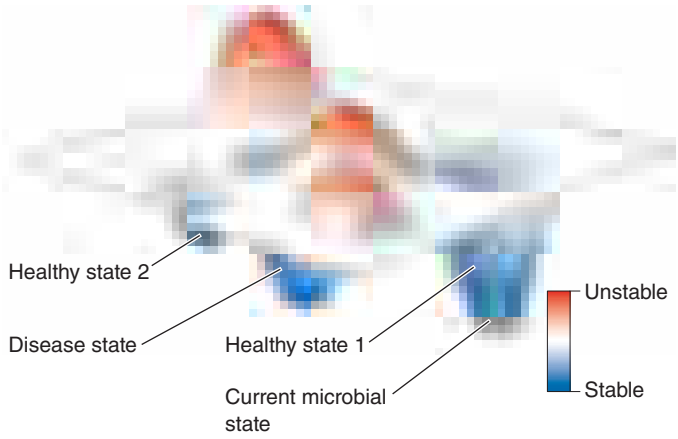


FIGURE 459-4 A stability landscape of the human microbial ecosystem. A stable state, illustrated as a depression in the landscape, can be associated with either a healthy state or a disease state. The topology of an individual's landscape reflects that person's genetics, age, diet, medications, medical history, and lifestyle. The position of the green ball represents the current microbial state. Clinical changes (e.g., administration of antibiotics, development of disease) can influence both the current state and the overall topology.

Diet Diet is a strong determinant of human health. The impact of diet is mediated, in part, by its effects on the composition of the gut microbiota. This makes intuitive sense, as the human diet provides nutrients needed not only by our own cells but also by the microbes living in the alimentary tract. In young children, this dietary influence is marked by major shifts (e.g., a decrease in *Bifidobacterium* species) in the intestinal microbiota that occur at weaning and with the introduction of solid food. In adults, long-term dietary patterns are associated with relatively stable microbial compositions. However, drastic changes in short-term macronutrient availability cause rapid (within 1 day) and reproducible fluctuations in the fecal microbiota that reflect the biological processes needed to degrade and metabolize the nutrients in the new diet. For example, vegetarian diets are associated with a microbiota that has an increased ability to metabolize plant polysaccharides (e.g., *Roseburia* species, *Eubacterium rectale*, *Ruminococcus bromii*), while animal-based diets result in an increased abundance of bile-tolerant organisms (e.g., *Alistipes*, *Bilophila*, and *Bacteroides* species). At the completion of dietary interventions and the resumption of the individual's normal dietary pattern, the microbial communities revert back to their previous states, probably because the individual resumes his or her typical diet. Taken together, dietary studies confirm that the microbiota is highly adaptable and varies in relation to changes in the diet. Of note, virtually all of these studies have focused on how the diet influences the fecal microbiota. It will be interesting to determine whether dietary changes similarly influence the microbiota at nonintestinal sites.

Drugs Virtually all drugs have the capacity to change the microbiota by altering the chemical landscape in which the microorganisms live (e.g., statins, bile acid sequestrants), modulating the host's ability to recognize and react to microbes (e.g., immunosuppressants), and/or directly interfering with the microbiota's constituents (e.g., antibiotics). These potential effects have made critical interpretation of microbiota studies much more difficult. A prominent study that claimed to identify a fecal microbiota signature associated with type 2 diabetes was later found actually to have identified a signature for patients taking metformin instead; the effects of this drug on the microbiota were far greater than the effects of the disease itself. These results highlight the importance of controlling for clinical variables in microbiota studies.

Antibiotics are the most obvious and best-studied class of drugs that modulate the microbiota. Multiple groups have demonstrated that antibiotics exert a considerable effect on the gut microbiota by depleting antibiotic-sensitive strains. What is more surprising is that many strains resistant to the antibiotic tested are also eliminated. This observation highlights the intricate microbe–microbe interactions that are fundamental to maintenance of the overall microbial community. For example, treatment with ciprofloxacin, which has little to no activity against clinically relevant anaerobes, leads to a loss of roughly one-third of the bacterial taxa in the gut. This broad effect is likely mediated by the depletion of certain "keystone" species that are required for the persistence of other, unrelated species. While many of the observed antibiotic effects (e.g., loss of specific taxa) are shared across many different individuals, some effects vary greatly among people. For example, studies found that microbiota recovery following antibiotic treatment differed significantly in terms of timing and degree. The microbiota of most healthy people who received ciprofloxacin for 5 days had completely recovered within 4 weeks, whereas microbiologic changes lasted up to 6 months in other individuals. Moreover, the degree of variation was compounded by repeated antibiotic administration, with fewer individuals reverting to their baseline microbiota after a second course of ciprofloxacin given 6 months after the first. These findings are consistent with those of microbial ecology experiments, which also showed that this type of repeated disturbance leads to less predictable results.

Lifestyle Many seemingly innocuous lifestyle decisions can impact the human microbiota. For example, a person's skin and fecal microbiotas are more similar to those of their household members, regardless of genetic relatedness, than to those of residents of different households. The degree of similarity in skin microbiotas is even greater if a dog also lives in the home; in contrast, the presence of a young child does not accentuate this microbial relatedness. The presumption is that the dog serves as a more effective "vector" for transmitting microbes during its frequent direct contact with adults in the household. The type of setting in which a person lives also impacts the microbiota. Living in a rural or farm setting leads to a different fecal microbiota than living in an urban environment. Similarly, the individual's country of residence affects the microbiota. An analysis of daily fecal samples from an individual who temporarily (i.e., for a couple of months) moved from the United States to Thailand demonstrated a large shift in the fecal microbiota that coincided with arrival in Thailand and a reversion in most respects to the "American" microbial configuration upon return to the United States. These geography-driven changes probably reflect a combination of environmental and dietary differences between locations.

Circadian Rhythms Many human biological processes follow a circadian clock; aspects of physiology are tuned by external cues, including the degree and timing of ambient light, temperature, and availability of nutrients. This endogenous biological clock enables animals to efficiently adapt to changing environmental conditions. Similarly, the microbiota maintains a circadian rhythm that is linked to the host's circadian clock. If circadian oscillations are disrupted in the host, they are similarly disrupted in the microbiota, and vice versa. These bacterial vacillations occur at the level of spatial localization within the intestine, relative species abundance, and bacterial metabolite secretion. Work in the 1960s showed that mice exhibited daily periodicity of susceptibility to infection with either *S. pneumoniae* or *E. coli* lipopolysaccharide. Although the fundamental basis for this difference was not known at the time, it is likely to be related, in part, to the microbial circadian clock. Derangements of these microbial oscillations have also been linked to the development of metabolic diseases and may underlie some of the health hazards associated with shift work and jet lag.

THE MICROBIOTA AND DISEASE

THE HYGIENE HYPOTHESIS

Over the past few decades, abundant epidemiologic data have revealed an inverse correlation between exposure to microbes and the incidence of autoimmune and/or atopic diseases (Fig. 459-5). This type of epidemiologic correlation led to the proposal of the "hygiene hypothesis" in 1989. Initially, this hypothesis focused on the development of atopic diseases in young children, with the idea that these epidemiologic observations could "be explained if allergic diseases were prevented by infection in early childhood, transmitted by unhygienic contact with older siblings, or acquired prenatally from a mother infected by contact

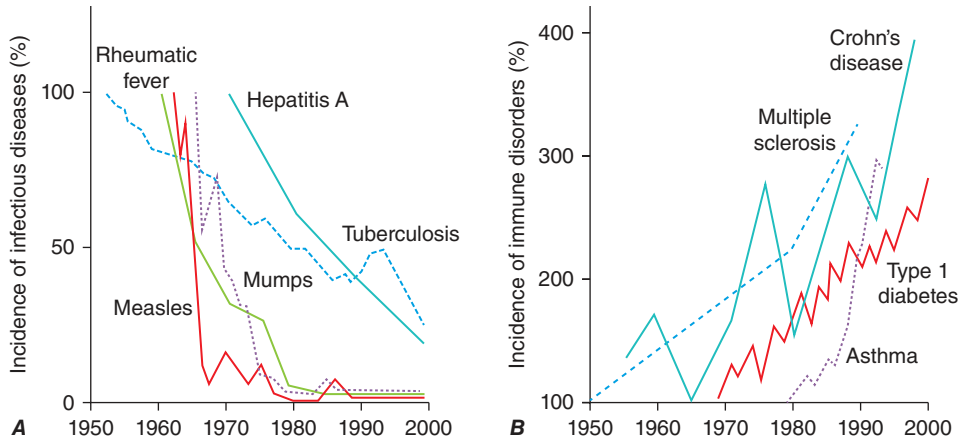


FIGURE 459-5 There was an inverse relationship between the incidence of select infectious diseases and the incidence of autoimmune disorders during the latter half of the twentieth century. *A*. Relative incidence of prototypical infectious diseases from 1950 to 2000. ***B***. Relative incidence of select autoimmune disorders from 1950 to 2000. *(From JF Bach: The effect of infections on susceptibility to autoimmune and allergic diseases. N Engl J Med 347:911, 2002. Copyright ©2002, Massachusetts Medical Society. Reprinted with permission from Massachusetts Medical Society.)*

with her older children."[1] In fact, this notion that differences in living conditions and environmental exposures contribute to susceptibility to hay fever (*summer catarrh*) dates back to at least the early nineteenth century. The hygiene hypothesis has continued to evolve over the past three decades and now posits that inadequacies in microbial exposure—in combination with genetic susceptibilities—lead to a collapse of the normally highly coordinated, homeostatic immune response. At its core, the hygiene hypothesis holds that specific early-life microbial exposures are required to prevent subsequent disease and that the "Westernization" of society has led to a decrease in such exposures. This concept is now being applied beyond atopic diseases to other inflammatory and autoimmune diseases and is thought to reflect processes that occur in later life as well.

RELATIONSHIP BETWEEN THE MICROBIOTA AND SPECIFIC DISEASE STATES

The ideas inherent in the hygiene hypothesis—in sum, that microbial exposure can affect long-term health outcomes—laid the theoretical foundation for translational microbiome studies. While most of the studies described earlier sought to describe how the microbiota responds to specific and often transient influences (e.g., a course of antibiotics, dietary interventions, travel), a multitude of studies have characterized the microbiota in patients with various diseases in the hope that a better understanding of the nature of disease-specific microbial communities will provide insight into disease pathogenesis and potentially uncover novel treatment modalities. Remarkably, virtually all of these studies have demonstrated differences between the microbiotas of healthy controls and patients, irrespective of the specific disease process examined. Although it is difficult to generalize across all studies, a couple of general themes have emerged. First, disease states are typically associated with microbiotas that are less diverse than those of healthy individuals. This loss of diversity can be measured either as a decrease in the number of species (*alpha diversity*; often measured as the number of operational taxonomic units, which are the bioinformatic equivalent of species) or as a reduction in the microbial relatedness of the species present (*beta diversity*). Often, both alpha and beta diversity decrease in the setting of disease. Second, states of inflammation—regardless of site or underlying disease process—are often associated with a relative increase in the abundance of the bacterial family Enterobacteriaceae and a decrease in the abundance of Lachnospiraceae.

Dissecting Correlation and Causality Given that most of these investigations have been designed as case–control studies, it is difficult to determine whether microbiologic findings are the cause or the effect of the disease. Even studies that examine treatment-naïve patients at the time of initial diagnosis are still confounded by this "chicken or egg" issue. Moreover, prospective, longitudinal clinical studies—still rare in the microbiome field—may simply yield correlations between the microbiome and subclinical disease rather than necessarily proving causality. Experiments in animals—specifically, studies using gnotobiotic mice (germ-free mice that have been colonized with specified microbial communities)—have been critical in this regard as they allow investigation of specific differences in microbial components while controlling for the host's genetics, diet, and housing conditions. Moreover, human microbes can be transplanted into gnotobiotic mice to permit in-depth mechanistic studies of how these microbial communities affect disease pathogenesis. This marriage of human samples and animal experiments has facilitated the identification of causal roles played by some microbes in disease pathogenesis; these findings provide a critical proof of concept for the interplay of the microbiota with human health. However, the vast majority of microbiome studies are still at the level of correlation. The next several sections describe the clinical and animal data for many different disease processes. Given the voluminous and rapidly changing nature of this field, it is impossible to cover all of the disease associations known to date; rather, the following discussion represents a combination of the leading exemplars of microbiome data and nascent areas of significant clinical interest. In all cases, the hope is that further study of the role of the microbiota will provide novel diagnostics, new therapeutic modalities, and/or additional insight into disease pathogenesis.

Gastrointestinal Diseases Given that the intestines harbor the largest number and greatest diversity of organisms in the body, much work has focused on how the microbiota impacts gastrointestinal diseases. Even though the luminal surface area of the gastrointestinal tract is 30–40 square meters (~90% of which is contained within the small intestine) and features marked anatomic and functional differences that result in many discrete macro- and micro-ecosystems, stool is often used as a surrogate for the intestinal microbiota given the relative ease of collecting samples. A few studies that have compared the microbial profile in stool with the mucosa-adherent organisms present in biopsy samples have demonstrated that stool is, in fact, a reasonable proxy for biopsy samples; however, the relative microbial "noise" present in stool can sometimes overwhelm the "signal," making biopsy samples more informative for some scientific questions. The key issue is to ensure that the biopsy samples evaluated represent relatively similar intestinal regions, as there are significant differences between the organisms present in the crypt and the tip of the villus and between microbes found in the ascending versus the descending colon.

OBESITY Obesity is a worsening epidemic throughout the world, and multiple studies have linked the composition of the intestinal microbiota to the development of obesity in animal models and in humans. Indeed, many of the initial translational microbiome studies

[1] D Strachan: BMJ 299:1259, 1989.

performed in mice at the beginning of the twenty-first century focused on obesity. These early studies suggested that the ratio of the relative abundance of Bacteroidetes to Firmicutes was lower in obese mice than in control animals. Moreover, a causal relationship between the microbiota and obesity was established by the finding that gnotobiotic mice colonized with the microbiota from obese individuals had more rapid and more extensive weight gain than gnotobiotic mice colonized with the microbiota from lean individuals. Biologically, it is posited on the basis of metagenomic surveys that the obesity-associated microbiome has an increased capacity to harvest energy from the diet. Notably, the relationship between the Bacteroidetes/Firmicutes ratio and obesity did not hold in initial human studies; however, the finding that this ratio increased in obese patients who lost weight while on a fat- or carbohydrate-restricted diet suggested some generalizability between mice and humans. Beyond this ratio of major bacterial phyla, obesity was linked to a microbiome with a lower alpha diversity. Over the past ~15 years, numerous human studies examining the relationship between the microbiome and obesity have been completed, all with mixed results. A recent meta-analysis of 10 studies including nearly 3000 individuals revealed an apparent lack of relationship between the Bacteroidetes/Firmicutes ratio and obesity, though there is ~2% lower diversity associated with obesity that is statistically significant but of unclear biological significance. This finding highlights a problem common to microbiome studies: i.e., there is no sense as to what magnitude of change is biologically meaningful. Ultimately, although murine studies have indicated a causal link between the microbiota and obesity, the human data are less convincing, and their significance may be limited by the studies' having primarily examined only high-level taxonomic information rather than also assessing transcriptional or metabolic differences.

The rise in obesity has elicited a plethora of ideas about the type of diet that might be most successful in leading to sustained weight loss. It has become clear that the same dietary ingredient can have highly diverse effects on blood glucose measurements in different people and that this effect is mediated largely by the microbiome. These observations suggest that the "optimal" diet needs to be individualized in the context of the person's microbiome, which itself may continue to change over the course of the diet. An intriguing parallel question is whether the microbiota may also influence dietary preferences; such an influence would suggest important feedback loops between the microbiome and diet.

MALNUTRITION Representing the other end of the metabolic spectrum from obesity, malnutrition is also linked to an altered microbiome. Analysis of Malawian twin pairs (≤3 years of age) who were discordant for kwashiorkor—a severe form of malnutrition—revealed that kwashiorkor is associated with a microbiologically "immature" fecal microbiota that resembles that of a chronologically younger child. Transplantation of the fecal microbiota from these discordant twins into gnotobiotic mice that were fed a diet similar in composition to a typical Malawian diet established that the kwashiorkor-associated microbiome is causally related to poor weight gain. Subsequent studies demonstrated these same general trends in malnourished Bangladeshi children. Investigators were able to identify five bacterial species (*Faecalibacterium prausnitzii*, *Ruminococcus gnavus*, *Clostridium nexile*, *Clostridium symbiosum*, and *Dorea formicigenerans*) that—when administered together as a "cocktail" to mice colonized with a kwashiorkor-associated microbiome—was able to prevent growth impairments. These results demonstrate that rationally designed modulation of the murine microbiota can lead to improved health outcomes. The clinical significance of these findings for humans remains to be clarified.

INFLAMMATORY BOWEL DISEASE Ulcerative colitis and Crohn's disease, the two predominant forms of inflammatory bowel disease (IBD), are chronic gastrointestinal inflammatory conditions that differ in their locations and patterns of inflammation **(Chap. 319)**. The following observations have led to the suggestion that IBD is the result of an immune response to a dysbiotic microbiota in a genetically susceptible individual: genes account for only ~20% of susceptibility to IBD (and many of the relevant genes are related to host–microbe interactions), antibiotic treatment reduces the clinical severity of disease, and relapses of Crohn's disease are prevented by diversion of the fecal stream. While the microbiota clearly is not the only driver of disease, it is considered to be an important element. Accordingly, numerous animal and clinical studies have been designed to tease out the nature of the relationship between the microbiota and IBD.

Most of these studies have focused on comparing the microbiome's composition in IBD patients with that in healthy controls, concentrating on microbial diversity and specific bacterial taxa that are associated with health or disease. Unfortunately, few, if any, results have been universally obtained, probably because of differences in study design, inclusion criteria, and methodology (e.g., the use of stool, rectal swabs, or biopsy samples; the choice of sequencing primers; the analysis pipeline). Even with these differences among studies, patients with IBD have been shown typically to have reduced alpha and beta diversity in their fecal microbiotas. Moreover, *Clostridium* clusters IV and XIVa, which are polyphyletic and encompass several different bacterial families, are generally reduced in patients with IBD. *F. prausnitzii* is a notable example from *Clostridium* cluster IV that is often underrepresented in the stool of patients who have Crohn's disease, with more mixed results in biopsy samples. The bacterial family Lachnospiraceae, which is largely contained in *Clostridium* cluster XIVa, and other butyrate-producing organisms are also reduced in the stool of patients with IBD. Some of these species produce butyrate by using acetate generated by other members of the microbiome, and some of these acetate-producing species are similarly reduced (e.g., *Ruminococcus albus*). These complex interactions and dependencies among bacterial species pose unique challenges to definitive ascertainment of the cause–effect relationships between microbes and disease. Even before researchers were able to assess the entire microbiome at once, they often noted that patients with Crohn's disease had a higher representation of adherent invasive *E. coli* in the ileal mucosa, an observation consistent with the increased abundance of Enterobacteriaceae seen in more recent microbiome studies. Beyond bacteria, burgeoning evidence supports a role for Caudovirales bacteriophages in IBD pathogenesis, though these findings may merely reflect the underlying dysbiosis related to the loss of bacterial diversity in IBD. Moreover, some data suggest that IBD is also associated with fungal dysbiosis; several studies have demonstrated an increased ratio of Basidiomycota to Ascomycota. It is still unclear whether any of these microbial associations reflect the cause of IBD or merely serve as biomarkers of disease.

Studies of antibiotic-treated mice and gnotobiotic mice colonized with IBD-associated microbiotas have been useful in confirming that the microbiota affects colitis severity. Several bacterial species have been identified as either promoting colitis in mice (e.g., *Klebsiella pneumoniae*, *Prevotella copri*) or protecting against it (e.g., *Bacteroides fragilis*, *Clostridium* species); however, these organisms do not always correlate with the taxa identified as differentially abundant across multiple clinical studies. In contrast, IgA-coated commensal organisms isolated from patients with IBD promote more severe colitis in mice than either IgA-uncoated bacteria from patients with IBD or IgA-coated bacteria from healthy controls. These data suggest that functional categorization of the microbiota based on immune recognition (e.g., IgA coating) may be a useful approach for identifying pathogenic organisms.

Cardiovascular Disease Inflammation helps drive the pathogenesis of atherosclerosis, and it has long been postulated that microbes are involved in the atherosclerotic process. Early work demonstrated that patients with cardiovascular disease have higher titers of antibody to *Chlamydia pneumoniae* than control patients, that *C. pneumoniae* is present within atherosclerotic lesions, and that *C. pneumoniae* can both initiate and exacerbate atherosclerotic lesions in animal models. This type of analysis has been extended to other bacteria, such as *Porphyromonas gingivalis*, with the idea that multiple different bacteria may play some role in the pathogenesis of atherosclerosis.

More recent studies have demonstrated clinical correlations between serum levels of trimethylamine N-oxide (TMAO) and atherosclerotic heart disease. Given that red meat, eggs, and dairy products are important sources of carnitine and choline (both precursors of TMAO), it is not

surprising that levels of TMAO are higher in omnivores than in vegans. Animal studies have confirmed that transfer of the gut microbiota from atherosclerosis-susceptible strains of mice to atherosclerosis-resistant animals leads to increased serum levels of TMAO and a dietary choline-dependent increase in atherosclerotic plaques; this observation confirms the role of the gut microbiota in the generation of TMAO and atherosclerosis. Moreover, treatment of atherosclerosis-susceptible strains of mice with a structural analogue of choline that inhibits the first enzymatic step in TMAO formation leads to decreased circulating TMAO levels and, more importantly, restrains macrophage foam-cell formation and atherosclerotic lesion development. In a study of more than 4000 patients, plasma TMAO levels were also predictive of incident thrombosis risk (myocardial infarction, stroke). Gnotobiotic animals were used to demonstrate that this risk was dependent on the microbiota; although eight bacterial taxa were identified as being associated with both plasma TMAO levels and thrombotic risk, organisms with choline-utilization genes that represent the first step of TMAO production were not more abundant in animals at greater risk for thrombosis. This discrepancy highlights the complexity of the microbiota and suggests that other aspects of the overall dynamics of the microbial community may be in play.

Oncology Recent studies exploring the link between the microbiota and cancer have demonstrated that specific members of the microbiota can affect treatment efficacy in both a positive and a negative manner. For example, therapy with antibody to programmed cell death ligand 1 (anti-PD-L1) has proven highly effective for a number of different cancers **(Chap. 69)**; however, a significant proportion of patients do not respond even when their tumors have high PD-L1 expression levels, a prerequisite for this type of checkpoint blockade inhibition. Using a murine melanoma model, investigators showed that variations in the microbiota resulted in differences in melanoma growth, an impact that was accentuated by anti-PD-L1 therapy. Ultimately, *Bifidobacterium* species were bioinformatically associated with improved anti-tumor responses, and administration of a "cocktail" of *Bifidobacterium* species (*B. bifidum*, *B. longum*, *B. lactis*, and *B. breve*) to melanoma-susceptible mice resulted in improved tumor-specific immunity and responses to anti-PD-L1 therapy. In a separate set of studies involving both patient data and animal experiments, the efficacy of therapy with antibody to cytotoxic T lymphocyte–associated antigen 4 (anti-CTLA-4) was associated with T-cell responses specific for either *Bacteroides thetaiotaomicron* or *B. fragilis*. In particular, administration of *B. fragilis* to germ-free or antibiotic-treated mice restored the normally absent anticancer response to anti-CTLA-4 therapy. While both of these examples demonstrate potentiation of anti-cancer therapies by the microbiota, other therapies can be antagonized. Some cancers, such as pancreatic ductal adenocarcinoma, contain intratumor bacteria, particularly Gammaproteobacteria, that can metabolize the chemotherapeutic agent gemcitabine and thereby contribute to the drug resistance of these tumors. Overall, these examples highlight the microbiota's critical impact—both direct and indirect—on the efficacy of drugs. Many other notable examples have been described (e.g., involving cyclophosphamide, digoxin, levodopa, and sulfasalazine), and many more likely remain to be discovered.

The application of microbiome science to hematopoietic stem cell transplantation (HSCT) is an area of expanding interest, particularly given the significant morbidity and mortality related to graft-versus-host disease (GVHD). In light of studies in the 1970s showing that germ-free mice developed less frequent and less severe gut GVHD than wild-type mice, clinicians began to use antibiotics to decontaminate the gut of patients undergoing HSCT. This decontamination approach yielded mixed results, probably because of differences in the antibiotic regimens used. The natural history of patients undergoing allogeneic HSCT includes a substantial loss of diversity in the fecal microbiota, with lower levels of bacterial diversity associated with increased mortality. Moreover, a retrospective analysis of ~850 patients undergoing allogeneic HSCT revealed that receipt of imipenem-cilastatin or piperacillin-tazobactam for neutropenic fever was associated with increased GVHD-related mortality at 5 years; this observation suggested that specific bacteria may help protect against GVHD-related mortality. More detailed analyses revealed an association between the abundance of *Blautia* species and protection against GVHD and mortality, though this correlation is still being examined with regard to its causal relationship. Despite significant interest in examining these microbial relationships with HSCT, little has yet been studied in the context of solid organ transplantation, which likely represents the next frontier of transplantation-related microbiome investigation.

Autoimmune Diseases The dramatic rise in the incidence of many autoimmune diseases over the past few decades has been far more rapid than can be explained simply by genetic factors (Fig. 459-5). It is increasingly thought that environmental triggers, including the microbiome, are partially responsible for the development of these autoimmune diseases.

TYPE 1 DIABETES Type 1 diabetes (T1D) is an autoimmune disorder characterized by T cell–mediated destruction of insulin-producing pancreatic islets **(Chap. 396)**. There is a clear genetic predisposition for the disease: ~70% of patients with T1D have human leukocyte antigen (HLA) risk alleles. However, only 3–7% of children with these risk alleles actually develop disease, an observation that suggests a role for other environmental factors. Studying a prospective, densely sampled, longitudinal cohort of at-risk, HLA-matched children from Finland and Estonia, investigators detailed changes in the microbiota prior to development of disease. Although only 4 of the 33 children studied developed T1D within the time frame of the study, a marked decrease of ~25% in alpha diversity occurred after seroconversion but before disease diagnosis. The low number of cases in this study unfortunately precluded identification of any specific disease-associated taxa. A follow-up study compared the microbiomes of a larger cohort of these high-risk northern European children with those of low-risk Russian children who lived in geographic proximity. *Bacteroides* species were more abundant in the high-risk group than in the low-risk group, particularly at early ages. This difference was postulated to be associated with an altered structure of the bacterial lipopolysaccharide to which children were exposed at a young age. It was further suggested that *Bacteroides*-derived lipopolysaccharide was not able to provide the immunogenic stimulus necessary to prevent T1D. These two studies offer attractive—though logistically complicated—options for future clinical investigations aimed at exploring the role of the microbiome. The first approach—longitudinally following individuals who are at high risk for a given disease—may provide insight into host–microbe relationships by mapping temporal changes in the microbiome with disease onset. An important caveat with this type of study, though, is that the associations identified may reflect preclinical disease rather than specifically indicating causality for any observed changes. The second approach illustrates how careful selection of study participants may offer an opportunity to uncover more meaningful associations that can subsequently be experimentally verified.

RHEUMATOID ARTHRITIS Similar to many other autoimmune diseases, rheumatoid arthritis (RA) is a multifactorial disease that comes to clinical attention after an environmental factor triggers symptoms in an individual with pre-existing autoantibodies. Multiple lines of evidence support the notion that RA pathogenesis is reliant on the microbiota, including the findings that germ-free mice do not develop symptoms in several RA models and that antibiotic treatment of mice mitigates against RA development. Several taxa (e.g., *Bacteroides* species, *Lactobacillus bifidus*, and segmented filamentous bacteria) have been implicated in promoting RA in murine models, and analysis of the fecal microbiota of patients with newly diagnosed RA have indicated that *P. copri* is a biomarker of disease. That this association with *P. copri* does not exist for chronic, treated RA or for psoriatic arthritis suggests some specificity for new-onset RA. A major limitation of this approach is that the identified association is shown to be a biomarker of disease (and, in this case, potentially of response to treatment) but no added insight is gained into a possible causal relationship between *P. copri* and RA. In fact, many of the patients with new-onset RA had no *Prevotella* detected, and several of the healthy controls had significant levels of

Prevotella. The lack of a strict concordance between the presence (or absence) of a specific taxon and a given disease state argues against a possible causal role.

MULTIPLE SCLEROSIS Epidemiologic studies of twin pairs and at-risk individuals moving between high- and low-risk geographic areas indicate that genetics plays a minor component in multiple sclerosis (MS) susceptibility relative to environmental factors. For example, in monozygotic twin pairs in which one sibling has MS, the other sibling also develops MS in only ~30% of cases. Although MS is a disease of the central nervous system (CNS), there is growing evidence of a link between MS and the microbiota, specifically that of the gut. Germ-free animals and antibiotic-treated animals display reduced disease incidence and severity in an MS model. Similarly, some clinical studies suggest improved disease outcomes in patients with MS who have been treated with minocycline, while patients treated with long-term penicillin appear to have an increased disease risk. Although several studies have compared the fecal microbiotas of healthy controls to those of patients with MS, these studies have all been relatively small and have yielded few results (if any) that are common throughout. Although work relating the microbiome to MS is ongoing, it has opened the door to exploring this link with other neurologic diseases. Already, there are animal data demonstrating links between the microbiota and both Parkinson's disease and autism, and there are clinical data assessing fecal microbiomes in relation to a variety of neurologic conditions. It is not quite clear how the gut microbiota is communicating with the CNS—i.e., whether communication takes place via bacterial metabolites that travel in the bloodstream and cross the blood–brain barrier, via migration of whole organisms into the CNS, or via feedback through the vagus nerve. Although our understanding of this brain–gut axis is still in its infancy, research in this area has elicited tremendous excitement as a tractable approach to potential treatments for these challenging diseases.

Atopic Diseases

The incidence and prevalence of allergic diseases continue to steadily increase, as do more severe clinical presentations. Life-threatening food allergies are now such a public health issue that nut-free classrooms are the norm in many cities. The development of allergic diseases often follows a stereotyped progression that begins with atopic dermatitis (AD) and continues, in order, with food allergy, asthma, and allergic rhinitis. The microbiome has been linked to all of these conditions and has the potential to modulate effects anywhere along this spectrum.

ATOPIC DERMATITIS The skin is the largest organ in the body, and its different anatomic sites (e.g., antecubital fossa, volar forearm, alar crease) represent distinct ecologic niches and harbor unique microbial communities. Moreover, given that the skin serves as a critical interface between the body and the external environment (e.g., microbes), it must be able to respond to unwanted microbes with an adequate immune response. AD is an inflammatory skin disorder involving immune dysfunction and a dysbiotic skin microbiota that is typically marked by greater abundances of *Staphylococcus aureus* and a lesser degree of bacterial diversity. Effective treatment of AD does not require complete elimination of *S. aureus* but is associated with restoration of the normal level of diversity. It is likely that this increase in bacterial diversity re-establishes normal immune homeostasis in the skin; specific members of the skin microbiota have been shown to induce protective skin-restricted immune responses. Coagulase-negative staphylococci (CoNS; primarily *S. epidermidis* and *S. hominis*) obtained from lesional and nonlesional skin of patients with AD were functionally screened and compared to CoNS from healthy controls; AD-lesional CoNS were much less often able to produce antimicrobial peptides (lantibiotics) directed against *S. aureus*. To demonstrate that these lantibiotic-producing CoNS were biologically relevant, they were incorporated into a lotion and applied to the arms of patients with AD. Surprisingly, a single application of the probiotic-laced lotion led to a decrease in the abundance of *S. aureus* recovered; no such decrease was observed when lantibiotic-negative strains were used. The authors of this study did not specifically comment on the clinical improvement of the AD lesions. Nevertheless, this is one of a limited number of studies that is beginning to extend microbiome-related findings into clinical trials.

ASTHMA Asthma is characterized by the clinical triad of airflow obstruction, bronchial hyperresponsiveness, and inflammation in the lower respiratory tract. Although the long-standing dogma was that the lungs are sterile, there is now convincing evidence for a constant ebb and flow of bacteria within the lower airways. In healthy states, the mucociliary escalator continually eliminates these bacteria soon after they land in the airways; in disease states (e.g., cystic fibrosis, chronic obstructive pulmonary disease), these bacteria establish long-term colonization of the airways and influence disease pathogenesis. In asthma specifically, both fecal and airway microbes have been linked to clinical outcomes.

Early studies of the microbiome's influence on asthma used culture-based methods to assess the hypopharyngeal microbiota of asymptomatic 1-month-old infants. Intriguingly, in one study, early-life colonization with *Streptococcus pneumoniae*, *Haemophilus influenzae*, *Moraxella catarrhalis*, or a combination of these organisms—but not *S. aureus*—was significantly associated with persistent wheeze and asthma at 5 years of age. Eosinophilia and total IgE levels at 4 years of age were also increased in children who were neonatally colonized with these organisms. Although this study examined a fairly focused set of bacteria, it laid the experimental groundwork indicating that early-life microbial exposures influence subsequent development of asthma. A later longitudinal investigation of the fecal microbiota in a general-population birth cohort of more than 300 children demonstrated that lower abundances of the genera *Lachnospira*, *Veillonella*, *Faecalibacterium*, and *Rothia* at 3 months of age were associated with an increased risk for development of asthma. The fact that these bacterial changes were no longer apparent when the children were 1 year of age is consistent with the notion that microbial exposures early in life are important to disease pathogenesis later in life. Transplantation of stool samples from 3-month-old children at risk for asthma into gnotobiotic mice resulted in significant airway inflammation in a murine model of asthma; pre- and postnatal exposure of mice to a four-species cocktail (*F. prausnitzii*, *Veillonella parvula*, *Rothia mucilaginosa*, and *Lachnospira multipara*) inhibited airway inflammation, with a marked reduction in neutrophil numbers in bronchoalveolar lavage fluid. These data suggest that early-life modulation of the microbiome may be an effective strategy to help prevent asthma, though the specific logistics (e.g., strains, dose, timing of exposure, patient selection) remain to be clarified.

Infectious Diseases

The increased susceptibility of antibiotic-treated mice to infection with a wide range of enteric pathogens was initially observed in the 1950s and led soon thereafter to the concept of *colonization resistance*, which holds that the normal intestinal microbiota plays a critical role in preventing colonization—and therefore disease production—by invading pathogens. Seminal work in the 1970s demonstrated that this protection is largely reliant on clostridial organisms, and the subsequent half-century has been spent trying to identify the specific microbes involved. Although much of the work relating the microbiota to infection has focused on enteric pathogens, the intestinal microbiota has also been clearly linked to bacterial pneumonia in mouse models, and changes in the microbial composition of the gut have been causally related to changes in the severity of disease. Although this gut–lung axis clearly exists in animals, its relevance in humans is still unclear. Several groups are beginning to study the human lung microbiome in the context of pneumonia and tuberculosis. Moreover, the relationships between the microbiota and both systemic infections (e.g., HIV infection, sepsis) and the response to vaccination are starting to be explored.

ENTERIC INFECTIONS *Clostridium difficile* infection (CDI) represents a growing worldwide epidemic and is the leading cause of antibiotic-associated diarrhea (Chap. 129). Roughly 15–30% of patients who are successfully treated for CDI end up with recurrent disease. The strong association between antibiotic exposure and CDI initially raised the idea that the microbiota is inextricably linked to acquisition of disease,

presumably because of the loss of colonization resistance. Consistent with the epidemiologic data, characterization of the fecal microbiota of patients with CDI revealed that it is a markedly less diverse, dysbiotic community. Fecal microbiota transplantation (FMT)—the "transplantation" of stool from a healthy individual into patients with disease—was successfully used in the 1950s to treat four patients with severe CDI and has recently been demonstrated in numerous studies to be an effective therapy for recurrent CDI, with clinical cure in 85–90% of patients (as detailed below). Thus, FMT for recurrent CDI has become the "poster child" for the idea that microbiome-based therapies may transform the management of many diseases previously considered to be refractory to medical therapy. Although FMT is agnostic as to the underlying mechanism of protection, work is ongoing to identify specific microbes and host pathways that can protect against CDI. Studying mice with differential susceptibilities to CDI due to antibiotic-induced changes in their microbiota, investigators identified a cocktail of four bacteria (*Clostridium scindens*, *Barnesiella intestihominis*, *Pseudoflavonifractor capillosus*, and *Blautia hansenii*) that conferred protection against CDI in a mouse model. Intriguingly, treatment of mice with just *C. scindens* offered significant, though not complete, protection in a bile acid–dependent manner. Clinical data from patients who underwent HSCT also associated *C. scindens* with protection from CDI, an observation that suggests the possibility of translating these findings from mice to humans. This study provides another example of the identification of relevant bacterial factors through examination of microbial differences in populations that differ in disease risk.

Microbiome-related changes associated with *Vibrio cholerae* infection include a striking loss of diversity (largely due to *V. cholerae*'s becoming the dominant member of the microbiota) and an altered composition that rapidly follows the onset of disease. These changes, which occur in a reproducible and stereotypical manner, are reversible with treatment of the disease. This recovery phase involves a microbial succession that is similar to the assembly and maturation of the microbiota of healthy infants. In addition to *V. cholerae*, streptococcal and fusobacterial species bloom during the early phases of diarrhea, and the relative abundances of *Bacteroides*, *Prevotella*, *Ruminococcus/Blautia*, and *Faecalibacterium* species increase during the resolution phase and mark the return to a healthy adult microbiota. Analysis of these microbial changes occurring in patients with cholera and in healthy children led to the selection of 14 bacteria that were transplanted into gnotobiotic mice, which were then challenged with *V. cholerae*. Bioinformatic analysis of specific taxa changing during cholera determined that *Ruminococcus obeum* restrained *V. cholerae* growth. Subsequently, this relationship was experimentally confirmed, and the *R. obeum* quorum-sensing molecule AI-2 (autoinducer 2) was found to be responsible for restricting *V. cholerae* colonization via an unclear mechanism. These studies highlight the potential for use of microbiome-based therapies to prevent and/or treat infectious diseases. Moreover, they suggest that temporal analysis of longitudinal microbiome data may be an effective strategy for identifying microbes with causal relationships to disease.

HIV INFECTION The augmentation of HIV pathogenesis by some viral, bacterial, and parasitic co-infections suggests that a patient's underlying microbial environment can influence the severity of HIV disease. Moreover, it has been hypothesized that the intestinal immune system plays a significant role in regulating HIV-induced immune activation; this seems particularly likely since the intestines are an early site for viral replication and exhibit immune defects before peripheral CD4+ T cell counts decrease. Several studies have examined the intestinal microbiotas of HIV-infected individuals. Initial studies performed in nonhuman primates infected with simian immunodeficiency virus found no alteration in the bacterial components of the fecal microbiota; however, there were profound changes in the enteric virome. In contrast, many recent studies exploring this issue in patients have identified substantial differences in the HIV-associated fecal microbiota that correlate with systemic markers of inflammation. Curiously, these microbial changes do not necessarily normalize with antiretroviral therapy; this finding suggests that the microbiota may have some "memory" of the previously high HIV loads and/or that HIV infection helps reset the "normal" microbiota. This memory-like capacity of the microbiota has been demonstrated in animal models in the context of other infections and in response to dieting.

Given that the majority of new HIV transmission events follow heterosexual intercourse, there has been significant interest in examining the relationship between the vaginal microbiota and HIV acquisition. A longitudinal study of South African adolescent girls who underwent high-frequency testing for incident HIV infection facilitated the identification of bacteria that were associated with reduced risk of HIV acquisition (*Lactobacillus* species other than *L. iners*) or with enhanced risk (*Prevotella melaninogenica*, *Prevotella bivia*, *Veillonella montpellierensis*, *Mycoplasma*, and *Sneathia sanguinegens*). In mice inoculated intravaginally with *Lactobacillus crispatus* or *P. bivia*, the latter organism induced a greater number of activated CD4+ T cells in the female genital tract, a result suggesting that the increased risk of HIV acquisition associated with *P. bivia* may be secondary to the increased presence of target cells. In a separate study, the composition of the vaginal microbiota was shown to modulate the antiviral efficacy of a tenofovir gel microbicide. Although tenofovir reduced HIV acquisition by 61% in women who had a *Lactobacillus*-dominant vaginal microbiota, it reduced HIV acquisition by only 18% in women whose vaginal microbiota comprised primarily *Gardnerella vaginalis* and other anaerobes. This difference in efficacy was due to the ability of *G. vaginalis* to metabolize tenofovir faster than the target cells can take up the drug and convert it into its active form, tenofovir diphosphate. These findings illustrate how microbial ecology can be an important consideration in choosing effective treatment regimens.

RESPONSE TO VACCINATION Second only to the provision of clean water, vaccination has been the most effective public health intervention in the prevention of serious infectious diseases. Its effects are mediated by antigen-specific antibodies and, in some cases, effector T-cell responses. Although vaccines are clearly effective on a population scale, the magnitude of the immune response to vaccines can vary among individuals by tenfold to a hundredfold. Although many factors (e.g., genetics, maternal antibody levels, prior antigen exposures) can affect vaccine immunogenicity, the microbiota is now recognized as another important factor. Analysis of the fecal microbiotas of ~50 Bangladeshi children identified specific taxa that exhibited positive associations (e.g., *Actinomyces*, *Rothia*, and *Bifidobacterium* species) and negative associations (e.g., *Acinetobacter*, *Prevotella*, and staphylococcal species) with responses to vaccines against polio, tuberculosis (bacille Calmette-Guérin), tetanus, and hepatitis B. A study of infants from Ghana revealed an inverse relationship between the fecal abundance of Bacteroidetes and a response to the rotavirus vaccine. Moreover, the nasal microbiota has been implicated as a factor that contributes to the IgA response to live, attenuated influenza vaccines. These correlations based on clinical data have been partially confirmed in animal studies. The best example is the demonstration that the responses to non-adjuvanted viral subunit vaccines (inactivated influenza and polio vaccines) are reliant on the microbiota, whereas the responses to live or adjuvanted vaccines (live attenuated yellow fever, Tdap/alum, an HIV envelope protein/alum vaccine) are not. An interesting note is that the antibody response to inactivated influenza vaccine is dependent on recognition of the microbiota by Toll-like receptor 5, presumably via flagellin-expressing microbes. These data suggest that the microbiota may serve as an adjuvant for certain vaccine types. Confirmation of these findings in clinical settings may suggest ways to improve vaccine efficacy in the future.

MECHANISMS OF MICROBIOME-MEDIATED EFFECTS

As highlighted in the examples above, numerous associations have been made between the microbiome and various disease states. These correlations have often been established at broad taxonomic levels, with little or no insight into causality. Given that most clinical studies of these relationships have a fairly small sample size (often <100) and are simultaneously comparing numerous variables (i.e., each of the bacterial species in the microbiota is effectively a different feature being

compared), many of these studies may not be adequately powered and therefore may yield false-positive results. Testing of these correlations in animal models of disease has been critical in demonstrating a causal relationship between microbes and specific phenotypes. Because microbiome-wide association studies typically result in a long list of bacterial taxa that are correlated with a disease, it has been challenging to know which organism to test further in mechanistic studies. Moreover, even if a specific bacterial species is identified in these analyses, there is potentially enough strain-to-strain variation that the "functional" isolate may need to be recovered from the individuals studied; a publicly available representative of the species may not confer the same phenotype.

Despite all these difficulties, a handful of specific microbes have now been linked to disease effects; some examples have been mentioned above. The next layer of challenges relates to identification of the specific mechanisms that underlie these causal relationships. Successes along these lines have been more limited, but approaches are being developed to tackle the issue of defining specific bacterial factors and metabolites that are responsible for the phenotypic changes. Complicating factors are that many organisms, particularly those in the phylum Firmicutes, are not readily genetically tractable and that many of the phenotypes are not easy to assess with high-throughput screening.

BACTERIAL FACTORS

B. fragilis polysaccharide A (PSA) is perhaps the best-studied commensal-derived molecule that has been demonstrated to influence disease outcomes in mouse models. PSA—one of at least eight capsular polysaccharides expressed by *B. fragilis*—has a unique zwitterionic structure that incorporates both a positive and a negative charge within each repeating unit. Studies in which mice have been treated either with isogenic strains of *B. fragilis* that differ in PSA expression or with purified PSA have shown that PSA confers protection—prophylactically and therapeutically—against experimental colitis and MS. PSA is recognized by Toll-like receptor 2 on antigen-presenting cells, particularly plasmacytoid dendritic cells, and—in the setting of inflammation—induces interleukin 10 (IL-10)–producing regulatory T cells (Tregs) that help restrain inflammation.

B. fragilis is also the source of the only other microbiota-based bacterial factor identified thus far: an immunomodulatory glycosphingolipid that affects the numbers of invariant natural killer T (iNKT) cells. It is not clear whether these glycosphingolipids activate or inhibit iNKT cells; results have been discordant, probably because different glycosphingolipid species have been tested. Analysis of a specific purified glycosphingolipid (Bf717) demonstrated that it inhibits endogenous iNKT cell agonists in vitro and in vivo. Treatment of neonatal mice with Bf717 leads to a decreased number of colonic iNKT cells in adulthood and to improved outcomes in a model of colitis.

BACTERIAL METABOLITES

The use of mass spectrometry to detect and profile tens of thousands of different metabolites present in different bodily fluids has offered the promise of deeper insight into microbially mediated processes that underlie disease susceptibility. However, the fact that the overwhelming majority of these metabolites are not annotated, coupled with the sheer volume of data generated, has so far limited the general utility of these untargeted approaches. Instead, interest in more targeted approaches has increased, with a current emphasis on examining the role of short-chain fatty acids (SCFAs) and bile acids.

Short-Chain Fatty Acids Several groups have demonstrated that SCFAs, the intestinal levels of which are largely determined by bacterial metabolism, are important for the induction of Tregs, though there is not agreement on which specific SCFA (propionate, acetate, or butyrate) is most relevant. Wild-type mice colonized with bacteria known to induce colonic Tregs have elevated cecal levels of SCFAs. Colonization with any of three *Bacteroides* species (*B. caccae*, *B. massiliensis*, and *B. thetaiotaomicron*) increases levels of acetate and propionate, whereas colonization with *Parabacteroides distasonis* or a mix of 17 human-derived *Clostridium* species elevates levels of all three SCFAs. In all of these cases, though, the SCFAs inhibit histone deacetylase, with a consequent increase in Foxp3 expression. Notably, microbe-induced SCFA production has not been shown to be critical for Treg induction by any of these organisms. In contrast, there appears to be no correlation between SCFA levels and Treg numbers in mice monocolonized with various Treg-inducing bacterial species. Taken together, these data suggest important heterogeneity in the mechanisms underlying Treg development and do not rule out the possibility of other, redundant mechanisms for Treg induction.

Bile Acids Bile acids are produced in the liver but then are metabolized by intestinal bacteria to form deconjugated and secondary bile acids. These microbially produced bile acid profiles act through complex signaling pathways to balance the metabolism of lipids and carbohydrates and to affect immune responses. Therefore, bile acids are now being investigated as microbial metabolites that are critical to maintaining human health. As mentioned above, *C. scindens* helps protect mice against CDI through a bile acid–dependent process. Alterations in bile acid profiles due to underlying microbial dysbiosis have also been associated with hepatic and colonic inflammation, hepatic cellular carcinoma, colorectal cancer, and impaired gut motility. Almost all of these relationships have been documented at the level of correlation and, at best, reflect a partial change in phenotype in the setting of bile acid sequestrants (e.g., cholestyramine). Work is ongoing to determine causal relationships between bacterial metabolism of bile acids and changes in host physiology.

Other Bacterial Metabolites Although most work has thus far focused on SCFAs and bile acids, a few notable examples of other bacterial metabolites have been implicated in maintaining health. Taurine enhances NLRP6 inflammasome–induced colonic IL-18 secretion, while histamine, spermine, and putrescine suppress IL-18 secretion. Desaminotyrosine produced by *Clostridium orbiscindens* confers protection from influenza by inducing type I interferon activity. In both of these cases, the microbiota was initially shown to influence the phenotype, with either untargeted metabolomics or a more targeted screen indicating a potential role for the indicated metabolites. Given the thousands of different bacterial metabolites throughout the body, many more metabolites will undoubtedly be linked to health and disease.

MOVING MICROBIOME SCIENCE FROM BENCH TO BEDSIDE

The numerous microbiome–disease associations identified thus far have generated a great deal of hope that understanding the relevant microbe–host interactions will open the door to unlimited therapeutic applications. Microbiome-based therapies offer several potential benefits. Patients often view such treatment as more "natural" than conventional drug therapy and are therefore more likely to comply with it. Biologically, microbiome-based therapies are more likely to address one of the root causes of disease (microbial dysbiosis) rather than simply affecting the downstream sequelae. Finally, a given microbiome-based therapy may serve as a "polypill" that is effective against several different diseases stemming from similar microbial changes. Despite tremendous interest in therapeutically exploiting the microbiome, there have thus far been few clinical successes along these lines.

The most successful therapeutic application of microbiome science has been the use of FMT, particularly for CDI. As mentioned earlier, FMT involves "transplanting" stool from a healthy individual to a diseased patient, with the idea that the "healthy" microbiota will correct whatever derangement may exist in the ill patient and therefore will alleviate symptoms. Fundamentally, this notion is agnostic as to the specific microbial dysbiosis and holds that any healthy microbiota will be curative. The idea of FMT dates back to at least the fourth century, when traditional Chinese doctors used a "yellow soup" (fresh human fecal suspension) to successfully treat food poisoning and severe diarrhea. The continued use of FMT through the centuries for the treatment of diarrheal illnesses in both humans and animals, along with the growing appreciation in recent years of the importance of the microbiota, laid the groundwork for using FMT to treat CDI. Since the first major prospective trial assessing FMT for recurrent CDI in 2013, most of the numerous

studies of FMT for CDI have demonstrated remarkable efficacy, with an average clinical cure rate of ~85%. The donor stool can be fresh or frozen (use of the latter allows biobanking of samples from a limited number of pre-screened donors) and can be administered via nasogastric tube, nasoduodenal tube, colonoscopy, enema, or oral capsules; the cure rate is slightly higher with lower-gastrointestinal administration than with upper-gastrointestinal treatment. The optimal screening, preparation, and concentration of infused donor stool have not yet been determined. The most common adverse effects of FMT include altered gastrointestinal motility (with constipation or diarrhea), abdominal cramps, and bloating, all of which are generally transient and resolve within 48 h. At least 80 immunosuppressed patients have undergone FMT with no serious adverse events noted during 3 months of follow-up.

The successful use and the favorable short-term safety profile of FMT for CDI have led to its expanded application for other indications. At the end of 2017, 190 trials (listed at *ClinicalTrials.gov*) were investigating the efficacy of FMT for a range of indications, including CDI, IBD (ulcerative colitis and Crohn's disease), obesity, eradication of multidrug-resistant organisms, anxiety and depression, cirrhosis, and type 2 diabetes. The few published studies regarding indications other than CDI have generally included small sample sizes and have offered mixed results. In contrast to the successes in CDI, the results have been more varied for patients with IBD, which is perhaps the second-best-studied indication. It is not clear whether these discrepancies are due to heterogeneity in recipients (e.g., in terms of underlying disease mechanisms or endogenous microbiotas), the donor material, and/or the logistical details of FMT administration (e.g., route, frequency, dose). However, these results demonstrate that—under the right circumstances—modulation of the microbiota can be an effective therapy for IBD.

Although FMT offers an important proof of concept that microbiome-based therapies can be effective, treatment is difficult to standardize across large populations because of variability among stool donors and among the endogenous microbiotas of recipients. In addition, FMT is fraught with safety concerns, and its mechanism(s) of action are unclear. FMT likely represents the first generation of microbiome-based therapies; subsequent generations will include the use of more refined bacterial cocktails, single strains of bacteria, or bacterial metabolites as the therapeutic intervention. The field of probiotics has a complicated history: many different strains have been tested against a multitude of diseases. Several meta-analyses have combined results across bacterial strains and/or disease indications and have generally concluded that the data are not yet convincing enough to support the use of the tested regimens. It should be noted that the tested organisms have been chosen mainly on the basis of their presumed safety profile rather than in light of a plausible biological link to disease. The hope is that more focused, mechanistic microbiome studies will identify specific commensal organisms—and their underlying mechanisms of action—that are involved in disease pathogenesis and that will serve as the basis for the next wave of rationally chosen probiotics, a few of which are currently in clinical trials. The main hurdle in this endeavor has been identifying specific microbes that are causally related to protection from disease.

PERSPECTIVE

The medical view of microbes has changed radically, moving from the early-twentieth-century notion that we are engaged in a constant struggle with microbes—an "us-versus-them" mentality that focused on the necessity of eradicating bacteria—to the more recent understanding that we live in a carefully negotiated state of détente with our commensal organisms. Instead of holding a simple view of microbes as enemies to be eliminated with antibiotics, scientists are increasingly recognizing the critical role these organisms play in maintaining human health; loss of these host–microbe interactions in the increasingly sterile environment typical of Western civilization may have predisposed to the increased incidence of autoimmune and inflammatory diseases. The field of microbiome research has made great strides over the past decade in cataloguing the normal microbiota and is now on the cusp of being able to identify clinically actionable microbe–host relationships.

The recent explosion of "–omics" technologies (e.g., metagenomics, metatranscriptomics, metabolomics) has enabled the generation of vast amounts of data, but it is not yet clear how best to integrate datasets in order to gain useful insights into host–microbe relationships. The use of FMT has demonstrated that modulation of an individual's microbiota can effectively treat certain diseases; however, models with which to predict specifically how a microbiota will change after modulation—and what potentially untoward effects these changes might have—are still lacking. Implicit in this limitation is our ignorance about what microbial configuration is optimal and how a given microbiota should be rationally altered to obtain an ideal outcome.

Despite initial hyperbolic hype and a few false starts, microbiome research now stands at the precipice of an ability to treat the fundamental basis of many diseases. As the field continues to mature, it will need to move beyond correlations and address causation. The identification of causal microbes and their mechanisms of action will create a "microbial toolbox" from which relevant bioactive strains can be chosen on a per-patient basis to correct specific underlying microbial dysbioses. In the near future, our knowledge base regarding the microbiome and its relationship to health and disease will be robust enough that this information can be applied in making important treatment decisions.

FURTHER READING

Goodrich JK et al: Conducting a microbiome study. Cell 158:250, 2014.

Human Microbiome Project Consortium: Structure, function and diversity of the healthy human microbiome. Nature 486:207, 2012.

Lloyd-Price J et al: Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 550:61, 2017. (Erratum: Nature 551:256, 2017.)

Rajilić-Stojanović M, de Vos WM: The first 1000 cultured species of the human gastrointestinal microbiota. FEMS Microbiol Rev 38:996, 2014.

Surana NK, Kasper DL: Deciphering the tête-à-tête between the microbiota and the immune system. J Clin Invest 124:4197, 2014.

460 Global Issues in Medicine

Joseph J. Rhatigan, Paul Farmer

WHY GLOBAL HEALTH?

Global health has emerged as an important field within medicine. Some scholars have defined global health as the field of study and practice concerned with improving the health of all people and achieving health equity worldwide, with an emphasis on addressing transnational problems. No single review can do much more than identify the leading problems in applying evidence-based medicine in settings of great poverty or across national boundaries. However, this is a moment of opportunity: only recently, persistent epidemics, improved metrics, and growing interest have been matched by an unprecedented investment in addressing the health problems of poor people in the developing world. To ensure that this opportunity is not wasted, the facts need to be laid out for specialists and laypeople alike. This chapter introduces the major international bodies that address health problems; identifies the more significant barriers to improving the health of people who to date have not, by and large, had access to modern medicine; and summarizes population-based data on the most common health problems faced by people living in poverty. Examining specific problems—notably HIV/AIDS **(Chap. 197)** but also tuberculosis **(Chap. 173)**, malaria **(Chap. 219)**, Ebola **(Chap. 205)**, and key "noncommunicable" chronic diseases (NCDs)—helps sharpen the discussion of barriers to prevention, diagnosis, and care as well as the means of overcoming them. This chapter closes by discussing global health equity, drawing on notions of social justice that once were central to international public health but had fallen out of favor during the last decades of the twentieth century.

A BRIEF HISTORY OF GLOBAL HEALTH INSTITUTIONS

Concern about illness across national boundaries dates back many centuries, predating the Black Plague and other pandemics. One of the first organizations founded explicitly to tackle cross-border health issues was the Pan American Sanitary Bureau, which was formed in 1902 by 11 countries in the Americas. The primary goal of what later became the Pan American Health Organization was the control of infectious diseases across the Americas. Of special concern was yellow fever, which had been running a deadly course through much of South and Central America and halted the construction of the Panama Canal. In 1948, the United Nations formed the first truly global health institution: the World Health Organization (WHO). In 1958, under the aegis of the WHO and in line with a long-standing focus on communicable diseases that cross borders, leaders in global health initiated the effort that led to what some see as the greatest success in international health: the eradication of smallpox. Naysayers were surprised when the smallpox eradication campaign, which engaged public health officials throughout the world, proved successful in 1979 despite Cold War tensions.

At the International Conference on Primary Health Care in Alma-Ata (in what is now Kazakhstan) in 1978, public health officials from around the world agreed on a commitment to "Health for All by the Year 2000," a goal to be achieved by providing universal access to primary health care worldwide. Critics argued that the attainment of this goal by the proposed date was impossible. In the ensuing years, a strategy for the provision of selective primary health care emerged. This strategy included four inexpensive interventions collectively known as GOBI: *g*rowth monitoring, *o*ral rehydration, *b*reast-feeding, and *i*mmunizations for diphtheria, whooping cough, tetanus, polio, tuberculosis, and measles. GOBI later was expanded to GOBI-FFF, which also included *f*emale education, *f*ood, and *f*amily planning. Some public health figures saw GOBI-FFF as an interim strategy to achieve "health for all," but others criticized it as a retreat from the bolder commitments of Alma-Ata.

The influence of the WHO waned during the 1980s. In the early 1990s, many observers argued that, with its vastly superior financial resources and its close—if unequal—relationships with the governments of poor countries, the World Bank had eclipsed the WHO as the most important multilateral institution working in the area of health. One of the stated goals of the World Bank was to help poor countries identify "cost-effective" interventions worthy of public funding and international support. At the same time, international financial institutions encouraged many of those nations to reduce public expenditures in health and education in order to stimulate economic growth as part of (later discredited) policies imposing restrictions as a condition for access to credit and assistance through the World Bank, the International Monetary Fund, and regional development banks. There was a resurgence of many diseases—including malaria, trypanosomiasis, and schistosomiasis—in Africa. Tuberculosis, an eminently curable disease, remained the world's leading infectious killer of adults. Half a million women per year died in childbirth during the last decade of the twentieth century, and few of the world's largest philanthropic or funding institutions focused on global health equity.

HIV/AIDS, first described in the medical literature in 1981, precipitated a change. In the United States, the advent of this newly described infectious killer marked the culmination of a series of events that dashed previous hopes of "closing the book" on infectious diseases. In Africa, which would emerge as the global epicenter of the pandemic, HIV disease strained tuberculosis control programs, and malaria continued to claim as many lives as ever: at the dawn of the twenty-first century, these three diseases alone killed nearly 6 million people each year. New research, new policies, and new funding mechanisms were called for. The past decade has seen the rise of important multilateral global health financing institutions such as the Global Fund to Fight AIDS, Tuberculosis, and Malaria; bilateral efforts such as the U.S. President's Emergency Plan for AIDS Relief (PEPFAR); and private philanthropic organizations such as the Bill & Melinda Gates Foundation. With its 193 member states and 147 country offices, the WHO remains important in matters relating to the cross-border spread of infectious diseases and other health threats. In the aftermath of the epidemic of severe acute respiratory syndrome in 2003, the WHO's International Health Regulations—which provide a legal foundation for that organization's direct investigation into a wide range of global health problems, including pandemic influenza, in any member state—were strengthened and brought into force in May 2007.

Even as attention to and resources for health problems in poor countries grow, the lack of coherence in and among global health institutions may undermine efforts to forge a more comprehensive and effective response. The WHO remains underfunded despite the ever-growing need to engage a wider and more complex range of health issues, such as the Ebola outbreak of 2014–2015 in West Africa. This may be what some have called "the golden age of global health," but leaders of major global health organizations must work together to design an effective architecture that will make the most of opportunities to link new resources for and commitments to global health equity with the emerging understanding of disease burden and the unmet need to create robust and resilient national health systems. To this end, new and old players in global health must invest heavily in *discovery* (relevant basic science), *development* of new tools (preventive, diagnostic, and therapeutic), and modes of *delivery* that will ensure the equitable provision of health products and services to all who need them.

The adoption of the Sustainable Development Goals (SDGs) in 2015 by the United Nations serves as an example of effective cooperation. The SDGs articulate 17 overarching goals across a number of domains to be achieved by 2030. Goal 3 specifically relates to global health and contains 13 distinct targets to be met, including reducing maternal and child mortality; ending the epidemics of HIV, tuberculosis, and malaria; and reducing the burden of NCDs.

Included in the SDGs is a commitment to achieve universal health coverage (UHC), providing universal access to high-quality essential health services at an affordable cost worldwide. Championed by the WHO, the World Bank, and many civil society organizations, Goal 3 will measure coverage of 16 essential health services and assess the financial burden of health spending by households in every country.

THE ECONOMICS OF GLOBAL HEALTH

Political and economic concerns have often guided global health interventions. As mentioned, early efforts to control yellow fever were tied to the completion of the Panama Canal. However, the precise nature of the link between economics and health remains a matter for debate. Some economists and demographers argue that improving the health status of populations must begin with economic development; others maintain that addressing ill health is the starting point for development in poor countries. In either case, there is increasing consensus that investments in health care delivery and the control of communicable diseases lead to increased productivity. The question is where to find the necessary resources to start the predicted "virtuous cycle."

During the past two decades, spending on health in poor countries has increased dramatically. According to a study from the Institute for Health Metrics and Evaluation (IHME) at the University of Washington, total development assistance for health worldwide grew to $36.4 billion in 2015—up from $5.6 billion in 1990. In 2015, the leading contributors included PEPFAR, the Global Fund, the GAVI Alliance, nongovernmental organizations (NGOs), the WHO, the Gates Foundation, and the World Bank, which have dramatically increased their development spending for health since 2014. It appears, however, that total development assistance for health plateaued in 2010, and it is unclear whether growth will continue in the future.

MORTALITY AND THE GLOBAL BURDEN OF DISEASE

Refining metrics is an important task for global health: only relatively recently have there been solid assessments of the global burden of disease. The first study to look seriously at this issue, conducted in 1990, laid the foundation for the first report on *Disease Control Priorities in Developing Countries* and for the World Bank's 1993 World Development Report *Investing in Health*. Those efforts represented a major advance in the understanding of health status in developing countries. *Investing in Health* has been especially influential: it familiarized a broad audience with cost-effectiveness analysis for specific health interventions and with the notion of disability-adjusted life years (DALYs). The DALY, which has become a standard measure of the impact of a specific health condition on a population, combines absolute years of life lost and years lost due to disability for incident cases of a condition. (See **Fig. 460-1** and **Table 460-1** for an analysis of the global disease burden by DALYs.)

In 2012, the IHME and partner institutions began publishing results from the Global Burden of Diseases, Injuries, and Risk Factors Study 2010 (GBD 2010). GBD 2010 is the most comprehensive effort to date to produce longitudinal, globally ambitious, and comparable estimates of the burden of diseases, injuries, and risk factors. This report reflects the expansion of the available data on health in the poorest countries and of the capacity to quantify the impact of specific conditions on a population. It measures current levels and recent trends for major diseases, injuries, and risk factors worldwide. The GBD 2010 team revised and improved the health-state severity weight system, collated published data, and used household surveys to enhance the breadth and accuracy of disease burden data. Updated reports were released in 2013 and 2015. As analytic methods and data quality improve, important trends can be identified in a comparison of global disease burden estimates from 1990 to 2015.

GLOBAL MORTALITY

Of the 55.8 million deaths worldwide in 2015, 20.2% (11.3 million) were due to communicable diseases, maternal and perinatal conditions, and nutritional deficiencies—a marked decrease compared with figures for 1990, when these conditions accounted for 34% of global mortality. Among the fraction of all deaths related to communicable diseases, maternal and perinatal conditions, and nutritional deficiencies, 74.2% occurred in sub-Saharan Africa and southern Asia. While the proportion of deaths due to these conditions has decreased significantly in the past decade, there has been a dramatic rise in the number of deaths from NCDs, which constituted the top four causes of death in 2015. The leading cause of death among adults in 2015 was ischemic heart disease, accounting for 8.9 million deaths (16% of total deaths) worldwide. In high-income countries ischemic heart disease accounted for 18.0% of total deaths, and in low- and middle-income countries it accounted for 15.5%. It is noteworthy that ischemic heart disease was responsible for just 4.7% of total deaths in sub-Saharan Africa **(Table 460-2)**. In second place—causing 11.3% of global mortality—was cerebrovascular disease (ischemic and hemorrhagic stroke), which accounted for 9.0% of deaths in high-income countries, 11.9% in low- and middle-income countries, and 5.0% in sub-Saharan Africa. Although the third leading cause of death in high-income countries was lung cancer (accounting for 5.8% of all deaths), this condition did not figure among the top 25 causes in sub-Saharan Africa. Among the 10 leading causes of death in sub-Saharan Africa, six were infectious diseases, with malaria and HIV/AIDS ranking as the dominant contributors to disease burden. In high-income countries, however, only one infectious disease—lower respiratory infection—ranked among the top 10 causes of death.

The GBD 2015 found that the worldwide mortality figure among children <5 years of age dropped from 16.39 million in 1970 to 12.1 million in 1990 and to 5.8 million in 2015—a decrease that surpassed predictions. Of childhood deaths in 2015, 2.6 million (45.0%) occurred in the neonatal period. About one-third of deaths among children <5 years old occurred in southern Asia and almost one-half in sub-Saharan Africa; ~1% occurred in high-income countries.

The global burden of death due to HIV/AIDS and malaria was on an upward slope until 2004; significant improvements have since been documented. Global deaths from AIDS fell from 1.7 million in 2006 to 1.0 million in 2016, while malaria deaths dropped from 1.2 million to 730,000 over the same period. Despite these improvements, malaria and HIV/AIDS continue to be major burdens in particular regions, with global implications. Although it has only a minor impact on mortality outside sub-Saharan Africa and Southeast Asia, malaria is the fifth leading cause of death of children <5 years of age worldwide. HIV infection ranked 42nd in global DALYs in 1990 but was the 10th leading cause of disease burden in 2015, with sub-Saharan Africa bearing the vast majority of this burden (Fig. 460-1).

The world's population is living longer: global life expectancy has increased significantly over the past 45 years from 58.8 years in 1970 to 71.8 years in 2015. This demographic change, accompanied by the fact that the prevalence of NCDs increases with age, is dramatically shifting the burden of disease toward NCDs, which have surpassed communicable, maternal, nutritional, and neonatal causes. By 2015, 71.3% of total deaths at all ages and 59.7% of all DALYs were due to NCDs. Increasingly, the global burden of disease comprises conditions and injuries that cause disability rather than death.

Worldwide, although both life expectancy and years of life lived in good health have risen, years of life lived with disability have also increased. Despite the higher prevalence of diseases common in older populations (e.g., dementia and musculoskeletal disease) in developed and high-income countries, best estimates from 2015 reveal that disability resulting from cardiovascular diseases, chronic respiratory diseases, and the long-term impact of communicable diseases was greater in low- and middle-income countries. In most developing countries, people lived shorter lives and experienced disability and poor health for a greater proportion of their lives. Indeed, 47.7% of the global burden of disease occurred in southern Asia and sub-Saharan Africa, which together account for only 37.4% of the world's population.

HEALTH AND WEALTH

Clear disparities in burden of disease (both communicable and noncommunicable) across country income levels are strong indicators that poverty and health are inherently linked. Poverty remains one of the most important root causes of poor health worldwide, and the global

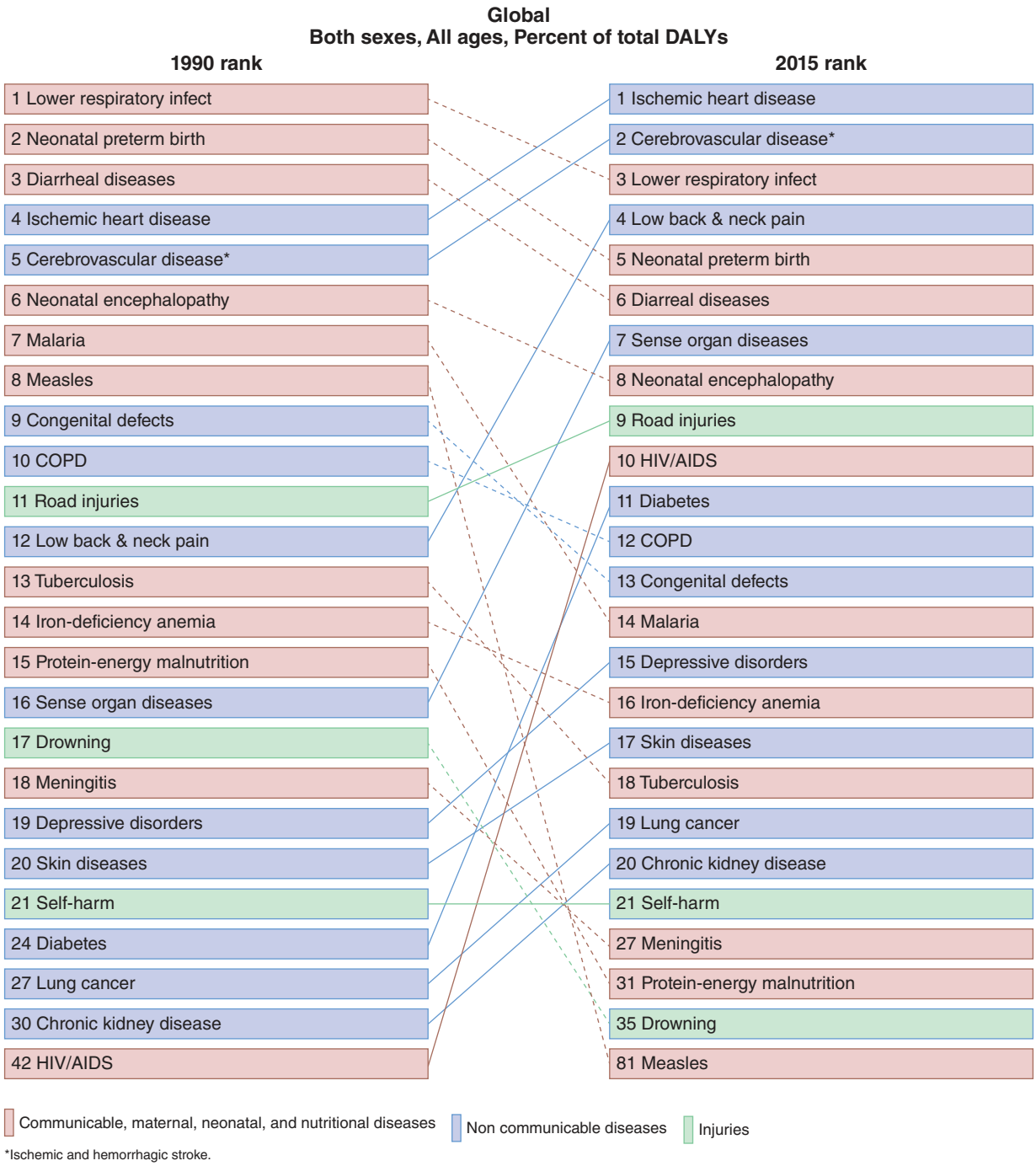


FIGURE 460-1 Global DALY (disability-adjusted life year) ranks for the top causes of disease burden in 1990 and 2015. COPD, chronic obstructive pulmonary disease. *(From the Institute for Health Metrics and Evaluation [IHME]. GBD Compare. Seattle, WA: IHME, University of Washington, 2016. Available at http://vizhub.healthdata.org/gbd-compare. Accessed July 18, 2017.)*

burden of poverty continues to be high. Among the 7.3 billion people alive in 2015, 10.7% (767 million) lived on <$1.90 (U.S.) per day—one standard measurement of extreme poverty—and another 1 billion lived on <$3.10 per day. Approximately 385 million children—19.5% of all children in low-income countries—lived in extreme poverty in 2013. Comparison of national health indicators with gross domestic product per capita among nations shows a clear relationship between higher gross domestic product and better health, with only a few outliers. Numerous studies have also documented the link between poverty and health within nations as well as across them.

RISK FACTORS FOR DISEASE BURDEN

The GBD study found that the three leading risk factors for global disease burden in 2015 were (in order of frequency) high blood pressure, tobacco smoking (including secondhand smoke), and high fasting plasma glucose—a substantial change from 1990, when childhood undernutrition was ranked first. Although ranking fifth in 2015, childhood undernutrition remains the leading risk factor for death worldwide among children <5 years of age. In an era that has seen obesity become a major health concern in many developed countries—and the fourth leading risk factor for disease burden worldwide—the persistence of undernutrition is cause for consternation. Low body weight is still the dominant risk factor for disease burden in sub-Saharan Africa. In its rural reaches, no health care initiative, however generously funded, will be effective or comprehensive without addressing undernutrition.

In a 2016 publication that examined how specific diseases and injuries are affected by environmental risk, the WHO estimated that 23% of all deaths and 26% of deaths among children <5 years of age were due to modifiable environmental factors: some 1.7 million children die every year from causes related to unhealthy environments, including

TABLE 460-1 Leading Causes of Burden of Disease (DALYs), 2015

DISEASE OR INJURY	DALYs (MILLIONS)	PERCENTAGE OF TOTAL DALYs
World	**2464.9**	**100**
1. Ischemic heart disease	164.0	6.7
2. Lower respiratory infections	103.0	4.3
3. Preterm birth complications	74.8	3.0
4. Hemorrhagic stroke	73.4	3.0
5. Diarrheal diseases	71.6	2.9
6. Neonatal encephalopathy	67.9	2.8
7. Diabetes mellitus	64.1	2.6
8. COPD	63.9	2.6
9. Low back pain	60.1	2.4
10. Malaria	55.8	2.3
Low- and Middle-Income Countries[a]		
1. Ischemic heart disease	138.2	6.4
2. Cerebrovascular disease	106.6	5.0
3. Lower respiratory infection	97.2	4.5
4. Preterm birth complications	72.7	3.4
5. Diarrheal disease	71.2	3.3
6. Hemorrhagic stroke	67.5	3.1
7. Neonatal encephalopathy	67.2	3.1
8. HIV/AIDS	65.8	3.1
9. Road injury	60.6	2.8
10. COPD	56.0	2.6
High-Income Countries[a]		
1. Ischemic heart disease	25.6	8.2
2. Low back pain	14.0	4.4
3. Lung cancer	11.5	3.7
4. Diabetes mellitus	10.8	3.4
5. Alzheimer's disease	9.5	3.0
6. Neck pain	8.3	2.6
7. Major depression	8.2	2.6
8. Hearing loss	8.0	2.6
9. COPD	7.8	2.6
10. Self-harm	6.9	2.2
Sub-Saharan Africa		
1. Malaria	49.8	9.4
2. HIV/AIDS	41.1	7.8
3. Lower respiratory infections	41.0	7.8
4. Diarrheal diseases	37.0	7.00
5. Neonatal sepsis	19.9	3.8
6. Neonatal encephalopathy	19.5	3.7
7. Preterm birth complications	17.6	3.3
8. Protein-energy malnutrition	14.7	2.8
9. Iron-deficiency anemia	11.2	2.1
10. Tuberculosis	10.9	2.1

[a]The World Bank classifies high-income countries as those whose gross national income per capita is ≥$12,476. Low- and middle-income countries are categorized as low-income (GNI per capita, <$1025), lower-middle income (GNI per capita, $1026–$4035), and upper-middle income (GNI per capita, $4036–$12,475) (*http://data.worldbank.org/about/country-classifications*).

Abbreviations: COPD, chronic obstructive pulmonary disease; DALYs, disability-adjusted life years.

Source: Institute for Health Metrics and Evaluation, University of Washington (2017). Data available at *http://www.healthmetricsandevaluation.org/gbd/visualizations/country.* Accessed July 18, 2017.

TABLE 460-2 Leading Causes of Death Worldwide, 2015

DISEASE OR INJURY	DEATHS (MILLIONS)	PERCENTAGE OF TOTAL DEATHS
World	**55.8**	**100**
1. Ischemic heart disease	8.9	16.0
2. Hemorrhagic stroke	3.3	6.0
3. COPD	3.2	5.7
4. Ischemic stroke	3.0	5.3
5. Lower respiratory infections	2.7	4.9
6. Alzheimer's disease	1.9	3.4
7. Lung cancer	1.7	3.1
8. Diabetes	1.5	2.7
9. Diarrheal diseases	1.3	2.4
10. Tuberculosis	1.1	2.0
Low- and Middle-Income Countries[a]		
1. Ischemic heart disease	7.0	15.5
2. Cerebrovascular disease	5.3	11.9
3. COPD	2.7	6.0
4. Lower respiratory infections	2.2	4.9
5. Diarrheal diseases	1.3	2.9
6. Diabetes	1.3	2.8
7. Road injury	1.2	2.7
8. HIV/AIDS	1.2	2.6
9. Tuberculosis	1.1	2.4
10. Lung cancer	1.1	2.4
High-Income Countries[a]		
1. Ischemic heart disease	1.9	18.0
2. Alzheimer's disease	0.9	8.6
3. Lung cancer	0.6	5.8
4. Ischemic stroke	0.6	5.4
5. Lower respiratory infection	0.5	4.9
6. COPD	0.5	4.5
7. Hemorrhagic stroke	0.4	3.6
8. Colon and rectum cancers	0.4	3.3
9. Diabetes mellitus	0.3	2.3
10. Breast cancer	0.2	1.8
Sub-Saharan Africa		
1. Lower respiratory infections	0.7	9.1
2. HIV/AIDS	0.7	8.8
3. Malaria	0.6	7.1
4. Diarrheal diseases	0.6	7.1
5. Ischemic heart disease	0.4	4.7
6. Tuberculosis	0.3	3.3
7. Hemorrhagic stroke	0.2	3.0
8. Neonatal sepsis	0.2	2.9
9. Neonatal encephalopathy	0.2	2.7
10. Preterm birth complications	0.2	2.5

[a]The World Bank classifies high-income countries as those whose gross national income per capita is ≥$12,476. Low- and middle-income countries are categorized as low-income (GNI per capita, <$1025), lower-middle income (GNI per capita, $1026–$4035), and upper-middle income (GNI per capita, $4036–$12,475) (*http://data.worldbank.org/about/country-classifications*).

Abbreviation: COPD, chronic obstructive pulmonary disease.

Source: Institute for Health Metrics and Evaluation, University of Washington (2017). Data available at *http://www.healthmetricsandevaluation.org/gbd/visualizations/country*. Accessed July 18, 2017.

the more than 360,000 deaths stemming from a lack of access to clean water and sanitation. Many of these modifiable factors lead to child and adult deaths from infectious pathologies; others lead to deaths from malignancies. Etiology and nosology are increasingly difficult to parse with regard to environmental harm. Risk factors such as indoor air pollution due to use of solid fuels, exposure to secondhand tobacco smoke, and outdoor air pollution account for 35% of lower respiratory infections globally. Various forms of unintentional injury and malaria top the list of health problems to which environmental factors contribute.

The third edition of *Disease Control Priorities in Developing Countries* (DCP3), published as a set of serial volumes based on content

area, provides evidence-based recommendations and cost-effectiveness analyses for numerous interventions, with attention to strategies for strengthening health systems. Cost-effectiveness analyses that compare relatively equivalent interventions in order to facilitate sound decisions under constraint are necessary; however, these analyses, as the DCP3 authors acknowledge, are unreliable when based on an incomplete knowledge of cost and evolving evidence of effectiveness. As both resources and objectives for global health initiatives grew, cost-effectiveness analyses (particularly those based on older evidence) sometimes steered policy makers and public health experts toward low-cost but ultimately ineffective interventions or away from higher-priced but effective ones. Thus we use the term *global health equity* to emphasize the need to ensure equitable access to high-value health interventions. To illustrate these points, it is instructive to look to HIV/AIDS, which in the course of the last four decades has become one of the world's leading infectious causes of adult death.

HIV INFECTION/AIDS

Chapter 197 provides an overview of the global HIV epidemic today. Approximately 38.8 million people worldwide were living with HIV infection in 2015; >10.3 million of those in sub-Saharan African countries were then receiving antiretroviral therapy (ART)—a two-fold increase over the corresponding figure for 2010, when ART was already the largest therapeutic rollout in the continent's history. Here the discussion will be limited to HIV/AIDS in the developing world. Lessons learned from tackling HIV/AIDS in low-resource settings are highly relevant to discussions of other chronic diseases, including NCDs, for which effective therapies have been developed. In the United States, after the mid-1990s, ART transformed HIV infection from an inescapably fatal disease into a manageable chronic illness. Across high-income countries, improved ART has dramatically prolonged life expectancy for people living with HIV infection, which now approaches that of the general population. This success rate exceeds that obtained with almost any treatment for adulthood cancer or for complications of coronary artery disease. In developing countries, treatment has been offered broadly only since 2003. Before 2003, many arguments were raised to justify not moving forward rapidly with ART programs for people living with HIV/AIDS in resource-limited settings. The standard litany included the price of therapy compared with the poverty of patients, the complexity of the intervention, the lack of infrastructure for laboratory monitoring, and the lack of trained health care providers. Narrow cost-effectiveness arguments that created false dichotomies—prevention *or* treatment rather than their synergistic integration—too often went unchallenged by policy makers, public health experts, and health economists. As a cumulative result of these delays in the face of health disparities both old and new, there were millions of premature deaths.

Disparities in access to HIV treatment did give rise to widespread moral indignation and a new type of health activism. In several middle-income countries, including Brazil, public programs have helped bridge the global access gap. Other innovative projects pioneered by international NGOs in diverse settings such as Haiti and Rwanda have established that a simple approach to ART based on intensive community engagement and social and economic support for patients and their village-based health workers can achieve remarkable results **(Fig. 460-2)**.

During the past decade, the availability of ART has increased sharply in the low- and middle-income countries that have borne the greatest burden of the HIV/AIDS pandemic. In 2000 very few people living with HIV/AIDS in these nations had access to ART, whereas by 2015 46% of people living with HIV infection were receiving ART, including 51% of people living with HIV in sub-Saharan Africa and 41% in southern Asia. It remains to be seen how many of these people are receiving ART regularly and with the requisite social support. In light of these dramatic gains, coverage targets have grown more ambitious; for example, the UNAIDS 90-90-90 Initiative aims to have 90% of people living with HIV know their status, 90% of those detected treated with ART, and 90% of those on treatment achieving viral load suppression by 2020.

This scale-up was made possible by a number of developments: a staggering drop in the cost of generically manufactured ART, the development of a standardized approach to treatment, substantial investments by funders, and the political commitment of governments to afford ART as a public good. Civil-society AIDS activists spurred many of these efforts.

Starting in the early 2000s, a combination of factors, including work by the Clinton Foundation HIV/AIDS Initiative and Médecins Sans Frontières, led to the availability of generic ART medications. While first-line ART cost more than $10,000 per patient per year in 2000, first-line regimens in low- and middle-income countries are now available for <$100 per year. At the same time, fixed-dose combinations made multidrug regimens easier to administer. Also around this time, the WHO began advocating a public health approach to the treatment of people with AIDS in low-resource settings; this approach promised—thanks to dropping viremia—to lower transmission rates and, if universally available, to end almost all mother-to-child transmission. Derived from models of care pioneered by the NGO Partners in Health and other groups, this approach proposed the use of standard first-line treatment regimens based on a simple five-drug formulary, with a more complex (and more expensive) set of second-line options in reserve. Clinical protocols were standardized, and intensive training packages for health professionals and community health workers were developed and implemented in many countries. Early rollout efforts were supported by new funding from the Global Fund and PEPFAR. In 2003, lack of access to ART was declared a global public health emergency by the WHO and UNAIDS, and those two agencies launched the 3 by 5 Initiative, setting an ambitious target: to have 3 million people in developing countries on treatment by the end of 2005. Many countries set corresponding national targets and have worked to integrate ART into their national AIDS programs and health systems and to harness the synergies between HIV/AIDS treatment and prevention activities. External funding to fight HIV/AIDS in such settings increased dramatically during this period and beyond, rising from $300 million in 1996 to over $10.8 billion in 2015. The integration of prevention and care led to a sharp drop in transmission—a 96% decline according to one review of the impact of ART rollout in heavily burdened countries in Africa and the Caribbean.

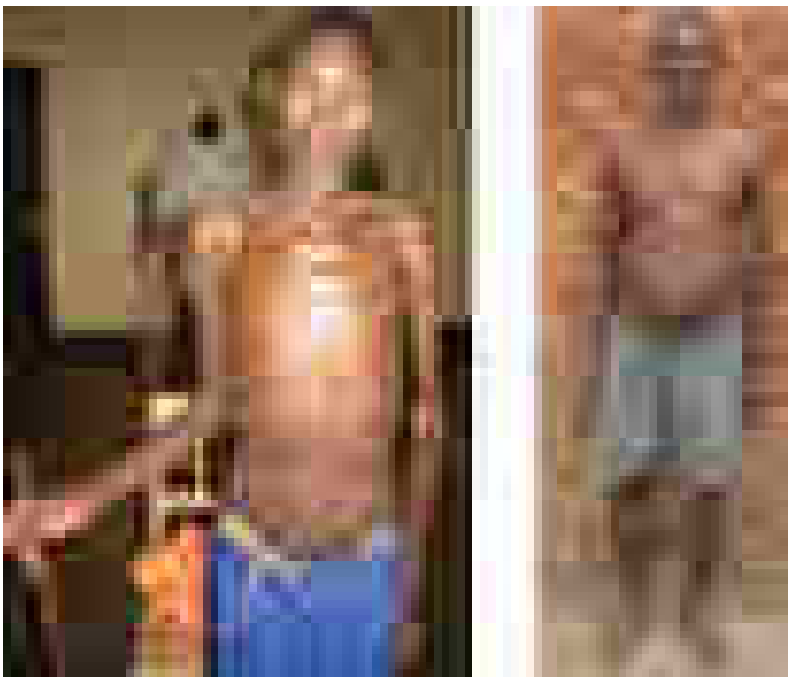


FIGURE 460-2 An HIV/TB-co-infected patient in Rwanda before (*left*) and after (*right*) 6 months of treatment.

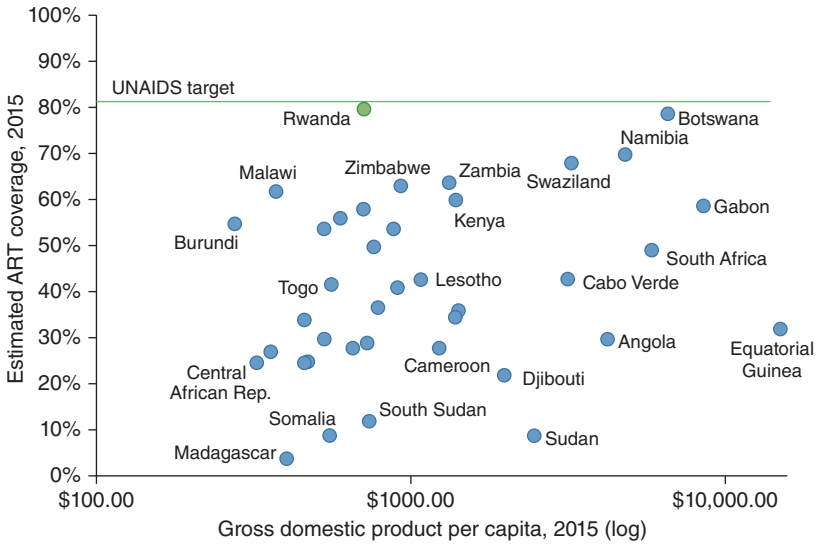


FIGURE 460-3 Antiretroviral therapy (ART) coverage (percentage of people living with HIV who received ART) in sub-Saharan Africa, 2015. *(Source: World Bank; data available through data.worldbank.org.)*

Further lessons with implications for policy and action have come from efforts now under way among lower-income countries. Rwanda provides an example: since 2000, mortality from HIV disease has fallen by >80% as the country—despite its relatively low gross national income **(Fig. 460-3)**—has provided almost universal access to ART. The reasons for this success include strong national leadership, evidence-based policy, cross-sector collaboration, community-based care, and a deliberate focus on a health-systems approach that embeds HIV/AIDS treatment and prevention in the primary health care service delivery platform. As we will discuss later in this chapter, these principles can be applied to other conditions, including NCDs.

TUBERCULOSIS

Chapter 173 provides a concise overview of the pathophysiology and treatment of tuberculosis. In 2015, an estimated 1.4 million people died from *Mycobacterium tuberculosis* infection; this figure made tuberculosis the leading infectious killer of adults globally. The disease is closely linked to HIV infection in much of the world: of the 10.4 million estimated new cases of tuberculosis in 2015, 1.2 million occurred among people living with HIV. A much more substantial proportion of the resurgence of tuberculosis registered in southern Africa is attributed to HIV co-infection. Even before the advent of HIV, however, it was estimated that fewer than one-half of all cases of tuberculosis in developing countries were ever diagnosed. Primarily because of the common failure to diagnose and treat tuberculosis, international authorities devised a single strategy to reduce the burden of disease. In the early 1990s, the World Bank, the WHO, and other international bodies promoted the DOTS strategy (*d*irectly *o*bserved *t*herapy using *s*hort-course isoniazid- and rifampin-based regimens) as highly cost-effective. Passive case-finding of smear-positive patients was central to the strategy, as was an uninterrupted drug supply.

DOTS was clearly effective for most uncomplicated cases of drug-susceptible tuberculosis, but a number of shortcomings were soon identified. First, the diagnosis of tuberculosis based solely on sputum smear microscopy—a method dating from the late nineteenth century—is not sensitive. Many cases of pulmonary tuberculosis and all cases of exclusively extrapulmonary tuberculosis are missed by smear microscopy, as are most cases of active disease in children. Second, passive case-finding relies on the availability of health care services, which is uneven in the settings where tuberculosis is most prevalent. Third, patients with multidrug-resistant tuberculosis (MDR-TB) are by definition infected with strains of *M. tuberculosis* resistant to isoniazid and rifampin; thus exclusive reliance on these drugs is unwarranted in settings in which drug resistance is an established problem.

The crisis of antibiotic resistance registered in U.S. hospitals is not confined to the industrialized world or to common bacterial infections. While the great majority of patients sick with and dying from tuberculosis are afflicted with strains susceptible to all first-line drugs, a substantial minority of patients with tuberculosis in some settings are infected with strains of *M. tuberculosis* resistant to at least one first-line antituberculosis drug. Globally in 2015, an estimated 3.9% of all patients with new *M. tuberculosis* infections and 21% of all previously treated patients were infected with MDR strains; most of these cases resulted from primary transmission. It was clear that poor infection control in hospitals and clinics in the face of delays in the initiation of effective therapy led to explosive and lethal epidemics due to these strains. To improve DOTS-based responses to MDR-TB, global health authorities adopted DOTS-Plus, which adds the diagnostics and drugs necessary to manage drug-resistant disease. Even as DOTS-Plus was being piloted in resource-constrained settings, however, new strains of extensively drug-resistant (XDR) *M. tuberculosis* (resistant to isoniazid and rifampin, any fluoroquinolone, and at least one injectable second-line drug) had already threatened the success of tuberculosis control programs in beleaguered South Africa, for example, where high rates of HIV infection had led to a doubling in the incidence of tuberculosis over the preceding decade. Gene probes of cultures of infected sputum and tissues suggest that patients may be infected by more than one strain. Despite the poor capacity for detection of MDR- and XDR-TB in most resource-limited settings, an estimated 580,000 cases of MDR-TB were thought to occur in 2015. Approximately 9.5% of these cases were caused by XDR strains.

TUBERCULOSIS AND AIDS AS CHRONIC DISEASES: LESSONS LEARNED

Strategies effective against MDR-TB have implications for the management of drug-resistant HIV infection and even drug-resistant malaria, which, through repeated infections and a lack of effective therapy, has become a chronic disease in parts of Africa (see "Malaria," below). As new therapies, whether for tuberculosis or for hepatitis C infection, become available, many of the problems encountered in the past will recur. Indeed, examining AIDS and tuberculosis as chronic diseases—instead of simply communicable ones—makes it possible to draw a number of conclusions, many of them pertinent to global health equity in general.

First, the chronic infections discussed here are best treated with multidrug regimens to which the infecting strains are susceptible. This is true of chronic infections due to many bacteria, fungi, parasites, or viruses; even acute infections such as those caused by *Plasmodium* species are not reliably treated with a single drug.

Second, charging fees for AIDS prevention and care poses insurmountable problems for people living in poverty, many of whom are unable to pay even modest amounts for services or medications. Like efforts to battle airborne tuberculosis, such services might best be seen as a public good promoting public health. Initially, a subsidy approach will require sustained donor contributions, but many African countries have set targets for increased national investments in health—a pledge that could render ambitious programs sustainable in the long run, as the Rwanda experience suggests. Meanwhile, as local investments increase, the price of AIDS care continues to decrease. The use of generic medications means that ART can now cost <$0.25 per day.

Third, the effective scale-up of pilot projects requires strengthening and sometimes rebuilding of health care systems, including those charged with delivering primary care. In the past, the lack of health care infrastructure has been cited as a barrier to providing ART in the world's poorest regions; however, AIDS resources, which are at last considerable, may be marshaled to rebuild public health systems in sub-Saharan Africa and other HIV-burdened regions—precisely the settings in which tuberculosis is resurgent. Failure to pursue such a

health-systems approach after civil wars ended in Sierra Leone and Liberia accounts for much of their extreme vulnerability to Ebola a decade later.

Fourth, the lack of trained health care personnel, most notably doctors and nurses, is incorrectly invoked as a primary reason for failure to treat AIDS in poor countries and must still be addressed. The WHO recommends a minimum of 4.45 physicians, nurses, and midwives per 1000 persons, but recent reports from that organization and others confirm that many countries, especially in sub-Saharan Africa, fall far short of those target numbers. Specifically, 44% of WHO member states reportedly have <1 physician per 1000 population. In contrast, the United States and Cuba report 2.55 and 7.52 doctors per 1000 population, respectively. Similarly, about 48% of WHO member states report having fewer than three nurses and midwives per 1000 population. Further inequalities in health care staffing exist within countries. Rural–urban disparities in health care personnel mirror disparities of both wealth and health. For instance, nearly 90% of Malawi's population lives in rural areas, but more than 95% of clinical officers work at urban facilities, and 47% of nurses work in urban tertiary-care facilities. Even community health workers trained to provide first-line services to rural populations often transfer to urban districts.

In what is termed the "brain drain," many physicians and nurses emigrate from their home countries to pursue opportunities abroad, leaving behind health systems that are understaffed and ill-equipped to deal with either emergencies like Ebola or the usual burden of disease. One reason doctors and nurses leave sub-Saharan Africa and other resource-poor areas is that they lack the tools to practice there. Funding for "vertical" (disease-specific) programs can be used not only to strengthen health systems but also to recruit and train physicians and nurses to underserved regions where they, in turn, can help to train and then work with community health workers in supervising care for patients with AIDS and many other diseases within their communities. Such training should be undertaken even where physicians are abundant, since close community-based supervision represents the highest standard of care for chronic disease, whether in developing or developed countries. The United States, which has a dearth of health care providers in many of its poor and rural communities, has much to learn from Rwanda in this regard.

Fifth, the many barriers to adequate health care and patient adherence that are raised by extreme poverty can be removed only with the deployment of "wrap-around services": food supplements for the hungry, help with transportation to clinics, child care, and housing. Extreme poverty makes it difficult for many patients to comply with therapy for chronic diseases, whether communicable or not. Experience shows, however, that these many barriers can be more readily surmounted than the extreme poverty itself to which chronic disease and acute infection contribute substantially. Indeed, poverty in its many dimensions is far and away the greatest obstacle to the scale-up of treatment and prevention services.

Finally, there is a need for a renewed basic-science commitment to the discovery and development of vaccines; more reliable, less expensive diagnostic tools; and new classes of therapeutic agents. This need applies not only to the three leading infectious killers—against none of which there is an effective vaccine—but also to most other neglected diseases of poverty.

MALARIA

Chapter 219 reviews the etiology, pathogenesis, and clinical treatment of malaria, the world's fifth-ranking infectious killer. In 2015, there were ~212 million cases of malaria, and the disease is thought to have killed 429,000 people; 70% of these deaths (~303,000) occurred among children <5 years old. The poor disproportionately experience the burden of malaria: >75% of estimated malaria deaths occur in just 13 countries, and mortality rates are highest in sub-Saharan Africa. The Democratic Republic of the Congo and Nigeria account for >36% of total estimated malaria deaths globally.

Malaria's human cost has been enormous, with the highest toll among children—especially African children—living in poverty. Macroeconomic analyses estimate that malaria may reduce the per capita gross national product of a disease-endemic country by 50% relative to that of a nonmalaria-endemic country. The causes of this drag include impaired cognitive development of children, decreased schooling, decreased savings, decreased foreign investment, and restriction of worker mobility. In light of this enormous cost, it is little wonder that an important review by the economists Sachs and Malaney concludes that "where malaria prospers most, human societies have prospered least."

Microeconomic analyses focusing on direct and indirect costs estimate that malaria may consume >10% of a household's annual income. A study in rural Kenya shows that mean direct-cost burdens vary between the wet and dry seasons (7.1% and 5.9% of total household expenditure, respectively) and that this proportion is >10% in the poorest households in both seasons. A Ghanaian study that categorized the population by income group highlighted the regressive nature of this cost: responding to malaria consumed only 1% of a wealthy family's income but 34% of a poor household's income.

In part because of differences in vector distribution and climate, resource-rich countries offer few blueprints for malaria control and treatment that are applicable in tropical (and resource-poor) settings. In 2001, African heads of state endorsed the WHO Roll Back Malaria (RBM) campaign, which prescribes strategies appropriate for sub-Saharan African countries. In 2008, the RBM partnership launched the Global Malaria Action Plan (GMAP). This strategy integrates prevention and care and calls for the avoidance of single-dose regimens and an awareness of existing drug resistance; the use of insecticide-treated bed nets (ITNs); indoor residual spraying; artemisinin-based combination therapy (ACT); intermittent preventive treatment during pregnancy; prompt diagnosis; and other vector control measures such as larviciding and environmental management.

Between 2000 and 2015, global malaria deaths were reduced by an estimated 62%, a figure equating to some 6.8 million deaths averted. Again, the experience in Rwanda is instructive: from 2000 to 2015, malaria deaths dropped by >85% for the same reasons mentioned earlier in recounting that nation's successes in battling HIV. A recent resurgence there has been linked to inadequately treated ITNs and other delivery failures, rising temperatures, and reintroduction from surrounding countries (e.g., by refugees from Burundi).

Meeting the challenge of malaria control will continue to require careful study of appropriate preventive and therapeutic strategies in the context of an increasingly sophisticated molecular understanding of pathogen, vector, and host. However, an appreciation of the economic and social devastation wrought by malaria—like that inflicted by diarrhea, AIDS, and tuberculosis—on the most vulnerable populations should heighten the level of commitment to critical analysis of ways to implement proven strategies for prevention and treatment.

Funding from the Global Fund, the Gates Foundation, the World Bank's International Development Association, and the U.S. President's Malaria Initiative, along with leadership from public health authorities, is critical to sustain the benefits of prevention and treatment. Building on the growing momentum of the last decade with adequate financial support, innovative strategies, and effective tools for prevention, diagnosis, and treatment, we may yet achieve the goal of a world largely free of malaria.

EBOLA

Chapter 205 provides an overview of the epidemiology, pathogenesis, and clinical manifestations of Ebola virus and Marburg virus infections. The 2013–2016 outbreak of Ebola virus disease in West Africa was the largest documented Ebola epidemic to date, with more than 28,000 recorded cases and 11,000 recorded deaths. Such figures underestimate the true burden, however, given inaccuracies in reporting and a growing body of evidence documenting the occurrence of minimally symptomatic Ebola virus infection.

Prior to the outbreak, the health systems of the three most affected countries—Liberia, Guinea, and Sierra Leone—were among the world's weakest. Histories of extractive colonial and postcolonial commerce, the conditional aid policies of international financial institutions, recent civil conflict, and under-resourced health ministries

left this part of West Africa bereft of the means to deliver modern medicine and promote public health. In 2013, Sierra Leone had the world's highest maternal mortality ratio, with 1460 deaths per 100,000 live births. According to one estimate, Liberia had just 51 physicians working in the entire country before the Ebola epidemic, or roughly one physician per 100,000 people. Clinics and hospitals were scarce across the region, especially in rural areas, and routinely lacked drugs, supplies, electricity, running water, laboratories, and personal protective equipment for the prevention of nosocomial infection. Such deficits were not surprising given these countries' meager public and private expenditures on health. In 2012, for example, the Guinean and Sierra Leonean governments allocated just 7 and 12% of their annual budgets, respectively, to health, according to WHO figures. Rwanda, in contrast, spent 22% of its national budget on health that year, the highest in the WHO African region.

The unprecedented scale of the recent West African Ebola epidemic was largely a symptom of these chronically weak health systems. As a result, clinicians, patients' families, and other caregivers—tasked with nursing the sick and interring the dead, but lacking the means to do so safely—faced disproportionately high risks of Ebola infection. Health facilities with poor infection control and unsafe burials served as amplifiers of transmission.

The quest to contain Ebola in West Africa was one of the largest global public health efforts in recent history, but it was far from ambitious clinically. As in previous Ebola outbreaks, preventing new infections was often prioritized over improving survival among those already infected, leading to substandard care for most West African patients and high case–fatality rates—by WHO estimates, ~70%. However, in settings in which quality supportive and critical care could be provided, clinical outcomes among Ebola-infected patients affirmed that Ebola virus disease is treatable, even in the absence of specific antiviral therapies and experimental drugs. Of the 27 Ebola patients treated in the United States and Europe in 2014 and 2015, more than 80% survived. This disparity of outcomes was also registered between European and African outbreaks of Marburg virus disease decades earlier.

As with efforts to combat AIDS and tuberculosis, the global response to Ebola reveals the unintended consequences of pitting preventive strategies against therapeutic ones—and the pull of debates about scarcity. Misguided (and often contradictory) public health messaging, distrust of disease-control and social mobilization teams, punitive containment measures, and the unavailability of safe Ebola treatment units capable of delivering effective clinical care deterred individuals from presenting to health facilities, reporting symptomatic patients and their contacts, and cooperating with epidemic response activities. The resulting epidemic of mistrust facilitated the further spread of new infections by impeding surveillance, timely diagnosis, contact tracing, and patient isolation.

The aftermath of this West African epidemic provides valuable opportunities to improve global responses to acute health crises while addressing the chronic ones that fuel them. Funds pledged to support short-term, outbreak-response operations in West Africa often went undisbursed, highlighting the paucity of measures to hold donors accountable.

Moreover, the withdrawal of humanitarian organizations and of funding earmarked solely for disease containment illustrates the inadequacies of emergency responses detached from long-term efforts to strengthen national health systems. In the absence of such systems and in the face of national economies burdened by high youth unemployment, the clinical, social, and economic needs of thousands of Ebola survivors remain vast and largely neglected. The epidemic also claimed sizeable proportions of the region's health workforce, shuttered or incapacitated health centers, and collapsed the delivery of even the modest clinical and public health interventions that had been delivered prior to Ebola's explosive spread. Yet the public sectors of Liberia, Guinea, and Sierra Leone continue to lack adequate resources to rebuild their economies and health systems: estimates as of June 2017 show that, of the roughly $9.1 billion required to finance the national and regional plans presented at a 2015 United Nations conference on Ebola recovery, only $4.5 billion was pledged, of which just 26.4% has been disbursed. It remains unclear to what extent disbursed funds have been channeled through recipient governments (or other national institutions) and whether donors and implementing partners will allow these governments the flexibility to deploy the funds to best correspond to a rapidly changing burden of disease.

"NONCOMMUNICABLE" CHRONIC DISEASES

Although the burden of communicable diseases—especially HIV infection, tuberculosis, and malaria—still accounts for the majority of deaths in resource-poor regions within sub-Saharan Africa and in the poorest reaches of several first-world cities, 71.3% of all deaths worldwide in 2015 were attributed to NCDs. Although we use this term to describe cardiovascular diseases, cancers, diabetes, and chronic lung diseases, this usage masks important distinctions. For instance, two significant NCDs in low-income countries, rheumatic heart disease (RHD) and cervical cancer, represent the chronic sequelae of infections with group A *Streptococcus* and human papillomavirus, respectively, and it is in these countries that the burden of disease due to NCDs is rising most rapidly: close to three-quarters of deaths attributable to NCDs occur in low- and middle-income countries, which also account for 82% of all early NCD-related deaths—a figure representing ~16 million people and exceeding the total number of deaths due to AIDS, tuberculosis, and malaria combined. By 2020, NCDs will account for 80% of the global burden of disease and for seven of every 10 deaths in developing countries. The recent increase in resources for and attention to communicable diseases is both welcome and long overdue, but developing countries are already carrying a "double burden" of communicable and noncommunicable diseases.

Diabetes, Cardiovascular Disease, and Cancer: A Global Perspective In contrast to tuberculosis, HIV infection, and malaria—diseases caused by single pathogens that damage multiple organs—cardiovascular diseases reflect injury to a single organ system downstream of a variety of insults, both infectious and noninfectious. Some of these insults result from rapid changes in diet and labor conditions; others are of a less recent vintage. The burden of cardiovascular disease in low-income countries represents one consequence of decades of neglect of health systems. Furthermore, cardiovascular research and investment have long focused on the ischemic conditions that are increasingly common in high- and middle-income countries.

Predictions of an imminent rise in the share of deaths and disabilities due to NCDs in developing countries have led to calls for preventive policies to improve diet, increase exercise, and restrict tobacco use, along with the prescription of multidrug regimens for persons at high-level vascular risk. Although this agenda could do much to prevent pandemic NCDs, it will do little to help persons with established heart disease stemming from nonatherogenic pathologies.

The misperception of cardiovascular diseases as a problem primarily of elderly populations in middle- and high-income countries has contributed to the neglect of these diseases by global health institutions, including regionally focused ones. Even in Eastern Europe and Central Asia, where the collapse of the Soviet Union was followed by a catastrophic surge in cardiovascular disease deaths (mortality rates from ischemic heart disease nearly doubled between 1991 and 1994 in Russia, for example), the modest flow of overseas development assistance to the health sector during these troubled years focused on the communicable causes that accounted for <1 in 20 excess deaths during that period. Even these focused investments were inadequate to halt resurgent tuberculosis (much of it caused by MDR strains), which is only now coming under control in much of the region.

DIABETES The International Diabetes Federation reports that the number of diabetic patients in the world is expected to increase from 415 million in 2015 to 642 million by 2040—nearly one in 10 adults. Already, a significant proportion of diabetic patients live in developing countries where, because those affected are far more frequently between ages 40 and 59, the complications of micro- and macrovascular disease take a far greater toll. Globally, these complications are a major cause of disability and reduced quality of life: a high fasting plasma glucose level ranks third among risks for disability and

global mortality. The GBD 2015 estimates that diabetes accounted for 1.5 million deaths in 2015; more than 80% of these deaths occurred in low- and middle-income countries.

CARDIOVASCULAR DISEASE Because systemic investigation of the causes of stroke and heart failure in sub-Saharan Africa has begun only recently, little is known about the impact of elevated blood pressure in this portion of the continent. Modestly elevated blood pressure in the absence of tobacco use in populations with low rates of obesity may confer little risk of adverse events in the short run. In contrast, persistently elevated blood pressure above 180/110 goes largely undetected, untreated, and uncontrolled in this part of the world. In the cohort of men assessed in the Framingham Heart Study, the prevalence of blood pressures above 210/120—severe hypertension—declined from 1.8% in the 1950s to 0.1% by the 1960s with the introduction of effective antihypertensive agents. Although debate continues about appropriate screening strategies and treatment thresholds, Africa's rural health centers, run largely by nurses, must quickly gain access to antihypertensive medications.

The epidemiology of heart failure also reflects inequalities in risk factor prevalence and in access to therapy. The reported burden of this condition has remained unchanged since the 1950s, but the causes of heart failure and the age of the people affected vary across the globe. Heart failure as a consequence of pericardial, myocardial, endocardial, or valvular injury accounts for as many as 5% of all medical admissions to hospitals around the world. In high-income countries, coronary artery disease and hypertension among the elderly account for most cases of heart failure. Among the world's poorest 1 billion people, however, heart failure reflects poverty-driven exposure of children and young adults to rheumatogenic strains of streptococci and cardiotropic microorganisms (e.g., HIV, *Trypanosoma cruzi*, enteroviruses, *M. tuberculosis*), untreated high blood pressure, and nutrient deficiencies. The mechanisms underlying other causes of heart failure common in these populations—such as idiopathic dilated cardiomyopathy, peripartum cardiomyopathy, and endomyocardial fibrosis—remain unclear.

In stark contrast to the extraordinary lengths to which clinicians in wealthy countries will go to treat ischemic cardiomyopathy among elderly patients, little attention has been paid to young patients with non-ischemic cardiomyopathies in resource-poor settings. Non-ischemic cardiomyopathies, such as those due to hypertension, RHD, and chronic lung disease, account for >90% of cases of cardiac failure in sub-Saharan Africa and include poorly understood entities such as peripartum cardiomyopathy (which has an incidence in rural Haiti of one per 300 live births) and HIV-associated cardiomyopathy. Lessons learned in the scale-up of chronic care for HIV infection and tuberculosis may be illustrative as progress is made in establishing the means to deliver heart-failure medications to these patients.

Some of the lessons learned from the chronic infections discussed above are, of course, relevant to cardiovascular disease, especially those classified as NCDs but caused by infectious pathogens. Integration of prevention and care remains as important today as in 1960 when Paul Dudley White and his colleagues found little evidence of myocardial infarction in the region near the Albert Schweitzer Hospital in Lambaréné, Gabon, but reported that "the high prevalence of mitral stenosis is astonishing." They termed it a duty to integrate prevention with penicillin prophylaxis and care, including medical management and surgery, when indicated. "The same responsibility," they agreed, "exists for those with correctable congenital cardiovascular defects."

RHD affects more than 30 million people worldwide, with more than 282,000 new cases each year. Among the 2.4 million cases of pediatric RHD, an estimated 42% occur in sub-Saharan Africa. This disease, which may cause endocarditis or stroke, leads to >233,000 deaths per year—almost all occurring in developing countries. A survey of acute heart failure among adults in sub-Saharan Africa showed that ~14.3% of these cases were due to RHD. Researchers in Ethiopia have reported annual death rates as high as 12.5% in rural areas. In part because the prevention of RHD has not advanced since the disease's disappearance in wealthy countries, no part of sub-Saharan Africa has eradicated RHD despite examples of success in Costa Rica, Cuba, and some Caribbean nations.

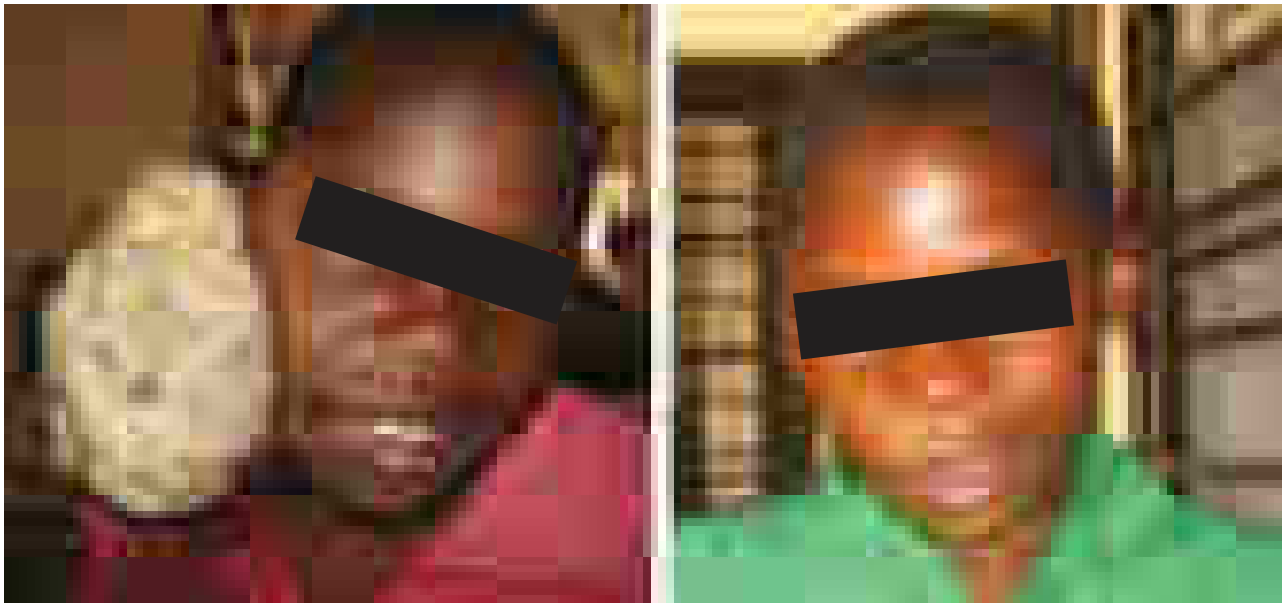
Strategies to eliminate RHD may depend on active case-finding, with confirmation by echocardiography, among high-risk groups as well as on efforts to expand access to surgical interventions among children with advanced valvular damage. Partnerships between established surgical programs and areas with limited or nonexistent facilities may help expand the capacity to provide lifesaving interventions to patients who otherwise would die early and painfully. Such partnerships can speed the further development of regional centers of excellence equipped to provide consistent, accessible, high-quality services to those now without them.

CANCER Low- and middle-income countries accounted for more than two-thirds of the 17.5 million cases and 8.7 million deaths due to cancer worldwide in 2015. By 2030, annual mortality from cancer is expected to increase by 4 million—with developing countries experiencing a sharper increase than developed nations. "Western" lifestyle changes may be responsible for the increased incidence of cancers of the breast, colon, and prostate among populations in low- and middle-income countries, but historic realities, sociocultural and behavioral factors, genetics, and poverty itself already have a profound impact on cancer-related mortality and morbidity rates. In 2012, 15.4% of the more than 14 million cases of cancer globally were attributable to infectious causes, which are responsible for <10% of cancers in developed countries but account for up to 25% of all malignancies in low- and middle-income countries. Infectious causes of cancer such as human papillomavirus, hepatitis B virus, and *Helicobacter pylori* will continue to have a much larger impact in developing countries. Environmental and dietary factors, such as indoor air pollution and high-salt diets, also contribute to increased rates of certain cancers (e.g., lung and gastric cancers). Tobacco use (both smoking and chewing) is the most important source of increased mortality rates from lung and oral cancers. In contrast to decreasing tobacco use in many developed countries, the number of smokers is growing in developing countries, especially among women and young persons.

For many reasons, outcomes of malignancies are far worse in developing countries than in developed nations. As currently funded, overstretched health systems in poor countries are not capable of early detection; at the time of tissue diagnosis, the majority of patients already have incurable malignancies. Treatment of cancers is available for only a very small number of mostly wealthy citizens in the majority of poor countries, and, even when treatment is available, the range and quality of services are often substandard. Yet this need not be the future. Fifteen years ago, MDR-TB and HIV infections were widely deemed untreatable in settings of great poverty. The feasibility of creating innovative programs that reduce technical and financial barriers to the provision of care for treatable malignancies among the world's poorest populations is now clear **(Fig. 460-4)**. Several middle-income countries, including Mexico, have expanded publicly funded cancer care to reach poorer populations. This commitment of resources has dramatically improved outcomes for cancers, from childhood leukemia to cervical cancer.

Prevention of Noncommunicable Diseases

False dichotomies, including those pitting prevention against care, persist in global health and reflect, in part, outmoded paradigms or a limited understanding of shifts in disease burden and causality as well as the dramatic variations in risk within a single nation. Moreover, such dichotomies or debates are sometimes politicized as a result of vested interests. Although globalization has had many positive effects, one negative effect has been the growth in both developed and developing countries of well-financed lobbies that have aggressively promoted unhealthy dietary changes and increased consumption of alcohol and tobacco. The WHO's 2003 Framework Convention on Tobacco Control represented a major advance, committing all of its signatories to a set of policy measures shown to reduce tobacco consumption. In 2004, the WHO released its Global Strategy on Diet, Physical Activity, and Health, which focused on population-level promotion of healthy diet and regular physical activity in an effort to reduce the growing global problem of obesity. Passing this strategy at the World Health Assembly proved difficult because of strong opposition from the food industry



A B

FIGURE 460-4 An 11-year-old Rwandan patient with embryonal rhabdomyosarcoma before (***left***) and after (***right***) 48 weeks of chemotherapy plus surgery. Eleven years later, she is healthy with no evidence of disease.

and from a number of WHO member states, including the United States. The strategy fails to focus on the NCD risk factors of the bottom billion.

The WHO estimates that 80% of all cases of cardiovascular disease and type 2 diabetes as well as 40% of all cancers can be prevented through healthier diets, increased physical activity, and avoidance of tobacco. These estimates mask large local variations. Although some evidence indicates that population-based measures can have some impact on these behaviors, it is sobering to note that increasing obesity levels have not been reversed in any population. Tobacco avoidance may be the most important and most difficult behavioral modification of all. In the twentieth century, 100 million people worldwide died of tobacco-related diseases; it is projected that >1 billion people will die of these diseases in the twenty-first century, with the vast majority of those deaths in developing countries. Today, >80% of the world's 1 billion smokers live in low- and middle-income countries. If trends continue, tobacco-related deaths will increase to 8 million per year by 2030, with 80% of those deaths in low- and middle-income countries. Investment in curbing NCDs remains disproportionately low despite the WHO's 2008–2013 Action Plan for the Global Strategy for the Prevention and Control of Noncommunicable Diseases.

MENTAL HEALTH

The WHO reports that some 676 million people worldwide suffer from depression and anxiety disorders; one in four patients visiting a health service has at least one mental, neurologic, or behavioral disorder, but most of these disorders are neither diagnosed nor treated. More than 800,000 people die by suicide every year, and major depression is the leading cause of years lost to disability in the world today. Most low- and middle-income countries devote <1% of their health expenditures to mental health.

Increasingly effective therapies exist for many of the major causes of mental disorders. One of the greatest barriers to delivery of such therapies is the paucity of skilled personnel. Most sub-Saharan African countries have only a handful of psychiatrists, for example; almost all of them practice in cities and are unavailable within the public sector or to patients living in poverty. Among the few patients who are fortunate enough to see a psychiatrist or neurologist, fewer still are able to adhere to treatment regimens: several surveys of already diagnosed patients ostensibly receiving daily therapy have revealed that, among the poor, multiple barriers prevent patients from taking their medications as prescribed. In one study from Kenya, no patients being seen in an epilepsy clinic had therapeutic blood levels of antiseizure medications, even though all had been prescribed these drugs. Moreover, many patients in this study had no detectable blood levels of these agents. The same barriers that prevent the poor from having reliable access to insulin or ART prevent them from benefiting from antidepressant, antipsychotic, and antiepileptic agents. To alleviate this problem, some authorities are proposing the training of health workers to provide community-based adherence support, counseling services, and referrals for patients in need of mental health services. One such program instituted in Goa, India, used lay counselors and resulted in a significant reduction in symptoms of common mental disorders among the target population.

CONCLUSION: TOWARD A SCIENCE OF IMPLEMENTATION

There is a long way to go before evidence-based internal medicine is applied effectively among the world's poor. Public health strategies draw largely on quantitative methods—epidemiology, biostatistics, and economics. Clinical practice, including the practice of internal medicine, draws on a rapidly expanding knowledge base and remains focused on individual patient care; clinical interventions are rarely population-based. However, global health equity depends on avoiding the false dichotomies of the past: neither public health nor clinical approaches alone are adequate to address the problems of global health. The integration of prevention and care, along with adequate funding, has shown that complex infectious diseases such as HIV/AIDS and tuberculosis are not impossible to manage, even though drug resistance and lack of effective health systems have complicated such work. Beyond what is usually termed communicable disease—i.e., in the arena of chronic diseases such as cardiovascular disease and mental illness—global health is a nascent endeavor. Efforts to address any one of these problems in settings of great scarcity need to be integrated into broader efforts to strengthen failing health systems and alleviate the growing personnel crisis within these systems. Such efforts must include the building of platforms for care delivery that are robust enough to incorporate new preventive, diagnostic, and therapeutic technologies rapidly in response to changes both in the burden of disease and in the needs not met by existing paradigms and systems of care delivery.

Academic medical centers have tried to address this "know–do" gap as new technologies are introduced and assessed through clinical trials, but the reach of these institutions into settings of poverty is limited

in rich and poor countries alike. When such centers link their capacities effectively to the public institutions charged with the delivery of health care to the poor, great progress can be made. For these reasons, scholarly work and practice in the field once known as "international health" and now often designated *global health equity* are changing rapidly. That work is still informed by the tension between clinical practice and population-based interventions, between analysis and action, and between prevention and care.

A number of university hospitals are developing training programs for physicians with an interest in global health. In medical schools across the United States and in other wealthy countries, interest in global health has exploded. One study has shown that more than 25% of medical students take part in at least one global health experience prior to graduation. Half a century or even a decade ago, such high levels of interest would have been unimaginable.

An estimated 400 million people lack access to essential health services; the consequence is millions of preventable deaths each year. An absolute majority of these premature deaths occur in Africa, with the poorer regions of Asia not far behind. They include deaths from vaccine-preventable illness, deaths during childbirth, deaths from infectious diseases that might be cured with access to antibiotics and other essential medicines, deaths from malaria that would have been prevented by ITNs and access to therapy, and deaths from waterborne illnesses. Other excess mortality is attributable to the inadequacy of efforts to develop new preventive, diagnostic, and therapeutic tools.

The development of tools must be followed quickly by their equitable distribution. Those funding the discovery and development of new tools typically neglect the concurrent need for strategies to make them available to the poor. Indeed, some would argue that the biggest challenge facing those who seek to address this outcome gap is the lack of practical means of delivery in the most heavily affected regions. When new preventive and therapeutic tools are developed without concurrent attention to delivery or implementation, one encounters what are sometimes termed "perverse effects": even as new tools are developed, inequalities of outcome—lower morbidity and mortality rates among those who can afford access, with sustained high morbidity and mortality among those who cannot—grow in the absence of an equity plan to deliver the tools to those most at risk. Preventing such a future is the most important goal of global health.

FURTHER READING

Dieleman JL et al: Development assistance for health: Past trends, associations, and the future of international financial flows for health. Lancet 387:2536, 2016.

Farmer P: Chronic infectious disease and the future of health care delivery. N Engl J Med 369:2424, 2013.

Farmer P et al (eds): *Reimagining Global Health: An Introduction.* Berkeley, University of California Press, 2013.

GBD 2015 Mortality and Cause of Death Collaborators: Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: A systematic analysis for the Global Burden of Disease Study 2015. Lancet 388:1459, 2016.

GBD 2015 DALYs and HALE Collaborators: Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE), 1990–2015: A systematic analysis for the Global Burden of Disease Study 2015. Lancet 388:1603, 2016.

Kim JY et al: Redefining global health-care delivery. Lancet 382:1060, 2013.

Mills A: Health care systems in low-and middle-income countries. N Engl J Med 370:552, 2014.

461 Worldwide Changes in Patterns of Infectious Disease

George W. Rutherford, Jaime Sepulveda

THE CONCEPT OF EMERGING INFECTIOUS DISEASES

Pathogenic organisms have been a constant companion of humans, their livestock, and their cultivated plants throughout evolution. Over the centuries, new organisms emerged as ecology changed or as humans managed to cross ecologic barriers such as oceans and mountains. Throughout history, there have been severe epidemics of infectious diseases, with devastating consequences to human populations over vast geographic regions. From the Plague of Justinian in Europe in the sixth century, to the Black Death in the fourteenth century, to the five cholera pandemics of the nineteenth century, to the 1918 Spanish influenza pandemic, to the current HIV/AIDS pandemic, the death toll in human populations has been enormous. The concept of emerging infectious diseases arose in the 1970s and 1980s with the recognition of several "new" diseases, such as legionellosis, HIV infection, Lyme disease, and toxic shock syndrome and was later expanded to include reemerging infectious diseases, such as tuberculosis. In 1991, the Institute of Medicine (IOM) convened a multidisciplinary committee to elucidate emerging microbial threats to health, with particular reference to the United States. In its report, the committee defined an emerging infectious disease as a disease "of infectious origin whose incidence in humans has either increased within the past two decades or threatens to increase in the near future." In the year following the publication of the committee's report, large outbreaks of *Escherichia coli* O157:H7 infection, cryptosporidiosis, and hantavirus pulmonary syndrome spurred the development of a national plan to recognize and interdict emerging and reemerging infectious disease threats by the Centers for Disease Control and Prevention (CDC). Since then, the list of emerging and reemerging viral, bacterial, fungal, and parasitic diseases has grown to include multiple infections and syndromes. Examples of emerging and reemerging infectious diseases, as of 2017, are shown in **Table 461-1**.

The reasons for the emergence of previously unrecognized diseases and the reemergence of diseases that have previously been largely under control are legion. At its core, however, emergence has to do with genetic changes in disease agents or changes in ecology, which includes human behavior. The IOM committee listed six primary reasons for disease emergence or reemergence: human demographics and behavior, technology and industry, economic development and land use, international travel and commerce, microbial adaptation and change, and breakdown of public health measures. A disease can emerge or reemerge for one or more of these reasons. For example, the worldwide spread of severe acute respiratory syndrome (SARS) began as a species cross-over, most likely involving transmission of a previously unknown coronavirus of horseshoe bats to Himalayan palm civets that were subsequently captured and transported to live-animal (i.e., "wet") markets in Guangzhou, China, for human consumption. The SARS coronavirus was then transmitted to humans—most likely by restaurant workers—and from them to medical personnel and eventually to individuals around the world. This spread had nothing to do with migratory patterns of bats or civets but was, instead, a consequence of human travel. The cities most effected by SARS—Hong Kong, Beijing, and Toronto—became involved because of rapid human movement via international passenger aircraft.

To this list, additional factors can now be added. Either therapeutic or acquired immunosuppression (e.g., in HIV infection) can render populations susceptible to infections that have not previously been recognized, such as that with human herpesvirus 8—the cause of Kaposi's sarcoma. Climate change, in particular, can expand the host

TABLE 461-1 Examples of Emerging and Reemerging Infectious Diseases

BACTERIAL AND RICKETTSIAL	VIRAL AND PRION	FUNGAL AND PARASITIC
Anaplasmosis	Chikungunya virus infection	Coccidioidomycosis
Anthrax	Variant Creutzfeldt-Jakob disease	*Cryptosporidium parvum* infection
Lyme disease	Dengue	*Cyclospora cayetanensis* infection
Vibrio cholerae O139 infection	Ebola virus, Marburg virus infection	Drug-resistant malaria
Diphtheria	Enterovirus D68 infection	
Ehrlichiosis	Hantavirus (Sin Nombre, Seoul) infection	
Escherichia coli O157:H7 infection	Hendra virus, Nipah virus infection	
Escherichia coli O154:H4 infection	Hepatitis C	
Legionella pneumophila infection	Hepatitis E	
Plague	HIV-1 and -2 infection	
Vancomycin-resistant *Staphylococcus aureus* infection	Human herpesvirus 6, 8 infection	
Tuberculosis	Human T-lymphotropic virus 1 and 2 infection	
	Influenza A H1N1pdm, H5N7, H7N7, H7N9	
	Lassa fever	
	Lyssavirus infection	
	Middle East respiratory syndrome (MERS)	
	Monkeypox	
	Rift Valley fever virus infection	
	Severe acute respiratory syndrome (SARS)	
	West Nile virus infection	
	Whitewater Arroyo virus infection	
	Yellow fever	
	Zika	

range of disease-transmitting vectors. In addition, the weaponization of pathogenic organisms for biological terrorism or warfare can lead, at least theoretically, to prolonged chains of human-to-human transmission. One factor is clear: the preponderance of emerging infectious diseases are zoonotic in origin. The authors of a 2008 review calculated that 60.3% of all emerging infectious disease events from 1940 to 2004 were zoonotic in origin, and 71.8% of these zoonotic events originated in wildlife.

In this chapter, we review the recent changing epidemiology of four emerging infectious viral diseases that exemplify some of the IOM's principles for emergence: infections caused by West Nile virus, dengue virus, Ebola virus, and Zika virus. This list is clearly not exhaustive but highlights a few prominent instances of the recent emergence of infectious diseases and their root causes.

EXAMPLES OF EMERGING INFECTIOUS DISEASES

WEST NILE VIRUS (WNV)

WNV is a flavivirus that was originally discovered in Uganda in 1937 and emerged as a cause of neurologic disease in humans and equines. WNV exists in nature in an enzootic cycle that involves certain birds and mosquitoes, particularly those of the genera *Culex* and *Aedes*. Humans, horses, and other vertebrates are incidental hosts and, except through blood transfusion, are unlikely to transmit WNV because levels of viremia are insufficiently high to infect mosquitoes. When originally described, WNV was believed to cause a mild febrile illness, but subsequent experience showed that it caused neuroinvasive disease in some cases. The first cases of neuroinvasive disease were described in an outbreak among elderly patients in Israel and subsequently in humans and horses in the Mediterranean basin, India, and South Africa. By the 1990s, outbreaks had been reported from Romania, Russia, and Central Asia; these outbreaks were probably a result of seasonal bird migrations from endemic Mediterranean countries, with introduction of infected mosquitoes and the establishment of infection in local bird species.

An explosive outbreak of WNV infection began in the United States in the summer of 1999 and initially involved infection of birds of the family Corvidae (e.g., the American crow and blue jay) that were susceptible to neuroinvasive disease. The first human cases appeared in New York City that same summer. Subsequently, sufficient numbers of birds more resistant to neuroinvasive disease and mosquitoes of the genus *Culex* became infected and an enzootic cycle was established in North America. Over the next 3 years, WNV spread across the continental United States, Canada, and Mexico and became an important cause of human and equine neurologic disease. The WNV clade causing the North American outbreak was the same (clade 1a) as that causing disease in the Middle East, Europe, North Africa, and parts of Asia.

In 2016, 2038 cases of WNV infection in humans, including 1140 cases of neuroinvasive disease, were reported in the United States; these figures are certainly gross underestimations of the actual number of cases. There were 94 deaths, primarily among the elderly. An additional 377 cases were reported in horses from 32 states despite the availability of a reasonably protective equine vaccine. Human cases were reported from 44 states, with only Alaska, Delaware, Hawaii, Maine, New Hampshire, and West Virginia free of the disease. Infected mosquito pools were even more widespread; Maine was the only state in the continental United States to be free of all WNV activity. Thus, from an initial introduction into New York City, WNV has successfully established itself across North America and infected an estimated 2.6–6.1 million people in the United States (1.1% of the population).

Why did this happen? First, microorganisms and larger organisms, such as plants and animals, have been exchanged between the Old and New Worlds since the initial voyages of exploration in the fifteenth and sixteenth centuries. However, it is the advent of modern high-speed transportation that allows vectors, such as mosquitoes, to move between continents in hours or days as opposed to months or years. In the most likely scenario for the introduction of WNV into North America, a single viremic mosquito was accidentally transported from an area endemic for clade 1a to New York City in the cargo hold of an airplane in 1999. The original strain associated with the 1999 outbreak (NY99) had caused outbreaks in Tunisia and Israel in 1997 and 1998, respectively; this information suggests that one of those countries was the source. The imported strain in turn infected crows, which in turn infected more competent mosquitoes, establishing an enzootic life cycle in North America with at least three *Culex* species and multiple species of birds involved. This scenario represents a successful invasion of WNV into a new ecologic niche.

The likelihood that WNV will gradually disappear is low. The virus has many avian hosts and more than one mosquito vector; it has undergone at least one successful mutation in the North American outbreak, thereby becoming infectious to *Culex piperans* and *Culex tarsalis*—mosquitoes with a broad range in the western United States. Moreover, the occurrence of outbreaks in 17 consecutive years in North America suggests that WNV has been successfully introduced onto the continent and will remain endemic for years to come.

DENGUE VIRUS

Dengue is the most important of the human arboviral infections, with almost half of the world's population at risk. Occurring in the range of *Aedes* mosquitoes, dengue virus infection imposes a heavy burden of morbidity and mortality worldwide, with as many as 50–200 million infections, 500,000 severe cases, and 20,000 deaths annually. Dengue virus is a flavivirus and exists in four serotypes (DENV1–4) that

circulate independently of one another; immunity to one serotype does not confer immunity to the others.

Dengue is transmitted primarily by *Aedes aegypti* and secondarily by *Aedes albopictus*. The original life cycle of dengue virus was most likely similar to that of yellow fever, consisting of sylvatic transmission from mosquitoes to nonhuman primates and back to mosquitoes; over the past few centuries, the virus has adapted to an urban and periurban mosquito–human–mosquito cycle as well. Dengue and its more severe manifestations, dengue hemorrhagic fever and dengue shock syndrome, were first described in outbreaks in Japan in 1943 and Hawaii in 1945. However, clinically similar diseases had been reported during the previous two centuries in a geographic band extending from India south to Queensland, Australia, and east through Polynesia; in addition, there had been occasional outbreaks in areas as disparate as Greece, Panama, and southern Texas.

The ecology of dengue changed dramatically in the second half of the twentieth century, led by the successful invasion by *A. aegypti* of the global tropics after World War II, with the postwar dispersion of troops and materiel. From its ancestral roots in Southeast Asia, all four dengue serotypes spread globally in the tropics. DENV2 had been introduced into West Africa by the 1960s and established both sylvatic enzootic nonhuman primate and urban endemic human cycles. Travel and commerce facilitated importations, probably through both viremic human hosts and infected mosquitoes. In the Americas in particular, a campaign to eradicate *A. aegypti*, which is also the principal vector of yellow fever, failed in the mid-1970s, and both *A. aegypti* and dengue virus, especially DENV2, rapidly reinvaded their prior habitat; thus dengue reemerged as a major arboviral disease extending from the southern United States in the north, through northern Argentina in the south. Recent outbreaks have occurred along the U.S.–Mexico border and in the state of São Paulo in Brazil, where DENV1, DENV2, and DENV4 are co-circulating.

Dengue's emergence and spread have been intimately linked to human activity. In particular, globalization, with the movement of viremic people and mosquitoes through modern transportation of both passengers and goods, has been critical to dengue's success. One particular adaptation has also facilitated its urban spread: *Aedes* is able to breed in standing water associated with human habitation, such as cisterns, ornamental ponds, puddles, and water trapped in abandoned tires. This ability of *Aedes* has allowed dengue to be one of the only two known arboviruses (the other being Zika) that are adapted to an urban environment and can replicate entirely in a mosquito-to-human cycle. Together, these factors have led to widespread dengue transmission in a band extending across the tropics worldwide.

EBOLA AND MARBURG VIRUSES

Ebola virus is a filovirus that most likely exists in a sylvatic cycle in bats in Central and West Africa. Four strains are known to cause human disease. The first outbreak was described in Zaire in 1976. Since then, 29 outbreaks have been reported across tropical Africa, ranging in size from tens of cases to tens of thousands of cases in the West African outbreak of 2013–2016.

The life cycle of Ebola virus in the wild is not fully understood. There is evidence for sustained transmission in fruit bats, with occasional nonhuman primate spillover infections. It has been speculated that humans become infected from contact with infected bats or nonhuman primates, but, once an index case has occurred, essentially all transmission is from human-to-human contact with blood and other body fluids. Preparing bodies for burial has been an especially efficient means of transmission. In addition, health care providers who do not wear adequate personal protective equipment while caring for Ebola patients are particularly vulnerable to acquiring infection. In the West African epidemic, there was only a single zoonotic introduction and all subsequent transmission was from human to human.

The principal cause of Ebola outbreaks prior to the West African outbreak was the migration of humans into sylvatic areas, with enzootic transmission and accidental infection. In West Africa, only a single case had been recognized in Côte d'Ivoire before the 2013–2016 outbreak in the Republic of Guinea, Liberia, and Sierra Leone. It has been speculated that cultivation of palm oil attracted fruit bats, who feed on palm fruit; if so, environmental modification from dense tropical forest to palm oil plantations may have been a contributory cause. Other evidence suggests that the index patient—a 2-year-old boy—was exposed to insectivorous free-tailed bats (*Mops condylurus*). Whatever the initial event, the explosive amplification that occurred in these countries and the seven countries to which cases were exported was due to an inadequate medical and public health infrastructure. In fact, when Ebola virus was imported to countries with more functional public health systems, such as Nigeria, transmission was extinguished within three generations.

Other filovirus outbreaks have involved the transport of infected primates for medical research. The original Marburg virus outbreak, which occurred in Marburg and Frankfurt, West Germany, and Belgrade, Yugoslavia, in 1967, was likely caused by the importation of African vervet monkeys (*Cercopithecus aethiops*) from Uganda for medical research. This outbreak resulted in 31 human cases and 7 deaths. In addition, an outbreak among five crab-eating macaques (*Macaca fascicularis*) imported from the Philippines and infected with Reston Ebola virus—a strain nonpathogenic for humans—led to an epizootic in northern Virginia in 1989, eventually resulting in the culling of more than 500 primates. This outbreak, however, had no human cases associated with it, although epidemiologic investigation identified a handful of asymptomatic primate handlers who were seropositive for Reston Ebola virus. Since 1989, four additional outbreaks have been recognized in *Cynomolgus* monkeys imported from the Philippines to the United States and Italy.

A new reservoir of Ebola virus infection has now been identified: the semen of patients who have survived Ebola infection. The occurrence of several small clusters of sexually transmitted cases developing up to 284 days after symptom onset indicates prolonged carriage of Ebola virus in the testes. Moreover, the virus may remain viable over the long term in the vitreous humor.

Thus, Ebola represents a spillover event to humans and nonhuman primates from their interaction with certain species of infected and infectious bats. Contact with either the bats themselves or an infected nonhuman primate leads to infection of an index patient, which leads in turn to ongoing transmission from humans to humans. Several factors clearly contribute to the continued transmission. First, medical and public health systems are weak in severely affected countries. As experience with Ebola grows and the capacity for surveillance and response improves, numbers of secondary cases can fall; for example, in five outbreaks in Uganda stretching from 2000 to 2012, the numbers of secondary cases and the geographic spread of the outbreaks decreased with each new introduction. Second, behavioral factors contribute, in particular funeral practices that bring mourners into close contact with infectious blood and tissues during preparation of a body for burial. Third, the areas in which the initial waves of transmission occur are often remote; thus, recognition of the outbreak can be delayed and, in the case of the West African outbreak, highly mobile populations can travel to larger cities to seek care.

ZIKA VIRUS

Zika virus is a flavivirus that is transmitted by *Aedes* mosquitoes and was originally described as an infection of nonhuman primates in Uganda in 1947. The first human cases were reported in Uganda in 1962 and 1963. Zika was thought to be an illness causing a mild rash and fever in humans in tropical Africa and southern Asia. The clinical and serologic similarity of Zika infection to dengue virus infection may have led to missed outbreaks. Since 2007, an Asian lineage of Zika virus has spread from the Western Pacific (initially, Yap Island) through Polynesia and on to Easter Island, Chile, where it was documented in 2014. From Polynesia, it also spread to Brazil, most likely through viremic travelers attending the world Va'a World Sprint Championships (Polynesian canoe racing) in Rio de Janeiro in the late summer of 2014. From there, Zika virus has spread hemisphere-wide, following the host range of *A. aegypti*. Forty-eight countries in the Americas have now reported autochthonously transmitted Zika virus infections.

In tropical Africa and Asia, Zika virus is most likely transmitted in a nonhuman primate–mosquito sylvatic cycle. Other animals may be

involved in Zika's life cycle as well. A number of *Aedes* species are competent vectors, although *A. aegypti* may be the source of the majority of infections worldwide.

As Zika virus spread through Latin America and the Caribbean, a parallel epidemic of fetal microcephaly appeared; this epidemic was both temporally and geographically associated with the spread of Zika virus. More than 1.6 million cases of Zika virus infection, including 41,473 cases in pregnant women and 1950 cases of Zika-associated microcephaly, were reported from Brazil alone in 2015 and 2016. Data from a large registry of Zika-exposed pregnancies in the U.S. territories show that the overall risk of microcephaly following confirmed Zika virus infection is ~5%, ranging from 8% for infection in the first trimester to 4% for infection in the third trimester. Other fetal complications include stillbirth, neural tube defects, eye abnormalities, and sensorineural deafness. Complications in adults occur in about one of every thousand cases and include Guillain-Barré syndrome, encephalitis, leukopenia, and thrombocytopenic purpura. Moreover, it is now recognized that Zika virus can be transmitted sexually and via blood transfusion.

Thus, the introduction of Zika into the Americas represents viral invasion of a new ecosystem already widely populated by a highly competent mosquito host with an established urban habitat and an immunologically inexperienced human population. The invasion by Zika virus is in many ways similar to the original dengue invasion in the Americas in the 1950s and to the introduction of WNV into North America in 1999. Both the original importation of Zika virus and its establishment of new foci in the Americas (e.g., Florida and the Caribbean) were consequences of modern travel. Zika's spread has also been linked to climate variations, deforestation, and urban poverty.

CONTROL OF EMERGING INFECTIOUS DISEASES

Humans will continue to experience outbreaks of emerging and reemerging infectious diseases. Emerging diseases will most likely come from two sources. The first source consists of organisms that have acquired new genetic materials from other strains of the same species or from different species altogether. An example of this is influenza A virus, in which strains can acquire new genetic material through a process called *reassortment*. If the new gene is a hemagglutinin gene, the resulting reassortant virus will have a new surface hemagglutinin that is unrecognized immunologically by most human populations. An interesting case is influenza A H1N1 virus, which emerged in 2009 from the reassortment of H1N1 swine influenza virus with human seasonal H3N2 influenza virus, North American avian influenza virus, and Eurasian avian-origin swine influenza viruses. Despite a worldwide pandemic, people born before 1950 were relatively spared because they had earlier exposure to an H1N1 strain sufficiently similar to provide them with cross-immunity. Another example of this is *E. coli* O157:H7, which acquired a virulence gene from *Shigella*, probably as the result of horizontal genetic exchange. The resulting organism and several other serotypes of *E. coli* that have acquired the gene are the leading cause of hemolytic–uremic syndrome worldwide. The second source for emergence consists of existing organisms entering new ecologic niches and spreading broadly, usually through insect vectors, to immunologically naïve humans—as occurred with Zika virus. A variation on this theme is humans entering new ecosystems and becoming infected with organisms to which they have no immunity. An organism's epidemic potential will be determined by whether it is largely incapable of leaving the human host to continue onward via human-to-human transmission (e.g., *Coccidioides*) or can be efficiently transmitted from human to human (e.g., HIV and Ebola virus).

In its 1994 strategic plan to address emerging infectious disease threats, the CDC listed four goals: (1) to detect, promptly investigate, and monitor emerging pathogens, the diseases they cause, and the factors influencing their emergence; (2) to integrate laboratory science and epidemiology in order to optimize public health practice; (3) to enhance communication of public health information about emerging diseases and ensure prompt implementation of prevention strategies; and (4) to strengthen local, state, and federal public health infrastructures in order to support surveillance and implement prevention and control programs. Much of this plan has been implemented. The concept of "emerging infectious diseases" has been broadly accepted, and molecular biological methods have improved to the point that, for example, the SARS coronavirus was completely sequenced in a matter of days. In addition, there has been an increasing recognition of the "one health" concept: the nexus among human, livestock, wildlife, and plant health and the development of surveillance systems to provide early warnings of emerging and reemerging infections. New vaccines and new vector-control agents are important promising weapons in the struggle to contain existing diseases; vaccines against both dengue and Ebola have been shown to be efficacious in phase 3 trials, and a new vector-control technique involving deliberate infection of the *Aedes* population with *Wolbachia*, a bacterial genus that inhibits the transmission of arbovirus from mosquito hosts, is being evaluated.

The World Health Organization has developed new international health regulations that are designed, in part, to facilitate the recognition and reporting of infectious disease threats. However, as evidenced by the 2013–2016 Ebola virus epidemic in West Africa, additional capacity and possibly new forms of global health governance and response may be required. What is clear is that humans will continue to experience new and reemerging infections, and we will need robust, flexible, and timely responses in order to control them.

FURTHER READING

Campbell-Lendrum D et al: Climate change and vector-borne diseases: What are the implications for public health research and policy? Philos Trans R Soc Lond B Biol Sci 370:20130552, 2015.

Heymann DL et al: Global health security: The wider lessons from the west African Ebola virus disease epidemic. Lancet 385:1884, 2015.

Jones KE et al: Global trends in emerging infectious diseases. Nature 451:990, 2008.

Lederberg J et al (eds): Committee on Emerging Microbial Threats to Health. Emerging infections. Microbial threats to health in the United States. Washington, DC, National Academy Press, 1992.

Morens DM, Fauci AS: Emerging infectious diseases in 2012: 20 years after the Institute of Medicine report. mBio 3:e00494, 2012.

462 Primary Care and Global Health

Tim Evans, Kumanan Rasanathan

The twentieth century witnessed the rise of an unprecedented global health divide. Industrialized or high-income countries experienced rapid improvement in standards of living, nutrition, health, and health care. Meanwhile, in low- and middle-income countries with much less favorable conditions, health and health care progressed much more slowly. The scale of this divide is reflected in the current extremes of life expectancy at birth, with Japan at the high end (85 years) and Chad at the low end (50 years). This 35-year difference reflects the daunting range of health challenges faced by low- and middle-income countries. These nations must deal not only with a complex mixture of diseases (both infectious and chronic) and illness-promoting conditions but also, and more fundamentally, with the fragility of the foundations underlying good health (e.g., sufficient food, water, sanitation, and education) and of the systems necessary for universal access to good-quality health care. In the last decades of the twentieth century, the need to bridge this global health divide and establish health equity was increasingly recognized. The Declaration of Alma Ata in 1978 crystallized a vision of justice in health, regardless of income, gender, ethnicity, or education, and called for "health for all by the year 2000" through primary health care. While much progress has been made since the declaration, at the end of the first decade and half of the

twenty-first century, much remains to be done to achieve global health equity.

This chapter looks first at the nature of the health challenges that underlie the health divide in low- and middle-income countries. It then outlines the values and principles of a primary health care approach, with a focus on primary care services. Next, the chapter reviews the experience of low- and middle-income countries in addressing health challenges through primary care and a primary health care approach. Finally, the chapter identifies how current challenges and global context provide an agenda and opportunities for the renewal of primary health care and primary care, allied to the movement to achieve universal health coverage.

PRIMARY CARE AND PRIMARY HEALTH CARE

The term *primary care* has been used in many different ways: to describe a level of care or the setting of the health system, a set of treatment and prevention activities carried out by specific personnel, a set of attributes for the way care is delivered, or an approach to organizing health systems that is synonymous with the term *primary health care*. In 1996, the U.S. Institute of Medicine encompassed many of these different usages, defining primary care as "the provision of integrated, accessible health care services by clinicians who are accountable for addressing a large majority of personal health care needs, developing a sustained partnership with patients, and practicing in the context of family and community."[1] We use this definition of *primary care* in this chapter. Primary care performs an essential function for health systems, providing the first point of contact when people seek health care, dealing with most problems, and referring patients onward to other services when necessary. As is increasingly evident in countries of all income levels, without strong primary care, health systems cannot function properly or address the health challenges of the communities they serve.

Primary care is only one part of a primary health care approach. The Declaration of Alma Ata, drafted in 1978 at the International Conference on Primary Health Care in Alma Ata (now Almaty in Kazakhstan), identified many features of primary care as being essential to achieving the goal of "health for all by the year 2000." However, it also identified the need to work across different sectors, address the social and economic factors that determine health, mobilize the participation of communities in health systems, and ensure the use and development of technology that was appropriate in terms of setting and cost. The declaration drew from the experiences of low- and middle-income countries in trying to improve the health of their people following independence. Commonly, these countries had built hospital-based systems similar to those in high-income countries. This effort had resulted in the development of high-technology services in urban areas while leaving the bulk of the population without access to health care unless they traveled great distances to these urban facilities. Furthermore, much of the population lacked access to basic public health measures. Primary health care efforts aimed to move care closer to where people lived, to ensure their involvement in decisions about their own health care, and to address key aspects of the physical and social environment essential to health, such as water, sanitation, and education.

After the Declaration of Alma Ata, many countries implemented reforms of their health systems based on primary health care. Most progress involved strengthening of primary care services; unexpectedly, however, much of this progress was seen in high-income countries, most of which constructed systems that made primary care available at low or no cost to their entire populations and that delivered the bulk of services in primary care settings. This endeavor also saw the reinforcement of family medicine as a specialty to provide primary care services. Even in the United States (an obvious exception to this trend), it became clear that the populations of states with more primary care physicians and services were healthier than those with fewer such resources.

[1]Institute of Medicine. Primary Care: America's Health in a New Era (1996).

Progress was also made in many low- and middle-income countries. However, the target of "health for all by the year 2000" was missed by a large margin. The reasons were complex but partly entailed a general failure to implement all aspects of the primary health care approach, particularly work across sectors to address social and economic factors that affect health and provision of sufficient human and other resources in order to make possible the access to primary care attained in high-income countries. Furthermore, despite the consensus in Alma Ata in 1978, the global health community rapidly became fractured in its commitment to the far-reaching measures called for by the declaration. Economic recession tempered enthusiasm for primary health care, and momentum shifted to programs concentrating on a few priority measures such as immunization, oral rehydration, breastfeeding, and growth monitoring for child survival. Success with these initiatives supported the continued movement of health development efforts away from the comprehensive approach of primary health care and toward programs that targeted specific public health priorities. This approach was reinforced by the need to address the HIV/AIDS epidemic. By the 1990s, primary health care had fallen out of favor in many global-health policy circles, and low- and middle-income countries were being encouraged to reduce public sector spending on health and to focus on cost-effectiveness analysis to provide a package of health care measures thought to offer the greatest health benefits.

HEALTH CHALLENGES IN LOW- AND MIDDLE-INCOME COUNTRIES

Low- and middle-income countries, defined by a per capita gross national income of <$12,476 (U.S.) per person per year, account for >80% of the world's population. Average life expectancy in these countries lags far behind that in high-income countries: whereas the average life expectancy at birth for a girl in high-income countries is 76 years, it is only 63 years for a girl in low-income countries. This discrepancy has received growing attention over the past 40 years. Initially, the situation in poor countries was characterized primarily in terms of high fertility and high infant, child, and maternal mortality rates, with most deaths and illnesses attributable to infectious or tropical diseases among remote, largely rural populations. With growing adult (and especially elderly) populations and changing lifestyles linked to global forces of urbanization, a new set of health challenges characterized by chronic diseases, environmental overcrowding, and road traffic injuries has emerged rapidly **(Fig. 462-1)**. The majority of tobacco-related deaths globally now occur in low- and middle-income countries, and the risk of a child's dying from a road traffic injury in Africa is more than twice that in Europe. Thus, low- and middle-income countries in the twenty-first century face a full spectrum of health challenges—infectious, chronic, and injury-related—at much higher incidences and prevalences than are documented in high-income countries and with many fewer resources to address these challenges.

Addressing these challenges, however, does not mean simply waiting for economic growth. Analysis of the association between wealth and health across countries reveals that, for any given level of wealth, there is a substantial variation in life expectancy at birth that has persisted despite overall global progress in life expectancy during the past 30 years **(Fig. 462-2)**. Health status in low- and middle-income countries varies enormously. Nations such as Cuba and Costa Rica have life expectancies and childhood mortality rates similar to or even better than those in high-income countries; in contrast, countries in sub-Saharan Africa and the former Soviet bloc have at times experienced significant reverses in these health markers in recent decades, particularly in the 1990s.

As Angus Deaton stated in the World Institute for Development Economics Research annual lecture on September 29, 2006, "People in poor countries are sick not primarily because they are poor but because of other social organizational failures, including health delivery, which are not automatically ameliorated by higher income." This analysis concurs with classic studies of the array of societal factors explaining good health in poor settings such as Cuba and Kerala

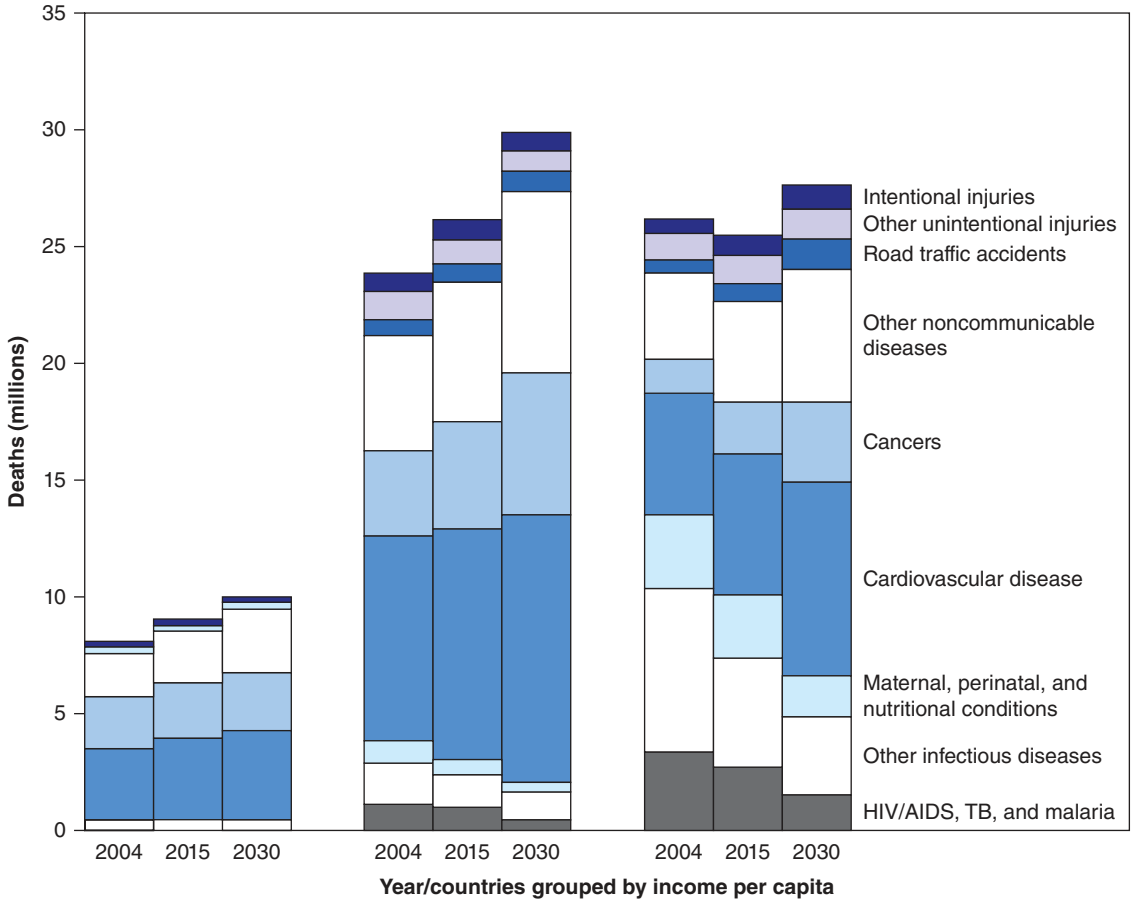


FIGURE 462-1 Projections of disease burden to 2030 for high-, middle-, and low-income countries (*left*, *center*, and *right*, respectively). TB, tuberculosis. *(Source: World Health Organization: The Global Burden of Disease 2004 Update, 2008.)*

State in India in the 1980s. Analyses conducted over the past three decades indeed show that rapid health improvement is possible in very different contexts. That some countries continue to lag far behind can be understood through a comparison of regional differences in progress in terms of life expectancy over this period **(Fig. 462-3)**.

As average levels of health vary across regions and countries, so too do they vary within countries **(Fig. 462-4)**. Indeed, disparities within countries are often greater than those between high-income and low-income countries. For example, if low- and middle-income countries could reduce their overall childhood mortality rate to that of the richest one-fifth of their populations, global childhood mortality could be decreased by 40%. Disparities in health are mostly a result of social and economic factors such as daily living conditions, access to resources, and ability to participate in life-affecting decisions. In most countries, the health care sector actually tends to exacerbate health inequalities (the "inverse-care law"); because of neglect and discrimination, poor and marginalized communities are much less likely to benefit from public health services than those that are better off. Reforming health systems toward people-centered primary care provides an opportunity to reverse these negative trends.

Health services have failed to make their contribution to reducing these pervasive social inequalities by ensuring universal access to existing, scientifically validated, low-cost interventions such as insecticide-treated bed nets for malaria, taxes on cigarettes, short-course chemotherapy for tuberculosis, antibiotic treatment for pneumonia, dietary modification and secondary prevention measures for high blood pressure and high cholesterol levels, and water treatment and oral rehydration therapy for diarrhea. Despite decades of "essential packages" and "basic" health campaigns, the effective implementation of what is already known to work appears (deceptively) to be difficult.

Recent analyses have begun to focus on "the how" (as opposed to "the what") of health care delivery, exploring why health progress is

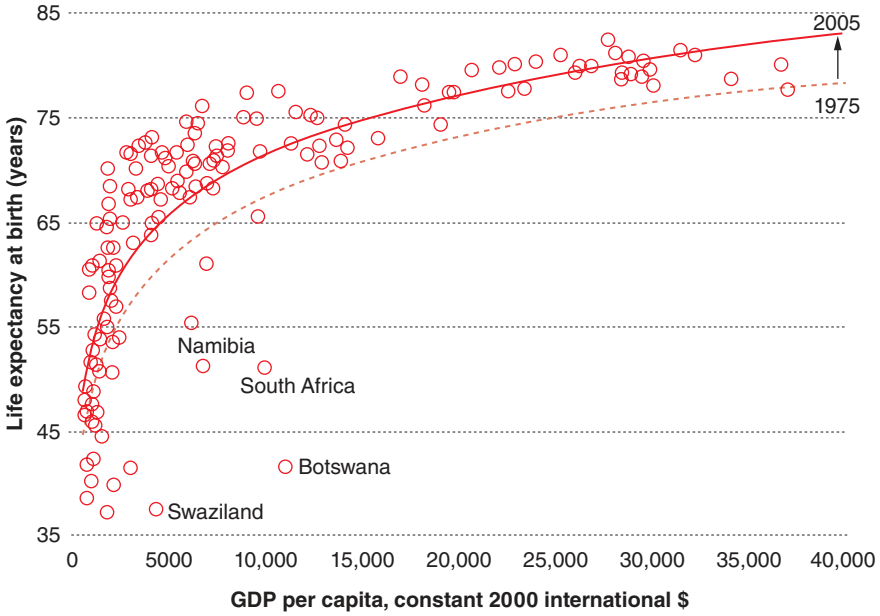


FIGURE 462-2 Gross domestic product (GDP) per capita and life expectancy at birth in 169 countries, 1975 and 2005. Only outlying countries are named. *(Source: World Health Organization: Primary Health Care: Now More Than Ever. World Health Report 2008.)*

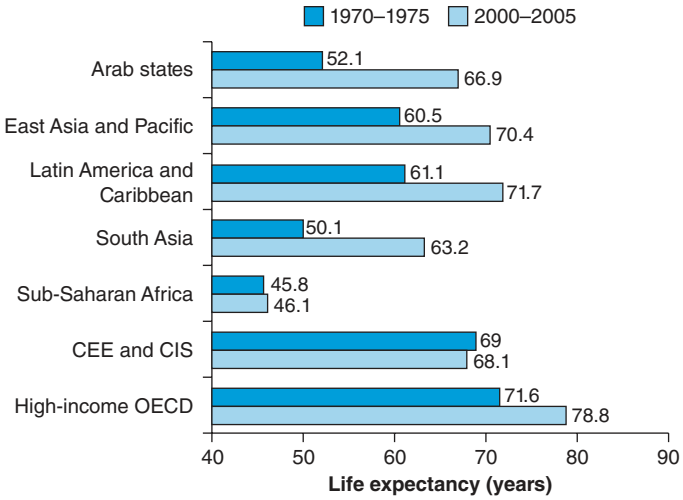


FIGURE 462-3 Regional trends in life expectancy. CEE, Central and Eastern Europe; CIS, the Commonwealth of Independent States; OECD, Organization for Economic Cooperation and Development. *(Source: World Health Organization: Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health. Commission on Social Determinants of Health Final Report, 2008.)*

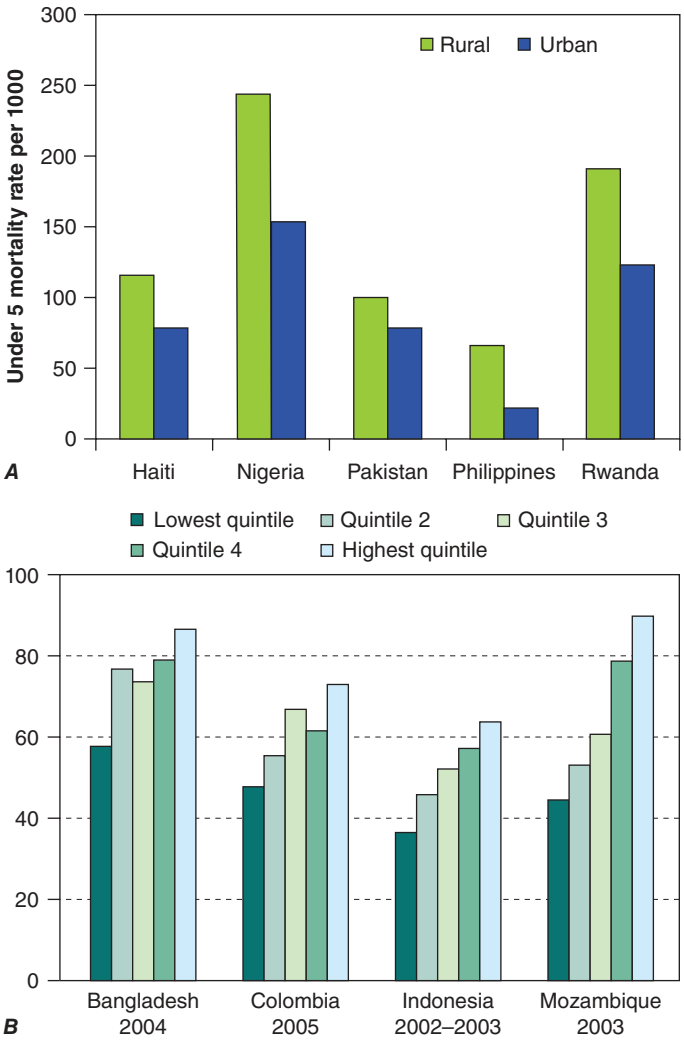


FIGURE 462-4 ***A.*** Mortality of children under 5 years old, by place of residence, in five countries. *(Source: Data from the World Health Organization.)* ***B.*** Full basic immunization coverage (%), by income group. *(Source: Primary Health Care: Now More Than Ever. World Health Report 2008.)*

slow and sluggish despite the abundant availability of proven interventions for health conditions in low- and middle-income countries. Three general categories of reasons are being identified: (1) shortfalls in performance of health systems; (2) stratifying social conditions; and (3) skews in science.

SHORTFALLS IN PERFORMANCE OF HEALTH SYSTEMS

Specific health problems often require the development of specific health interventions (e.g., tuberculosis requires short-course chemotherapy). However, the delivery of different interventions is often facilitated by a common set of resources or functions: money or financing, trained health workers, and facilities with reliable supplies fit for multiple purposes. Unfortunately, health systems in most low- and middle-income countries are largely dysfunctional at present.

In the large majority of low- and middle-income countries, the level of public financing for health is woefully insufficient: whereas high-income countries spend, on average, 7% of the gross domestic product on health, middle-income countries spend <4%, and low-income countries <3%. External financing for health through various donor channels grew rapidly in the first decade of the twenty-first century but has grown more slowly in the second decade to its current level of $37 billion. While these funds for health are significant, they represent <2% of total health expenditures in low- and middle-income countries and therefore are neither a sufficient nor a long-term solution to chronic underfinancing. In Africa, 70% of health expenditures come from domestic sources. The predominant form of health care financing—charging patients at the point of service—is the least efficient and the most inequitable, tipping millions of households into poverty annually.

Health workers, who represent another critical resource, are often inadequately trained and supported in their work. Recent estimates indicate a shortage of >7 million health workers, constituting a crisis that is greatly exacerbated by the migration of health workers from low- and middle-income countries to high-income countries. Sub-Saharan Africa carries 24% of the global disease burden but has only 3% of the health workforce **(Fig. 462-5)**. The International Organization for Migration estimated in 2006 that there were more Ethiopian physicians practicing in Chicago than in Ethiopia itself.

Critical diagnostics and drugs often do not reach patients in need because of supply-chain failures. Moreover, facilities fail to provide safe care: new evidence suggests much higher rates of adverse events among hospitalized patients in low- and middle-income countries than in high-income countries. Weak government planning, regulatory, monitoring, and evaluation capacities are associated with rampant, unregulated commercialization of health services and chaotic fragmentation of these services as donors "push" their respective priority programs. With such fragile foundations, it is not surprising that low-cost, affordable, validated interventions are not reaching those who need them.

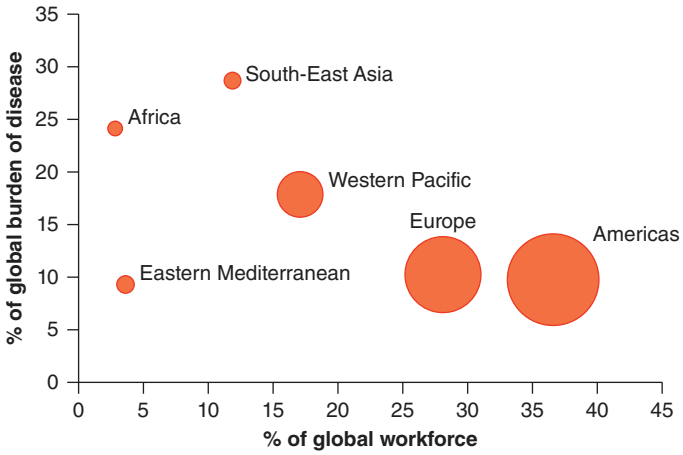


FIGURE 462-5 Global burden of disease and health workforce. *(Source: World Health Organization: Working Together for Health, 2006.)*

STRATIFYING SOCIAL CONDITIONS

Health care delivery systems do not exist in a vacuum but rather are embedded in a complex of social and economic forces that often stratify opportunities for health unfairly. Most worrisome are the pervasive forces of social inequality that serve to marginalize populations with disproportionately large health needs (e.g., the urban poor; illiterate mothers). Why should a poor slum dweller with no income be expected to come up with the money for a bus fare needed to travel to a clinic to learn the results of a sputum test for tuberculosis? How can a mother living in a remote rural village and caring for an infant with febrile convulsions find the means to get her child to appropriate care? Shaky or nonexistent social security systems, dangerous work environments, isolated communities with little or no infrastructure, and systematic discrimination against minorities are among the myriad forces with which efforts for more equitable health care delivery must contend.

SKEWS IN SCIENCE

While science has yielded enormous breakthroughs in health in high-income countries, with some spillover to low- and middle-income countries, many important health problems continue to affect primarily low- and middle-income countries whose research and development investments are deplorably inadequate. The past decade has seen growing efforts to right this imbalance with research and development investment for new drugs, vaccines, and diagnostics that effectively cater to the specific health needs of populations in low- and middle-income countries. For example, the Medicines for Malaria Venture has revitalized a previously "dry" pipeline for new malaria drugs. This is but one of many such efforts, but much more needs to be done.

As discussed above, the primary constraint on better health in low- and middle-income countries is related less to the availability of health technologies and more to their effective delivery. Underlying these systems and social challenges to greater equity in health is a major bias regarding what constitutes legitimate "science" to improve health equity. The lion's share of health research financing is channeled toward the development of new technologies—drugs, vaccines, and diagnostics; in contrast, virtually no resources are directed toward research on how health care delivery systems can become more reliable and overcome adverse social conditions. The complexity of systems and social context is such that this issue of delivery requires an enormous investment in terms not only of money but also of scientific rigor, with the development of new research methods and measures and the attainment of greater legitimacy in the mainstream scientific establishment.

These common challenges to low- and middle-income countries partly explain the resurgence of interest in the primary health care approach and the emergence of a global movement toward universal health coverage, now enshrined as one of the Sustainable Development Goal targets adopted in Agenda 2030 by all countries at the United Nations in September 2015. In some countries (mostly middle-income), significant progress has been made in expanding coverage by health systems based on primary care and even in improving indicators of population health. More countries are embarking on the creation of primary care services despite the challenges that exist, particularly in low-income countries. Even when these challenges are acknowledged, there are many reasons for optimism that low- and middle-income countries can accelerate progress in building primary care as a key vehicle toward achieving universal health coverage.

PRIMARY HEALTH CARE IN THE TWENTY-FIRST CENTURY

The new millennium has seen a resurgence of interest in primary health care as a means of addressing global health challenges. This interest has been driven by many of the same issues that led to the Declaration of Alma Ata: rapidly increasing disparities in health between and within countries, spiraling costs of health care at a time when many people lack quality care, dissatisfaction of communities with the care they are able to access, and failure to address changes in health threats, especially noncommunicable disease epidemics. These challenges require a comprehensive approach and strong health systems with effective primary care. Global health development agencies have recognized that sustaining gains in public health priorities such as HIV/AIDS requires not only robust health systems but also the tackling of social and economic factors related to disease incidence and progression. Weak health systems have proved a major obstacle to delivering new technologies, such as antiretroviral therapy, to all who need them. Changing disease patterns have led to a demand for health systems that can treat people as individuals whether or not they present to a health facility with the public health "priority" (e.g., HIV/AIDS or tuberculosis) to which that facility is targeted. We discuss experiences in low- and middle-income countries in relation to primary care in greater detail below. First, we consider the features of primary health care and primary care as currently understood.

REVITALIZATION OF PRIMARY HEALTH CARE

At the 2009 World Health Assembly (an annual meeting of all countries to discuss the work of the World Health Organization [WHO]), a resolution was passed reaffirming the principles of the Declaration of Alma Ata and the need for national health systems to be based on primary health care. This resolution did not suggest that nothing had changed in the intervening 30 years since the declaration, nor did it dispute that its prescription needed reframing in light of changing public health needs. The 2008 WHO World Health Report describes how a primary health care approach is necessary "now more than ever" to address global health priorities, especially in terms of disparities and new health challenges. As discussed below, this report highlights four broad areas in which reform is required **(Fig. 462-6)**. One of these areas—the need to organize health care so that it places the needs of people first—essentially relates to the necessity for strong primary care in health systems and what this requirement entails. The other three areas also relate to primary care. All four areas require action to move health systems in a direction that will reduce disparities and increase the satisfaction of those they serve. The World Health Report's recommendations present a vision of primary health care that is based on the principles of Alma Ata, but that differs from many attempts to implement primary health care in the 1970s and 1980s.

Universal Coverage Reforms to Improve Health Equity Despite progress in many countries, most people in the world can receive health care services only if they can pay at the point of service. Disparities in health are caused not only by a lack of access

FIGURE 462-6 The four reforms of primary health care renewal. *(Source: World Health Organization: Primary Health Care: Now More Than Ever. World Health Report 2008.)*

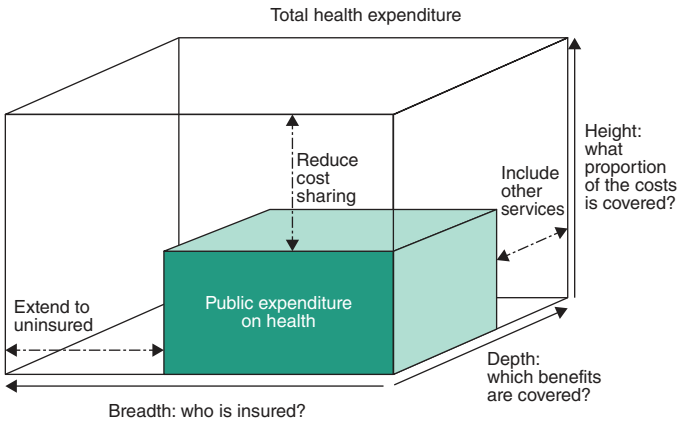


FIGURE 462-7 Three ways of moving toward universal coverage.

to necessary health services but also by the impact of expenditure on health. More than 100 million people are driven into poverty each year by health care costs, with countless others deterred from accessing services at all. Moving toward prepayment financing systems for universal coverage, which ensure access to a comprehensive package of services according to need without precipitating economic ruin, has therefore emerged as a major priority in low- and middle-income countries. Increasing coverage of health services can be considered in terms of three axes: the proportion of the population covered, the range of services underwritten, and the percentage of costs paid (Fig. 462-7). Moving toward universal health coverage requires ensuring the availability of health care services to all, eliminating barriers to access, and organizing pooled financing mechanisms, such as taxation or insurance, to remove user fees at the point of service. It also requires measures beyond financing, including expansion of health services in poorly served areas, improvement in the quality of services provided to marginalized communities, and increased coverage of other social services that significantly affect health (e.g., education).

Service Delivery Reforms to Make Health Systems People-Centered Health systems have often been organized around the needs of those who provide health care services, such as clinicians and policy makers. The result is a centralization of services or the provision of vertical programs that target single diseases. The principles of primary health care, including the development of primary care, reorient care around the needs of the people to whom services cater. This "people-centered" approach aims to provide health care that is both more effective and appropriate.

The increase in noncommunicable diseases in low- and middle-income countries offers a further stimulus for urgent reform of service delivery to improve chronic disease care. As discussed above, large numbers of people currently fail to receive relatively low-cost interventions that have reduced the incidence of these diseases in high-income countries. Delivery of these interventions requires health systems that can address multiple problems and manage people over a long period within their own communities, yet many low- and middle-income countries are only now starting to adapt and build primary care services that can address noncommunicable diseases and communicable diseases requiring chronic care. Even some countries (e.g., Iran) that have had significant success in reducing communicable diseases and improving child survival have been slow to adapt their health systems to rapidly accelerating noncommunicable disease epidemics.

People-centered care requires a safe, comprehensive, and integrated response to the needs of those presenting to health systems, with treatment at the first point of contact or referral to appropriate services. Because no discrete boundary separates people's needs for health promotion, curative interventions, and rehabilitation services across different diseases, primary care services must address all presenting problems in a unified way. Meeting people's needs also involves improved communication between patients and their clinicians, who must take the time to understand the impact of the patients' social context on the problems they present with. This enhanced understanding is made possible by improvements in the continuity of care so that responsibility transcends the limited time people spend in health care facilities. Primary care plays a vital role in navigating people through the health system; when people are referred elsewhere for services, primary care providers must monitor the resulting consultations and perform follow-up. All too often, people do not receive the benefit of complex interventions undertaken in hospitals because they lose contact with the health care system once discharged. Comprehensiveness and continuity of care are best achieved by ensuring that people have an ongoing personal relationship with a care team.

Public Policy Reforms to Promote and Protect the Health of Communities Public policies in sectors other than health care are essential to reduce disparities in health and to make progress toward global public health targets. The 2008 final report of the WHO Commission on Social Determinants of Health provides an exhaustive review of the multisectoral policies required to address health inequities at the local, national, and global levels. Advances against major challenges such as HIV/AIDS, tuberculosis, emerging pandemics, cardiovascular disease, cancers, and injuries require effective collaboration with sectors such as transport, housing, labor, agriculture, urban planning, trade, and energy. While tobacco control provides a striking example of what is possible if different sectors work together toward health goals, the lack of implementation of many evidence-based tobacco control measures in most countries just as clearly illustrates the difficulties encountered in such multisectoral work and the unrealized potential of public policies to improve health. At the local level, primary care services can help enact health-promoting public policies in other sectors.

Leadership Reforms to Make Health Authorities More Reliable The Declaration of Alma Ata emphasized the importance of participation by people in their own health care. In fact, participation is important at all levels of decision-making. Contemporary health challenges require new models of leadership that acknowledge the role of government in reducing disparities in health but that also recognize the many types of organizations that provide health care services. Governments need to guide and negotiate among these different groups, including nongovernmental organizations (NGOs) and the private sector, and to provide strong regulation where necessary. This difficult task requires a massive reinvestment in leadership and governance capacity, especially if action by different sectors is to be effectively implemented. Moreover, disadvantaged groups and other actors are increasingly expecting that their voices and health needs will be included in the decision-making process. The complex landscape for leadership at the national level is mirrored in many ways at the international or global level. The transnational character of health and the increasing interdependence of countries with respect to outbreak diseases, climate change, security, migration, and agriculture place a premium on more effective global health governance.

EXPERIENCES WITH PRIMARY CARE IN LOW- AND MIDDLE-INCOME COUNTRIES

Aspects of the primary health care approach described above, with an emphasis on primary care services, have been implemented to varying degrees in many low- and middle-income countries over the past half-century. As discussed above, some of these experiences inspired and informed the Declaration of Alma Ata, which itself led many more countries to attempt to implement primary health care. This section describes the experiences of a selection of low- and middle-income countries in improving primary care services that have enhanced the health of their populations.

Before Alma Ata, few countries had attempted to develop primary care on a national level. Rather, most focused on expanding primary care services to specific communities (often rural villages), making use of community volunteers to compensate for the absence of facility-based care. In contrast, in the post–World War II period, China invested in primary care on a national scale, and life expectancy doubled within roughly 20 years. The Chinese expansion of primary care

services included a massive investment in infrastructure for public health (e.g., water and sanitation systems) linked to innovative use of community health workers. These "barefoot doctors" lived in and expanded care to rural villages. They received a basic level of training that enabled them to provide immunizations, maternal care, and basic medical interventions, including the use of antibiotics. Through the work of the barefoot doctors, China brought low-cost universal basic health care coverage to its entire population, most of which had previously had no access to these services.

In 1982, the Rockefeller Foundation convened a conference to review the experiences of China along with those of Costa Rica, Sri Lanka, and the state of Kerala in India. In all of these locations, good health care at low cost appeared to have been achieved. Despite lower levels of economic development and health spending, all of these jurisdictions, along with Cuba, had health indicators approaching—or in some cases exceeding—those of developed countries. Analysis of these experiences revealed a common emphasis on primary care services, with expansion of care to the entire population free of charge or at low cost, combined with community participation in decision-making about health services and coordinated work in different sectors (especially education) toward health goals. During the more than three decades since the Rockefeller meeting, some of these countries have built on this progress, while others have experienced setbacks. Recent experiences in developing primary care services show that the same combination of features is necessary for success. For example, Brazil—a large country with a dispersed population—has made major strides in increasing the availability of health care in the past quarter century. In this millennium, the Brazilian Family Health Program has expanded progressively across the country, with almost all areas now covered. This program provides communities with free access to primary care teams made up of primary care physicians, community health workers, nurses, dentists, obstetricians, and pediatricians. These teams are responsible for the provision of primary care to all people in a specified geographic area—not only those who access health clinics. Moreover, individual community health workers are responsible for a named list of people within the area covered by the primary care team. Problems with access to health care persist in Brazil, especially in isolated areas and urban slums. However, solid evidence indicates that the Family Health Program has already contributed to impressive gains in population health, particularly in terms of childhood mortality and health inequities. In fact, this program has already had an especially marked impact on childhood mortality reduction in less developed areas **(Fig. 462-8)**.

Chile has also built on its existing primary care services in the past decade, aiming to improve the quality of care and the extent of coverage in remote areas, above all for disadvantaged populations. This effort has been made in concert with measures aimed at reducing social inequalities and fostering development, including social welfare benefits for families and disadvantaged groups and increased access to early-childhood educational facilities. As in Brazil, these steps have improved maternal and child health and have reduced health inequities. In addition to directly enhancing primary care services, Brazil and Chile have instituted measures to increase both the accountability of health providers and the participation of communities in decision-making. In Brazil, national and regional health assemblies with high levels of public participation are integral parts of the health policy-making process. Chile has instituted a patient's charter that explicitly specifies the rights of patients in terms of the range of services to which they are entitled.

Other countries that have made recent progress with primary health care include Bangladesh, once one of the poorest countries in the world and still relatively poor. Since achieving its independence from Pakistan in 1971, Bangladesh has seen a dramatic increase in life expectancy, and childhood mortality rates are now lower than those in neighboring nations such as India and Pakistan. The expansion of access to primary care services has played a major role in these achievements. This progress has been spearheaded by a vibrant NGO community that has focused its attention on improving the lives and livelihoods of poor women and their families through innovative and integrated microcredit, education, and primary care programs.

The above examples, along with others from the past 30 years in countries such as Thailand, Malaysia, Portugal, and Oman, illustrate how the implementation of a primary health care approach, with a greater emphasis on primary care, has led to better access to health care services—a trend that has not been seen in many other low- and middle-income countries. This trend, in turn, has contributed to improvements in population health and reductions in health inequities. However, as these nations have progressed, other countries have shown how previous gains in primary care can easily be eroded. In sub-Saharan Africa, undermining of primary care services contributed to catastrophic reversals in health outcomes catalyzed by the HIV/AIDS epidemic. Countries such as Botswana and Zimbabwe implemented primary health care strategies in the 1980s, increasing access to care and making impressive gains in child health. Both countries were severely affected by HIV/AIDS, with pronounced decreases in life expectancy. However, Zimbabwe has also seen political turmoil, a decline of health and other social services, and the flight of health personnel, whereas Botswana has maintained primary care services to a greater extent and has managed to organize widespread access to antiretroviral therapy for people living with HIV/AIDS. Zimbabwe's health situation has therefore become more desperate than that in Botswana.

China provides a particularly striking example of how changes in health policy relevant to the organization of health systems **(Fig. 462-9)** can have rapid, far-reaching consequences for population health. Even as the 1982 Rockefeller conference was celebrating China's achievements in primary care, its health system was unraveling. The decision to open up the economy in the early 1980s led to rapid privatization of the health sector and the breakdown of universal health coverage. As a result, by the end of the 1980s, most people, especially the poorer segments of the population, were paying directly out of pocket for health care, and almost no Chinese had insurance—a dramatic transformation. The "barefoot doctor" schemes collapsed, and the population either turned to care paid for at hospitals or simply became unable to access care. This undermining of access to primary care services in the Chinese system and the resulting increase in impoverishment due to illness contributed to the stagnation of progress in health in China at the same time that incomes in that country increased at an unprecedented rate. Reversals in primary care have meant that China now increasingly faces health care issues similar to those faced by India, although the country has more recently implemented measures to restore universal health coverage, with significant success. In both countries, rapid economic growth has been linked to lifestyle changes and noncommunicable disease epidemics. The health care systems of the two nations share

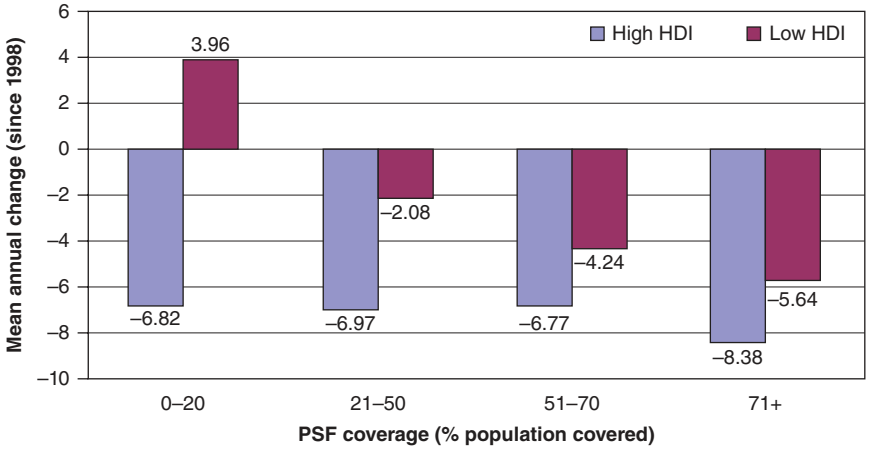


FIGURE 462-8 Improvements in childhood mortality following the Family Health Program in Brazil. HDI, Human Development Index; PSF, Program Saúde da Família (Family Health Program). *(Source: Ministry of Health, Brazil.)*

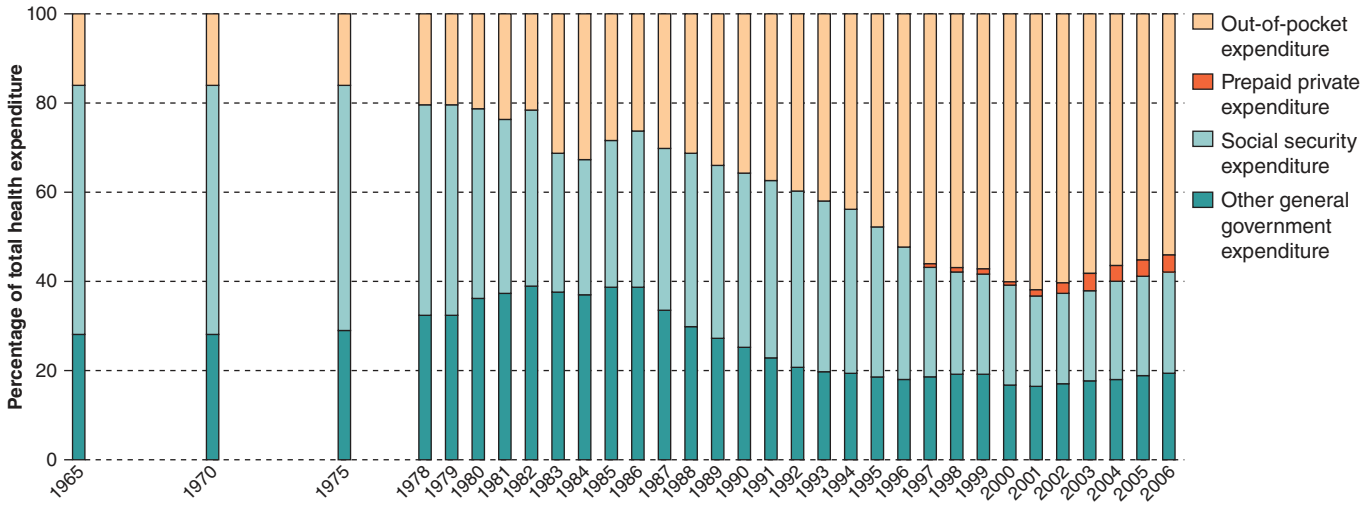


FIGURE 462-9 Changes in source of health expenditure in China over the past 40 years. *(Source: World Health Organization: Primary Health Care: Now More Than Ever. World Health Report 2008.)*

two negative features that are common when primary care is weak: a disproportionate focus on specialty services provided in hospitals and unregulated commercialization of health services. China and India have both seen expansion of private hospital services that cater to middle-class and urban populations who can afford care; at the same time, hundreds of millions of people in rural areas now struggle to access the most basic services. Even in the former groups, a lack of primary care services has been associated with late presentation with illness and with insufficient investment in primary prevention approaches. This neglect of prevention poses a risk of large-scale epidemics of cardiovascular disease, which could endanger continued economic growth. In addition, the health systems of both countries now depend for the majority of their funding on out-of-pocket payments by people when they use services. Thus substantial proportions of the population must sacrifice other essential goods as a result of health expenditure and may even be driven into poverty by this cost. The commercial nature of health services with inadequate or no regulation has also led to the proliferation of charlatan providers, inappropriate care, and pressure for people to pay for expensive and sometimes unnecessary care. Commercial providers have limited incentives to use interventions (including public health measures) that cannot be charged for or that are what the person who is paying can afford.

Faced with these problems, China and India have implemented measures to strengthen primary health care. China has increased government funding of health care, has taken steps toward restoring health insurance, and has enacted a target of universal access to primary care services. India has similarly mobilized funding to greatly expand primary care services in rural areas and is now duplicating this process in urban settings. Both countries are increasingly using public resources from their growing economies to fund primary care services.

These encouraging trends are illustrative of new opportunities to implement a primary health care approach and strengthen primary care services in low- and middle-income countries. Brazil, India, China, and Chile are being joined by many other low- and middle-income countries, including Indonesia, Mexico, the Philippines, Turkey, Rwanda, Ethiopia, South Africa, and Ghana, in ambitious initiatives mobilizing new resources to move toward universal health coverage—the provision of quality health services in a timely manner at affordable cost.

OPPORTUNITIES TO BUILD PRIMARY CARE IN LOW- AND MIDDLE-INCOME COUNTRIES

Global public health targets will not be met unless health systems are significantly strengthened. More money is currently being spent on health than ever before. In 2013, global health spending totaled $7.8 trillion (U.S.)—more than double the amount spent a decade earlier. Although most expenditure occurs in high-income countries, spending in many emerging middle-income countries has rapidly accelerated, as has the allocation of monies for this purpose by both governments in and donors to low-income countries. These twin trends—greater emphasis on building health systems based on primary care and allotment of more money for health care—provide opportunities to address many of the challenges discussed above in low- and middle-income countries.

Accelerating progress requires a better understanding of how global health initiatives can more effectively facilitate the development of primary care in low-income countries. A review by the WHO Maximizing Positive Synergies Collaborative Group looked at programs funded by the Global Fund to Fight AIDS, Tuberculosis and Malaria; the Global Alliance for Vaccines and Immunisation (GAVI); the U.S. President's Emergency Plan for AIDS Relief (PEPFAR); and the World Bank (HIV/AIDS). This group found that global health initiatives had improved access to and quality of the targeted health services and had led to better information systems and more adequate financing. The review also identified the need for better alignment of global health initiatives with other national health priorities and systematic exploitation of potential synergies. If global health initiatives implement programs that work in tandem with other components of national health systems without undermining staffing and procurement of supplies, they have the potential to contribute substantially to the capacity of health systems to provide comprehensive primary care services.

Even in the aftermath of the global financial crisis, global health initiatives continued to draw significant funding. In 2009, for example, U.S. President Barack Obama announced increasing development assistance from the United States for global health, earmarking $63 billion over the period 2009–2014 for a Global Health Initiative. New funding is also promised through a range of other initiatives focusing particularly on maternal and child health in low-income countries. The general trend is to coordinate this funding in order to reduce fragmentation of national health systems and to concentrate more on strengthening these systems. Comprehensive primary care in low-income countries must inevitably deal with the rapid emergence of chronic diseases and the growing prominence of injury-related health problems; thus international health development assistance must become more responsive to these needs. More recent political currents that threaten global health funding from traditional sources only underscore the need to take a more comprehensive and integrated health systems approach to the use of these funds.

Beyond funding for health services, other opportunities exist. Increased social participation in health systems can help build primary care services. In many countries, political pressure from community advocates for more holistic and accountable care as well as entrepreneurial initiatives to scale up community-based services through

NGOs have accelerated progress in primary care without major increases in funding. Participation of the population in the provision of health care services and in relevant decision-making often drives services to cater to people's needs as a whole rather than to narrow public health priorities.

Participation and innovation can help address critical issues related to the health workforce in low- and middle-income countries by establishing effective people-centered primary care services. Many primary care services do not need to be delivered by a physician or a nurse. Multidisciplinary teams can include paid community workers who have access to a physician if necessary but who can provide a range of health services on their own. In Ethiopia, more than 38,000 community health workers have been trained and deployed to improve access to primary care services, and there is increasing evidence that this measure is contributing to better health outcomes. In India, more than 600,000 community health advocates have been recruited as part of expanded rural primary care services. In Niger, the deployment of community health workers to deliver essential child health interventions (as a component of integrated community case management) has had impressive results in reducing childhood mortality and decreasing disparities. After the Declaration of Alma Ata, experiences with community health workers were mixed, with particular problems about levels of training and lack of payment. Current endeavors are not immune from these concerns. However, with access to physician support and the deployment of teams, some of these concerns may be addressed. Growing evidence from many countries indicates that shifting appropriate tasks to primary care workers who have had shorter, less expensive training than physicians will be essential to address the human resources crisis.

Finally, recent improvements in information and communication technologies, particularly mobile phone and internet systems, have created the potential for systematic implementation of e-health, telemedicine, and improved health data initiatives in low- and middle-income countries. These developments raise the tantalizing possibility that health systems in these countries, which have long lagged behind those in high-income countries but are less encumbered by legacy systems that have proved hard to modernize in many settings, could leapfrog their wealthier counterparts in exploiting these technologies. Although the challenges posed by poor or absent infrastructure and investment in many low- and middle-income countries cannot be underestimated and will need to be addressed to make this possibility a reality, the rapid rollout of mobile networks and their use for health and other social services in many low-income countries where access to fixed telephone lines was previously very limited offer great promise in building primary care services in low- and middle-income countries.

CONCLUSION

As concern continues to mount about glaring inequities in global health, there is a growing commitment to redress these egregious shortfalls, as exemplified by the central place of equity in the United Nations' Sustainable Development Goals adopted in 2015, including a specific target on the achievement of universal health coverage in all countries by 2030. This commitment begins first and foremost with a clear vision of the fundamental importance of health in all countries, regardless of income. The values of health and health equity are shared across all borders, and primary health care provides a framework for their effective translation across all contexts.

The translation of these fundamental values has its roots in four types of reform that reflect the distinct but interlinked challenges of (re)orienting a society's resources on the basis of its citizens' health needs: (1) organizing health care services around the needs of people and communities; (2) harnessing services and sectors beyond health care to promote and protect health more effectively; (3) establishing sustainable and equitable financing mechanisms for universal health coverage; and (4) investing in effective leadership of the whole of society. This common primary health care agenda highlights the striking similarity, despite enormous differences in context, in the nature and direction of the reforms that national health systems must undertake to promote greater equity in health. This shared agenda is complemented by the growing reality of global health interconnectedness due, for example, to shared microbial threats, bridging of ethnolinguistic diversity, flows in migrant health workers, and mobilization of global funds to support the neediest populations. Embracing solidarity in global health while strengthening health systems through a primary health care approach is fundamental to sustained progress in global health.

FURTHER READING

Aquino R et al: Impact of the family health program on infant mortality in Brazilian municipalities. Am J Public Health 99:87, 2009.

Commission on Social Determinants of Health: *Closing the Gap in a Generation: Health Equity through Action on the Social Determinants of Health: Commission on Social Determinants of Health Final Report*. Geneva, World Health Organization, 2008.

Kruk ME et al: The contribution of primary care to health and health systems in low- and middle-income countries: A critical review of major primary care initiatives. Soc Sci Med 70:904, 2010.

Li X et al: The primary health-care system in China. Lancet 390:2584, 2017.

Macinko J et al: The impact of primary healthcare on population health in low- and middle-income countries. J Ambul Care Manage 32:150, 2009.

Patel V et al: Assuring health coverage for all in India. Lancet 386:2422, 2015.

Rasanathan K et al: Primary health care and the social determinants of health: Essential and complementary approaches for reducing inequities in health. J Epidemiol Community Health 65:656, 2011.

Starfield B et al: Contribution of primary care to health systems and health. Milbank Q 83:457, 2005.

Tangcharoensathien V et al: *Primary Health Care: Now More Than Ever. World Health Report 2008*. Geneva, World Health Organization, 2008.

Tangcharoensathien V et al: Health systems development in Thailand: A solid platform for successful implementation of universal health coverage. Lancet pii:S0140-6736(18)30198-3, 2018.

463 The Biology of Aging

Rafael de Cabo, David G. Le Couteur

THE IMPACT OF AGING ON MEDICINE

Aging and old age are among the most significant challenges facing medicine this century. The aging process is the major risk factor underlying disease and disability in developed nations, while older people respond differently to therapies developed for younger adults (usually with less effectiveness and more adverse reactions). Modern medicine and healthier lifestyles have increased the likelihood that younger adults will now achieve old age. However, this has led to rapidly increasing numbers of older people, often encumbered with age-related disorders that are predicted to overwhelm health care systems. Improved health in old age and further extension of human health span are now likely to be generated primarily from increased understanding of the biology of aging, age-related susceptibility to disease, and modifiable factors that influence the aging process.

Definitions of Aging Aging is easy to recognize but difficult to define. Most definitions of aging indicate that it is a progressive process associated with declines in structure and function, impaired maintenance and repair systems, increased susceptibility to disease and death, and reduced reproductive capacity. There are both statistical and phenotypic components to aging. As recognized by Gompertz in the nineteenth century, aging in humans is associated with an exponential risk of mortality with time **(Fig. 463-1)**, although it is now realized that this plateaus in extreme old age because of healthy survivor bias. The phenotypic components of aging include structural and functional changes that are separated, somewhat artificially, into either primary aging changes (e.g., sarcopenia, grey hair, oxidative stress, increased peripheral vascular resistance) or age-related disease (e.g., dementia, osteoporosis, arthritis, insulin resistance, hypertension).

Definitions of aging rarely acknowledge the possibility that some of those biological and functional changes with aging might be adaptive or even reflect improvement and gain. Nor do they emphasize the impact of aging on responses to medical treatments. Old age is associated with increased vulnerability to many perturbations, including therapeutic interventions. This is a critical issue for clinicians—aging would be a less difficult problem if our disease-specific therapies retained their balance of risk to benefit into old age.

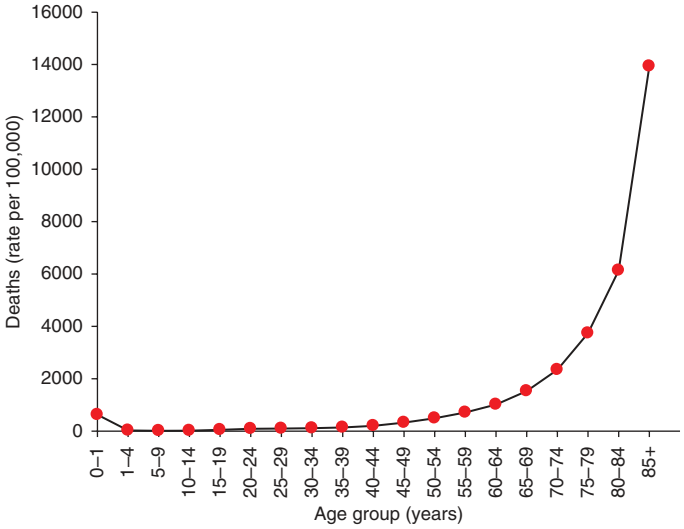


FIGURE 463-1 The rates of death in the United States (2010) showing exponential increase in mortality risk with chronological age.

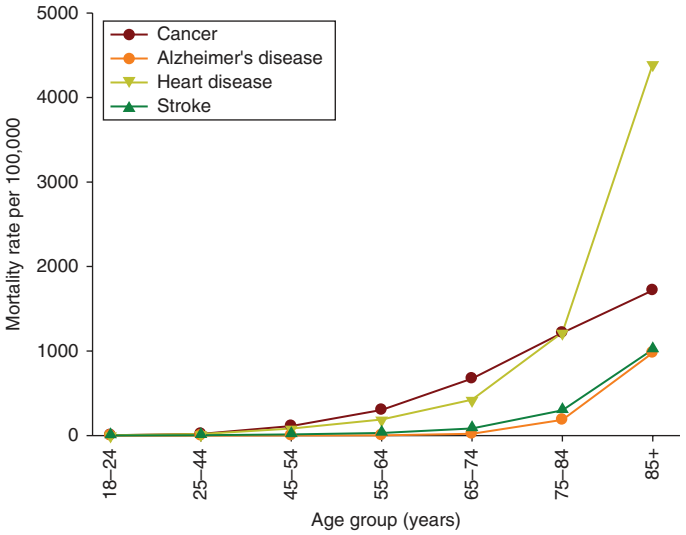


FIGURE 463-2 The rates of most common chronic diseases and related mortality increase with old age. *(Data from USA 2008-2010 CDC.)*

Aging and Disease Susceptibility Old age is the major independent risk factor for chronic diseases (and associated mortality) that are most prevalent in developed countries such as cardiovascular disease, cancers, and neurodegenerative disorders **(Fig. 463-2)**. Consequently, older people have multiple comorbidities, usually in the range of 5–10 illnesses per person.

Disease in older people is typically multifactorial with a strong component related to the underlying aging process. For example, in younger patients with dementia, Alzheimer's disease is a single disorder confirmed by examining brain tissue for plaques and tangles containing amyloid and tau proteins. However, the vast majority of people with dementia are elderly, and here the association between typical Alzheimer's neuropathology and dementia becomes less definitive. In the oldest-old, the prevalence of Alzheimer-type brain pathology is similar in people with and without clinical features of dementia. On the other hand, brains of older people with dementia usually show mixed pathology with evidence of Alzheimer's pathology along with features of other dementias such as vascular lesions, Lewy bodies, and non-Alzheimer's tauopathy. Many typical aging changes, such as microvascular dysfunction, oxidative injury, and mitochondrial impairment underlie many of the pathological changes.

The Longevity Dividend "Compression of morbidity" refers to the concept that the burden of lifetime illness might be compressed by medical interventions into a shorter period before death without necessarily increasing longevity. However, continuing development of successful therapeutic and preventive interventions focusing on individual diseases is less effective in older people because of multiple comorbidities, complications of overtreatment, and competing causes of death. Therefore, it has been proposed that further gains in health span and life expectancy will be achieved by a single intervention that delays aging and age-related disease susceptibility, rather than multiple treatments each targeting different individual age-related illnesses. This is called the "longevity dividend" and is driving an explosion of research into aging biology and, more importantly, interventions—genetic, pharmaceutical, and nutritional—that influence the rate of aging and delay age-related disease.

EVOLUTIONARY MECHANISMS FOR AGING

At the most basic level, living things have only two approaches to maintain their existence: immortality or reproduction. In a changing environment, reproduction combined with a finite life span has proved to be the successful strategy. Of course, finite life span is not the same as aging—although aging, by definition, contributes to a finite life span.

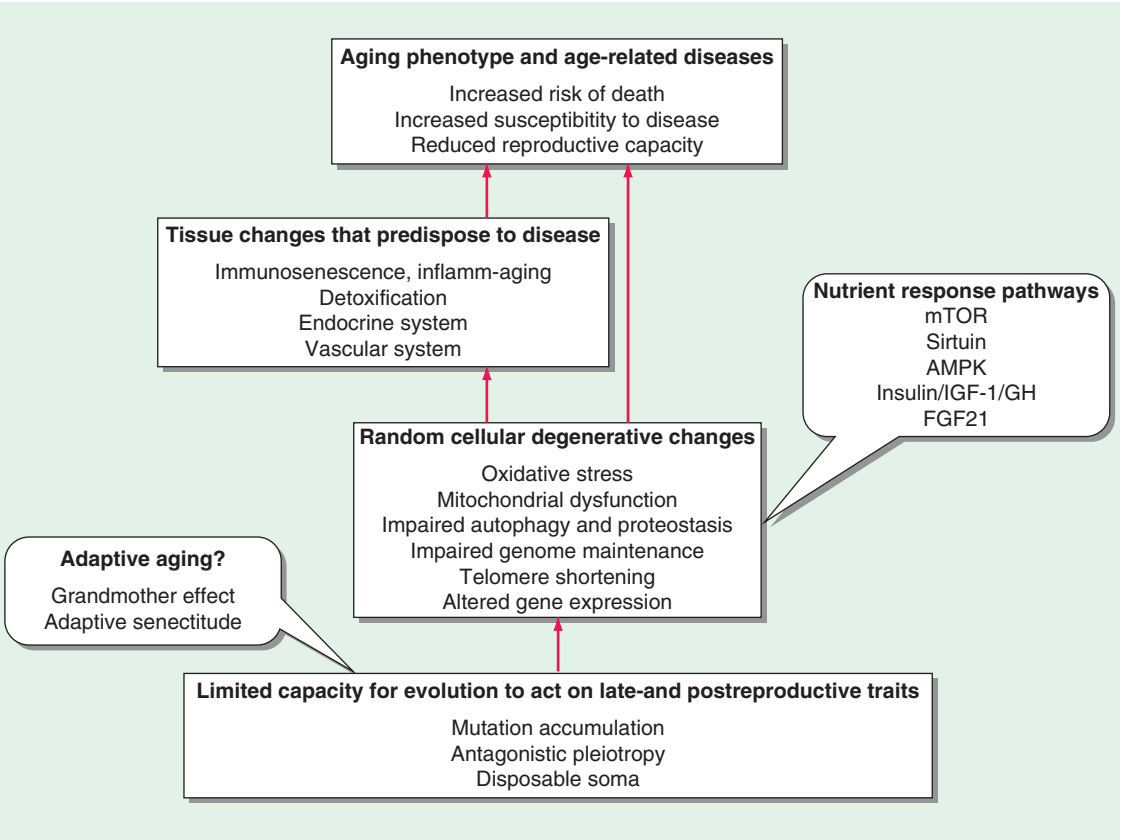


FIGURE 463-3 Schema linking evolution, cellular and tissue changes with aging. The call-out blue boxes indicate factors that might delay the aging process including nutrient response pathways and, possibly, adaptive evolutionary effects.

Many evolutionary theories related to aging are linked by their attempts to explain this interaction between reproduction and longevity **(Fig. 463-3)**. Most mainstream aging theories stem from the fact that evolution is driven by early reproductive success, whereas there is minimal selection pressure for late-life reproduction or postreproductive survival. Aging is seen as the random degeneration resulting from the inability of evolution to prevent it, i.e., the nonadaptive consequence of evolutionary "neglect." This conclusion is supported by studies that restricted reproduction to later life in the fruit fly, *Drosophila melanogaster*, thus permitting natural selection to operate on later life traits and leading to an increase in longevity. Likewise, many of the interventions that delay aging are associated with reduced reproductive capacity.

There are some species of plants and animals that do not appear to age, or at least they undergo an extremely slow aging process, termed "negligible senescence." The mortality rates of these species are relatively constant with time and they do not display any obvious phenotypic changes of aging. Conversely, some living things undergo programmed death immediately after reproduction such as annual plants and semelparous animals **(Fig. 463-4)**. However, many other living things from yeast to humans undergo a gradual aging process leading to death that is surprisingly similar at the cellular and biochemical level across taxa.

Some of the major classical evolutionary theories of aging include:

- *Programmed death*. The first evolutionary theory of aging was proposed by Weismann in 1882. This theory states that aging and death are programmed and have evolved to remove older animals from the population so that environmental resources such as food and water are freed up for younger members of the species.
- *Mutation accumulation*. This theory was proposed by Medawar in 1952. Natural selection is most powerful for those traits that influence reproduction in early life, and therefore, the ability of evolution to shape our biology declines with age. Germline mutations that are deleterious in later life can accumulate simply because natural selection cannot act to prevent them.
- *Antagonistic pleiotropy*. George C. Williams extended Medawar's theory when he proposed that evolution can allow for the selection of genes that are pleiotropic, i.e., beneficial for survival and reproduction in early life, but harmful in old age. For example, genes for sex hormones are necessary for reproduction in early life but contribute to the risk of cancer in old age.
- *Life history theory*. Evolution is influenced by the way that limited resources are allocated to all aspects of life including development, sexual maturation, reproduction, number of offspring, and senescence and death. Therefore "trade-offs" occur between these phases of life. For example, in a hostile environment, survival is highest

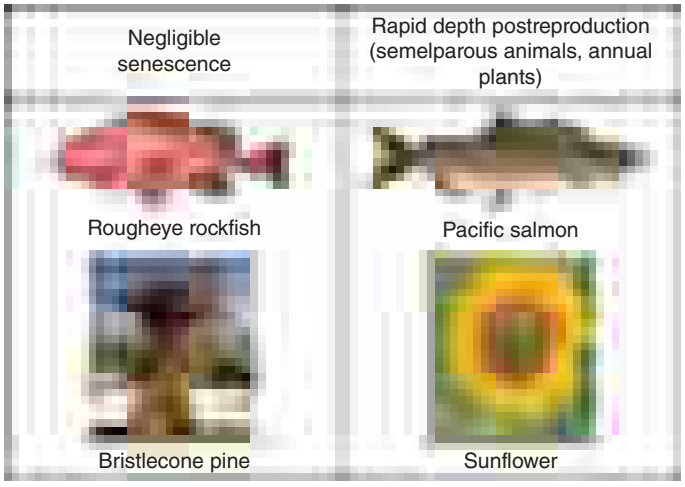


FIGURE 463-4 The typical features of aging (aging phenotype and exponential increase in risk of death) are not universal findings in living things. Some living things (e.g., Rockfish and the Bristlecone pine, sometimes called the Methuselah tree) undergo negligible senescence while others die almost immediately after first reproductive effort is completed (e.g., semelparous animals and annual plants).

for those species that have large numbers of offspring and short life span while in a safe and abundant environment, survival is highest for those species that invest resources in a smaller number of offspring and a longer life.

- *Disposable soma theory*. Kirkwood and Holliday in 1979 combined many of these ideas in the disposable soma theory of aging. There are finite resources available for the maintenance and repair of both germ and soma cells so there must be a trade-off between germ cells (i.e., reproduction) versus soma cells (i.e., longevity and aging). The soma cells are disposable from an evolutionary perspective, so they accumulate damage that causes aging while resources are preferentially diverted to the maintenance and repair of the germ cells. For example, the longevity of nematode worm, *Caenorhabditis elegans* is increased when its germ cells are ablated early in life.

All of these theories assume that natural selection has negligible or negative influences on aging. Some ideas propose that aspects of aging might be adaptive and raise the possibility that evolution can act on the aging process in a positive way. These include:

- *Grandmother hypothesis*. The grandmother hypothesis proposed by Hamilton in 1966 describes how evolution can enhance old age. In some animals, including humans, the survival of multiple, dependent offspring is beyond the capacity and resources of their mother. In this situation, the presence of a long-lived grandmother who shares in care of her grandchildren can have a major impact on their survival. These children share some of the genes of their grandmother including those that promoted their grandmother's longevity.
- *Mother's curse*. Mitochondrial dysfunction is a key component of the aging process. Mitochondria contain their own DNA and are only passed on from mother to child, because sperm cells contain almost no mitochondria. Therefore, natural selection can only act on the evolution of mitochondrial DNA in females. The "mother's curse" of the maternal inheritance of mitochondrial DNA might explain why females live longer and age more slowly than males.
- *Adaptive senectitude*. Many traits that are harmful in younger humans such as obesity, hypertension, oxidative stress, and declines in growth hormone and insulin-like growth factor type I (IGF-1) paradoxically appear to be associated with greater survival and function in old age. Perhaps driven by the grandmother effect, this might represent "adaptive senectitude" or "reverse antagonistic pleiotropy" whereby some traits that are harmful in young people become beneficial in older people.

CELLULAR PROCESSES THAT ACCOMPANY AGING

Many cellular processes change with aging. These are generally considered to be degenerative and stochastic or random changes that reflect some sort of time-dependent damage (Fig. 463-3) and have been called the hallmarks and pillars of aging. Whether any of these is the root cause of aging is unknown but they all contribute to the aging phenotype and disease susceptibility.

Oxidative Stress and the "Free Radical Theory of Aging" Free radicals are chemical species that are highly reactive because they contain unpaired electrons. Oxidants are oxygen-based free radicals that include the hydroxyl free radical, superoxide, and hydrogen peroxide. Most cellular oxidants are waste products generated by mitochondria during the production of ATP from oxygen. More recently, the role of oxidants in cellular signaling and inflammatory responses has been recognized. Unchecked, oxidants can generate chain reactions leading to widespread damage to biological molecules. Cells contain numerous antioxidant defense mechanisms to prevent such oxidative stress including enzymes (superoxide dismutase, catalase, glutathione peroxidase) and chemicals (uric acid, ascorbate). In 1956, Harman proposed the "free radical theory of aging" whereby oxidants generated by metabolism or irradiation are responsible for age-related damage. It is now well established that old age in most species is associated with increased oxidative stress, for example to DNA (8-hydroxyguanosine derivatives), proteins (carbonyls), lipids (lipoperoxides, malondialdehydes), and prostaglandins (isoprostanes). Conversely, many of the cellular antioxidant defense mechanisms including the antioxidant enzymes decline in old age. The free radical theory of aging has spawned numerous studies of supplementation with antioxidants such as vitamin E to delay aging in animals and humans. However, meta-analyses of human clinical trials performed to treat and prevent various diseases with antioxidant supplements indicate that they have no effect on, or may even increase, mortality.

Mitochondrial Dysfunction Aging is characterized by altered mitochondrial production of ATP and oxygen-derived free radicals. This leads to a vicious cycle mediated by accumulation of oxidative injury to mitochondrial proteins and DNA. With age, the number of mitochondria in cells decreases and there is an increase in their size (megamitochondria) associated with other structural changes including vacuolization and disrupted cristae. These morphological aging changes are linked with decreased activity of mitochondrial complexes I, II, and IV and decreased ATP production. Of all of the complexes involved in ATP production, the activity of complex IV (COX) is usually reported to be most impaired in old age. Reduced energy production is linked with generation of hydrogen peroxide and superoxide radicals leading to oxidative injury to mitochondrial DNA and accumulation of carbonylated mitochondrial proteins and mitochondrial lipoperoxides. As well as being implicated in the aging process, common geriatric syndromes including sarcopenia, frailty, and cognitive impairment are associated with mitochondrial dysfunction.

Telomere Shortening and Replicative Senescence Cells that are isolated from animal tissue and grown in culture only divide for a certain number of times before entering a senescent phase. This number of divisions is known as the Hayflick limit and tends to be less in cells isolated from older animals compared to younger animals. It has been suggested that aging in vivo might in part be secondary to some cells ceasing to divide because they have reached their Hayflick limit. Senescent cells produce a variety of cytokines, chemokines, and proteases, termed the senescence-associated secretory phenotype (SASP), that are major drivers of age-related inflammation. Eliminating senescent cells delayed aging in mice. On the other hand, cellular senescence may have a role in preventing proliferation of cells at risk of malignant transformation. One mechanism for replicative senescence relates to telomeres. Telomeres are repeat sequences of DNA at the end of linear chromosomes that shorten by around 50–200 base pairs during each cell division by mitosis. Once telomeres become too short, cell division can no longer occur. This mechanism contributes to the Hayflick limit and has been called the cellular clock. There are some studies that suggest that the length of telomeres in circulating leukocytes (leukocyte telomere length, LTL) decreases with age in humans. However, the aging process also occurs in tissues that do not undergo repeated cell division such as neurons.

Altered Gene Expression, Epigenetics, and microRNA The expression of many genes and proteins changes during the aging process. These changes are complicated and vary between species and tissue. Such heterogeneity reflects increasing dysregulation of gene expression with age while appearing to exclude a programmed and/or uniform response. With old age, reductions in the expression of genes and proteins associated with mitochondrial function and increased expression of those involved with inflammation, genome repair, and oxidative stress are noted. Several factors control the regulation of gene and protein expression that change with aging. These include the epigenetic state of the chromosomes (e.g., DNA methylation and histone acetylation) and microRNAs (miRNAs). DNA methylation correlates with age, although the pattern of change is complex. Histone acetylation is regulated by many enzymes including Sirtuin 1 (SIRT1), a protein that has marked effects on aging and the response to dietary restriction in many species. miRNAs are a very large group of noncoding lengths of RNA (18–25 nucleotides) that inhibit translation of multiple different mRNAs through binding their 3′ untranslated regions (3′UTRs). The expression of miRNAs usually decreases with aging and is altered in some age-related diseases. Specific miRNAs

linked with aging pathways include miR-21 (associated with target of rapamycin pathway) and miR-1 (associated with insulin/insulin-like growth factor 1 pathway).

Impaired Autophagy and Proteostasis Cells can remove damaged macromolecules and organelles in a number of ways, often generating cellular energy as a by-product. Intracellular degradation is undertaken by the lysosomal system and the ubiquitin proteasomal system. Both are impaired with aging, leading to the accumulation of waste products that alter cellular functions. Such waste products include lipofuscin, a brown auto-fluorescent pigment found within lysosomes of most cells in old age and often considered to be one of the most characteristic histological features of aging cells. Lysosomes are organelles that contain proteases, lipases, glycases, and nucleotidases that degrade intracellular macromolecules, membrane components, organelles, and some pathogens through a process called autophagy. The lysosomal process most impaired with aging is macroautophagy, which is regulated by numerous autophagy-related genes. Proteostasis refers to the maintenance of protein quality through regulation of protein folding and protein degradation. Chaperones orchestrate appropriate folding of proteins while degradation involves ubiquitin tagging, proteases, and the unfolded protein response. With aging damaged, aggregated and misfolded proteins increase because of age-related changes in proteostasis. This may contribute to the aggregation of proteins such as tau, β-amyloid, α-synuclein in age-related neurodegenerative diseases such as dementia and Parkinson's disease.

AGING CHANGES IN SPECIFIC TISSUES THAT PREDISPOSE TO DISEASE

Aging changes in some tissues increase susceptibility to age-related disease as a secondary or downstream phenomenon (Fig. 463-3). In humans, this includes, but is not limited to, the immune system (leading to increased infections and autoimmunity), hepatic detoxification (leading to increased exposure to disease-inducing endobiotics and xenobiotics), the endocrine system (leading to hypogonadism and bone disease), and the vascular system (leading to segmental or global ischemic changes in many tissues).

Inflamm-aging and Immunosenescence Old age is associated with increased background levels of inflammation including blood measurements of C reactive protein (CRP), erythrocyte sedimentation rate (ESR), and cytokines such as interleukin 6 (IL-6) and tumor necrosis factor alpha (TNFα). This has been termed "inflamm-aging" and elevated IL-6 in particular has been associated with frailty and dementia. T cells are less numerous because of age-related atrophy of the thymus, while B cells overproduce autoantibodies, leading to the age-related increase in autoimmune diseases and gammopathies. Thus, older people are generally considered to be immunucompromised and have reduced responses to infection (fever, leukocytosis) with increased mortality.

Detoxification and the Liver Old age is associated with impaired detoxification of various disease-causing endobiotics (e.g., lipoproteins) and xenobiotics (e.g., neurotoxins and carcinogens) leading to increased systemic exposure. In humans, the liver is the major organ for the clearance of such toxins. Hepatic clearance of many substrates is reduced in old age as a consequence of reduced hepatic blood flow, impaired hepatic microcirculation and, in some cases, reduced expression of xenobiotic metabolizing enzymes. These changes in hepatic detoxification also increase the likelihood of increased blood levels of, and adverse reactions to, medications.

Endocrine System Hormonal changes with aging have been a focus for aging research for over a century, partly because of the erroneous belief that supplementation with sex hormones will delay aging and rejuvenate older people. There are age-related reductions in sex steroids, growth hormone, IGF-1, and dehydroepiandrosterone (DHEA). These hormonal changes may contribute to some features of aging such as sarcopenia and osteoporosis but also provide protection against cancer and cardiovascular disease. Adverse effects of long-term hormonal supplementation outweigh any potential beneficial effects on life span.

Vascular Changes There is a continuum from vascular aging through to atherosclerotic disease, present in many, but not all, older people. Vascular aging changes overlap with the early stages of hypertension and atherosclerosis, with increasing arterial stiffness and vascular resistance. This contributes to myocardial ischemia and strokes but also appears to be associated with geriatric conditions such as dementia, sarcopenia, and osteoporosis. In these conditions, impaired exchange between blood and tissues is a common pathogenic factor. For example, the risk of Alzheimer's disease and dementia is increased in patients with risk factors for vascular disease, and pathological evidence for microvascular changes is seen in postmortem studies of brains of people with established Alzheimer's disease. Similarly, strong epidemiological links have been found between osteoporosis and standard vascular risk factors, while significant age-associated changes are in the microcirculation of osteoporotic bone. Sarcopenia might also be related to the effects of age on the muscle vasculature, which is altered in old age. The sinusoidal microcirculation of the liver becomes markedly altered during aging (pseudocapillarization), which influences hepatic uptake of lipoproteins, insulin, and other substrates. In fact, it has often been overlooked that in his original exposition of the free radical theory of aging, Harman proposed that the primary target of oxidative stress was the vasculature and that many aging changes were secondary to impaired exchange across the damaged blood vessels.

GENETIC INFLUENCES ON AGING

There is variability in aging and life span in populations of genetically identical species such as mice that are housed in the same environment. The heritability of life span in human twin studies is estimated to be only 25% (although there is stronger hereditary contribution to extreme longevity). These two observations indicate that the cause of aging is unlikely to lie only within the DNA code. On the other hand, genetic studies initially undertaken in the nematode worm *C. elegans* and, more recently, in models from yeast to mice have shown that manipulating genes can have profound effects on the rate of aging. Perhaps surprisingly, this can often be generated by variability in *single* genes, and for some genetic mechanisms, there is very strong evolutionary conservation.

Genetic Progeroid Syndromes There are a few very rare, genetic premature aging conditions that are called progeroid syndromes. These conditions recapitulate some, but not all, age-related diseases and senescent phenotypes. They are mostly caused by impairment of genome and nuclear maintenance. These syndromes include the following:

- *Werner's syndrome*. This is an autosomal recessive condition caused by a mutation in the *Werner's (WRN)* gene. This gene codes for a RecQ helicase, which unwinds DNA for both repair and replication. It is typically diagnosed in teen years and is associated with premature onset of atherosclerosis, osteoporosis, cancers, and diabetes with death by age of 50 years.
- *Hutchinson–Gilford progeria syndrome (HGPS)*. This usually occurs as a de novo, noninherited mutation in the lamin A gene (*LMNA*) leading to an abnormal protein called progerin. LMNA is required for the nuclear lamina, which provides structural support to the nucleus. Marked development changes are obvious in infancy with subsequent onset of atherosclerosis, kidney failure, and scleroderma-like features and death during the teen years. Lamin A-dependent nuclear defects have been reported in normal human aging.
- *Cockayne syndrome*. This includes a number of autosomal recessive disorders with features such as impaired neurological growth, photosensitivity (xeroderma pigmentosum), and death during childhood years. These are caused by mutations in the genes for DNA excision repair proteins, *ERCC-6* and *ERCC-8*.

Gene Studies in Long-Lived Humans The main genes that have been consistently associated with increased longevity in human candidate gene studies are *APOE* and *FOXO3A*. ApoE is an apoprotein

found in chylomicrons while the ApoE4 isoform is a risk factor for Alzheimer's disease and cardiovascular disease, which might explain its association with reduced life span. FOXO3A is a transcription factor involved in the insulin/IGF-1 pathway, and its homolog in *C. elegans*, daf16, has a marked impact on aging in these nematodes. Genome-wide association studies (GWASs) of centenarians have confirmed the association of longevity with *APOE*. GWAS has been used to identify a range of other single nucleotide polymorphisms (SNPs) that might be associated with longevity including SNPs in the sirtuin genes and the progeroid syndrome genes, *LMNA* and *WRN*. Gene set analysis of GWAS studies has shown that both the insulin/IGF-1 signaling pathway and the telomere maintenance pathway are associated with longevity.

One group of particular interest are people with Laron-type dwarfism. These people have mutations in the growth hormone receptor that causes severe growth hormone resistance. In mice, similar knockout of the growth hormone receptor (GHRKO mice, Methuselah mice) is associated with extremely long life. Therefore, subjects with Laron syndrome have been carefully studied and it was found that they have very low rates of cancer and diabetes mellitus, and possibly, longer lives.

Nutrient-Sensing Pathways Many living things have evolved to respond to periods of nutritional shortage and famine by increasing cellular resilience and delaying reproduction until food supply becomes abundant once again. This increases the chances of reproductive success and survival of offspring. Lifelong food shortage, often termed "caloric restriction (CR)" (or "dietary restriction"), increases life span and delays aging in many animals, probably as a side effect of this famine response. Many of the genes and pathways that regulate the way that cells respond to nutritional undersupply have been identified, initially in yeast and *C. elegans*. In general, manipulation of these pathways (through genetic knockout or overexpression, or pharmacological agonists and antagonists) alters the aging benefits of CR, and in some cases, the life span of animals on normal diets. These pathways are all very influential cellular "switches" that control a wide range of key functions including protein translation, autophagy, mitochondrial function and bioenergetics, and the cellular metabolism of fats, proteins, and carbohydrates. The discovery of these nutrient-sensing pathways has led to targets for pharmacological extension of life span. The main nutrient-sensing pathways that influence aging and responses to CR include:

- *SIRT1*. The activity of SIRT1 is regulated by levels of reduced nicotinamide adenine dinucleotide (NAD^+), which are increased when cellular energy stores are depleted. Important downstream targets include PGC-1α and Nrf2, which act on mitochondrial biogenesis.
- *Mechanistic target of rapamycin (mTOR)*. mTOR is activated by branched chain amino acids, providing a link to dietary protein intake. It has two main components, mTORC1 and mTORC2, of which mTORC1 is most relevant to aging. Key downstream targets of mTOR of relevance to aging include the tuberous sclerosis protein (TSC) and 4EBP1, which influences protein production.
- *5′ adenosine monophosphate-activated protein kinase (AMPK)*. AMPK is activated by increased levels of AMP, which reflect cellular energy status.
- *Insulin signaling and IGF-1/growth hormone (IIS)*. These two pathways are usually considered together because they overlap in lower animals and have diverged only in higher animals. They respond to carbohydrate and protein intake. An important downstream target for this pathway is a transcription factor called FOXO.
- *Fibroblast growth factor 21 (FGF21)*. FGF21 is produced mostly by the liver in response to CR and reduced protein intake.

Mitochondrial Genes Mitochondrial function is influenced by genes located both in the mitochondria (mtDNA) and the nucleus. mtDNA is considered to have a prokaryotic origin and is highly conserved across taxa. It forms a circular loop of 16,569 nucleotides in humans. Aging is associated with increased frequency of mutations in mtDNA as a consequence of its high exposure to oxygen-derived free radicals and relatively inefficient DNA repair machinery. Nuclear DNA encodes ~1000–1500 genes for mitochondrial function including genes involved with oxidative phosphorylation, mitochondrial metabolic pathways, and enzymes required for biogenesis. These genes are thought to have originated in mtDNA but subsequently translocated to the nucleus and, unlike mtDNA genes, their sequence is stable with aging.

Genetic manipulation of mitochondrial genes in animals influences aging and life span. In *C. elegans*, many mutants with defective electron transfer chain function have increased life span. The mtDNA "mutator" mice which lack the mtDNA proofreading enzyme have increased mtDNA mutations and premature aging, while overexpression of mitochondrial uncoupling proteins leads to longer life span. In humans, hereditary variability in mtDNA is associated with diseases (mitochondriopathies such as Leigh's disease) and aging. For example, in Europeans, mitochondrial DNA haplogroup J (haplogroups are combinations of genetic variants that exist in specific populations) is associated with longevity, and haplogroup D is overrepresented in Asian centenarians.

Nuclear DNA Genomic DNA damage accumulates in cells with aging while genetic progeroid conditions such as Werner's syndrome and HGPS are associated with impaired DNA maintenance and repair. Age-related DNA changes include mutations, chromosomal aneuploidy, copy number variations, and telomere shortening. Apart from telomere attrition, these changes are random and vary between cells, and may contribute to age-related cancers.

STRATEGIES THAT INCREASE HEALTH SPAN AND DELAY AGING

Aging is an intrinsic feature of human life whose manipulation has fascinated humans ever since becoming conscious of their own existence. Several long-term experimental interventions (e.g., resveratrol, rapamycin, spermidine, and metformin) may open doors for corresponding pharmacological strategies. Surprisingly, most of the effective aging interventions proposed converge on only a few molecular pathways: nutrient signaling, mitochondrial proteostasis, and the autophagic machinery.

Life span is inevitably accompanied by functional decline, steady increase of a plethora of chronic diseases, and ultimately death. For millennia, it has been a dream of mankind to prolong both life span and health span. Developed countries have profited from the medical improvements and their transfer to public health care systems—as well as from better living conditions derived from their socioeconomic power—to achieve remarkable increases in life expectancy during the last century. In the United States, the percentage of the population aged ≥65 years is projected to increase from 13% in 2010 to 19.3% in 2030. However, old age remains the main risk factor for major life-threatening disorders, and the number of people suffering from age-related diseases is anticipated to almost double over the next two decades. The prevalence of age-related pathologies represents a major threat as well as an economic burden that urgently needs effective interventions.

Molecules, drugs, and other interventions that might decelerate aging processes continue to raise interest among both the general public and scientists of all biological and medical fields. Over the past two decades, this interest has taken root in the fact that many of the molecular mechanisms underlying aging are interconnected and linked with pathways that cause disease, including cancer, cardiovascular and neurodegenerative disorders. Unfortunately, among the many proposed aging interventions, only a few have reached a certain age themselves. Results often lack reproducibility because of a simple inherent problem: interventions in aging research take a lifetime to assess. Experiments lasting the lifetime of animal models are prone to develop artifacts, increasing the possibilities and time windows for experimental discrepancies. Some inconsistencies in the field arise from overinterpreting life span-shortening models and scenarios as being accelerated aging.

Many substances and interventions have been claimed to be antiaging throughout history and into the present. In the following sections, interventions will be restricted to those that meet the following highly selective criteria: (1) promotion of life span and/or health span, (2) validation in at least three model organisms, and (3) confirmation by at least three different laboratories. These include: (1) CR and fasting regimens, (2) some pharmacotherapies (resveratrol, rapamycin, spermidine, and metformin), and (3) exercise.

Caloric Restriction One of the most important and robust interventions that delays aging is CR. This outcome has been recorded in rodents, dogs, worms, flies, yeasts, monkeys, and prokaryotes. CR is defined as a reduction in the total caloric intake, usually of about 30% and without malnutrition. CR reduces the release of growth factors such as growth hormone, insulin, and IGF-1, which are activated by nutrients and have been shown to accelerate aging and enhance the probability for mortality in many organisms. Yet the effects of CR on aging were first discovered by McCay in 1935 long before the effects of such hormones and growth factors on aging were recognized. The cellular pathways that mediate this remarkable response have been explored in many experimental models. These include the nutrient sensing pathways (mTOR, AMPK, insulin/IGF-1, and sirtuins) as well as transcription factors (FOXO in *D. melanogaster* and daf-16 in *C. elegans*). The transcription factor Nrf2 appears to confer most of the anticancer properties of CR in mice, even though it is dispensable for life span extension.

The effects of CR in monkeys have been assessed in two studies with different outcomes: one study observed prolonged life while the other did not. However, both studies confirmed that CR increases health span by reducing the risk for diabetes, cardiovascular disease, and cancer. In humans, CR is associated with increased life and health span. This is most convincingly demonstrated in Okinawa, Japan, where one of the most long-lived human populations resides. In comparison to the rest of the Japanese population, Okinawan people usually combine an above-average amount of daily exercise with a below-average food intake. However, when Okinawan families move to Brazil, they adopt a Western lifestyle that affects both exercise and nutrition, causing a rise in weight and a reduction in life expectancy by nearly two decades. In the Biosphere II project, where volunteers lived together for 24 months undergoing an unforeseen severe CR, there were improvements in insulin, blood sugar, glycated hemoglobin, cholesterol levels, and blood pressure—all outcomes that would be expected to benefit life span. CR changes many aspects of human aging that might influence life span such as the transcriptome, hormonal status (especially IGF-1 and thyroid hormones), oxidative stress, inflammation, mitochondrial function, glucose homeostasis, and cardiometabolic risk factors. Epigenetic modifications are an emerging target for CR.

It must be noted that maintaining CR and avoiding malnutrition is not only arduous in humans but is also linked with substantial side effects. For instance, prolonged reduction of calorie intake may decrease fertility and libido, impair wound healing, reduce the potential to combat infections, and lead to amenorrhea and osteoporosis.

While extreme obesity (body mass index [BMI] >35) leads to a 29% increased risk of dying, people with BMI in the overweight range seem to have reduced mortality, at least in population studies of middle-aged and older subjects. People with a BMI in the overweight range seem more able to counteract and respond to disease, trauma, and infection, whereas CR impairs healing and immune responses. On the other hand, BMI is an insufficient denominator of body and body fat composition. A well-trained athlete may have an equivalent BMI compared to a fat person because of the higher muscle mass density. The waist:hip ratio is a much better indicator for body fat and an excellent and stringent predictor for the risk to die from cardiovascular disease: the lower the waist:hip ratio, the lower the risk.

PERIODIC FASTING How can CR be translated to humans in a socially and medically feasible way? A whole series of periodic fasting regimens are asserting themselves as suitable strategies, among them the alternate-day fasting diet, the "five:two" intermittent fasting diet, and a 48-h fast once or twice each month. Periodic fasting is psychologically more viable, lacks some of the negative side effects and is only accompanied by minimal weight loss.

It is striking that many cultures implement periodic fasting rituals, for example Buddhists, Christians, Hindus, Jews, Muslims, and some African animistic religions. It could be speculated that a selective advantage of fasting versus nonfasting populations is conferred by health-promoting attributes of religious routines that periodically limit caloric intake. Indeed, several lines of evidence indicate that intermittent fasting regimens exert antiaging effects. For example, improved morbidity and longevity were observed among Spanish home nursing residents who underwent alternate-day fasting. Even rats subjected to alternate-day fasting live up to 83% longer than normally fed control animals and one 24-h fasting period every 4 days is sufficient to generate life span extension.

Repeated fasting and eating cycles may circumvent the negative side effects of sustained CR. This strategy may even yield effects despite extreme overeating during the nonfasting periods. In a spectacular experiment, mice fed a high-fat diet in a time-restricted manner, i.e., with regular fasting breaks, showed reduced inflammation markers, no fatty liver and were slim in comparison to mice with equivalent total calorie consumption but ad libitum. From an evolutionary point of view, this kind of feeding pattern may reflect mammalian adaptation to food availability: overeating in times of nutrient availability (e.g., after a hunting success) and starvation in between. This is how some indigenous peoples who have avoided Western lifestyles live today; those who have been investigated show limited signs of age-induced diseases such as cancer, neurodegeneration, diabetes, cardiovascular disease, and hypertension.

Fasting exerts beneficial effects on health span by minimizing the risk of developing age-related diseases including hypertension, neurodegeneration, cancer, and cardiovascular diseases. The most effective and rapid repercussion of fasting is reduction in hypertension. Two weeks of water-only fasting resulted in a blood pressure below 120/80 mmHg in 82% of subjects with borderline hypertension. Ten days of fasting cured all hypertensive patients who had been taking antihypertensive medication previously.

Periodic fasting dampens the consequences of many age-related neurodegenerative diseases (Alzheimer's disease, Parkinson's disease, Huntington's disease, and frontotemporal dementia but not amyotrophic lateral sclerosis in mouse models). Fasting cycles are as effective as chemotherapy against certain tumors in mice. In a combination with chemotherapy, fasting protected mice against the negative side effects of chemotherapeutic drugs, while it enhances their efficacy against tumors. Combining fasting and chemotherapy rendered 20–60% mice cancer-free when inoculated with highly aggressive tumors like glioblastoma or pancreatic tumors, which have 100% mortality even with chemotherapy.

Pharmacological Interventions to Delay Aging and Increase Life Span Virtually all obese people know that stable weight reduction will reduce their elevated risk of cardiometabolic disease and enhance their overall survival, yet only 20% of overweight individuals are able to lose 10% weight for a period of at least 1 year. Even in the most motivated people (such as the "Cronies" who deliberately attempt long-term CR in order to extend their lives), long-term CR is extremely difficult. Thus, focus has been directed at the possibility of developing medicines that replicate the beneficial effects of CR without the need for reducing food intake ("CR-mimetics," **Fig. 463-5**):

- *Resveratrol*. Resveratrol, an agonist of SIRT1, is a polyphenol that is found in grapes and in red wine. The potential of resveratrol to promote life span was first identified in yeast, and it has gathered fame since, at least in part because it has been suggested to be responsible for the so-called French paradox whereby wine reduces some of the cardiometabolic risks of a high fat diet. Resveratrol has been reported to increase life span in many lower order species such as yeast, fruit flies, worms as well as mice on high-fat diets. In

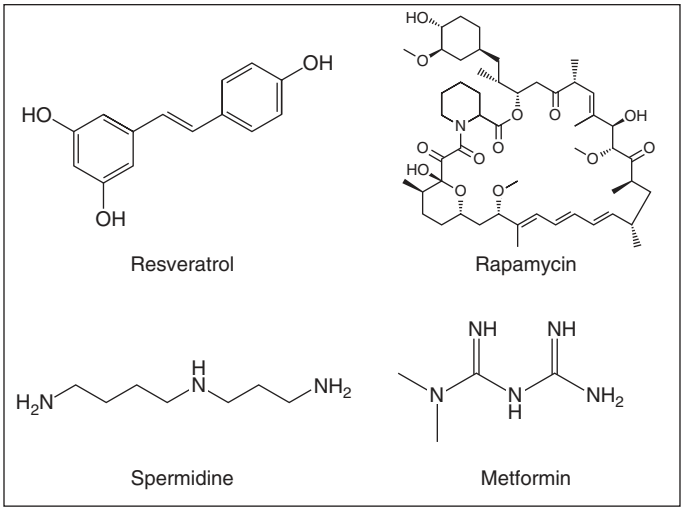


FIGURE 463-5 Chemical structure of four agents (resveratrol, rapamycin, spermidine, and metformin) that have been shown to delay aging in experimental animal models.

monkeys fed with a diet high in sugar and fat, resveratrol had beneficial outcomes related to inflammation and cardiometabolic parameters. Some studies in humans have also shown improvements in cardiometabolic function while others have not. Gene expression studies in animals and humans reveal that resveratrol mimics some of the metabolic and gene expression changes of CR.

- *Rapamycin*. Rapamycin, an inhibitor of mTOR, was originally discovered on the Easter Island (Rapa Nui, hence its name) as a bacterial secretion with antibiotic properties. Before its immersion in the antiaging field, rapamycin was known as an immunosuppressant and cancer chemotherapeutic in humans. Rapamycin extends life span in all organisms tested so far, including yeast, flies, worms, and mice. However, the potential utility of rapamycin for human life span extension is likely to be limited by adverse effects related to immunosuppression, wound healing, proteinuria, and hypercholesterolemia, among others. An alternative strategy may be intermittent rapamycin feeding, which was found to increase mouse life span.
- *Spermidine*. Spermidine is a physiological polyamine that induces autophagy-mediated life span extension in yeast, flies, and worms. Spermidine levels decrease during life of virtually all organisms including humans, with the stunning exception of centenarians. Oral administration of spermidine or upregulation of bacterial polyamine production in the gut both lead to life span extension in short-lived mouse models. Spermidine has also been found to have beneficial effects on neurodegeneration probably by increasing transcription of genes involved in autophagy.
- *Metformin*. Metformin, an activator of AMPK, is a biguanide first isolated from the French lilac that is widely used for the treatment of type 2 diabetes mellitus. Metformin decreases hepatic gluconeogenesis and increases insulin sensitivity. Metformin has other actions including inhibition of mTOR and mitochondrial complex I, and activation of the transcription factor SKN-1/Nrf2. Metformin increases life span in different mouse strains including female mouse strains predisposed to high incidence of mammary tumors. At a biochemical level, metformin supplementation is associated with reduced oxidative damage and inflammation and mimics the some of the gene expression changes seen with CR.

Exercise and Physical Activity In humans and animals, regular exercise reduces the risk of morbidity and mortality. Given that cardiovascular diseases are the dominant cause of aging in humans but not in mice, the effects on human health may be even stronger than those seen in mouse experiments. An increase in aerobic exercise capacity, which declines during aging, is associated with favorable effects on blood pressure, lipids, glucose tolerance, bone density, and depression in older people. Likewise, exercise training protects against aging disorders such as cardiovascular diseases, diabetes mellitus, and osteoporosis. Exercise is the only treatment that can prevent or even reverse sarcopenia (age-related muscle wasting). Even moderate or low levels of exercise (30 min walking per day) have significant protective effects in obese subjects. In older people, regular physical activity has been found to increase the duration of independent living.

While clearly promoting health and quality of life, regular exercise does not extend life span. Furthermore, the combination of exercise with CR has no additive effect on maximal life span in rodents. On the other hand, alternate-day fasting with exercise is more beneficial for the muscle mass than single treatments alone. In nonobese humans, exercise combined with CR has synergistic effects on insulin sensitivity and inflammation. From the evolutionary perspective, the responses to hunger and exercise are linked: when food is scarce, increased activity is required to hunt and gather.

Hormesis The term hormesis describes the, at first sight paradoxical, protective effects conferred by the exposure to low doses of stressors or toxins (or as Nietzsche stated "What does not kill me makes me stronger"). Adaptive stress responses elicited by noxious agents (chemical, thermal, or radioactive) precondition an organism rendering it resistant to subsequent higher and otherwise lethal doses of the same trigger. Hormetic stressors have been found to influence aging and life span presumably by increasing cellular resilience to factors that might contribute to aging such as oxidative stress.

Yeast cells that have been exposed to low doses oxidative stress exhibit a marked antistress response that inhibits death following exposure to lethal doses of oxidants. During ischemic preconditioning in humans, short periods of ischemia protect the brain and the heart against a more severe deprivation of oxygen and subsequent reperfusion-induced oxidative stress. Similarly, the lifelong and periodic exposure to various stressors can inhibit or retard the aging process. Consistent with this concept, heat or mild doses of oxidative stress can lead to life span extension in *C. elegans*. CR can also be considered as a type of hormetic stress that results in the activation of antistress transcription factors (Rim15, Gis1, and Msn2/Msn4 in yeast and FOXO in mammals) that enhance the expression of free radical-scavenging factors and heat shock proteins.

CONCLUSIONS

Clinicians need to understand aging biology in order to better manage those people who are elderly. Moreover, there is an urgent need to develop strategies based on aging biology that delay aging, reduce the onset of age-related disorders, and increase health span for future generations. Interventions related to nutrition and those drugs that act on nutrient-sensing pathways are being developed and, in some cases, are already being tested in humans.

FURTHER READING

Campisi J: Aging, cellular senescence, and cancer. Annu Rev Physiol 75:685, 2013.

de Cabo R et al: The search for antiaging interventions: From elixirs to fasting regimens. Cell 157:1515, 2014.

Kaeberlein M et al: Healthy aging: The ultimate preventative medicine. Science 350:1191, 2015.

Kennedy BK et al: Geroscience: Linking aging to chronic disease. Cell 159:709, 2014.

Longo VD et al: Interventions to slow aging in humans: Are we ready? Aging Cell 14:497, 2015.

López-Otín C et al: The hallmarks of aging. Cell 153:1194, 2013.

López-Otín C et al: Metabolic control of longevity. Cell 166:802, 2016.

464 Clinical Problems Associated with the Aging Process

Joseph G. Ouslander, Bernardo Reyes

AGING AND GERIATRIC CARE

Demographics of Aging and its Implications for Geriatric Care The United States and other countries will continue to experience a rapid increase in the number of older adults who seek health care. The most rapidly growing segment of the population in the United States and many other developed countries is those aged >80 years (Fig. 464-1). Sex composition of the aging population around the world is also expected to change. Although females outlive males, an improvement in survival of the oldest males could result in more balanced sex distribution in the geriatric population in the future.

Based on the United Nations' 2015 World Population Aging Report, in high-income countries, consumption of health care resources will be most affected by the shift in the age distribution of the population over the next several decades. The World Health Organization continues to work actively to raise awareness of the changes necessary in current health care systems beyond increments in their budgets. Planning is increasingly being based on expected levels of disability and comorbidity. As lifespan increases, efforts should continue to focus on promoting healthy aging to reduce the burden of disability in health care systems all over the world.

Implications of the Aging Population for Health Care Systems and System-Based Practice The geriatric population requires different approaches to care for several reasons. For example, acute illnesses are most often not treated in isolation, but in the context of multiple co-morbidities. Close to half of those aged >80 have three, and about one-third have four or more chronic conditions (Fig. 464-2). Functional disabilities are prevalent (Fig. 464-3), which require careful attention in the evaluation of the older patient, along with assessment of social supports available for assistance when needed for independent and safe living.

Effectively caring for the geriatric population requires consideration of several key principles:

1. Aging is not a disease; normal aging changes generally do not cause symptoms, but do increase susceptibility to many diseases and conditions due to diminished physiologic reserve (which has been termed "homeostenosis"). Aging is also associated with greater heterogeneity in virtually every measureable variable. Lab values outside the "normal" range are more common and may not reflect pathology.
2. Medical conditions are commonly multiple ("multi-morbidity") and multifactorial in origin, requiring a comprehensive approach to evaluation and management.
3. Many potentially reversible and treatable conditions are underdiagnosed and under-evaluated in this population, such as fall risk, urinary incontinence, and elder abuse and neglect; simple screening tools can help detect them.
4. Similarly, cognitive and affective disorders (e.g. mild cognitive impairment, dementia, depression, anxiety) are common and may be undiagnosed in early stages; simple screening tools can help detect them.
5. Iatrogenic illnesses are common, especially related to adverse drug reactions and immobility and related deconditioning and other complications.
6. Functional ability and quality of life, as opposed to cure, are key goals of care.
7. Social history, social support, and patient preferences are critical to treat older people in a safe and person-centered manner.
8. Effective geriatric care requires inter-professional collaboration among many different disciplines.
9. Geriatric care is provided largely outside the hospital—at home, in skilled nursing and assisted living settings; and attention to care transitions between settings is essential for effective care.
10. Ethical issues, palliative care, and end-of-life care are critical aspects of caring for the geriatric population.

In this chapter, these key principles are highlighted. The reader is referred to textbooks of geriatric medicine for more details on each of

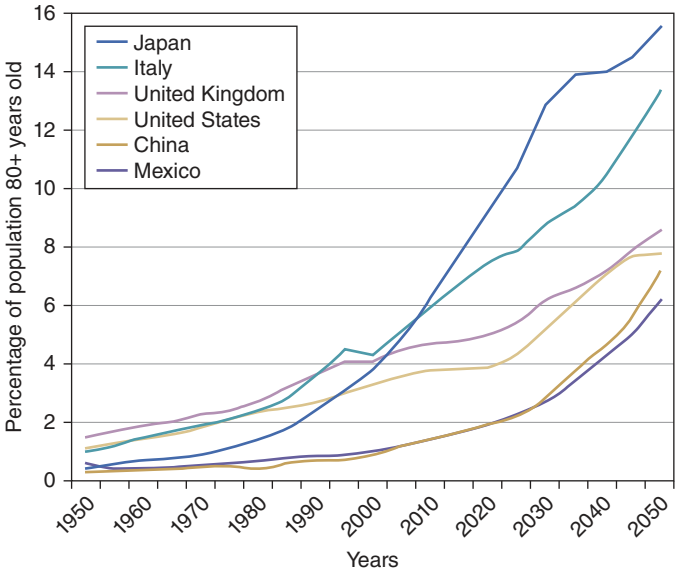


FIGURE 464-1 Percentage of the population age >80 years from 1950 to 2050 in representative nations. *(Updated data available at https://esa.un.org/unpd/wpp/Graphs/DemographicProfiles/. Accessed December 30, 2016.)*

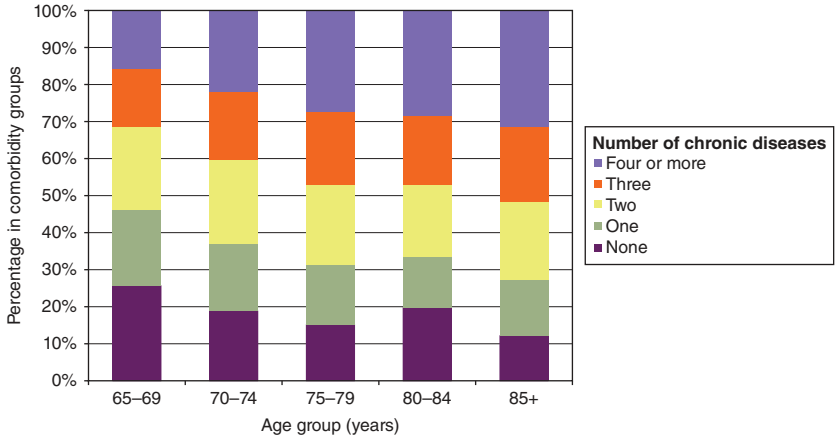


FIGURE 464-2 Prevalence of comorbidity by age group in persons ≥65 years old living in the United States and enrolled in Medicare parts A and B in 1999. (*From JL Wolff et al: Arch Intern Med 162:2269, 2002.*)

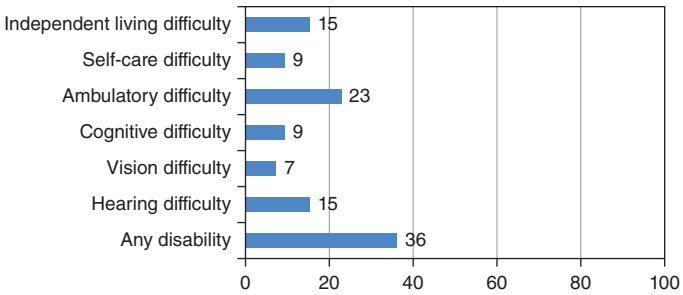


FIGURE 464-3 Percentage of people age 65+ with various disabilities. (*Source: US Census Bureau, American Community Survey, 2013. Available at https://aoa.acl.gov/Aging_Statistics/Profile/2014/index.aspx. Accessed December 30, 2016.*)

the principles, and the management of common diseases and conditions in this population.

Models of Geriatric Care and Care Transitions Several innovative models of care have been developed over the last three decades designed to provide high quality and effective care for the burgeoning geriatric population with multi-morbidity, functional and cognitive impairment, and challenges with social support. These include outpatient comprehensive geriatric assessment programs, inpatient acute care for the elderly (ACE) units and consultation services, and home-based programs such as Geriatric Resources for Assessment and Care of Elders (GRACE), Home Based Primary Care (in the VA system), and Independence at Home. These models of care are assuming greater importance in the emerging era of value-based purchasing for health care services. While they may be challenging and inefficient to implement in the Medicare fee-for-service system, they may also result in improved care and lower costs overall as Medicare shifts from fee-for-service to other models of reimbursement, such as accountable care organizations, bundled payment programs, and increasing the number of older people enrolled in Medicare managed care.

Improving transitions of care between settings has become a major focus of the federal government, health systems, hospitals, post-acute (PAC) and long-term care (LTC) organizations and programs, physicians, and other health care professionals. Geriatric patients are especially vulnerable to complications at the time of discharge from an acute medical or psychiatric hospital, as well as at the time of discharge from a PAC facility (skilled nursing facility [SNF]; acute rehabilitation or long-term hospital) or home care program. With the increasing role of hospitalists and physicians who specialize in SNF care, medical care for geriatric patients has become fragmented at the time transitions, creating opportunities for communication problems and medical errors. Changes in reimbursement and financial penalties for high rates of hospital readmissions have driven the development of many care transition interventions (Table 464-1). These interventions involve inter-professional collaboration and a variety of strategies targeted at making care transitions safer, and reducing unnecessary return visits to the emergency department, hospital readmissions, and related complications and costs.

Inter-Professional Teams and Co-Managed Care The complexity of caring for the aging population is more evident during a hospitalization due to a new acute illness or exacerbation of pre-existing chronic conditions. Inter-professional teams integrate different areas of expertise with the aim of providing patient-centered care. Physicians should understand and respect the roles of nurses, physical, occupational, and speech therapists, nutritionists, pharmacists, psychologists, social workers, clergy, and other direct care staff. The evolution of inter-professional teams has resulted in a comprehensive approach to care by opening channels of communication between these health professionals from different disciplines.

TABLE 464-1 Examples Care Transitions Interventions

INTERVENTION	WEBSITE	CORE INTERVENTIONS
Re-Engineered Discharge (Project RED) (Jack et al: 2009)	*https://www.bu.edu/fammed/projectred/*	"Discharge advocate" performs the following: · Facilitates patient education and understanding · Performs medication reconciliation · Coordinates post-discharge appointments and communication with primary care provider (PCP) · Calls patient 2–3 days post-discharge
Transitional Care Model (Naylor et al: 2004; Naylor et al: 1999)	*https://www.nursing.upenn.edu/ncth/transitional-care-model/*	Advanced practice nurse performs the following: · Coordinates patient care pre- and post-discharge · Assesses each patient's needs; engages and activates the patient and family · Facilitates communication among patient, family, and health care providers · Conducts regular home visits and telephone support after discharge
Care Transitions Program® (Coleman et al: 2004)	*http://www.caretransitions.org*	"Transition Coach" performs the following: · Facilitates improved self-management skills including medication management and how to respond to warning signs/symptoms. · Makes post-discharge home visits and phone calls.
Better Outcomes for Older Adults through Safe Transitions (BOOST) (Hansen et al: 2013)	*http://www.hospitalmedicine.org/Web/Quality_Innovation/Implementation_Toolkits/Project_BOOST/Web/Quality___Innovation/Implementation_Toolkit/Boost/Overview.aspx?hkey=09496d80-8dae-4790-af72-efed8c3e3161*	Includes toolkit facilitating the following: · Comprehensive identification and assessment of high-risk patients · Patient/caregiver education · Enhanced communication with post-hospitalization care providers · Follow-up phone call with patient post-discharge
Interventions to Reduce Acute Care Transfers (INTERACT) (Ouslander et al: 2013)	*https://interact.fau.edu*	Includes tools for skilled nursing, assisted living and home health care including: · Quality Improvement · Communication · Decision support · Advance care planning

"Huddles" are a mechanism of enhanced communication for inter-professional teams. The implementation of efficient huddles has been associated with improved safety and better utilization of resources by predicting patient needs and making appropriate changes in staffing and care plans. Huddles can also help identify potential threats to patient care, such as socioeconomic challenges that can make care plans ineffective of even harmful.

Another strategy for enhanced communication and collaboration in the care of complex geriatric patients is "Co-Managed Medicine." In this model, internists serve as part of a multispecialty team of physicians (that often include surgeons) that provides daily assessments, addresses medical comorbidities, and facilitates transitions of care; thereby enhancing the typical consultant model. Co-managed medicine is another example of how enhanced communication between different providers improves outcomes, avoids common complications, and saves resources. In the era of person-centered care and value-based medicine, effective co-managed medicine appears to deliver consistently high quality care at a lower cost. Since the rise of hospitalist-based care, the use of co-managed care has increased significantly. Hip fracture co-management, as well as trauma co-management, and collaborations between internists and geriatricians are examples of this strategy.

FUNDAMENTALS OF GERIATRIC CARE

Person-Centered Care Person-centered care is a critical concept in caring for older people because of the complexity of their medical, functional, and psychosocial problems, and in many instances the lack of rigorous data on the most effective strategies for caring for specific conditions in patients with multi-morbidity. Thus, decision-making on goals and approaches to care must account for patient and family preferences and goals, values, perception of risk, prognosis, and other individual factors. For almost any condition, from common disorders such as hypertension and diabetes, to geriatric syndromes such as fall risk and urinary incontinence, the answer to how best to treat medical conditions in an older patient with multi-morbidity does not only depend on evidence-based medicine—it also depends on careful weighing of the factors listed above. In everyday practice with complex older patients, a focus on improving or maintaining function and independence, quality of life, comfort, and dignity will be consistent with patient and family goals.

The American Geriatrics Society (AGS) identifies the following elements as key to person-centered care: (1) an individualized, goal-oriented care plan based on the person's preferences; (2) ongoing review of the person's goals and care plan; (3) continual information sharing and integrated communication; (4) education and training for providers and, when appropriate, the person and those important to the person; and (5) performance measurement and quality improvement using feedback from the person and caregivers. Several tools are available to assist in implementing person-centered care, including estimation of prognosis (e.g., "ePrognosis"), and "choosing wisely" recommendations from the AGS and AMDA—The Society for Post-Acute and Long-Term Care Medicine. Examples of these recommendations that are relevant to internal medicine practice are illustrated in Table 464-2.

TABLE 464-2 Examples of Choosing Wisely Recommendations Helpful in Implementing Person-Centered Care in Complex Geriatric Patients

- Don't recommend percutaneous feeding tubes in patients with advanced dementia; instead offer oral assisted feeding.
- Don't use antipsychotics as the first choice to treat behavioral and psychological symptoms of dementia.
- Avoid using medications other than metformin to achieve hemoglobin A1c <7.5% in most older adults; moderate control is generally better.
- Don't use benzodiazepines or other sedative-hypnotics in older adults as first choice for insomnia, agitation or delirium.
- Don't use antimicrobials to treat bacteriuria in older adults unless specific urinary tract symptoms are present.
- Don't prescribe cholinesterase inhibitors for dementia without periodic assessment for perceived cognitive benefits and adverse gastrointestinal effects.
- Don't recommend screening for breast, colorectal, prostate or lung cancer without considering life expectancy and the risks of testing, over-diagnosis and overtreatment.
- Don't routinely prescribe lipid-lowering medications in individuals with a limited life expectancy.
- Don't obtain a *Clostridium difficile* toxin test to confirm "cure" if symptoms have resolved.
- Don't recommend aggressive or hospital-level care for a frail elder without a clear understanding of the individual's goals of care and the possible benefits and burdens.

Adapted from *http://www.choosingwisely.org/societies/american-geriatrics-society/* and *http://www.choosingwisely.org/societies/amda-the-society-for-post-acute-and-long-term-care-medicine/amda-choosing-wisely-list/*. Accessed December 31, 2016.

Evaluation of the Geriatric Patient • GERIATRIC ASSESSMENT A series of screening questions can be useful as a "geriatric review of systems" in clinical practice with older patients because of the importance and high prevalence of functional impairments and disabilities, limited social support to assist with functional limitations, cognitive and affective disorders, and geriatric conditions that may go undetected and cause patient safety issues and complications (Table 464-3). These questions may be especially helpful in conducting annual Medicare "Wellness Visits." Positive responses to one or more of the screening questions for each item should prompt consideration of further assessments, many of which can be accomplished using standard and validated tools available on the internet, such as activities of daily living scales, depression scales, sleep questionnaires, and mental status examinations.

EVALUATION OF MEDICAL DECISION-MAKING CAPACITY Key aspects of decision-making in older adults are illustrated in Fig. 464-4. Including the patient in the consent process for any treatment is the foundation of patient autonomy and person-centered care. Because aging is associated with an increasing potential to develop cognitive impairment, determination of decision-making capacity is important not only to protect the patients against potential abuse, but also to preserve autonomy when possible and when it is not, that an appropriate surrogate decision-making process is followed. Assessing for capacity is usually triggered by specific circumstances (e.g., the need for invasive diagnostic testing or surgery). Determination of decision-making capacity limited to medical circumstances should be differentiated from declaring a patient "incompetent" to make all decisions. Declaring someone incompetent is a legal definition and usually is reserved for court settings. Another caveat about evaluating decision-making capacity is distinguishing lack of capacity from poorly presented information, sensory impairment, and/or low level of literacy. The clinician should corroborate that the patient has received all the necessary information, comprehends the information provided, and there are no major auditory or visual impairments. For geriatric patients, it is important to determine if the patient uses hearing aids of prescription glasses and they are available for their use.

Standard tests of cognitive function such as the Mini Mental State Examination correlate poorly with capacity to consent for specific interventions. Several standardized tools have been validated to determine decision-making capacity. The MacArthur Competence Assessment Tool-Treatment (MacCAT-T) is a structured tool that has been validated, but it is lengthy and can be difficult to administer in some patients. The Capacity to Consent to Treatment Instrument (CCTI) is another tool that has been validated in patients with mild to moderate Alzheimer's disease. It is structured in two different vignettes and the patient is asked to answer a series of questions. The test has high inter-rater reliability and validity.

EVALUATION OF THE OLDER DRIVER For many older adults in the United States, driving is essential for maintaining independence and driving cessation is associated with negative outcomes including social isolation and depression. On the other hand, older adults have the highest risk of being involved in fatal crashes with up to a nine-time higher risk for those 85 years old compared to younger people. Older people should be routinely assessed for their driving status and if they have been in

TABLE 464-3 Examples of Screening Questions and Tools and Strategies for Further Evaluation of Social Support, Functional Status, Geriatric Syndromes, and Cognition and Affect

GERIATRIC ASSESSMENT DOMAINS		RECOMMENDED SCREENS	FURTHER ASSESSMENT FOR POSITIVE SCREEN*
SOCIAL	**Social Support**	Do you live alone? Do you have a caregiver? Are you a caregiver?	• Consider referral to a social worker • Refer to Area Agency on Aging
	Elder Neglect/Abuse	Do you ever feel unsafe where you live? Has anyone ever threatened or hurt you? Has anyone been taking your money without your permission?	• Consider referral to a social worker and/or Adult protective services
	Advance Directives	Would you like information or forms for a power of attorney for health care? Would you like information on a living will?	• Discussion on advance directives • Physician Orders for Life-Sustaining Treatment (POLST) (or MOLST or POST)
FUNCTIONAL	**Functional Status**	Do you need assistance with shopping or finances? Do you need assistance with bathing or taking a shower?	• Instrumental Activities of Daily Living (ADL) Scale • Basic ADL Scale
	Driving	Do you still drive? If yes: While driving, have you had an accident in the past 6 months? Driving concerns by family member?	• Vision testing • Consider occupational therapy and/or formal driving evaluation
	Vision	Do you have trouble seeing, reading, or watching TV? (with glasses, if used)	• Vision testing • Consider referral for eye exam
	Hearing	Do you have difficulty hearing conversation in a quiet room? Unable to hear whisper test 6 inches away?	• Check for cerumen in ear canals and remove if impacted • Hearing Handicap Inventory • Consider audiology referral
GERIATRIC SYNDROMES	**Medications**	Do you take 5 or more routine medications? Do you understand the reason for each of your medications?	• Match medications with diagnoses • Consider reducing doses, stopping drugs, adherence aides, and/or consultation with a pharmacist
	Fall Risk	Have you fallen in the past year? Are you afraid of falling? Do you have trouble climbing stairs or rising from chairs?	• "Get Up and Go" test • Consider full Fall Assessment • Consider Physical Therapy Evaluation • Consider Home Safety Assessment
	Continence	Do you have any trouble with your bladder? Do you lose urine or stool when you do not want to? Do you wear pads or adult diapers?	• Consider full continence Assessment • 3 IQ Questionnaire (women) • American Urological Association (AUA) 7 symptom inventory (men)
	Weight Loss	Weight <100 pounds or Unintentional weight loss >10 pounds over 6 months?	• Assess for common risk factors for malnutrition • Consider referral to dietician for nutritional evaluation
	Sleep	Do you often feel sleepy during the day? Do you have difficulty falling asleep at night?	• Epworth Sleepiness Scale or Pittsburgh Sleep Index • Consider referral for sleep evaluation
	Pain	Are you experiencing pain or discomfort?	• Pain Assessment
	Alcohol Abuse	Do you drink > 2 drinks / day?	• AUDIT-C
COGNITION AND AFFECT	**Depression**	Do you often feel sad or depressed? Have you lost pleasure in doing things over the past few months?	• PHQ – 9 or Geriatric Depression Scale • Screen for suicide risk
	Cognition	Self-reported memory loss? Cognitive screen positive? (3-item recall and Clock Draw test "Mini-Cog") Confusion Assessment (CAM) for delirium	• Montreal Cognitive Assessment or Mini Mental State Examination • If diagnosis is unclear, consider neuropsychological testing

Source: Adapted from RL Kane et al: *Essentials of Clinical Geriatrics*, 8th ed. New York, McGraw-Hill, 2018.

any car crashes, as well as for sensory, functional, and cognitive impairments that can make driving unsafe (Table 464-3). Like many geriatric conditions described below, many different types of drugs can impair various aspects of driving performance, and should be carefully considered in older people who continue to drive **(Table 464-4)**.

Suspected driving impairment can be a source of conflict between the patient (who wants to maintain independence), the family (who may want their relative to continue driving due to lack of other transportation; or may be concerned about their safety, or both) and the physician (who is concerned about the patient's, passengers', and other drivers' safety). There is liability involved in these decisions, since any states do not require driving re-testing for all older drivers, and many require physicians to report older people who they believe are unsafe drivers. Evaluation of driving should be inter-professional and aimed to first try to correct any reversible causes of losing driving skills, such as vision and hearing impairment. Although tests of executive function such as the Trails B have been associated with poor driving performance, no single screening test predicts unsafe driving.

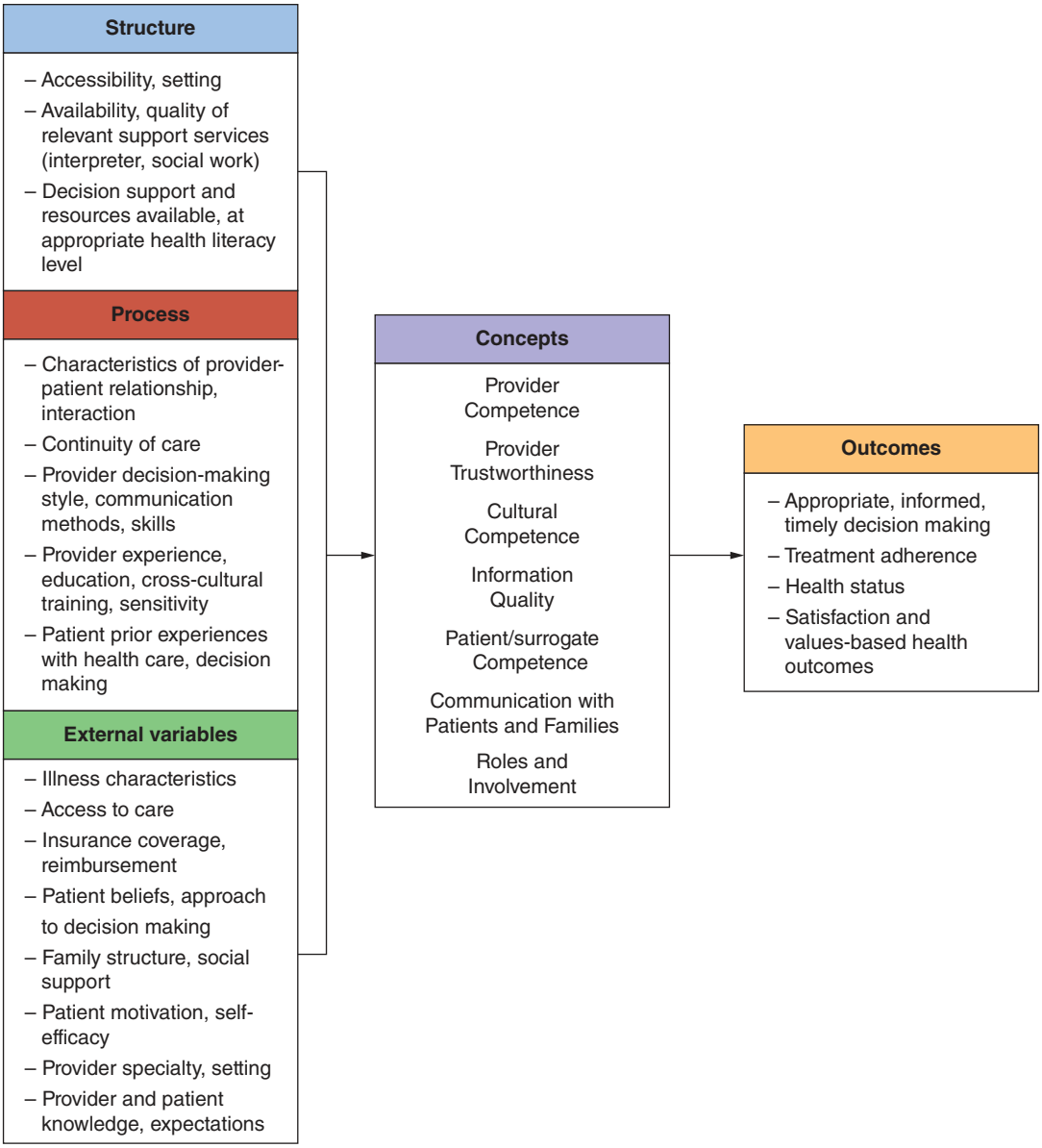


FIGURE 464-4 Key aspects of decision-making in older adults. *(From SM Dy, SP Tanjala: Key concepts relevant to quality of complex and shared decision-making in health care: A literature review. Soc Sci Med 74:582, 2012.)*

A combination of neuropsychological testing by a psychologist, and on-road testing by a trained occupational therapist can provide the physician with essential input in making the difficult decision on driving cessation. The AGS and the U.S. Department of Transportation's National Highway Traffic Safety Administration have updated a "Physician's Guide to Assessing and Counseling Older Drivers," which can be helpful to practicing clinicians and is available on the AGS website (see "Further Reading").

TABLE 464-4 Medications with Strong Potential to Affect Driving

Anticholinergics
Anticonvulsants
Antidepressants
Antiemetics
Antihistamines
Antihypertensives
Antiparkinsonian agents
Antipsychotics
Benzodiazepines and other sedatives/anxiolytics/hypnotics
Muscle relaxants
Narcotic analgesics
Stimulants
Other agents with significant anticholinergic properties

INTERPRETATION OF DIAGNOSTIC TESTS Atypical presentations of medical conditions are a common feature of geriatric medicine. Physiologic changes associated with aging can affect the results of common diagnostic tests as well. The large variation of many physiologic measures that is associated with normal aging makes establishing what is "normal" for many tests challenging. For this reason, the results of several diagnostic tests must be interpreted with caution. Examples include creatinine clearance, pulmonary function, and sedimentation rate (which can confound the diagnosis of polymyalgia). Ambulatory cardiac monitoring may identify a variety of arrhythmias, but they have a high incidence in older people and must be linked with symptoms before considering potentially toxic or invasive treatment. Musculoskeletal imaging, such as an MRI of the spine, may reveal multiple abnormalities that may or may not be related to symptoms. For the most part, however, abnormal diagnostic tests require further evaluation in older patients, unless further evaluation would not lead to a change in the goals of care and treatment plan. Examples include low hemoglobin levels, abnormal

thyroid function tests, age/sex/weight adjusted creatinine clearance, and elevated liver function tests; these examples do not result from normal aging and generally indicate a physiologic abnormality resulting from a disease or disorder that may or may not be reversible.

Prevention in Older Adults • AGE-APPROPRIATE SCREENING

Screening tests for specific diseases, as opposed to screening for geriatric conditions requires a careful person-centered approach. The focus of preventive medicine depends heavily on the ability to identify those who are at risk for specific conditions **(see Chap. 4)**. Several professional societies have provided guidance regarding specific tests in older adults **(Table 464-5)**. An important caveat about screening to prevent disease in older patients (e.g., colonoscopy for colon cancer; PAP smears; PSA testing) is that abnormal results may lead to subsequent testing and treatment among individuals who will not suffer morbidity or mortality from the disease because of limited life expectancy. Thus, geriatric patients pose a significant challenge for deciding what screening tests could offer a reasonable ratio of benefit and risk as well as being cost-effective. The *ePrognosis.com* website is a very helpful tool in these determinations.

VACCINATIONS The use of vaccines in older adults is aimed at creating immunity against common infections that could lead to serious complications, as well as for rebuilding previously obtained immunity. Currently, the CDC recommends routine vaccination against influenza, pneumococcus and shingles as they are prevalent in this age group. Other countries in Europe and Asia have similar trends on vaccinations with small variances.

SEXUALLY TRANSMITTED DISEASES (STDS) Although most STDs occur in younger people **(see Chap. 131)**, a portion of older adults have high-risk sexual behavior. Most Americans remain sexually active in their sixties and seventies, and up to a quarter of individuals in their eighties consider themselves sexually active. Sexually active older adults may have a lower awareness of the need for safe sexual practices, such as the risks of multiple sexual partners and condom use. The incidence of STDs in older people is still relatively low. Individuals born in the United States between 1945 and 1965 are at higher risk of having hepatitis C due to lack of awareness of the disease and lack of implementation of universal precautions before the 1980s for blood transfusions. Other factors that could affect such risk are use of intravenous drugs and unprotected sex with multiple partners. The prevalence of tertiary syphilis is higher than newly contracted syphilis in older adults. The incidence of gonococcal infection decreases with age. Nonetheless, patients presenting with symptoms compatible with syphilis or gonococcal infection (cervicitis, urethritis, proctitis, epididymitis) should be screened for high-risk sexual behavior and educated if necessary. Clinical symptoms of herpes simplex infection and the possibility of becoming contagious also decrease with age. As ulcerative lesions are less frequent, HSV-2 specific serologic testing should be considered for patients with recurrent nonspecific genital symptoms. Therapy should not be started unless the patients are symptomatic.

Almost 3 million patients aged >50 live with HIV. Since the introduction of HAART, life expectancy of patients with HIV has increased, resulting in a significant increase in the number of older adults living with the disease. De novo infections have also contributed to the rising number of HIV cases in older adults. The low rate of condom use and lack of knowledge of the disease play a key role in the transmission rate. Age is an independent predictor of HIV progression and associated mortality. There are no specific age-specific guidelines for treating HIV. Like all other conditions, there is a higher incidence of medication-related side effects in older patients, especially those with other comorbidities and on multiple other medications, and this should be considered in treatment decisions.

Treatment of Common Diseases in the Geriatric Population • HYPERTENSION

There have been several clinical trials demonstrating the benefits of hypertension treatment on the reduction of risk of cardiovascular events in older people. Nonetheless, blood pressure targets remain controversial. The balance between the cardiovascular protective benefits versus the risk of treatment-related adverse events must be considered in individual patients based on their comorbidities. For example, hypotension and postural hypotension related to antihypertensive therapy are common causes of near-syncope and falls and related injuries in the geriatric population, especially those with multi-morbidity. On the other hand, control of systolic blood pressure, in addition to preventing cardiovascular events, may reduce the burden of white matter changes in the brain, which are associated with gait abnormalities and falls and cognitive decline. However, no studies in older patients with multi-morbidity to date have documented any beneficial effects of tight control of hypertension on the incidence of falls and cognitive decline.

Two large studies (HYVET and SPRINT) have shed some light on these issues. HYVET was a multicenter study conducted in several countries involving ~3,800 patients ≥80 years old. The study demonstrated that active treatment of hypertension with a target of ≤150 mmHg not only significantly reduced the risk of stroke and heart failure, but also the mortality risk. As with other large hypertension studies like ALLHAT, there was a linear association between blood pressure and stroke reduction. Nonetheless, in the HYVET study, this association was less prominent as age increased. SPRINT was another large randomized trial targeting lowering systolic blood pressure to targets of <140 vs 120 mmHg (measured with an automated device) with a subgroup analysis in those aged ≥75 years. There were significant reductions in the primary end-point—a composite of cardiovascular disease events (including myocardial infarction, acute coronary syndrome, heart failure, stroke, or death from cardiovascular causes). However, it is critical to recognize that patients with diabetes, history of stroke or heart failure, and systolic blood pressure <110 mmHg after 1 min of standing were excluded from the SPRINT trial.

Overall, these data strongly suggest a person-centered approach to hypertension in the heterogeneous older population. For older patients with minimal comorbidity, no postural hypotension, and low risk of falls and volume depletion, the benefit/risk ratio favors lower targets for systolic blood pressure (<130 mmHg measured by a hand sphygmomanometer). However, for those with diabetes, heart failure, history of stroke, postural hypotension, careful treatment of blood pressure with higher systolic targets (< 150 mmHg) is probably a safer approach.

DIABETES The prevalence of diabetes in the older adult population is now over 25% and expected to increase due to adverse lifestyle changes and an increased incidence of obesity. Due to a lack of data on patients with multi-morbidity and those aged ≥80, and the high incidence of hypoglycemia in this population when treated with multiple hypoglycemic agents, the approach to management of diabetes requires a person-centered approach like that described for hypertension. Older diabetic patients are at significant risk of hypoglycemia because of potential medication errors, progressive decline in renal function, and inconsistent oral intake among other reasons. Hypoglycemic episodes are associated with progressive cognitive decline in older adults, especially those with existing cognitive impairment. On the other hand, uncontrolled diabetes is associated with an increased risk of all-cause dementia.

Data from randomized clinical trials that have largely excluded those aged ≥80 suggest that intensive glycemic control does not reduce major macrovascular events in older adults for at least 10 years or result in improved microvascular outcomes for at least 8 years, and at the same time increases the risk of severe hypoglycemia 1.5- to 3-fold. Thus, the AGS guideline on diabetes in older adults and the Choosing Wisely recommendations (Table 464-2) suggest that in most adults >65 years, the harms associated with a hemoglobin A1c (HbA1c) target <7.5% or >9% are likely to outweigh the benefits. These recommendations are consistent with the American Diabetes Association. Thus, the goals of treating diabetes in the geriatric population should be tailored to the patient's functional and medical status, social support, personal goals, perception of risk, and life expectancy. For specifics of treatment options, refer to **Table 464-6**. Regardless of the therapeutic goals for HbA1c, older diabetic patients should be regularly examined for the development of neuropathy, which can lead to the development of

TABLE 464-5 Recommendations for Primary Prevention Screening for Specific Diseases in Older Adults from Different Professional Societies

TYPE OF SCREENING	TEST	FREQUENCY	PROFESSIONAL SOCIETY ISSUING RECOMMENDATIONS		
			USPSTF[a]	ACS[b]	ACP[c]
Colorectal	Fecal occult blood test or fecal immunochemical test (FIT) or Sigmoidoscopy or Colonoscopy	Annual Every 5 y Every 10 y	Screen all adults age 50–75; prognosis may support screening individuals of age 76–85; not recommended for adults over age 85	Screen all adults age >50; discontinuing screening is reasonable in people with severe comorbidity that would preclude treatment	Screen all adults age 50–75 People with life expectancy <10 y should not be screened
Breast	Mammography	Every 1–2 y	Biennial screening all women age 50–74; evidence of benefits and harms is insufficient for women age >75	Annual screening starting at age 40; continue while in good health	ECOG[d] Annual screening starting at 40 y/o
Cervical	Pap smear HPV test	Pap only, every 3 y HPV + Pap, every 5 y	Screen women age 21–65; discontinue at age 65 if adequate prior screening	Screen women age 21–65; discontinue at age 65 if regular screening normal	Screening should stop at age 65 if evidence of negative adequate prior screening.
Lung	Low-dose CT scan	Annual	Screen age 55–80 current and former smokers with a 30+ pack-year smoking history; discontinue screening once a person has not smoked for 15 years or develops a health problem that limits their ability or willingness to have curative surgery	Screen 55–74-year-old current and former smokers in good health with a 30+ pack-year smoking history	ACCP[e] In settings that can deliver the comprehensive care provided to National Lung Screening Trial participants, offer screening to people age 55–74 current and former smokers with 30+ pack-year smoking history
Prostate	Prostate Specific Antigen (PSA)	1–2 y	Do not screen men for prostate cancer with PSA test	Screen men aged ≥50 with a life expectancy >10 y after discussion about the risks, benefits, and uncertainties of PSA screening. Follow-up screening should occur annually if PSA >2.5 ng/mL or biennially if PSA <2.5 ng/mL	AUA[f] Biennial PSA screening in men of 55–69 y with life expectancy >10–15 y, after shared decision-making discussions accounting for values and preferences
Osteoporosis	Dual-energy x-ray absorptiometry (DEXA) Measure height, preferably with a wall mounted stadiometer	Perform BMD testing 1 to 2 years after initiating medical therapy for osteoporosis and every 2 years annually thereafter	AAFP[g] Screening women age 65 or men age 70	NOF[h] Screen women age 65 and older and men age 70 and older; postmenopausal women and men age 50–69, based on risk factor profile; postmenopausal women and men age 50 and older who have had an adult age fracture	
Carotid Disease	Carotid ultrasound	Once	Society of Vascular Surgery Age >65, coronary artery disease, need for coronary bypass, symptomatic lower extremity arterial occlusive disease, history of tobacco use and high cholesterol would be appropriate risk factors to prompt ultrasound in patients with a bruit		
Coronary Artery Disease	Coronary Calcium Score (CCS)	Once	SCCT[i] Do not use CCS for patients with known CAD.	AHA/ACC[j] CCS of 0 may have a strong negative predictive value for coronary events in older adults.	
Abdominal Aortic Aneurysm	Abdominal Ultrasound	Once	USPSTF[a] Insufficient evidence to assess the balance of benefits and harms of screening women aged 65–75 years who have ever smoked.	AAFP[g] Recommended for men aged 65–75 years who have ever smoked.	
Diabetes	Fasting blood glucose, glucose tolerance test, or HbA1C	Annually	USPSTF[a] Abnormal blood glucose as part of cardiovascular risk assessment in adults aged 40–70 years who are overweight or obese	ADA[k] Screening people 45 years and older	

[a]United States Prevention Service Task Force; [b]American Cancer Society; [c]American College of Physicians; [d]Eastern Cooperative Oncology Group; [e]American College of Chest Physicians; [f]American Urology Association; [g]American Academy of Family Physicians; [h]National Osteoporosis Foundation; [i]Society of Computed Tomography; [j]American Heart Association/American College of Cardiology; [k]American Diabetes Association.

TABLE 464-6 Recommendations and Considerations for Pharmacologic Therapy of Diabetes in Older Adults[a]

MEDICATION	RECOMMENDATIONS AND CONSIDERATIONS
Metformin	• Metformin is the first-line agent for older adults with type 2 diabetes • Recent studies suggest it may be used safely in patients with estimated glomerular filtration rate ≥30 mL/min/1.73 m^2 • Contraindicated in patients with advanced renal insufficiency or significant heart failure • May be temporarily discontinued before procedures or during hospitalizations and when acute illness compromises renal or liver function
Sodium–Glucose Cotransporter 2 Inhibitors	• Offer an oral route, which may be convenient for older adults • Long-term experience is limited despite the initial efficacy and safety data
Thiazolidinediones	• If used at all, should be used very cautiously in those with, or at risk for, congestive heart failure and those at risk for falls or fractures
Secretagogues	• Associated with hypoglycemia and should be used with caution • Shorter-duration sulfonylureas such as glipizide are preferred • Glyburide is a longer-duration and contraindicated in older adults
Incretin-Based Therapies	• Few side effects and minimal hypoglycemia, but costs may be a barrier • No evidence of increase in major adverse cardiovascular events • Glucagon-like peptide 1 receptor agonists are injectable, which require visual, motor, and cognitive skills • Associated with nausea, vomiting, diarrhea, and weight loss, which may not be desirable in some older patients, particularly those with cachexia
Insulin Therapy	• Requires that patients or their caregivers have good visual and motor skills and cognitive ability • Insulin doses should be titrated to meet individualized glycemic targets and to avoid hypoglycemia • Once-daily basal insulin injection therapy is associated with minimal side effects and may be a reasonable option in many older patients • Multiple daily injections of insulin may be too complex for the older patient with advanced diabetes complications, life-limiting comorbid illnesses, or limited functional status

[a]Based on recommendations from the American Diabetes Association 2017.

lesions on the feet that could become infected, as well as for retinopathy and vision loss that may require ophthalmologic intervention.

HYPERLIPIDEMIA While good evidence exists regarding the benefits of statins on primary cardiovascular risk prevention in patients ≤75 years old, for those aged >75 the data are very limited. The use of statins in those aged >75 or 80 for prevention of cardiovascular events and mortality is the subject of ongoing debate in the geriatric literature. No evidence from randomized controlled trials exists to guide statin initiation after age 80 years; treatment of hypercholesterolemia for patients at risk of atherosclerotic cardiovascular disease should start before they turn 80 years old. There are two other factors that make the use of statins in older adults controversial. First, the major benefits of statins have been demonstrated over long-term use; thus, life expectancy is a limiting factor to observe any meaningful change in outcomes. A substantial proportion of patients are maintained on statins at the end of life, but they can be safely discontinued. On the other hand, statins are well tolerated in older adults especially at moderate to low doses. Although many older adults on statins complain of muscle pain, the risk of myositis and rhabdomyolysis is increased mostly with the use of high doses; adverse effects of statins on cognitive function appear to be uncommon. Thus, some relatively healthy adults aged >75 years with life expectancy of >10 years may benefit, and the approach to hyperlipidemia should be person-centered in this population, as discussed for both hypertension and diabetes.

OSTEOARTHRITIS (OA) The approach to the management of symptomatic OA in the geriatric population differs from the approach in younger patients **(see Chaps. 363 and 364)** because of the risks of toxicity of nonsteroidal anti-inflammatory drugs (NSAIDs) in older patients. Nonpharmacologic interventions should be the first line of treatment. While some patients aged >65 years can tolerate NSAID use with concomitant protection from gastrointestinal (GI) bleeding with a proton pump inhibitor (PPI), this regimen exposes patients to two drugs with numerous potential adverse drug effects. NSAIDs are well known to be associated with GI bleeding; they are also associated with worsening renal function based on multiple potential mechanisms, and with sodium and fluid retention and exacerbation of hypertension and congestive heart failure. In addition, a substantial number of older patients are on anticoagulants or platelet aggregation inhibitors, which could further increase the risk of bleeding from NSAIDs. PPIs are associated with a higher incidence of pneumonia, osteoporosis, and *C. difficile*-associated diarrhea; and possibly with a higher risk of dementia.

Topical NSAIDs are better tolerated; lidocaine patches and nonprescription creams may also be effective. The AGS guideline on the management of chronic pain recommends that routine acetaminophen in doses up to 1 g four times daily should be the basis of pharmacologic treatment. Failure to respond could be followed up with careful trials of tramadol or a narcotic agent (started in a short-acting preparation) with appropriate attention to avoiding narcotic-induced constipation. Although prescription of narcotics is getting increasingly cumbersome because of high rates of abuse, this should not deter prescription of these agents to relieve pain and disability in older patients.

Many older patients respond well to a variety of non-pharmacologic interventions, including stretching, strengthening, timely and appropriate use of heat and ice, massage, swimming and whirlpool therapy, bracing, acupuncture, and therapeutic electrical stimulation. These interventions are best carried out under the supervision of physical therapists or other professionals with appropriate expertise to avoid injury. Surgical interventions, including replacement of major joints, have improved over the last several years, and even older patients with multi-morbidity may experience improved function and quality of life. Total knee replacement, for example, has been shown to be effective in generally healthy older patients, and should be considered in selected higher risk patients. "Pre-habilitation" with targeted strengthening and endurance exercises, and willingness to go through several weeks of post-operative physical therapy should be prerequisites for referring older patients for joint replacement.

CANCER More than half of new cases of cancer and mortality associated with it occur after the age of 65. There are limited data regarding older adults with multiple comorbid conditions and their response to cancer treatment. While only ~10% of clinical trials have had age-stratification analyses, the available evidence suggests that age alone is not a predictor of harm. Nonetheless, making treatment decisions is challenging due to both shorter life expectancy in older adults and the cumulative effect of multiple comorbidities. Thus, a person-centered approach is essential.

Older adults generally experience decreases in functional status after receiving chemotherapy. Most of this negative effect appears to be related to comorbidity and baseline functional status, rather than due

to age alone. For this reason, specialists in geriatric oncology have proposed using comprehensive geriatric assessment, including many of the issues addressed in Table 464-3 as a strategy to better predict which older adults will tolerate and benefit most from cancer treatment. Other considerations before making decisions about treatment plans should include socioeconomic factors. Lack of social support has been associated with poor outcomes after radiation and chemotherapy, especially in older women. Other important issues in cancer treatment planning include availability of transportation for treatments, economic and insurance status, the patient's ability to follow treatment plans, and family and social support available during therapy, when adverse effects and functional decline may occur.

GERIATRIC SYNDROMES AND CONDITIONS

In this section selected geriatric syndromes and conditions likely to be encountered by internists in hospital, clinic, office, PAC and LTC settings are discussed.

Falls • **EPIDEMIOLOGY AND IMPACT** Among all geriatric syndromes, falls are probably the most common that internists will encounter. Falls are responsible for potentially devastating consequences for function and quality of life, as well as mortality, in the geriatric population. About one in three older community-dwelling, and one in two older LTC residents fall annually; many more are at risk for falls. The impacts of falls include fear of falling with adverse effects on quality of life, painful injuries including hip and wrist fractures, subdural hematomas, and death. Falls are associated with loss of function and death within the year after a fall. For these reasons, internists should regularly screen older people for falling using questions such as: "Have you fallen in the past year?" "Are you afraid of falling?" "Do you have trouble climbing stairs or rising from chairs?" (Table 464-3).

EVALUATION The risks and causes of falls are multifactorial. Most older people at risk for a fall or who have suffered a fall have more than one potential underlying risk factor or cause. Many falls are labeled as "mechanical" and attributed to simply tripping or slipping. It is essential to recognize, however, that older people who trip or slip may have a variety of underlying reversible conditions that could have contributed to the event. Thus, a thorough evaluation of all falls is warranted. In addition to evaluating the patient who has fallen for injury, it is critical to determine, to the extent possible, whether the patient had a syncopal episode or a seizure, which dictate a very different approach to evaluation and management. As many as half of "unexplained falls" in older people with dementia (e.g., found on the floor) may be due to near-syncope or syncope related to postural hypotension.

Figure 464-5 illustrates an overview of the approach to an older person who reports a history of one or more fall in the past six months, and Table 464-7 provides more detail on the immediate evaluation of an older person who has fallen. **Chapter 23 provides more detail on the evaluation of gait and balance disorders.**

MANAGEMENT Table 464-8 illustrates approaches to the management of falls. Immediately after a fall injuries and underlying acute illnesses should be identified and treated. It is common practice for older patients who come to an Emergency Department with a history of a fall to have a brain imaging study. While this is understandable from a potential liability standpoint, it is also reasonable to avoid such studies if there is no history or signs of head trauma, neurologic symptoms or signs, or anticoagulation, and monitor the patient carefully over the

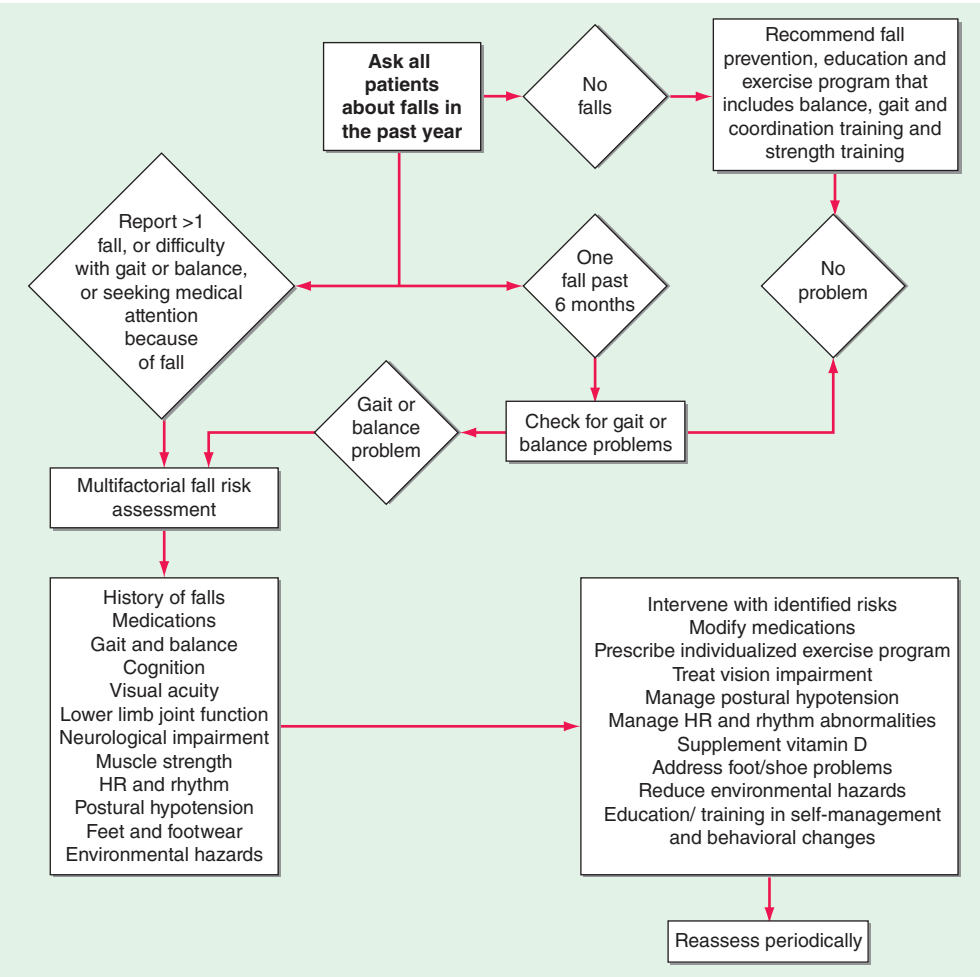


FIGURE 464-5 Algorithm depicting assessment and management of falls in older patients. HR, heart rate. *(From American Geriatrics Society and British Geriatrics Society: Clinical Practice Guideline for the Prevention of Falls in Older Persons. New York, American Geriatric Society, 2010.)*

TABLE 464-7 Evaluating the Older Person Who Falls—Immediate Post-Fall Evaluation

History

- Circumstances surrounding the fall
  - Relationship to changes in posture, turning of head, after a meal or medication intake, rushing to the toilet, nocturia, straining to urinate or defecate
  - Accidental trip or slip (note that many correctable factors can contribute to a reported "mechanical" fall – see text)
  - Hazards in the living environment (loose rugs; cords; unsafe steps; slippery floors; etc.)
- Premonitory or associated symptoms
  - Dizziness (lightheadedness vs vertigo); cardiovascular (postural lightheadedness, palpitations, chest pain, shortness of breath); focal neurological symptoms suggestive of stroke or transient ischemic attack (weakness, sensory disturbance, dysarthria, ataxia, aphasia); symptoms of a seizure (witnessed clinic movements; incontinence of urine or stool; tongue biting)
  - Symptoms over the previous few days that may have led to volume depletion (poor food/fluid intake; nausea/vomiting; diarrhea; urinary frequency/polyuria)
- Exclude loss of consciousness or seizure (may be difficult without a witness)
- Medications – chronic and within the few hours before the fall
  - Diuretics and other antihypertensive drugs
  - Nitrates
  - Drugs that cause bradycardia – beta blockers; cholinesterase inhibitors
  - Psychotropic - antipsychotics; hypnotics; sedatives; antidepressants
  - Antiparkinsonian
  - Hypoglycemic drugs
- Excessive alcohol intake

Physical Examination

- Exclude physical injury
  - Head trauma; hip range of motion; pubic bone tenderness; wrist pain; other signs of trauma
  - Bruising in patients on anticoagulants/platelet inhibitors
- Exclude acute illness
  - Vital signs
  - Postural vital signs (if feasible/safe)
  - Finger-stick glucose in diabetics
  - Poor skin turgor suggesting volume depletion (over chest; other areas unreliable)
  - Signs of an acute respiratory, cardiovascular, abdominal conditions
  - Focal neurological signs suggestive of stroke
- Signs of conditions that increase risk for falls
  - Poor visual acuity; use of bifocals
  - Limited range of motion of neck (to detect possible cervical arthritis/disk disease)
  - Cardiovascular—arrhythmias; carotid bruits; aortic stenosis; mitral insufficiency; heart failure
  - Degenerative joint disease in lower extremities causing pain, limited range of motion, and/or deformity
  - Podiatric conditions (calluses; bunions; ulcerations; poorly fitted, inappropriate, or unsafe shoes)
  - Neurological signs—lower extremity muscle weakness; peripheral neuropathy; tremor, rigidity, and/or bradykinesia suggestive of undiagnosed Parkinson's disease; cerebellar signs (abnormal heel to shin or heel tapping); abnormal reflexes that could reflect upper motor neuron disorder such as spinal cord compression or subdural hematoma; cognitive deficits that can result in poor judgement
  - Observation of gait and balance – simple get up and go test (see text) with observation for short steps, poor foot elevation, wide-based gait, multiple steps to turn 180 degrees; other abnormalities that might suggest normal pressure hydrocephalus (especially in combination with symptoms of incontinence and/or cognitive impairment)

Laboratory and/or Imaging Studies

- Should be guided by history and physical examination – common examples include:
  - Complete blood count, basic metabolic panel to exclude/verify acute illness
  - Urinalysis (only when additional symptoms of urinary tract infection present)
  - Electrocardiogram (in patients suspected of acute coronary syndrome or with significant known cardiovascular disease)
  - X-rays to exclude fractures
  - Brain imaging if signs present to exclude subdural hematoma, stroke
  - Cardiac monitoring in patients with history suggestive of syncope or near-syncope
  - Electroencephalography in patients with history suggestive of seizure

next 48–72 h for the development of specific indications for a brain imaging study.

Because the causes of and risk factors for falls are often multifactorial, management commonly requires multiple interventions in the same patient. Among the most common and effective interventions are physical therapy for strengthening and balance; Tai Chi has also been shown to be effective in multiple trials. Although many older people who fall are vitamin D–deficient, the role of vitamin D replacement in preventing falls, or preventing injuries from falls when combined with interventions such as strength and balance training is not clear. The risk/benefit ratio probably favors vitamin D replacement with at least 800 IU per day, but high dose vitamin D (60,000 IU in one oral dose monthly) has been associated with an increase in risk of falls. Patients who suffer a fracture after a fall should be investigated and treated for osteoporosis. Patients at high risk for recurrent falls and injuries should be encouraged to use a fall alert system; selected patients may benefit from hip protectors.

Polypharmacy • **EPIDEMIOLOGY AND IMPACT** Polypharmacy has been defined as the prescription of multiple medications using various thresholds (generally ranging from five up to nine simultaneous drugs), and has been identified as a major challenge in the geriatric

TABLE 464-8 Examples of Management for Underlying Causes of Falls in Older Patients

CAUSES	EXAMPLES OF TREATMENT
Cardiovascular	
Arrhythmias	Antiarrhythmic medication, ablation, pacemaker (depending on nature of arrhythmia)
Aortic stenosis with syncope or near syncope	Valve surgery (trans-catheter procedure if appropriate)
Postural hypotension	Reduce or eliminate hypotensive drugs Hydration, support stockings Medication (proamitine, fludrocortisone, droxidopa) Adaptive behaviors (e.g., pausing and getting up slowly)
Hypertension	Manage carefully to avoid hypotension and near syncope; control may be important in patients with periventricular white matter changes in preventing further gait disturbance
Neurologic	
Autonomic dysfunction with postural hypotension	As above
Cervical spondylosis (with spinal cord compression)	Neck brace; physical therapy; consider surgery
Parkinson disease	Antiparkinsonian drugs
Visual impairment	Ophthalmological/optometric evaluation and specific treatment
Seizure disorder	Anticonvulsants
Normal-pressure hydrocephalus	Surgery (ventricular-peritoneal shunt)
Dementia	Supervised activities Hazard-free environment
Benign positional vertigo	Habituation exercises Anti-vertiginous medication
Others	
Foot disorders	Podiatric evaluation and treatment
Gait and balance disorders	Properly fitted shoes Physical therapy Exercise with balance training (including Tai Chi where available)
Muscle weakness, deconditioning	Lower extremity strength training
Drug overuse (eg, sedatives, alcohol, other psychotropic drugs, antihypertensive)	Elimination of drug(s) when feasible
Vitamin D deficiency	Vitamin D supplementation
Recurrent falls	Fall alert system for those who live alone; hip protectors in selected patients

Source: Adapted from RL Kane et al: *Essentials of Clinical Geriatrics*, 8th ed. New York, McGraw-Hill, 2018.

population for decades. About 40% of the U.S. population aged ≥65 take 5–9 medications, and ~20% take 10 or more. Polypharmacy is an increasingly complex challenge because of the rising prevalence of multi-morbidity, a plethora of clinical practice guidelines, proliferation of medications that can effectively treat common geriatric conditions, and rising patient and family demand for medications due in part to television advertising and information available on the internet. For example, based on several condition specific clinical practice guidelines (which do not account for multi-morbidity), an 80-year-old person with multi-morbidity including diabetes, chronic obstructive lung disease, hypertension, osteoporosis, and degenerative joint disease might be prescribed an extremely complicated non-pharmacologic regimen and over a dozen medications with the potential for multiple drug-drug and drug-disease interactions.

Polypharmacy increases the risks associated with age-related changes in the pharmacology of many drugs, and the risk of adverse drug events (ADEs). ADEs cause over 100,000 hospitalizations per year; the main culprits are warfarin and other antiplatelet agents, and insulin and other hypoglycemic agents. Other categories of drugs are also involved, including cardiovascular drugs that can cause electrolyte and volume disturbances and hypotension, falls, and syncope; central nervous system drugs associated with altered mental status and falls; and antimicrobials which cause allergic reactions, diarrhea, and other ADEs.

EVALUATION All older patients should have careful medication reconciliation at each office or clinic visit, and especially at the time of care transitions, including acute hospitalization, hospital discharge, admission to a PAC facility or home health program, and discharge from a PAC facility to home. At each transition, all medications should be considered in terms of unclear diagnosis or indication, uncertain dose or route of administration, stop date, hold parameters, lab tests needed for monitoring, dosages different than the last care setting, medication duplication, medications that should be restarted, and the potential for drug-drug and drug-disease interactions. At each clinic or office visit for community-dwelling older people, possible ADEs, effectiveness of drug therapy, and adherence should be evaluated.

MANAGEMENT Table 464-9 lists several general recommendations for geriatric prescribing that should help make drug therapy more

TABLE 464-9 General Recommendations for Geriatric Prescribing

1. Evaluate geriatric patients thoroughly to identify all conditions that could (a) benefit from drug treatment; (b) be adversely affected by drug treatment; and (c) influence the efficacy of drug treatment
2. Manage medical conditions without drugs as often as possible
3. Know the pharmacology of the drug(s) being prescribed
4. Consider how the clinical status (e.g., renal function, hydration) of each patient could influence the pharmacology of the drug(s)
5. Avoid potentially serious adverse drug-drug interactions
6. For drugs or their active metabolites eliminated predominantly by the kidney, use a formula to approximate age-related changes in renal function and adjust dosages accordingly—the Cockcroft-Gault formula (below) is probably safer as it tends to underestimate creatinine clearance

$$\text{Creatinine clearance} = \frac{(140 - \text{age}) \times \text{body weight (kg)}}{72 \times \text{serum creatinine level}} \quad (\times 0.85 \text{ for women})$$

7. If there is a question about drug dosage, start with smaller doses and increase gradually until the drug is effective or intolerable side effects are observed
8. Drug blood concentrations can be helpful in monitoring several potentially toxic drugs used in the geriatric population
9. Help to ensure adherence by:
   a. Making drug regimens and instructions as simple as possible
   b. Using the same dosage schedule for all drugs whenever feasible (e.g., once or twice per day)
   c. Timing the doses in conjunction with a daily routine
   d. Paying attention to impaired cognitive function, diminished hearing, and poor vision when instructing patients and labeling prescriptions
   e. Instructing relatives and caregivers on the drug regimen
   f. Enlisting other health professionals (e.g., home health aides, pharmacists) to help ensure compliance
   g. Making sure the older patient can get to a pharmacist (or vice versa), can afford the prescriptions, and can open the container
   h. Using aids (such as special pillboxes and drug calendars) whenever appropriate
   i. Performing careful medication adjudication and patient/family education at the time of every hospital discharge
   j. Keeping updated medication records and review them at each visit
   k. Reviewing knowledge of and adherence with drug regimens regularly
10. Monitor older patients frequently for adherence, drug effectiveness, and adverse effects, and adjust drug therapy accordingly

Source: Adapted from RL Kane et al: *Essentials of Clinical Geriatrics*, 8th ed. New York, McGraw-Hill, 2018.

effective and safer in older patients, especially those with multimorbidity. Chapter 63 also provides information on general principles of clinical pharmacology. Because these patients often see multiple specialists, the internist should serve as the "quarterback" for all prescribing to help ensure adherence and minimize the potential for ADEs. In hospital, PAC, and LTC settings clinical pharmacists can be extremely helpful in achieving these recommendations and goals.

While there may be undertreatment of certain conditions in older people (such as osteoporosis, depression, and overactive bladder), more attention is now being paid to "de-prescribing." De-prescribing must be done carefully, especially at the time care transitions, when indications for specific drugs and patient preferences may not be clear. The American Geriatrics Society's updated Beers Criteria includes a comprehensive list of drugs that may be inappropriate in older people and the rationale for this rating. The STOPP criteria are also useful in identifying drugs that should be reconsidered on older people.

Several commonly prescribed drugs should be considered for "de-prescribing" efforts, including: (1) diuretics and hypotensive agents when patients have systolic hypotension or postural hypotension that can precipitate near-syncope and falls; (2) over-reliance on antianxiety and hypnotic medications, especially benzodiazepines; (3) psychotropic and other drugs with anticholinergic activity that can cause dry mouth, constipation, and increase the long-term risk of cognitive impairment; (4) PPIs with unclear indications because of numerous reported potential ADEs including increased risk of pneumonia, osteoporosis, and dementia; (5) cholinesterase inhibitors and memantine in patients with severe cognitive impairment who have been on them for years: and (6) hypoglycemic agents in patients with multi-morbidity who should not have tightly controlled blood sugar with increased risk of hypoglycemia; and (7) statins in patients with severe chronic illness who are near the end of life.

Careful de-prescribing is a critical aspect of person-centered care in the geriatric population. Several general principles, including some in Table 464-9, may assist with de-prescribing efforts, including: (1) ascertain all drugs the patient is currently taking and the reasons for each one; (2) consider overall risk of drug-induced harm in individual patients in determining the required intensity of deprescribing intervention; (3) assess each drug as to its current or future benefit potential compared with current or future harm or burden potential; (4) prioritize drugs for discontinuation that have the lowest benefit-harm ratio and lowest likelihood of adverse withdrawal reactions or disease rebound syndromes; and (5) implement a discontinuation regimen based on the pharmacology of the drug being discontinued, and monitor patients closely for improvement in outcomes or onset of adverse effects.

Cognitive Impairment—Delirium and Dementia

The reader is referred to other chapters in this text (Chapters 423–426) for detailed information on delirium and dementia in the overall population and among older patients specifically.

EPIDEMIOLOGY AND IMPACT Delirium occurs in up to 40% of hospitalized older patients, and is associated with increased morbidity, need for institutional care, and mortality in this population. While most episodes of delirium clear within a few days if the underlying cause(s) are identified and treated, delirium may persist for weeks, and in a few cases for months, after an acute hospitalization.

Normal aging does not cause impairment of cognitive function of sufficient severity to render an individual dysfunctional, which is the hallmark of a dementia syndrome. Slowed thinking and reaction time, mild recent memory loss, and impaired executive function can occur with increasing age and may or may not progress to dementia. Figure 464-6 illustrates the prevalence of memory impairment with increasing age. Just over 20% of people aged >70 in the United States have cognitive impairment without dementia (generally referred to as "Mild cognitive impairment [MCI])." Up to 15 to 20% of those diagnosed with MCI will progress to dementia over the course of a year; thus, most people with MCI will progress to dementia within 5 years. Therapeutic implications of MCI are subjects of intensive research.

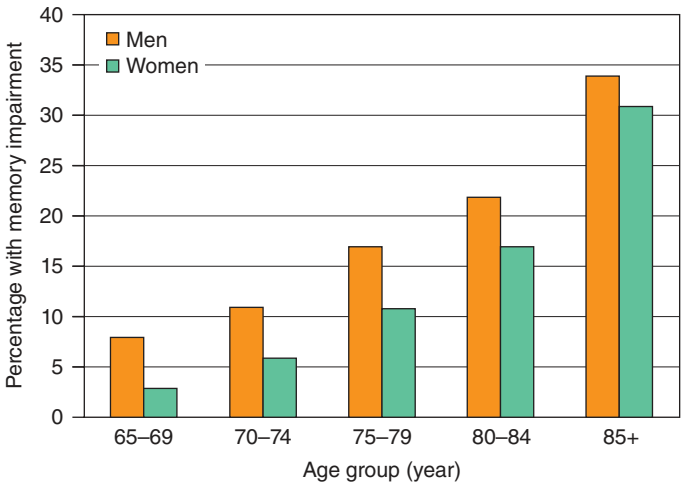


FIGURE 464-6 Rates of memory impairment in different age groups. The definition of "moderate or severe memory impairment" is 4 or fewer words recalled out of 20. *(Source: Health and Retirement Survey; details can be found at: http://hrsonline.isr.umich.edu/. Accessed December 30, 2016.)*

No non-pharmacological or pharmacological intervention has been shown to prevent progression to dementia.

The definitions of dementia syndromes (now also termed "major neurocognitive disorder") have been updated by the American Psychiatric Association. The prevalence of dementia increases with age; by age 85 between 30 and 40% have a dementia syndrome. Alzheimer's disease and vascular dementia, which often occur together based on pathologic studies, account for most dementias in older people. Dementia with Lewy bodies accounts for up to 25% of dementia, and is characterized by Parkinsonian features early in the disease (as opposed to dementia in Parkinson's disease, which generally occurs years after the onset of Parkinson's), personality changes, alterations in alertness and attention, and visual hallucinations that can cause paranoia. Most dementia syndromes are slowly progressive over several years; however, dementia is a terminal illness among patients who do not succumb to other comorbidities, and results in devastating loss of cognition and function in the later stages.

EVALUATION Regardless of setting, the new onset of delirium should be treated as a medical emergency because it can be the manifestation of an underlying critical illness. Figure 464-7 illustrates an overview of the assessment and management of impaired mental status and delirium in older hospitalized patients. The most validated evaluation for delirium is the Confusion Assessment Method, which requires an acute onset and fluctuating course *and* inattention *and* disorganized thinking *or* altered level of consciousness. Because the causes and risk factors for delirium are multifactorial, evaluation requires a careful history, physical examination, and selected laboratory studies based on the findings.

The benefits of screening older adults for cognitive impairment are controversial, but there are many interventions that may benefit patients and families early in the course of the disease (see below). Older patients in outpatient settings with complaints (or family reports of) early signs of cognitive impairment benefit from neuropsychological testing, which can help differentiate MCI and dementia, and identify concomitant factors such as depression and anxiety. The "Mini-Cog" is a sensitive screening tool for cognitive impairment, and consists of a 3-item recall test and clock drawing. Further evaluation of dementia includes a comprehensive history and physical examination, functional status assessment (since the diagnosis depends on impaired function), a brain imaging study, and selected laboratory tests, including a complete blood count, comprehensive metabolic panel, thyroid function tests, vitamin B12 level, and, if suspected tests for syphilis and human immunodeficiency virus antibodies.

MANAGEMENT Table 464-10 lists non-pharmacologic management strategies for various underlying risk factors and causes of delirium. Every attempt should be made to avoid or discontinue any medication

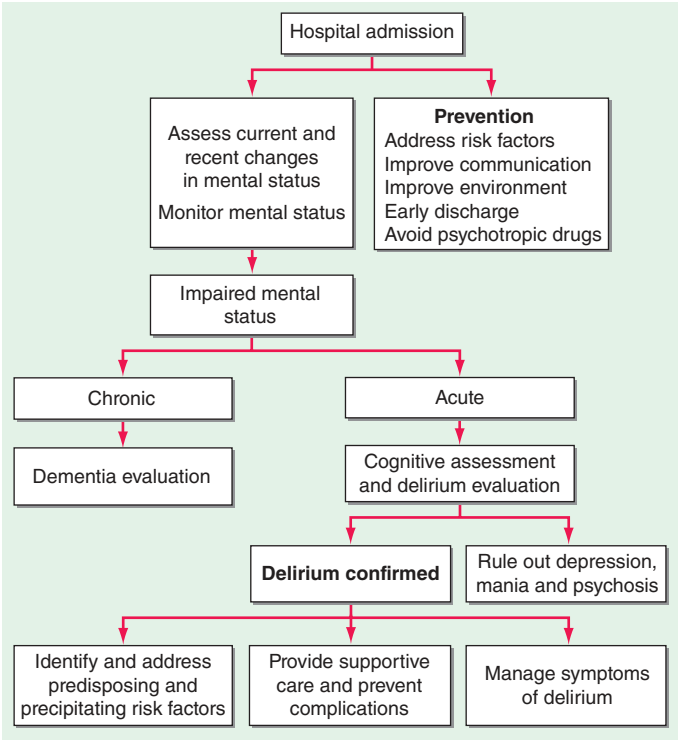


FIGURE 464-7 Algorithm depicting assessment and management of delirium in hospitalized older patients. *(Modified from SK Inouye: N Engl J Med 354:1157, 2006.)*

TABLE 464-10 Interventions for Risk Factors for Delirium

RISK FACTOR	INTERVENTION PROTOCOL
Cognitive impairment	• Orienting communication, including orientation board • Therapeutic activities program
Immobilization	• Early mobilization (e.g., ambulation or bedside exercises) • Minimizing immobilizing equipment (e.g., restraints, bladder catheters)
Psychoactive medications	• Restricted use of PRN sleep and psychoactive medications (e.g., sedative–hypnotics, narcotics, anticholinergic drugs) • Nonpharmacological protocols for management of sleep and anxiety
Sleep deprivation	• Noise-reduction strategies • Scheduling of night time medications, procedures, and nursing activities to allow uninterrupted period of sleep.
Vision impairment	• Provision of vision aids (e.g., magnifiers, special lighting) • Provision of adaptive equipment (e.g., illuminated phone dials, large-print books)
Hearing impairment	• Provision of amplifying devices; repair hearing aids • Instruct staff in communication methods
Dehydration	• Early recognition and volume repletion

Source: Data from SK Inouye et al: A clinical trial of a multicomponent intervention to prevent delirium in hospitalized older patients. N Engl J Med 340:669, 1990.

that may be worsening cognitive function in a delirious geriatric patient. This may not be possible, and in some patients, psychotropic drugs may be needed to treat delirium if the patient is a danger to themselves or others. Low dose haloperidol (0.25–0.5 mg) is generally recommended; more sedating antipsychotics and benzodiazepines should be avoided unless the goal is to put the patient to sleep for a short time. If a benzodiazepine is used, it should be short-acting and in a low dose.

Although the benefits of screening for cognitive impairment in older people are controversial, there are many non-pharmacologic interventions of older patients, their families, and other caregivers that may be beneficial **(Table 464-11)**. There are four basic approaches to the pharmacological treatment of dementia: (1) avoid drugs that can worsen cognitive function, mainly those with strong anticholinergic activity; (2) agents that enhance cognition and function; (3) drug treatment of coexisting depression, which is common throughout the course of dementia; and (4) pharmacological treatment of complications such as paranoia, delusions, psychosis, and behavioral symptoms such as agitation (verbal and physical). The use of antipsychotics to treat the neuropsychiatric symptoms of dementia is highly controversial. Most experts and guidelines recommend avoiding these drugs and using nonpharmacological strategies unless patients are a danger to themselves and others or if nonpharmacological interventions have failed. Patients with new or worsening behavioral symptoms associated with dementia should have a medical evaluation to identify potentially treatable precipitating conditions. Pain may be especially hard to detect, and if suspected, a therapeutic trial of acetaminophen should be considered.

The effectiveness of cholinesterase inhibitors and memantine in improving function and quality of life in patients with various types of dementia is controversial, and the potential benefits of these drugs versus their risks and costs must be weighed carefully to provide optimal person-centered care. The best evidence for effectiveness of cholinesterase inhibitors is in delaying progression of Alzheimer's disease and increasing the time before institutional placement is needed. Gastrointestinal side effects can be problematic and include nausea, vomiting, and diarrhea; nightmares can be bothersome as well. In addition to these bothersome side effects, cholinesterase inhibitors can cause bradycardia, and have been associated with syncope, injurious falls, and pacemaker placement. Memantine can cause dizziness, headache, confusion, and constipation. In one study, vitamin E was more effective than memantine in preventing functional decline in patients with Alzheimer's disease.

Urinary Incontinence • **EPIDEMIOLOGY AND IMPACT** Urinary incontinence is curable or controllable in many geriatric patients, especially those who have adequate mobility and mental functioning. Even when it is not curable, incontinence can be managed in a manner that keeps people comfortable, makes life easier for caregivers, and minimizes the costs of caring for the condition and its complications. Approximately one in three women and 15 to 20% of men aged >65 years have some degree of urinary incontinence. Between 5 and 10% of community-dwelling older adults have incontinence more often than weekly and/or use a pad for protection from urinary accidents. The prevalence is as high as 60–80% in many nursing homes, where residents often have both urinary and stool incontinence. Many older people (~40%) suffer from "overactive bladder," which may or may not include symptoms of incontinence. Symptoms of overactive bladder include urinary urgency (with or without incontinence), urinary frequency (voiding every two hours or more often), and nocturia (awakening at night to void). If nocturia alone is the predominant symptom, the patient should be asked about sleep disorders (see section that follows). The pathophysiology, evaluation, and management of overactive bladder are essentially the same as for urge urinary incontinence.

Incontinence is associated with social isolation and depression, and can be a precipitating factor in the decision to seek nursing home care when it cannot be managed in a manner that maintains hygiene and safety. In addition to predisposing to skin irritation and pressure ulcers, the most important potential complication of urinary incontinence and overactive bladder are falls and resultant injuries related to rushing to get to a toilet. Older people with gait disorders, especially those who have multiple episodes of nocturia or nocturnal incontinence, are at especially high risk for injuries. In addition to the bother of the condition to the older person or a caregiver, fall risk is a compelling reason for undertaking a diagnostic evaluation and specific treatment for incontinence and overactive bladder in the geriatric population.

TABLE 464-11 Key Principles in the Management of Dementia

Optimize the patient's physical and mental function through physical activity and mind plasticity principles and activities
Treatment underlying medical and other conditions (e.g., hypertension, Parkinson disease, depression
Avoid use of drugs with central nervous system side effects unless required for management of psychological or behavioral disturbances
Assess the environment and suggest alterations, if necessary
Encourage physical and mental activity
Avoid situations stressing intellectual capabilities; use memory aids whenever possible
Prepare the patient for changes in location
Emphasize good nutrition
Identify and manage behavioral symptoms and complications
Driving (consider a formal driving evaluation)
Wandering
Dangerous driving
Behavioral disorders
Depression
Agitation or aggressiveness
Psychosis (delusions, hallucinations)
Malnutrition
Incontinence
Provide ongoing care
Reassessment of cognitive and physical function
Treatment of medical conditions
Provide information to patient and family
Nature of the disease
Extent of impairment
Prognosis
Provide social service information to patient and family
Local Alzheimer's Association
Community health-care resources (day centers, homemakers, home health aides)
Legal and financial counseling
Use of advance directives
Provide family counseling for:
Setting realistic goals and expectations
Identification and resolution of family conflicts
Handling anger and guilt
Decisions on respite or institutional care
Legal concerns
Ethical concerns
Consideration of palliative and hospice care
Protect the caregiver from effects of caregiver stress

Source: Adapted from RL Kane et al: *Essentials of Clinical Geriatrics*, 8th ed. New York, McGraw-Hill, 2018.

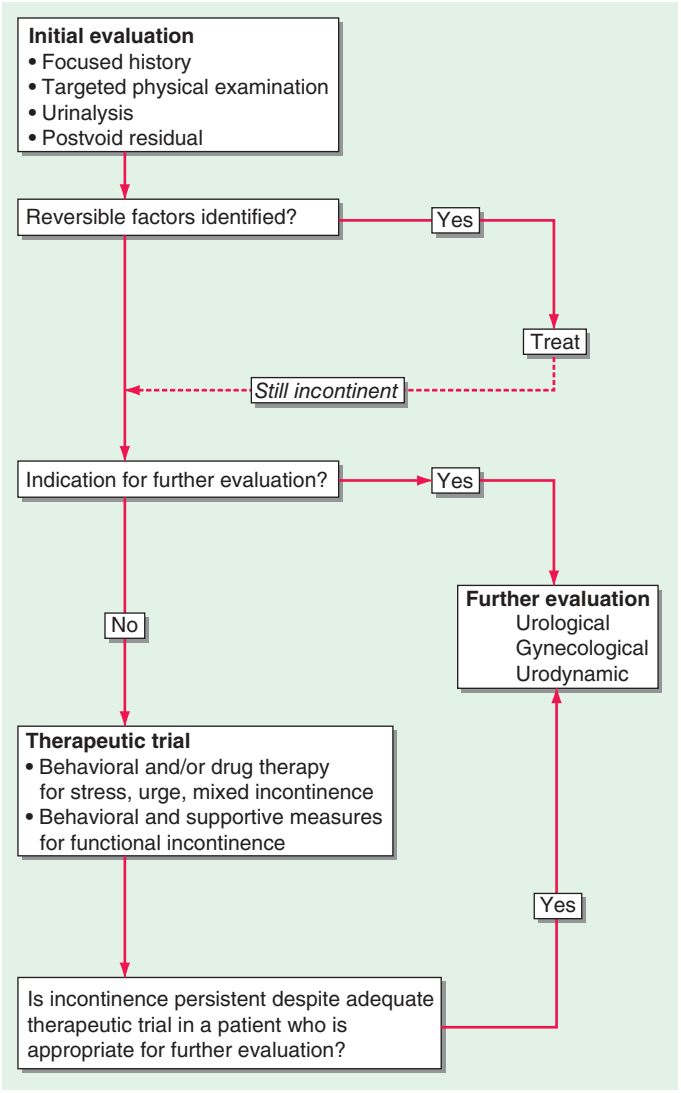


FIGURE 464-8 Algorithm for the basic evaluation and management of geriatric urinary incontinence in outpatient practice. *(Adapted from RL Kane et al: Essentials of Clinical Geriatrics, 8th ed. New York, McGraw-Hill, 2018.)*

EVALUATION Internists should ask older people about symptoms of urinary incontinence because these symptoms are often hidden out of embarrassment or fear. Simple questions can help identify incontinent patients, such as: "Do you have trouble with your bladder?" "Do you ever lose urine when you don't want to?" "Do you ever wear padding to protect yourself in case you lose urine?" (Table 464-3). A substantial number of older people will respond "no" to the first two questions, but "yes" to the third one.

Figure 464-8 illustrates an overall approach to the evaluation of the older patient with symptoms of incontinence and overactive bladder. The history and physical examination should focus on identifying potentially reversible causes and contributing factors (Table 464-12) and identifying the specific lower urinary tract symptoms. A simple, 3-item validated questionnaire can assist in distinguishing between the most common types of incontinence (Fig. 464-9). Among older women, the most common symptoms are a mixture of urge and stress incontinence (Fig. 464-10); urge is usually the more bothersome. Stress incontinence can often be objectively observed during a physical examination with a comfortably full bladder by having the patient cough in the standing position; leakage of urine simultaneously with coughing indicates that stress incontinence is present. Older men commonly have symptoms associated with overactive bladder and/or symptoms of voiding difficulty (hesitancy, poor or intermittent urinary stream, post-void dribbling); the overactive bladder symptoms are usually more bothersome. These symptoms overlap with those of both benign and malignant disorders of the prostate, and many internists may choose to consult a urologist for further management (see Chap. 83) because a urinary flow rate and post-void residual determination (PVR), and further evaluation if malignancy is suspected, are helpful in determining therapy.

Most older patients with symptoms of incontinence should have a PVR, especially men, diabetics, those with neurological disorders, and those with symptoms of voiding difficulty, because incomplete bladder emptying is common in older patients and is difficult to detect by history and physical examination alone. There is no specific cutoff for an abnormal PVR; the test must be done with a full bladder and straining during the test can alter the results. In older patients, a PVR between 0 and 100 mL is normal, between 100 and 200 mL must be interpreted based on symptoms, and a PVR higher than 200 is abnormal and usually influences treatment.

MANAGEMENT Patients who meet certain criteria should be referred for further urological, gynecological, and/or urodynamic evaluation

TABLE 464-12 Reversible Conditions That Cause or Contribute to Urinary Incontinence and Overactive Bladder Symptoms in Older People

CONDITION	MANAGEMENT
Lower urinary tract conditions	
Urinary tract infection (symptomatic with frequency, urgency, dysuria, etc.)	Antimicrobial therapy
Atrophic vaginitis/urethritis	Topical estrogen (not a primary treatment for incontinence but may help prevent recurrent infections and ameliorate symptoms of overactive bladder; oral estrogens can cause or worsen incontinence)
Stool impaction with irritation of bladder/urethral innervation and/or partial bladder outlet obstruction	Dis-impaction; appropriate use of stool softeners, bulk-forming agents, and laxatives if necessary; implement
Increased urine production	
Metabolic (hyperglycemia, hypercalcemia)	Better control of diabetes mellitus
	Therapy for hypercalcemia depends on underlying cause
Excess caffeine or fluid intake	Reduction in intake caffeinated beverages; reduction in fluid intake (most older people with incontinence or overactive bladder self-restrict fluid intake)
Volume overload with increased urine production at night	Support stockings
Venous insufficiency with edema	Leg elevation
	Sodium restriction
	Diuretic therapy (late afternoon dose may be effective)
Congestive heart failure	Medical therapy
Impaired ability or willingness to reach a toilet	
Delirium	Diagnosis and treatment of underlying cause(s)
Chronic illness, injury, or restraint that interferes with mobility	Regular toileting
	Use of toilet substitutes
	Environmental alterations (eg, bedside commode, urinal)
	Remove restraints if possible
Psychological (depression, anxiety)	Appropriate non-pharmacologic and/or pharmacologic treatment
Drug Side Effects	Remove offending drug(s) if feasible; modification of dose, frequency or timing may also reduce symptoms for some drugs:
	Diuretics (polyuria, frequency, urgency)
	Anticholinergics (constipation, incomplete bladder emptying)
	Psychotropic drugs
	Tricyclic antidepressants (anticholinergic effects)
	Antipsychotics (immobility, sedation)
	Sedative-hypnotics (immobility, sedation)
	Narcotic analgesics (constipation, incomplete bladder emptying)
	Alpha-Adrenergic blockers (urethral relaxation)
	Alpha-Adrenergic agonists (urethral contraction and potential incomplete bladder emptying)
	Cholinesterase inhibitors (urinary frequency, urgency)
	Angiotensin-converting enzyme inhibitors (cough precipitating stress incontinence)
	Calcium channel blockers, gabapentin, pregabalin, glitazones (edema with nocturia)
	Alcohol (polyuria, frequency, urgency, sedation, delirium, immobility)
	Caffeine (polyuria, bladder irritation)

Source: Adapted from RL Kane et al: *Essentials of Clinical Geriatrics*, 8th ed. New York, McGraw-Hill, 2018.

before initiating specific therapy. Examples include history of lower urinary tract surgery or radiation or recurrent symptomatic urinary tract infections, marked pelvic prolapse on physical examination of a woman, suspected prostate cancer, and sterile hematuria.

Potentially reversible conditions should be addressed, including the many types of medications that can affect bladder function, which should be eliminated if possible (Table 464-12). Table 464-13 lists treatments for different types of incontinence. Many patients respond well to properly taught and adhered to behavioral interventions. Physical therapists and nurses who specialize in treating lower urinary tract symptoms can be very helpful and should be consulted if available. Pharmacologic treatment of incontinence and overactive bladder is dictated by the innervation of the lower urinary tract. Alpha-adrenergic stimulation increases tone in the smooth muscle of the urethra, thus alpha agonists have been used to treat stress incontinence in women, and alpha blocker are used to decrease urethral tone in men with overactive bladder associated with prostate enlargement. Anticholinergic/antimuscarinic agents and beta-3 stimulation inhibit bladder contraction and are used for overactive bladder and urge incontinence. Patients with severe cognitive impairment and/or immobility can generally be managed effectively by prompted voiding and/or incontinence undergarments, as long comfort, dignity, and safety are maintained.

Sleep Disorders Sleep disorders are discussed in more detail for the general adult population in Chap. 27. Because they are so common and have some unique features in older patients, they are discussed briefly here.

EPIDEMIOLOGY AND IMPACT Aging is associated with multiple changes in sleep architecture as well as multiple diseases and disorders that can disrupt sleep. Thus, complaints of sleep difficulty are common in older adults. Consequences of sleep difficulty include lower health-related quality of life, increased medication use, more cognitive decline, and greater health care utilization. Four types of primary sleep disorders are common in the geriatric population: insomnia, sleep disordered breathing due to obstructive sleep apnea (OSA), restless leg syndrome (RLS), periodic leg movements in sleep (PLMS). Complaints of bothersome insomnia—the inability to fall asleep or stay asleep despite a conducive environment—increase with age and occur in ~30% of people aged >65. Insomnia is commonly associated with depression, anxiety, alcohol

The 3IQ is a patient questionaire that helps your doctor distinguish urge incontinence from stress incontinence. It should take no more than a couple of minutes. Complete the quiz and bring it to your next appointment.

1. During the last 3 months, have you leaked urine (even a small amount)?

☐ **Yes** ☐ **No** (if this response is marked, the 3IQ test is complete)

2. During the last 3 months, did you leak urine (check all that apply):

☐ When you were performing some physical activity, such as coughing, sneezing, litting, or exercising?

☐ When you had the urge or the feeling that you needed to empty your bladder, but you could not get to the toilet fast enough?

☐ Without physical activity and without sense of urgency?

3. During the last 3 months, did you leak urine most often (check only one):

☐ When you were performing some physical activity, such as coughing, sneezing, litting, or exercising?

☐ When you had the urge or the feeling that you needed to empty your bladder, but you could not get to the toilet fast enough?

☐ Without physical activity and without sense of urgency?

☐ About equally as often with physical activity as with a sense of urgency?

Definitions of type of urinary incontinence are based on responses to question 3.

Response to Question 3	Type of incontinence
Most often with physical activity	Stress only or stress predominant
Most often with the urge to empty the bladder	Urge only or urge predominant
Without physical activity or sense of urgency	Other cause only or other cause predominant
About equally with physical activity and sense of urgency	Mixed

FIGURE 464-9 The 3 Incontinence Questions (3IQ) Assessment Tool. *(From JS Brown et al: The sensitivity and specificity of a simple test to distinguish between urge and stress urinary incontinence. Ann Intern Med 144:715, 2006.)*

intake, and ingestion of caffeinated beverages later in the day. OSA occurs in ~10% of older adults, but is probably under-reported and under-diagnosed. It is associated with medical comorbidities, such as obesity and congestive heart failure. RLS occurs in 5–10% of adults, and its prevalence increases in those aged >70. It is almost twice as common in women than men; family history, iron deficiency, and intake of antihistamines and most antidepressants are risk factors. PLMS can be found in up to 45% of older people, but is often of unknown clinical consequence and remains undiagnosed.

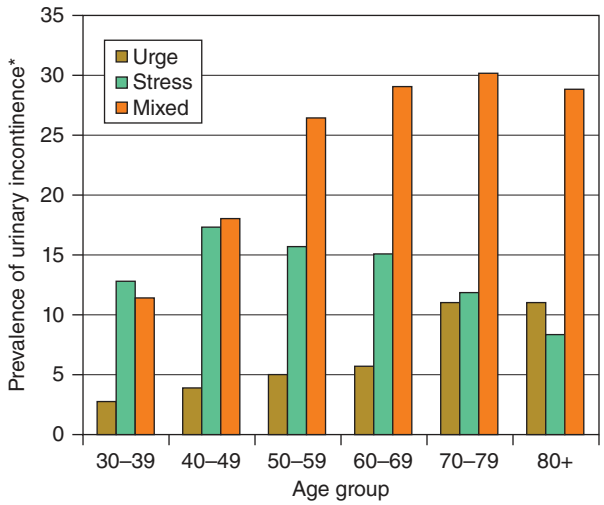


FIGURE 464-10 Rates of urge, stress, and mixed incontinence, by age group, in a sample of 3552 women. *Based on a sample of 3553 participants. *(From JL Melville et al: Arch Intern Med 165:537, 2005.)*

EVALUATION Older people should be screened for sleep difficulty with questions such as: "Do you often feel sleepy during the day?" "Do you have difficulty falling asleep at night?" Further evaluation of the nature and impact of the complaints can be accomplished with standardized questionnaires (Table 464-3). Patients with significant sleep complaints should be asked about conditions that can interrupt sleep, such as nocturia, gastroesophageal reflux, chronic pain, and caffeine and alcohol intake. Specific questions characterizing the complaints should include inquiring about loud snoring (for OSA), the urge to move legs associated with uncomfortable sensations (RLS), and leg movements during sleep (PLMS; which may result in kicking a bed partner).

MANAGEMENT Patients suspected for OSA, RLS, or PLMS should be referred for formal sleep evaluation. While hypnotics are among the most commonly prescribed drugs in the geriatric population, non-pharmacologic management of sleep should be the initial and primary approach, as many patients can benefit from properly taught and adhered to interventions (Table 464-14). Benzodiazepine hypnotics should be avoided whenever feasible because they are associated with next-day hangover effects, which may manifest as cognitive impairment and can precipitate falls and car crashes; and rebound insomnia.

Elder Abuse and Neglect • EPIDEMIOLOGY AND IMPACT

The incidence of elder abuse and neglect, and self-neglect are unknown

TABLE 464-13 Primary Treatments for Different Types of Geriatric Urinary Incontinence

TYPE OF INCONTINENCE	PRIMARY TREATMENTS
Stress	Pelvic muscle (Kegel) exercises Other behavioral interventions including timed voiding and double voiding to avoid residual urine α-Adrenergic agonist (none are FDA approved for this purpose) Topical estrogen to strengthen periurethral tissue (not effective alone; oral estrogens contraindicated) Periurethral injections to provide bulking and support Surgical bladder neck suspension or sling for severe incontinence, based on patient preference
Urge and overactive bladder symptoms	Pelvic muscle (Kegel) exercises Other behavioral interventions—timed voiding and double voiding to avoid residual urine Antimuscarinic and beta-3 adrenergic drugs
Incontinence with incomplete bladder emptying	α-Adrenergic antagonists (men) Bladder training, double voiding Intermittent catheterization Indwelling catheterization in selected patients in whom risks and discomforts of urinary retention outweigh risks of a chronic indwelling catheter
Incontinence with impaired physical and/or cognitive function	Behavioral interventions (prompted voiding, habit training) Environmental manipulation including use of urinal or bedside commode, safe lit path to bathroom Incontinence undergarments and pads

Source: Adapted from RL Kane et al: *Essentials of Clinical Geriatrics*, 8th ed. New York, McGraw-Hill, 2018.

TABLE 464-14 Non-Pharmacologic Management of Insomnia in Older Adults

Sleep Hygiene Rules
Check effect of medication on sleep and wakefulness
Avoid caffeine, alcohol, and cigarettes after lunch
Limit liquids in the evening
Keep a regular bedtime–waketime schedule
Avoid naps or limit to 1 nap a day, no longer than 30 min
Spend time outdoors (without sunglasses), particularly in the late afternoon or early evening
Exercise—but limit exercise immediately before bedtime
Instructions for Stimulus-Control Therapy
Only go to bed when tired or sleepy
If unable to fall asleep within 20 min, get out of bed (and bedroom if possible); while out of bed, do something quiet and relaxing
Only return to bed when sleepy
If unable to fall asleep within 20 min, again get out of bed
Repeat these behaviors until able to fall asleep within a few minutes
Get up at the same time each morning (even if only a few hours of sleep)
Avoid naps

Source: Adapted from JB Halter et al (eds): *Hazzard's Geriatric Medicine and Gerontology*, 7th ed. New York, McGraw-Hill, 2017.

because it is often not asked about or reported. The best data suggest that the incidence over 12 months is at least 8–10%. Abuse and neglect can result in physical injuries and related pain, worsening of chronic medical conditions, dehydration and pressure ulcers, emotional distress, and loss of income and savings and related consequences.

EVALUATION Because abuse and neglect are so under-reported, may be unsuspected, and have such devastating consequences, older adults should be screened (in the absence of caregivers) with questions such as: "Do you ever feel unsafe where you live? "Has anyone ever threatened or hurt you?" "Has anyone been taking your money without your permission?" (Table 464-3). Table 464-15 outlines the definitions, symptoms and signs, and key aspects of evaluating suspected abuse and neglect.

MANAGEMENT In addition to treating the physical, medical, and emotional consequences, patients suspected of elder abuse or neglect should be reported to the appropriate local or state agency to investigate and ensure the patient's safety. The reader is referred to two reviews of this topic for further information on specific aspects of management.

TABLE 464-15 Elder Abuse and Neglect

CATEGORY	DEFINITION AND EXAMPLES	SYMPTOMS AND SIGNS	KEY ASPECTS OF EVALUATION
Physical Abuse	Acts of violence that may result in pain, injury, or impairment • Pushing, slapping, hitting, force-feeding • Improper positioning or use of restraints • Improper use of medications	Abrasions Lacerations Bruises Fractures Use of restraints Burns Pain Depression Delirium or onset or worsening of dementia-related behavioral symptoms	The interview should be conducted alone with the patient; it may reveal discordant histories or findings inconsistent with the history provided by the caregiver. Ankles and wrists should be examined for abrasions suggestive of the use of restraints. Findings that are discordant with the mechanism of injury reported or multiple injuries in various stages of healing should raise the suspicion of abuse. Injuries to the head, neck, and upper arms occur in victims of physical elder abuse, but must be distinguished from accidental injuries. Jaw and zygomatic fractures are more likely to be sustained from a punch than from a fall, which more typically result in fractures to orbital and nasal bones.
Psychological or Verbal Abuse	Conduct that causes mental or emotional distress • Verbal harassment or intimidation • Threats of punishment or deprivation • Isolation	Direct observation of verbal abuse Subtle signs of intimidation, such as deferring questions to a caregiver or potential abuser Evidence of isolation Depression, anxiety, or both	Assess the size and quality of the patient's social network (beyond the suspected abuser). Conduct standardized assessments of depression, anxiety, and cognition, directly or through referral. Ask specifically about verbal or psychological abuse with questions such as "Does your relative/caregiver ever yell or curse at you?"; "Have you been threatened with being put into a nursing home?"; or "Are you ever prevented from seeing friends and family members whom you wish to see?"

(*Continued*)

TABLE 464-15 Elder Abuse and Neglect (Continued)

CATEGORY	DEFINITION AND EXAMPLES	SYMPTOMS AND SIGNS	KEY ASPECTS OF EVALUATION
Financial Abuse	Misuse of the person's income or resources for the financial or personal gain of a caregiver or advisor • Stealing money or possessions • Denying a home • Coercing to sign contracts or spend money	Inability to pay for medicine, medical care, food, rent, or other necessities Failure to renew prescriptions, adhere to medication regimens or other treatments, or keep medical appointments Malnutrition, weight loss, or both, without an obvious medical cause Evidence of poor financial decision making Firing of home care or other service providers by abuser Unpaid utility bills Initiation of eviction proceedings	Ask about financial exploitation with questions such as "Has money or property been taken from you without your consent?"; "Have your credit cards or automated-teller-machine card been used without your consent?"; and "At the end of the month, do you have enough money left for food and other necessities?" Abrupt changes in financial circumstances of the caregiver in either direction may herald an increased risk of financial exploitation or exploitation already under way. Abuse of the power of attorney; if the person with power of attorney or health care proxy is suspected of not acting in the best interest of the patient, documents necessary to ensure that the assumption of fiduciary responsibilities is authorized.
Sexual Abuse	Sexual coercion or assault	Bruising, abrasions, lacerations in the genital or anal areas or abdomen Newly acquired sexually transmitted diseases, especially in nursing home Urinary tract infection	Inquire directly about sexual assault or coercion. For patients with dementia, direct queries to caregivers about hypersexual behavior as part of a larger history regarding dementia-related behaviors and assess patient's capacity for decision-making about sexual activity If indicated, refer to an emergency department for assessment for sexual assault and collection of specimens (forensic evidence should be collected by experienced professionals, such as nurses who have undergone Sexual Assault Nurse Examiners (SANE) training.
Neglect (by caregiver or self-neglect)	Failure to provide the materials, supplies, food and drink or services necessary for optimal functioning or to avoid harm	Malnutrition Dehydration Poor hygiene Pressure ulcers Nonadherence to medication regimen or other treatments Worsening of dementia-related behavioral symptoms	Interview primary caregiver about his or her understanding of the nature of the patient's care needs and how well care is being rendered. Neglect by a caregiver may be intentional or unintentional Assess hygiene, cleanliness, and appropriateness of dress. Examine the skin for pressure ulcers, infections, and infestations. Assess nutrition and hydration, including measuring body-mass index And blood urea nitrogen and creatinine to assess hydration.

Source: Adapted from JB Halter et al (eds): *Hazzard's Geriatric Medicine and Gerontology*, 7th ed. New York, McGraw-Hill, 2017 and MS Lachs, KA Pillemer: Elder abuse. N Engl J Med 373:1947, 2015.

END-OF-LIFE AND PALLIATIVE CARE

As stated earlier, end-of-life and palliative care are critical aspects of caring for the geriatric population, and require a comprehensive, person-centered approach. End-of-life and palliative care are addressed in detail in Chap. 9, and pain management is addressed in Chap. 10. For geriatric patients, limited life expectancy is a critical factor to consider when making end-of-life care decisions. General principles of decision-making are especially relevant when considering palliative and/or end-of-life care in older patients (Fig. 464-4). Decision-making becomes complicated, however, among older patients with multimorbidity. Without a clear terminal diagnosis, when to start palliative care/end-of-life care could be challenging. While it is sometimes clear when an older patient has a terminal condition, such as end-stage congestive heart failure or chronic obstructive pulmonary disease, many older patients with multi-morbidity have combinations of conditions of varying severity. Moreover, neurogenerative disorders, including most forms of dementia, Parkinson's disease, and patients with multiple strokes commonly have a gradually progressive course, and it can be challenging to determine when discussions about palliative and end-of-life care should be initiated. Dementia, however, should be considered a terminal illness in the advanced stages.

Internists should play a pivotal role in making the decision when to initiate these discussions, and should be proactive in encouraging patients and their families to execute advance directives before a health care crisis occurs. The survivability of cardiopulmonary resuscitation in hospitalized patients aged ≥65 is <20%; among the old-old with multimorbidity it is much lower. The survivability of cardiopulmonary resuscitation (CPR) in nursing home residents is almost zero, making it a futile intervention for most in this setting. Enteral feeding tubes should not be placed in patients with end-stage dementia (Table 464-2). Estimation of prognosis using tools such as ePrognosis and careful attention to other factors that contribute to person-centered care will assist internists in dealing with these challenging issues in geriatric care.

FURTHER READING

American Geriatrics Society 2015 Beers Criteria Update Expert Panel. American Geriatrics Society 2015 updated Beers criteria for potentially inappropriate medication use in older adults. J Am Geriatr Soc 63:2227, 2015.

American Geriatrics Society Expert Panel on the Care of Older Adults with Diabetes Mellitus: Guidelines Abstracted from the American Geriatrics Society Guidelines for Improving the Care of Older Adults with Diabetes Mellitus: 2013 Update. J Am Geriatr Soc 61:2020, 2013.

American Geriatrics Society Panel on Pharmacologic Management of Persistent Pain in Older Persons: Pharmacological management of persistent pain in older persons. J Am Geriatr Soc 57:1331, 2009.

American Geriatrics Society: Choosing Wisely: Ten things clinicians and patients should question. *http://www.choosingwisely.org/societies/american-geriatrics-society/*. Accessed January 14, 2017.

AMDA—The Society for Post-Acute and Long-Term Care Medicine: Ten things clinicians and patients should question. *http://www.choosingwisely.org/societies/amda-the-society-for-post-acute-and-long-term-care-medicine/*. Accessed January 14, 2017.

CDC Recommended Adult Immunization Schedule United States—2016 *https://www.cdc.gov/vaccines/schedules/downloads/adult/adult-combined-schedule.pdf*. Accessed January 12, 2017.

Clinician's Guide to Assessing and Counseling Older Drivers. *http://www.michigan.gov/documents/sos/Clinicians_Guide_To_OlderDrivers_3rd_edition_523147_7.pdf*. Accessed January 12, 2017.

Halter J et al (eds): *Hazzard's Geriatric Medicine and Gerontology*, 7th ed. New York, McGraw Hill, 2016.

Kane RL et al: *Essentials of Clinical Geriatrics*, 8th ed. New York, McGraw Hill, 2018.

Malone ML et al (eds): *Geriatric Models of Care*. New York, Springer, 2015.

465 Approach to Medical Consultation

Jack Ende, Jeffrey Berns

Effective health care requires teams of generalists and specialists with complementary expertise. Many clinical conditions require the input of more than one clinical provider, either because the diagnosis and recommended treatment is uncertain or because a patient may have multiple diseases that may be best managed by involving multiple specialists.

To *consult* is to seek advice from someone with expertise in a particular area, whereas *consultation* refers to the meeting or comparable outcome arising from that request. Medical consultation takes several forms. Its most traditional forms include in-hospital consultation in which physicians provide recommendations or perform procedures for a hospitalized patient, and out-patient consultations, in which patients are seen in the office setting. More contemporary forms of consultation include e-consultations, telemedicine evaluations (see "Consultation Involving Telemedicine," below), and remote medical second opinions. In these forms, the consultant may not actually see the patient but, nonetheless, assumes the responsibility of evaluating the patient's clinical condition, assessing and analyzing pertinent clinical data, and offering a synthesis and appropriate recommendations.

While forms of medical consultation evolve, basic responsibilities associated with medical consultation endure. These responsibilities can be divided into those that fall to the requesting physician or nonphysician practitioner; the consultant, who provides the consultation; and the health system, hospital, or organization that must support this important medical encounter **(Table 465-1)**.

TABLE 465-1 Stakeholder Responsibilities in the Medical Consultation Process

REFERRING PHYSICIAN OR PROVIDER	CONSULTANT PHYSICIAN	HEALTH SYSTEM, HOSPITAL, OR CARE ORGANIZATION
• Ensure patient participation and engagement • Be specific regarding clinical question and desired outcome • Communicate level of urgency • Avoid consulting for nonclinical purposes	• Maintain standards of professionalism, including those pertaining to availability, communication, respect, and collegiality • Appreciate levels of urgency and respond appropriately • Assemble and develop one's own database • Be specific in synthesis and recommendations • Understand desired outcomes, including arrangements for follow-up • Communicate with referring provider in whatever manner is mutually desirable	• Maintain adequate specialty workforce to enable appropriate access • Support systems for efficient exchange of clinical information • Develop culture of collegiality and team-based care

RESPONSIBILITIES OF THE REQUESTING PRACTITIONER

Before requesting a consultation, the provider should ensure that the patient endorses the purpose of the consultation, understands the role of the consultant, and anticipates the likely outcomes of the encounter. Further responsibilities of the requesting practitioner include being specific and communicating clearly the reason for the consultation. Vague messages such as, "Please evaluate" are not as helpful as more specific inquiries such as, "What is the cause of the declining kidney function?" or, "How should this asymptomatic pulmonary nodule be evaluated?" To the extent possible, the requesting practitioner should provide the relevant clinical information, summarized as succinctly as possible. Urgency should be clearly conveyed, typically with a phone call or other direct communication.

The requesting practitioner should be explicit regarding the intended outcome of the consultation, i.e., is this for a single evaluation or ongoing co-management? Communication between the requesting and the consulting providers is paramount. Whether this communication includes direct contact is less important than that the relevant information and desired outcome be explicit and clear, regardless of communication medium. Consultations should be requested for clinical purposes and always directed to qualified consultants; they should not be driven by entrepreneurial or relationship-building purposes. Another responsibility of the referring provider is not to "overconsult." Medical care should be focused on value, not volume.

RESPONSIBILITIES OF THE CONSULTANT

Just as the referring provider should attend to clear and explicit communication, so too should the consultant follow the precepts of effective interactions between professionals, which include courtesy, availability, and clarity. Particularly on the inpatient service, where consultants may receive several requests each day, it is important that the incoming consultations are triaged and dispatched as clinically appropriate. Consultants also need to determine the requested level of involvement going forward and not assume that long-term co-management is being sought. While consultants can and should make use of available clinical data, they should also assemble independently their own database, including taking a history, performing a physical exam, and reviewing pertinent clinical studies. Absent that, they may be unable to provide an independent and actionable synthesis. Just as the referring provider needs to be clear and concise, so too should the consultant be specific and focused in the recommendations provided. "Possible malignant ascites" is less helpful than, "I will arrange for paracentesis to exclude the possibility of malignant ascites." For the most part, recommendations to "consider" some diagnosis or test are less helpful than more specific and concrete advice. Some referring practitioners wish to be called after a patient is seen; others prefer that communication be handled as part of the medical record. How this communication is handled must also align with the complexity and urgency of the consultation and clinical circumstances.

RESPONSIBILITIES OF HEALTH SYSTEMS, HOSPITALS, AND MEDICAL ORGANIZATIONS

Health systems, hospitals, and medical organizations also have responsibilities in the consultation process. This responsibility includes ensuring that qualified consultants are accessible and available on the medical staff. Consultations within a single system are aided by common shared electronic medical records, particularly when consultations originate in the hospital, but can also involve care in the outpatient setting. Finally, health care entities should strive to foster a culture of team-based care and collegiality.

SPECIAL ISSUES IN MEDICAL CONSULTATION

Curbside Consults Curbside consults are requests from one practitioner to another for an informal and unwritten opinion about a

specific patient care matter. They are typically limited in scope, mostly regarding management or questions regarding procedures, and developed from information provided by the consulting practitioner and perhaps the medical record (such as labs and imaging studies), but without a comprehensive review of the record or any direct contact with the patient. Although often viewed as convenient, efficient, and a common aspect of clinical care, by their very nature, curbside consults have been found to often be incomplete or even flawed. It is not uncommon for the question being asked to be deemed too complex for a curbside consult, or for it not to be the actual or only issue the consultant feels needs to be addressed. As a general rule, curbside consults should be avoided. While medicolegal liability is often cited as a reason to limit curbside consults, the risk is actually negligible as U.S. courts have ruled that curbside consults do not establish a doctor-patient relationship necessary for creating the basis for medical malpractice litigation. An important exception, however, is when a curbside consult is provided by a resident or fellow in training; in this circumstance the trainee's supervising physician, whether aware of the curbside consult or not, is responsible for the recommendations of the trainee.

Second Opinions Physicians may find themselves providing consultations requested by patients who have already been evaluated for the same problem by another physician. Not a "consult" in the usual context of one physician referring a patient to another, the service provided by the consultant here is, nonetheless, very much aligned with a physician-referred consult. Second opinions, which often are encouraged by the patient's physician, may be sought by patients for reassurance that a diagnosis and treatment recommendation is correct, out of dissatisfaction with the initial physician, or with the hope of an entirely different opinion and recommendation. The physician providing the second opinion should strive to understand the patient's motivations for seeking the additional opinion. While a second opinion may have been initiated by the patient rather than referral from another physician, it is recommended that the consulting physician communicate with the patient's primary physician or specialist as would be done following a standard consultation unless the patient insists otherwise. In addition, professional behavior in how the consulting physician refers to the recommendations or actions of previously consulted physicians is important, even when there is disagreement. Likewise, it is important that a transfer of care from prior consultants to the one providing a second opinion be enacted only if specifically requested by the patient or the physician who encouraged the second opinion.

Consults Involving Mid-Level Providers Increasingly, specialist physicians may find themselves being consulted by nurse practitioners and physician assistants rather than other physicians. Whether the quality of the information provided to the consultant physician by a mid-level provider is different from physician-to-physician referrals has not been studied. Consulting physicians should know whether they should respond back to the mid-level provider or to the supervising physician. As with physician-to-physician consults, it is also important for the consultant to know whether the individual calling for the consult has an ongoing role in the care of the patient or is simply covering for a limited period of time. Finally, the consultant, if responding back to the mid-level provider, should make sure that the information provided meets the needs of that provider, and that questions are answered as they would be if responding back to another physician.

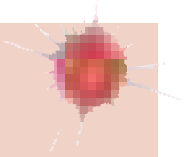
Consultation Involving Telemedicine Consultations making use of electronic health records, patient portals, and various forms of telecommunication technology, including video conferencing or cell phone communication, can improve access to care, reduce cost, and improve outcomes. This is particularly true when employed in geographic areas of health care shortage and when the clinical issues can be handled without direct contact with the patient, e.g., radiology or dermatology. However, the absence of direct contact between patient and consultant introduces special issues related to diagnostic accuracy and physician-patient relationship. Regulatory issues, liability, security, and confidentiality issues arise as well. Consultation via telemedicine holds considerable promise, but the aforementioned concerns will need to be better understood.

FURTHER READING

Daniel H, Sulmasy LS: Policy recommendations to guide the use of telemedicine in primary care: An American College of Physicians Position Paper. Ann Intern Med 163:787, 2015.

Pearson SD: Principles of generalist-specialist relationships. J Gen Intern Med 14(Suppl 1):S13, 1999.

466 Medical Disorders During Pregnancy



Robert L. Barbieri, John T. Repke

Each year, approximately 4 million births occur in the United States, and more than 130 million births occur worldwide. A significant proportion of births are complicated by medical disorders. Advances in medical care and fertility treatment have increased the number of women with serious medical problems who attempt to become pregnant. Medical problems that interfere with the physiologic adaptations of pregnancy increase the risk for poor pregnancy outcome; conversely, in some instances, pregnancy may adversely impact an underlying medical disorder.

HYPERTENSION

(See also Chap. 271) In pregnancy, cardiac output increases by 40%, with most of the increase due to an increase in stroke volume. Heart rate increases by ~10 beats/min during the third trimester. In the second trimester, systemic vascular resistance decreases, and this decline is associated with a fall in blood pressure. During pregnancy, a blood pressure of 140/90 mmHg is considered to be abnormally elevated and is associated with an increase in perinatal morbidity and mortality. In all pregnant women, the measurement of blood pressure should be performed in the sitting position, because the lateral recumbent position may result in a lower blood pressure. The diagnosis of hypertension requires the measurement of two elevated blood pressures at least 4 h apart. Hypertension during pregnancy is usually caused by preeclampsia, chronic hypertension, gestational hypertension, or renal disease.

PREECLAMPSIA

Approximately 5–7% of all pregnant women develop *preeclampsia*, the new onset of hypertension (blood pressure >140/90 mmHg) and proteinuria (either a 24 h urinary protein >300 mg/24 h, or a protein-creatinine ratio ≥0.3) after 20 weeks of gestation. Recent revisions to the diagnostic criteria include: proteinuria is no longer an absolute requirement for making the diagnosis; the terms mild and severe preeclampsia have been replaced; and the disease is now termed preeclampsia either with or without severe features and fetal growth restriction is no longer a defining criterion for preeclampsia with severe features. Although the precise pathophysiology of preeclampsia remains unknown, recent studies show excessive placental production of antagonists to both vascular endothelial growth factor (VEGF) and transforming growth factor β (TGF-β). These antagonists to VEGF and TGF-β disrupt endothelial and renal glomerular function resulting in edema, hypertension, and proteinuria. The renal histological feature of preeclampsia is glomerular endotheliosis. Glomerular endothelial cells are swollen and encroach on the vascular lumen. Preeclampsia is associated with abnormalities of cerebral circulatory autoregulation, which increase the risk of stroke at mildly and moderately elevated blood pressures. Risk factors for the development of preeclampsia include nulliparity, diabetes mellitus, a history of renal disease or chronic hypertension, a prior history of preeclampsia, extremes of maternal age (>35 years or <15 years), obesity, antiphospholipid antibody syndrome, and multiple gestation. Low-dose aspirin (81 mg daily, initiated at the end of the first trimester) modestly reduces the risk of preeclampsia in pregnant women at high risk of developing the disease.

Preeclampsia with severe features is the presence of new-onset hypertension and proteinuria accompanied by end-organ damage. Features may include severe elevation of blood pressure (>160/110 mmHg), evidence of central nervous system (CNS) dysfunction (headaches, blurred vision, seizures, coma), renal dysfunction (oliguria or creatinine >1.5 mg/dL), pulmonary edema, hepatocellular injury (serum alanine aminotransferase level more than twofold the upper limit of normal), hematologic dysfunction (platelet count <100,000/L or disseminated intravascular coagulation [DIC]). The *HELLP syndrome* (*h*emolysis, *e*levated *l*iver enzymes, *l*ow *p*latelets) is a special subtype of severe preeclampsia and is a major cause of morbidity and mortality in this disease. Platelet dysfunction and coagulation disorders further increase the risk of stroke.

TREATMENT

Preeclampsia

Preeclampsia resolves within a few weeks after delivery. For pregnant women with preeclampsia prior to 37 weeks of gestation, delivery reduces the mother's morbidity but exposes the fetus to the risk of premature birth. The management of preeclampsia is challenging because it requires the clinician to balance the health of the mother and fetus simultaneously. In general, prior to term, women with preeclampsia without severe features may be managed conservatively with limited physical activity, although bed rest is not recommended, close monitoring of blood pressure and renal function, and careful fetal surveillance. For women with preeclampsia with severe features, delivery is recommended unless the patient is eligible for expectant management in a tertiary hospital setting. Expectant management of preeclampsia with severe features remote from term affords some benefits for the fetus, but significant risks for the mother. Postponing delivery beyond 34 weeks gestation in this group of patients is not recommended. In preeclampsia without severe features delivery at 37 weeks is recommended.

The definitive treatment of preeclampsia is delivery of the fetus and placenta. For women with preeclampsia with severe features, aggressive management of blood pressures >160/105 mmHg reduces the risk of cerebrovascular accidents. IV labetalol or hydralazine is most commonly used to acutely manage severe hypertension in preeclampsia; labetalol is associated with fewer episodes of maternal hypotension. Elevated arterial pressure should be reduced slowly to avoid hypotension and a decrease in blood flow to the fetus.

Magnesium sulfate is the preferred agent for the prevention and treatment of eclamptic seizures. Large, randomized clinical trials have demonstrated the superiority of magnesium sulfate over phenytoin and diazepam in reducing the risk of seizure and, possibly, the risk of maternal death. Magnesium may prevent seizures by interacting with *N*-methyl-D-aspartate (NMDA) receptors in the CNS. The universal use of magnesium sulfate for seizure prophylaxis in preeclampsia without severe features is no longer recommended by most experts. There is consensus that magnesium sulfate should be used in all cases of preeclampsia with severe features, or in cases of eclampsia. Women who have had preeclampsia appear to be at increased risk of cardiovascular and renal disease later in life.

CHRONIC ESSENTIAL HYPERTENSION

Pregnancy complicated by chronic essential hypertension is associated with intrauterine growth restriction and increased perinatal mortality. Pregnant women with chronic hypertension are at increased risk for superimposed preeclampsia and abruptio placentae. Women with chronic hypertension should have a thorough prepregnancy evaluation, both to identify remediable causes of hypertension and to ensure that the prescribed antihypertensive agents (e.g., angiotensin-converting enzyme [ACE] inhibitors, angiotensin-receptor blockers) are not associated with an adverse outcome of pregnancy. Labetalol and nifedipine are the most commonly used medications for the treatment of chronic hypertension in pregnancy. The target blood pressure is in the range of 130–150 mmHg systolic and 80–100 mmHg diastolic. Should hypertension worsen during pregnancy, baseline evaluation of renal function (see below) is necessary to help differentiate the effects of chronic hypertension from those of superimposed preeclampsia. There are no convincing data that the treatment of mild chronic hypertension improves perinatal outcome.

GESTATIONAL HYPERTENSION

The development of elevated blood pressure after 20 weeks of pregnancy or in the first 24 h post-partum in the absence of preexisting chronic hypertension or proteinuria is referred to as *gestational hypertension*. Mild gestational hypertension that does not progress to preeclampsia has not been associated with adverse pregnancy outcome or adverse long-term prognosis.

RENAL DISEASE

Normal pregnancy is characterized by an increase in glomerular filtration rate and creatinine clearance. This increase occurs secondary to a rise in renal plasma flow and increased glomerular filtration pressures. Patients with underlying renal disease and hypertension may expect a worsening of hypertension during pregnancy. If superimposed preeclampsia develops, the additional endothelial injury results in a capillary leak syndrome that may make management challenging. In general, patients with underlying renal disease and hypertension benefit from aggressive management of blood pressure. Preconception counseling is also essential for these patients so that accurate risk assessment and medication changes can occur prior to pregnancy. In general, a prepregnancy serum creatinine level <133 μmol/L (<1.5 mg/dL) is associated with a favorable prognosis. When renal disease worsens during pregnancy, close collaboration between the internist and the maternal-fetal medicine specialist is essential so that decisions regarding delivery can be weighed to balance the sequelae of prematurity for the neonate versus long-term sequelae for the mother with respect to future renal function.

CARDIAC DISEASE

VALVULAR HEART DISEASE

(See also Chaps. 256–263) Valvular heart disease is the most common cardiac problem complicating pregnancy.

Mitral Stenosis This is the valvular disease most likely to cause death during pregnancy. The pregnancy-induced increase in blood volume, cardiac output, and tachycardia can increase the transmitral pressure gradient and cause pulmonary edema in women with mitral stenosis. Women with moderate to severe mitral stenosis (mitral valve area ≤1.5 cm^2) who are planning pregnancy and have either symptomatic disease or pulmonary hypertension should undergo valvuloplasty prior to conception, preferably with percutaneous balloon valvotomy (PBV). Pregnancy associated with long-standing mitral stenosis may result in pulmonary hypertension. Sudden death has been reported when hypovolemia occurs. Careful control of heart rate, especially during labor and delivery, minimizes the impact of tachycardia and reduced ventricular filling times on cardiac function. Pregnant women with mitral stenosis are at increased risk for the development of atrial fibrillation and other tachyarrhythmias. The immediate postpartum period is a time of particular concern secondary to rapid volume shifts. Careful monitoring of cardiac and fluid status should be observed.

Mitral Regurgitation and Aortic Regurgitation and Stenosis The pregnancy-induced decrease in systemic vascular resistance reduces the risk of cardiac failure with these conditions, especially in women with chronic lesions. Acute onset of mitral or aortic regurgitation may not be well tolerated during pregnancy. For women with severe aortic stenosis, treatment before pregnancy should be considered for a peak-to-peak valve gradient >50 mmHg. In women with aortic stenosis and a mean valve gradient <25 mmHg, pregnancy is likely to be well tolerated. For women with mitral or aortic regurgitation and left ventricular dysfunction (LVEF <30%) pregnancy should be avoided.

CONGENITAL HEART DISEASE

(See also Chap. 264) Reparative surgery has markedly increased the number of adult women with surgically repaired congenital heart disease. Maternal morbidity and mortality are greater among these women than among those without surgical cardiac repair. When pregnant, these patients should be jointly managed by a cardiologist and an obstetrician familiar with these problems. The presence of a congenital cardiac lesion in the mother increases the risk of congenital cardiac disease in the newborn. Prenatal screening of the fetus for congenital cardiac disease with ultrasound is recommended.

OTHER CARDIAC DISORDERS

Supraventricular tachycardia (Chap. 241) is a common cardiac complication of pregnancy. Treatment is the same as in the nonpregnant patient, and fetal tolerance of medications such as adenosine and calcium channel blockers is acceptable. When necessary, pharmacologic or electric cardioversion may be performed to improve cardiac performance and reduce symptoms. This intervention is generally well tolerated by mother and fetus.

Peripartum cardiomyopathy (Chap. 254) is an uncommon disorder of pregnancy and its etiology remains unknown. Approximately 10% of women with peripartum cardiomyopathy carry a truncating mutation in the gene encoding the titin sarcomere protein. Treatment is directed toward symptomatic relief and improvement of cardiac function. Many patients recover completely; others are left with progressive dilated cardiomyopathy. Recurrence in a subsequent pregnancy has been reported, and women who do not have normal baseline left-ventricular function after an episode of peripartum cardiomyopathy should be counseled to avoid pregnancy.

SPECIFIC HIGH-RISK CARDIAC LESIONS

Marfan Syndrome (See also Chap. 406) This autosomal dominant disease is associated with an increased risk of aortic dissection and rupture. An aortic root diameter <40 mm is associated with a favorable outcome of pregnancy; conversely, an aortic root diameter >40 mm is associated with an increased risk of aortic dissection. Prophylactic therapy with beta blockers has been advocated to reduce aortic dilation and the risk of dissection. A "cardiac delivery" with reduced pushing and early intervention with operative delivery is often recommended to reduce increases in aortic wall stress caused by the Valsalva maneuver.

Ehlers-Danlos syndrome (*EDS*) may be associated with premature labor, and in type IV EDS there is increased risk of organ or vascular rupture that may cause death. For women with vascular EDS, pregnancy is relatively contraindicated because of the high risk of vascular and uterine rupture.

Pulmonary Hypertension (See also Chap. 277) Maternal mortality in the setting of severe pulmonary hypertension is high, and primary pulmonary hypertension is a contraindication to pregnancy. Termination of pregnancy may be advisable in these circumstances to preserve the life of the mother. In the *Eisenmenger syndrome*, i.e., the combination of pulmonary hypertension with right-to-left shunting due to congenital abnormalities (Chap. 264), maternal and fetal deaths occur frequently. Systemic hypotension may occur after blood loss, prolonged Valsalva maneuver, or regional anesthesia; sudden death secondary to hypotension is a dreaded complication. Management of these patients is challenging, and invasive hemodynamic monitoring during labor and delivery is recommended in severe cases.

In patients with pulmonary hypertension, vaginal delivery is less stressful hemodynamically than cesarean section, which should be reserved for accepted obstetric indications.

DEEP VENOUS THROMBOSIS AND PULMONARY EMBOLISM

(See also Chap. 273) Pregnancy is associated with venous stasis, endothelial injury and a hypercoagulable state. Inherited thrombophilias and the presence of antiphospholipid antibodies increase the risk of venous thromboembolism (VTE) in pregnancy. Deep venous thrombosis (DVT) or pulmonary embolism (PE) occurs in about 1 in 500 pregnancies, with DVT being three times more common than PE. VTE occurs more commonly in the 6 weeks post-partum than antepartum. In pregnant women, most unilateral DVTs occur in the left leg because the left iliac vein is compressed by the right iliac artery and the uterus compresses the inferior vena cava.

TREATMENT

Deep Venous Thrombosis

Aggressive diagnosis and management of DVT and suspected pulmonary embolism optimize the outcome for mother and fetus. In general, all diagnostic and therapeutic modalities afforded that the nonpregnant patient should be utilized in pregnancy except for D-dimer measurement, in which values are elevated in normal pregnancy. Anticoagulant therapy with low-molecular-weight heparin (LMWH) or unfractionated heparin is indicated in pregnant women with DVT. LMWH may be associated with an increased risk of epidural hematoma in women receiving an epidural anesthetic in labor and must be discontinued at least 24 h before placement of an epidural catheter. Warfarin therapy is contraindicated in the first trimester due to its association with fetal chondrodysplasia punctata. In the second and third trimesters, warfarin may cause fetal optic atrophy and mental retardation. In pregnancy the use of warfarin is restricted to women with mechanical heart valves. Warfarin is not contraindicated in breast-feeding women. For women at moderate or high risk of DVT who have a cesarean delivery, mechanical and/or pharmacologic prophylaxis is warranted.

ENDOCRINE DISORDERS

DIABETES MELLITUS

(See also Chaps. 396–398) In pregnancy, the fetoplacental unit induces major metabolic changes, the purpose of which is to shunt glucose and amino acids to the fetus while the mother uses ketones and triglycerides to fuel her metabolic needs. These metabolic changes are accompanied by maternal insulin resistance caused in part by placental production of steroids, a growth hormone variant, and placental lactogen. Although pregnancy has been referred to as a state of "accelerated starvation," it is better characterized as "accelerated ketosis." In pregnancy, after an overnight fast, plasma glucose is lower by 0.8–1.1 mmol/L (15–20 mg/dL) than in the nonpregnant state. This difference is due to the use of glucose by the fetus. In early pregnancy, fasting may result in circulating glucose concentrations in the range of 2.2 mmol/L (40 mg/dL) and may be associated with symptoms of hypoglycemia. In contrast to the decrease in maternal glucose concentration, plasma hydroxybutyrate and acetoacetate levels rise to two to four times normal after a fast.

TREATMENT

Diabetes Mellitus in Pregnancy

Pregnancy complicated by diabetes mellitus is associated with higher maternal and perinatal morbidity and mortality rates. Preconception counseling and treatment are important for the diabetic patient contemplating pregnancy and can reduce the risk of congenital malformations and improve pregnancy outcome. Folate supplementation reduces the incidence of fetal neural tube defects, which occur with greater frequency in fetuses of diabetic mothers. In addition, optimizing glucose control during key periods of organogenesis reduces other congenital anomalies, including sacral agenesis, caudal dysplasia, renal agenesis, and ventricular septal defect.

Once pregnancy is established, glucose control should be managed more aggressively than in the nonpregnant state. In addition to dietary changes, this enhanced management requires more frequent blood glucose monitoring and often involves additional injections of insulin or conversion to an insulin pump. Fasting blood glucose levels should be maintained at <5.8 mmol/L (<105 mg/dL), with avoidance of values >7.8 mmol/L (140 mg/dL). Sequential

measurement of hemoglobin A1c is of minimal value for monitoring glucose control during pregnancy because of the higher rate of red blood cell turnover during pregnancy. Commencing in the third trimester, regular surveillance of maternal glucose control as well as assessment of fetal growth (obstetric sonography) and fetoplacental oxygenation (fetal heart rate monitoring or biophysical profile) optimize pregnancy outcome. Pregnant diabetic patients without vascular disease are at greater risk for delivering a macrosomic fetus, and attention to fetal growth via clinical and ultrasound examination is important. Fetal macrosomia is associated with an increased risk of maternal and fetal birth trauma, including permanent injury to the brachial plexus. Pregnant women with diabetes have an increased risk of developing preeclampsia, and those with vascular disease are at greater risk for developing intrauterine growth restriction, which is associated with an increased risk of fetal and neonatal death. Excellent pregnancy outcomes in patients with diabetic nephropathy and proliferative retinopathy have been reported with aggressive glucose control and intensive maternal and fetal surveillance.

As pregnancy progresses, glycemic control may become more difficult to achieve due to an increase in insulin resistance. In pregnant women with Type 1 diabetes, closed-loop insulin delivery with both continuous interstitial glucose monitoring and sensor-augmented insulin pump therapy is helpful in normalizing circulating glucose with few episodes of hypoglycemia. In general, efforts to control glucose and avoid preterm delivery result in the best overall outcome for both mother and newborn. Preterm delivery is generally performed only for the usual obstetric indications (e.g., preeclampsia, fetal growth restriction, non-reassuring fetal testing) or for worsening maternal renal or active proliferative retinopathy.

GESTATIONAL DIABETES (GDM)

GDM occurs in ~4% of pregnancies. Because about 90% of women have at least one risk factor for GDM, all pregnant women should be screened for GDM. A typical two-step strategy for establishing the diagnosis of GDM is performed at 24–28 weeks of gestation and involves administration of a 50-g oral glucose challenge with a single serum glucose measurement at 60 min. If the plasma glucose is <7.8 mmol/L (<130 mg/dL) the test is considered normal. Plasma glucose >7.8 mmol/L (>130 mg/dL) warrants administration of a 100-g oral glucose challenge with plasma glucose measurements obtained in the fasting state and at 1, 2, and 3 h. Normal plasma glucose concentrations at these time points are <5.3 mmol/L (<95 mg/dL), <10 mmol/L (<180 mg/dL), <8.6 mmol/L (<155 mg/dL), and <7.8 mmol/L (<140 mg/dL) as the upper norms for a 3-h glucose tolerance test. Two elevated glucose values indicate a positive test. Adverse pregnancy outcomes for mother and fetus appear to increase with glucose as a continuous variable; thus it is challenging to define the optimal threshold for establishing the diagnosis of GDM.

Pregnant women with GDM are at increased risk of stillbirth, preeclampsia, and delivery of infants who are large for their gestational age, with resulting birth lacerations, shoulder dystocia, and birth trauma including brachial plexus injury. These fetuses are at risk of hypoglycemia, hyperbilirubinemia, and polycythemia. Tight control of blood sugar during pregnancy and labor can reduce these risks.

TREATMENT

Gestational Diabetes

Treatment of GDM with a two-step strategy—dietary intervention followed by insulin injections if diet alone does not adequately control blood sugar (fasting glucose <5.6 mmol/L [<100 mg/dL] and 2-h postprandial glucose <7.0 mmol/L [<126 mg/dL])—is associated with a decreased risk of birth trauma for the fetus. Oral hypoglycemic agents such as glyburide and metformin have become more commonly utilized for managing GDM refractory to nutritional management, but most experts favor insulin therapy. For women with GDM, there is a 40% risk of being diagnosed with diabetes within the 10 years after the index pregnancy. All women with GDM should have a formal glucose tolerance test (GTT) to screen for T2DM at ~6 weeks post-partum. In women with a history of GDM, exercise, weight loss, and treatment with metformin reduce the risk of developing diabetes. Lactation also reduces the risk of GDM progressing to T2DM. All women with a history of GDM should be counseled about prevention strategies and evaluated regularly for diabetes.

OBESITY

(See also Chap. 395) Pregnant women who are obese have an increased risk of stillbirth, congenital fetal malformations, GDM, preeclampsia, urinary tract infections, preterm and post-date delivery, and cesarean delivery. Women contemplating pregnancy should attempt to attain a healthy weight prior to conception. For morbidly obese women who have not been able to lose weight with lifestyle changes, bariatric surgery reduces the risks for GDM, macrosomia, and preterm delivery. Following bariatric surgery, women should delay conception for 1 year to avoid pregnancy during an interval of rapid metabolic changes. The National Academy of Medicine guidelines for weight gain during pregnancy recommend that for BMI ranges of <18.5, 18.5–24.9, 25.0–29.9, and ≥30 kg/m^2, weight gain targets should be 12.5–18 kg, 11.5–16 kg, 7–11.5 kg, and 5–9 kg, respectively.

THYROID DISEASE

(See also Chap. 375) In pregnancy, the estrogen-induced increase in thyroxine-binding globulin increases circulating levels of total T_3 and total T_4. Placental human chorionic gonadotropin (hCG) directly stimulates the thyroid causing an increase in free T_3 and T_4. Interpretation of the measurement of free T_4, free T_3, and thyroid-stimulating hormone (TSH) should use trimester-specific ranges.

TREATMENT

Hyperthyroidism in Pregnancy

HYPERTHYROIDISM

Methimazole crosses the placenta to a greater degree than propylthiouracil and has been associated with fetal aplasia cutis. However, propylthiouracil can be associated with liver failure. Some experts recommend propylthiouracil in the first trimester and methimazole thereafter. Radioiodine should not be used during pregnancy, either for scanning or for treatment, because of effects on the fetal thyroid. In emergent circumstances, additional treatment with beta blockers may be necessary. Hyperthyroidism is most difficult to control in the first trimester of pregnancy and easiest to control in the third trimester. In women with high-titer thyroid stimulating antibodies, the newborn may be born with neonatal Graves' disease.

HYPOTHYROIDISM

The goal of therapy for hypothyroidism is to maintain the serum TSH in the normal range, and thyroxine is the drug of choice. During pregnancy, the dose of thyroxine required to keep the TSH in the normal range rises. In one study, the mean replacement dose of thyroxine required to maintain the TSH in the normal range was 0.1 mg daily before pregnancy and increased to 0.15 mg daily during pregnancy. Since the increased thyroxine requirement occurs as early as the fifth week of pregnancy, one approach is to increase the thyroxine dose by 30% (two additional pills weekly) as soon as pregnancy is diagnosed and then adjust the dose by serial measurements of TSH.

HEMATOLOGIC DISORDERS

Pregnancy has been described as a state of physiologic anemia. Part of the reduction in hemoglobin concentration is dilutional, but iron and folate deficiencies are major causes of correctable anemia during pregnancy.

In populations at high risk for hemoglobinopathies **(Chap. 94)**, hemoglobin electrophoresis should be performed as part of the prenatal screen. Hemoglobinopathies can be associated with increased maternal and fetal morbidity and mortality. Management is tailored

to the specific hemoglobinopathy and is generally the same for both pregnant and nonpregnant women. Prenatal diagnosis of hemoglobinopathies in the fetus is readily available and should be discussed with prospective parents either prior to or early in pregnancy.

Thrombocytopenia occurs commonly during pregnancy. The majority of cases are benign gestational thrombocytopenias, but the differential diagnosis should include immune thrombocytopenia **(Chap. 111)**, preeclampsia, and thrombotic thrombocytopenic purpura. Benign gestational thrombocytopenia is unlikely if the platelet count is <100,000 per μL.

NEOPLASIA

Cancer complicates ~1 in every 1000 pregnancies. Of all the cancers that occur in women, <1% complicate pregnancies. The four cancers that occur most commonly in pregnancy are cervical cancer, breast cancer, melanoma, and lymphomas (particularly Hodgkin's lymphoma); however, virtually every form of cancer has been reported in pregnant women **(Table 466-1)**. In addition to cancers developing in other organs of the mother, gestational trophoblastic tumors can arise from the placenta.

Managing cancer in a pregnant woman is complex. One must take into account (1) the possible influence of the pregnancy on the natural history of the cancer, (2) effects on the mother and fetus of complications from the malignancy (e.g., anorexia, nausea, vomiting, malnutrition), (3) potential effects of diagnostic and staging procedures, and (4) potential effects of cancer treatments on both the mother and the developing fetus. Generally, the management that optimizes maternal physiology is also best for the fetus. The dilemma occasionally arises that what is best for the mother may be harmful to the fetus, and what is best for the fetus may compromise the ultimate prognosis for the mother. The best way to approach management of a pregnant woman with cancer is to ask, "What would one do in this clinical situation if she was not pregnant? Then, which, if any aspect of those plans need to be modified because she is pregnant?"

TREATMENT

Special Therapeutic Considerations in Pregnancy

Exposure of developing fetuses to ionizing radiation may cause adverse fetal effects; awareness of this potential toxicity has resulted in a disproportionate aversion to diagnostic imaging in pregnancy. The fetus is most sensitive to teratogenesis during organogenesis in the first trimester. Imaging that uses ionizing radiation should not be done without a compelling reason and due consideration to obtaining the necessary information by alternative imaging modalities. Exposure to diagnostic and therapeutic radionuclides, especially radioactive iodine, poses unique risks, but a full discussion of these is beyond the scope of this chapter.

Generally, toxic chemotherapy should be avoided during pregnancy, if at all possible. It should virtually never be given in the first trimester. A variety of single agents and combinations have been administered in the second and third trimesters, without a high frequency of toxic effects to the pregnancy or the fetus, but data on safety are sparse. A database on the risks associated with individual chemotherapy agents is available (*http://ntp.niehs.nih.gov/ntp/ohat/cancer_chemo_preg/chemopregnancy_monofinal_508.pdf*). If the malignancy is slowly progressive, and if the patient is near her delivery date, and if waiting until delivery to begin treatment is not anticipated to compromise maternal prognosis, then delaying treatment until after delivery to avoid fetal exposure to chemotherapy is desirable. If there is a greater sense of urgency to begin definitive treatment to avoid compromising maternal prognosis, and the patient is beyond 24 weeks of gestation but remote from her delivery date, then treatment (surgical, medical, or both) might be initiated during pregnancy and plans made to deliver the fetus early to avoid exposure to more chemotherapy than absolutely necessary. Since neonatal prognosis is most closely linked to gestational age at delivery, decisions regarding timing of delivery should include input from Maternal-Fetal Medicine, Neonatology, and Oncology. Finally, if the patient is in her first trimester and toxic chemotherapy must be initiated promptly to avoid a very poor maternal outcome, then it may be necessary to consider therapeutic abortion to avoid maternal disaster and fetal survival with injury resulting in long-term morbid sequelae. In general, pregnancy has relatively little or no impact on the natural history of malignancies, despite the hormonal influences. Spread of the mother's cancer to the fetus (so-called *vertical transmission*) is exceedingly rare.

TABLE 466-1 Incidence of Malignant Tumors During Gestation

TUMOR TYPE	INCIDENCE PER 10,000 PREGNANCIES[a]	% OF CASES[b]
Breast cancer	1–3	25%
Cervical cancer	1.2–4.5	25%
Thyroid cancer	1.2	15%
Hodgkin's disease	1.6	10%
Melanoma	1–2.6	8%
Ovarian cancer	0.8	2%
All sites	10	100%

[a]These are estimates based on extrapolations from a review of more than 3 million pregnancies (LH Smith et al: Am J Obstet Gynecol 184:1504, 2001).
[b]Based on accumulating case reports from the literature; the precision of these data is not high.

NEUROLOGIC DISORDERS

For women with epilepsy planning pregnancy, consideration should be given to switching from valproate, a known teratogen, to another medication. If valproate is continued during pregnancy, folic acid supplementation should be increased to 4 mg daily.

Patients with preexisting *multiple sclerosis* **(Chap. 436)** experience a gradual decrease in the risk of relapses as pregnancy progresses and, conversely, an increase in attack risk during the postpartum period. Disease-modifying agents, including interferon β, should *not* be administered to pregnant multiple sclerosis patients, but moderate or severe relapses can be safely treated with pulse glucocorticoid therapy. Finally, certain tumors, particularly pituitary adenoma and meningioma **(Chap. 373)**, may manifest during pregnancy because of accelerated growth, possibly driven by hormonal factors.

Peripheral nerve disorders associated with pregnancy include *Bell's palsy* (idiopathic facial paralysis) **(Chap. 438)**, which is approximately threefold more likely to occur during the third trimester and immediate postpartum period than in the general population. Therapy with glucocorticoids should follow the guidelines established for nonpregnant patients. Entrapment neuropathies are common in the later stages of pregnancy, presumably as a result of fluid retention. *Carpal tunnel syndrome* (median nerve) presents first as pain and paresthesia in the hand (often worse at night) and later with weakness in the thenar muscles. Treatment is generally conservative; wrist splints may be helpful, and glucocorticoid injections or surgical section of the carpal tunnel can usually be postponed. *Meralgia paresthetica* (lateral femoral cutaneous nerve entrapment) consists of pain and numbness in the lateral aspect of the thigh without weakness. Patients are usually reassured to learn that these symptoms are benign and can be expected to remit spontaneously after the pregnancy has been completed. *Restless leg syndrome* is the most common peripheral nerve and movement disorder in pregnancy. Disordered iron metabolism is the suspected etiology. Management is expectant in most cases.

GASTROINTESTINAL AND LIVER DISEASE

Up to 90% of pregnant women experience nausea and vomiting during the first trimester of pregnancy. *Hyperemesis gravidarum* is a severe form that prevents adequate fluid and nutritional intake and may require hospitalization to prevent dehydration and malnutrition.

Crohn's disease may be associated with exacerbations in the second and third trimesters. Ulcerative colitis is associated with disease exacerbations in the first trimester and during the early postpartum period.

Medical management of these diseases during pregnancy is similar to management in the nonpregnant state (Chap. 319).

Exacerbation of gallbladder disease is common during pregnancy. In part, this aggravation may be due to pregnancy-induced alteration in the metabolism of bile and fatty acids. Intrahepatic cholestasis of pregnancy is generally a third-trimester event. Profound pruritus may accompany this condition, and it may be associated with increased fetal mortality. Placental bile salt deposition may contribute to progressive uteroplacental insufficiency. Therefore, regular fetal surveillance should be undertaken once the diagnosis of intrahepatic cholestasis is made, and delivery should be planned once the fetus reaches about 37 weeks of gestation. Favorable results with ursodiol have been reported.

Acute fatty liver is a rare complication of pregnancy. Frequently confused with the HELLP syndrome (see "Preeclampsia" above) and severe preeclampsia, the diagnosis of acute fatty liver of pregnancy may be facilitated by imaging studies and laboratory evaluation. Acute fatty liver of pregnancy is generally characterized by markedly increased serum levels of bilirubin and ammonia and by hypoglycemia. Management of acute fatty liver of pregnancy is supportive; recurrence in subsequent pregnancies has been reported.

All pregnant women should be screened for hepatitis B. This information is important for pediatricians after delivery of the infant. All infants receive hepatitis B vaccine. Infants born to mothers who are carriers of hepatitis B surface antigen should also receive hepatitis B immune globulin as soon after birth as possible and preferably within the first 72 h. Screening for hepatitis C is recommended for individuals at high risk for exposure.

INFECTIONS

BACTERIAL INFECTIONS

Other than bacterial vaginosis, the most common bacterial infections during pregnancy involve the urinary tract (Chap. 130). Many pregnant women have asymptomatic bacteriuria, most likely due to stasis caused by progestational effects on ureteral and bladder smooth muscle and later in pregnancy due to compression effects of the enlarging uterus. In itself, this condition is not associated with an adverse outcome of pregnancy. If asymptomatic bacteriuria is left untreated, symptomatic pyelonephritis may occur. Indeed, ~75% of pregnancy-associated pyelonephritis cases are the result of untreated asymptomatic bacteriuria. All pregnant women should be screened with a urine culture for asymptomatic bacteriuria at the first prenatal visit. Subsequent screening with nitrite/leukocyte esterase strips is indicated for high-risk women, such as those with sickle cell trait or a history of urinary tract infections. All women with positive screens should be treated. Pregnant women who develop pyelonephritis need inpatient IV antibiotic administration due to the elevated risk of urosepsis and acute respiratory distress syndrome in pregnancy. Pregnant women with recurrent urinary tract infections, or one episode of pyelonephritis, should be considered for daily antibiotic suppressive treatment throughout the remainder of their pregnancy.

All pregnant patients are screened prenatally for syphilis, gonorrhea, and chlamydial infections, and the detection of any of these should result in prompt evaluation and treatment (Chaps. 151 and 184).

VIRAL INFECTIONS

Zika Virus (ZV) ZV can be transmitted from mother to fetus throughout gestation and often results in fetal death, severe microcephaly, or other malformations of the central nervous system. Pregnant symptomatic women with relevant epidemiologic exposure within 2 weeks of symptom onset should have serum and urine tested for ZV ribonucleic acid by real-time reverse transcriptase-polymerase chain reaction (RT-PCR). Testing 2–12 weeks after symptom onset utilizes serum measurement of Zika and dengue virus IgM. Sequential obstetrical ultrasound is recommended to assess for fetal growth and anomalies. Couples considering pregnancy should avoid travel to areas with known mosquito transmission of ZV.

Influenza (See also Chap. 195) Pregnant women with influenza are at increased risk of serious complications and death. All women who are pregnant or plan to become pregnant in the near future should receive inactivated influenza vaccine. The prompt initiation of antiviral treatment is recommended for pregnant women in whom influenza is suspected. Treatment can be reconsidered once the results of high-sensitivity tests are available. Prompt initiation of treatment lowers the risk of admission to an intensive care unit and death.

Cytomegalovirus Infection The most common cause of congenital viral infection in the United States is cytomegalovirus (CMV) (Chap. 190). As many as 50–90% of women of childbearing age have antibodies to CMV, but only rarely does CMV reactivation result in neonatal infection. More commonly, primary CMV infection during pregnancy creates a risk of congenital CMV. No currently accepted treatment of CMV infection during pregnancy has been demonstrated to protect the fetus effectively. Moreover, it is difficult to predict which fetus will sustain a life-threatening CMV infection. Severe CMV disease in the newborn is characterized most often by petechiae, hepatosplenomegaly, and jaundice. Chorioretinitis, microcephaly, intracranial calcifications, hepatitis, hemolytic anemia, and purpura may also develop. CNS involvement, resulting in the development of psychomotor, ocular, auditory, and dental abnormalities over time, has been described. Women with a primary CMV infection should delay conception for 6 months.

Rubella (See also Chap. 201) Rubella virus is a known teratogen; first-trimester rubella carries a high risk of fetal anomalies, though the risk significantly decreases later in pregnancy. Congenital rubella may be diagnosed by percutaneous umbilical-blood sampling with the detection of IgM antibodies in fetal blood. All pregnant women and all women of childbearing age should be tested for their immune status to rubella. All women who might become pregnant and who are not immune to rubella should be vaccinated at least 3 months before conception.

Herpesvirus Infection (See also Chap. 187) The acquisition of genital herpes during pregnancy is associated with spontaneous abortion, prematurity, and congenital and neonatal herpes. A cohort study of pregnant women without evidence of previous herpesvirus infection demonstrated that ~2% acquired a new herpesvirus infection during the pregnancy. Approximately 60% of the newly infected women had no clinical symptoms. Infection occurred with equal frequency in all three trimesters. If herpesvirus seroconversion occurred early in pregnancy, the risk of transmission to the newborn was very low. In women who acquired genital herpes shortly before delivery, the risk of transmission was high. The risk of active genital herpes lesions at term can be reduced by prescribing acyclovir for the last 4 weeks of pregnancy to all women who had an episode of genital herpes during the pregnancy.

Herpesvirus infection in the newborn can be devastating. Disseminated neonatal herpes carries with it high mortality and morbidity rates from CNS involvement. It is recommended that pregnant women with active genital herpes lesions at the time of presentation in labor be delivered by cesarean section.

Parvovirus Infection (See also Chap. 192) Parvovirus infection (caused by human parvovirus B19) may occur during pregnancy. It rarely causes sequelae, but susceptible women infected during pregnancy may be at risk for fetal hydrops secondary to erythroid aplasia and profound anemia.

HIV Infection (See also Chap. 197) The predominant cause of HIV infection in children is transmission of the virus from mother to newborn during the perinatal period. All pregnant women should be screened for HIV infection. Factors that increase the risk of mother-to-newborn transmission include high maternal viral load, low maternal CD4+ T cell count, prolonged labor, prolonged duration of membrane rupture, and the presence of other genital tract infections, such as syphilis or herpes. Prior to the widespread use of antiretroviral treatment, the perinatal transmission rate was in the range of 20%. In women with a good response to antiretroviral treatment, the transmission rate is about 1%. Measurement of maternal plasma HIV RNA copy number guides the decision for vaginal versus cesarean delivery. For women with

<1000 copies of plasma HIV RNA/mL who are receiving combination antiretroviral therapy, the risk of transmission to the newborn is ~1% regardless of mode of delivery or duration of membrane rupture. These women may elect to attempt a vaginal birth following the spontaneous onset of labor. For women with a viral load of ≥1000 copies/mL prior to 38 weeks of gestation, a scheduled prelabor cesarean at 38 weeks is recommended to reduce the risk of HIV transmission to the newborn.

VACCINATIONS

(See also Chap. 118) For rubella-nonimmune individuals contemplating pregnancy, measles-mumps-rubella vaccine should be administered, ideally at least 3 months prior to conception, but otherwise in the immediate postpartum period. In addition, pregnancy is not a contraindication for vaccination against influenza, tetanus, diphtheria, and pertussis (Tdap), and these vaccines are recommended for appropriate individuals.

MATERNAL MORTALITY

Maternal death is defined as death occurring during pregnancy or within 42 days of completion of pregnancy from a cause related to or aggravated by pregnancy, but not due to accident or incidental causes. The maternal mortality ratio is the number of maternal deaths per 100,000 live births. From 1935 to 2007, the U.S. maternal mortality ratio decreased from nearly 600/100,000 births to 12.7/100,000 births. Since 2007, the U.S. maternal mortality ratio has increased to 21.5/100,000 births. There are significant health disparities in the maternal mortality ratio. In the United States, in the period from 2005 to 2014, the maternal mortality ratios (per 100,000 live births) by race were 11.3 among Hispanic women, 14.1 among non-Hispanic white women, and 40.2 among non-Hispanic black women. The most common causes of maternal death in the United States today are pulmonary embolism, obstetric hemorrhage, hypertension, sepsis, cardiovascular conditions (including peripartum cardiomyopathy and stroke), and ectopic pregnancy. Specialists in internal medicine play an important role in national efforts to reduce the maternal mortality ratio.

As stated above, the maternal mortality ratio in the United States is about 21.5/100,000 live births. In some countries in sub-Saharan Africa and southern Asia, the maternal mortality ratio is >500/100,000 live births. The most common causes of maternal death in these countries are maternal hemorrhage, hypertensive disorders, infection, obstructed labor, and complications from unsafe pregnancy termination. The health interventions that would have the greatest impact on maternal health include improving the following components of the health system: (1) access to contraceptive services in order to space births and limit total family size; (2) access to safe pregnancy termination; (3) presence of trained birth attendants at all deliveries; and (4) transportation to emergency obstetrical centers that can provide intensive medical and surgical services, including cesarean delivery. Maternal death is a global public-health tragedy that could be mitigated with the application of modest resources.

SUMMARY

With improved diagnostic and therapeutic modalities as well as advances in the treatment of infertility, more patients with serious medical complications will be seeking to become pregnant and will require complex obstetric care. Improved outcomes of pregnancy in these women will be best attained by a team of internists, maternal-fetal medicine (high-risk obstetrics) specialists, pediatricians and anesthesiologists assembled to counsel these patients about the risks of pregnancy and to plan their treatment prior to, and following, conception. The importance of preconception counseling cannot be overstated. It is the responsibility of all physicians caring for women in the reproductive age group to assess their patients' reproductive plans as part of their overall health evaluation.

Acknowledgement

The authors are grateful to Michael F. Greene and Dan L. Longo for their contributions to the content on neoplasia in pregnancy based upon material from previous editions of Harrison's.

FURTHER READING

Brasil P et al: Zika virus infection in pregnant women in Rio de Janeiro-preliminary report. N Engl J Med 375:2321, 2016.

Esposito S et al: Chemotherapy against cancer during pregnancy: A systematic review on neonatal outcomes. Medicine (Baltimore) 95:e4899, 2016.

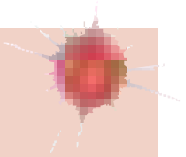
LeFevre ML et al: Low-dose aspirin use for the prevention of morbidity and mortality from preeclampsia: U.S. Preventive Services Task Force recommendations statement. Ann Int Med 161:819, 2014.

Moaddab A et al: Health care disparities and state-specific pregnancy-related mortality in the United States, 2005–2014. Obstet Gynecol 126:869, 2016.

Stewart ZA et al: Closed-loop insulin delivery during pregnancy in women with Type 1 Diabetes. N Engl J Med 375:644, 2016.

Ware JS et al: Shared genetic predisposition in peripartum and dilated cardiomyopathies. N Engl J Med 374:233, 2016.

467 Medical Evaluation of the Surgical Patient



Prashant Vaishnava, Kim A. Eagle

Cardiovascular and pulmonary complications continue to account for major morbidity and mortality in patients undergoing noncardiac surgery. Emerging evidence-based practices dictate that the internist should perform an individualized evaluation of the surgical patient to provide an accurate preoperative risk assessment and stratification that will guide optimal perioperative risk-reduction strategies. This chapter reviews cardiovascular and pulmonary preoperative risk assessment, emphasizing the goal-directed management of patients at elevated risk for adverse cardiovascular outcomes in the perioperative period. In addition, perioperative management of diabetes mellitus and prophylaxis of endocarditis and for venous thromboembolism are reviewed.

EVALUATION OF INTERMEDIATE- AND HIGH-RISK PATIENTS

Simple, standardized preoperative screening questionnaires, such as the one shown in Table 467-1, have been developed for the purpose of identifying patients at intermediate or high risk who may benefit from a more detailed clinical evaluation. Evaluation of such patients for surgery should always begin with a thorough history and physical examination and with a 12-lead resting electrocardiogram, in accordance with the American College of Cardiology/American Heart Association guidelines. The history should focus on symptoms of occult cardiac or pulmonary disease. The urgency of the surgery should be determined, as true emergency procedures are associated with unavoidably higher morbidity and mortality risk. Preoperative laboratory testing should be carried out only for specific clinical conditions, as noted during clinical examination. Thus, healthy patients of any age who are undergoing elective surgical procedures without coexisting medical conditions should not require any testing unless the degree of surgical stress may result in unusual changes from the baseline state.

PREOPERATIVE CARDIAC RISK ASSESSMENT

A stepwise approach to cardiac risk assessment and stratification in patients undergoing noncardiac surgery is illustrated in Fig. 467-1. The evaluation begins with characterization of the combined surgical and clinical risk into categories of low (<1%) and elevated risk for major adverse cardiovascular events (MACE). Select surgeries are associated with very low risk for MACE; these surgeries and procedures include select ophthalmologic surgeries (e.g., cataract surgery), select endoscopic procedures, and select superficial procedures. Patients undergoing these low-risk procedures should proceed to surgery without

TABLE 467-1 Standardized Preoperative Questionnaire[a]

1. Age, weight, height
2. Are you:
   Female and 55 years of age or older or male and 45 years of age of older?
   If yes, are you 70 years of age or older?
3. Do you take anticoagulant medications ("blood thinners")?
4. Do you have or have you had any of the following heart-related conditions?
   Heart disease
   Heart attack within the last 6 months
   Angina (chest pain)
   Irregular heartbeat
   Heart failure
5. Do you have or have you ever had any of the following?
   Rheumatoid arthritis
   Kidney disease
   Liver disease
   Diabetes
6. Do you get short of breath when you lie flat?
7. Are you currently on oxygen treatment?
8. Do you have a chronic cough that produces any discharge or fluid?
9. Do you have lung problems or diseases?
10. Have you or any blood member of your family ever had a problem other than nausea with any anesthesia?
    If yes, describe:
11. If female, is it possible that you are pregnant?
    Pregnancy test:
    Please list date of last menstrual period:

[a]University of Michigan Health System patient information report. Patients who answer yes to any of questions 2–9 should receive a more detailed clinical evaluation.

Source: Adapted from KK Tremper, P Benedict: Anesthesiology 92:1212, 2000; with permission.

further testing. Clinical risk may be estimated with the American College of Surgeons National Surgical Quality Improvement Program (NSQIP) risk calculator (*http://www.riskcalculator.facs.org*) or with calculation of the Revised Cardiac Risk Index (RCRI).

Previous studies have compared several cardiac risk indices. The American College of Surgeons' National Surgical Quality Improvement Program prospective database has identified five predictors of perioperative myocardial infarction (MI) and cardiac arrest based on increasing age, American Society of Anesthesiologists class, type of surgery, dependent functional status, and abnormal serum creatinine level. However, given its accuracy and simplicity, the RCRI **(Table 467-2)** is often the favored risk index. The RCRI relies on the presence or absence of six identifiable predictive factors: high-risk surgery, ischemic heart disease, congestive heart failure, cerebrovascular disease, diabetes mellitus treated with insulin, and renal insufficiency with a creatinine >2.0 mg/dL. Each of these predictors is assigned one point. The risk of major cardiac events—defined as MI, pulmonary edema, ventricular fibrillation or primary cardiac arrest, and complete heart block—can then be predicted. Based on the presence of none, one, two, three, or more of these clinical predictors, the rate of development of one of these four major cardiac events is estimated to be 0.4, 0.9, 7, and 11%, respectively **(Fig. 467-2)**. An RCRI score of 0 signifies a 0.4–0.5% risk of cardiac events; RCRI 1, 0.9–1.3%; RCRI 2, 4–7%; and RCRI ≥3, 9–11%. The clinical utility of the RCRI is to identify patients with three or more predictors who are at very high risk (≥11%) for cardiac complications and who may benefit from further risk stratification with noninvasive cardiac testing, initiation of preoperative preventive medical management, or avoidance of surgery.

For patients at elevated combined clinical and surgical risk for MACE, the stepwise perioperative cardiac assessment for coronary artery disease (CAD) proceeds with consideration of functional capacity. Participation in activities of daily living offers an expression of functional capacity, often expressed in terms of metabolic equivalents (METs). For predicting perioperative events, poor exercise tolerance has been defined as the inability to walk four blocks or climb two flights of stairs at a normal pace or to meet a MET level of 4 (e.g., carrying objects of 15–20 lb or playing golf or doubles tennis) because of the development of dyspnea, angina, or excessive fatigue **(Table 467-3)**. Patients with moderate or greater (≥4 METs) functional capacity (e.g., climbing up a flight of stairs, walking up a hill, or walking on level ground at 4 mph) generally should not undergo further non-invasive cardiac testing prior to elective non-cardiac surgery. Those patients with poor (<4 METs) or unknown functional capacity should undergo pharmacological stress testing if the results of such testing would impact decision-making or perioperative care.

PREOPERATIVE NONINVASIVE CARDIAC TESTING FOR RISK STRATIFICATION

There is little evidence to support widespread application of preoperative noninvasive cardiac testing for all patients undergoing major surgery. The current paradigm to guide the need for noninvasive cardiac testing is to perform such testing in patients with poor or unknown capacity if it would alter clinical management or modify perioperative care. Options for pharmacological stress testing include dobutamine stress echocardiography or myocardial perfusion imaging with coronary vasodilator stress (dipyridamole, adenosine, or regadenoson) with thallium-201 and/or technetium-99m. Routine screening with noninvasive stress testing is not recommended in patients at low risk for noncardiac surgery. Furthermore, coronary revascularization before noncardiac surgery is not recommended for the express purpose of reducing perioperative cardiac events. That said, revascularization before noncardiac surgery should be considered in patients if it would be indicated regardless of the surgery planned and instead according to clinical practice guidelines. In the Coronary Artery Revascular Prophylaxis trial, there were no differences in perioperative and long-term cardiac outcomes with or without preoperative coronary revascularization; of note, patients with left main disease were excluded.

RISK MODIFICATION: PREVENTIVE STRATEGIES TO REDUCE CARDIAC RISK

Perioperative Coronary Revascularization *Prophylactic* coronary revascularization with either coronary artery bypass grafting (CABG) or percutaneous coronary intervention (PCI) provides no short- or mid-term survival benefit for patients *without* left main CAD or three-vessel CAD in the presence of poor left ventricular systolic function and is *not recommended* for patients with stable CAD before noncardiac surgery. Although PCI is associated with lower procedural risk than is CABG in the perioperative setting, the placement of a coronary artery stent soon before noncardiac surgery may increase the risk of bleeding during surgery if dual antiplatelet therapy (DAPT) (aspirin and thienopyridine) is administered; moreover, stent placement shortly before noncardiac surgery increases the perioperative risk of MI and cardiac death due to stent thrombosis if such therapy is withdrawn prematurely **(Chap. 270)**. It is recommended that, if possible, elective noncardiac surgery be delayed 30 days after placement of a bare metal intracoronary stent and ideally for 6 months after deployment of a drug-eluting stent. Contemporary stent platforms allow for greater flexibility in the earlier interruption of DAPT; current clinical practice guidelines do suggest consideration of elective noncardiac surgery 6 months after drug eluting stent (DES) implantation if the risk of further delaying surgery exceeds the risk of stent thrombosis/myocardial ischemia. For patients who *must* undergo noncardiac surgery early (>14 days) after PCI, balloon angioplasty without stent placement appears to be a reasonable alternative because DAPT is not necessary in such patients.

PERIOPERATIVE PREVENTIVE MEDICAL THERAPIES The goal of perioperative preventive medical therapies with β-adrenergic antagonists, hydroxymethylglutaryl-coenzyme A (HMG-CoA) reductase inhibitors (statins), and antiplatelet agents is to reduce perioperative adrenergic stimulation, ischemia, and inflammation, all of which are heightened during the perioperative period.

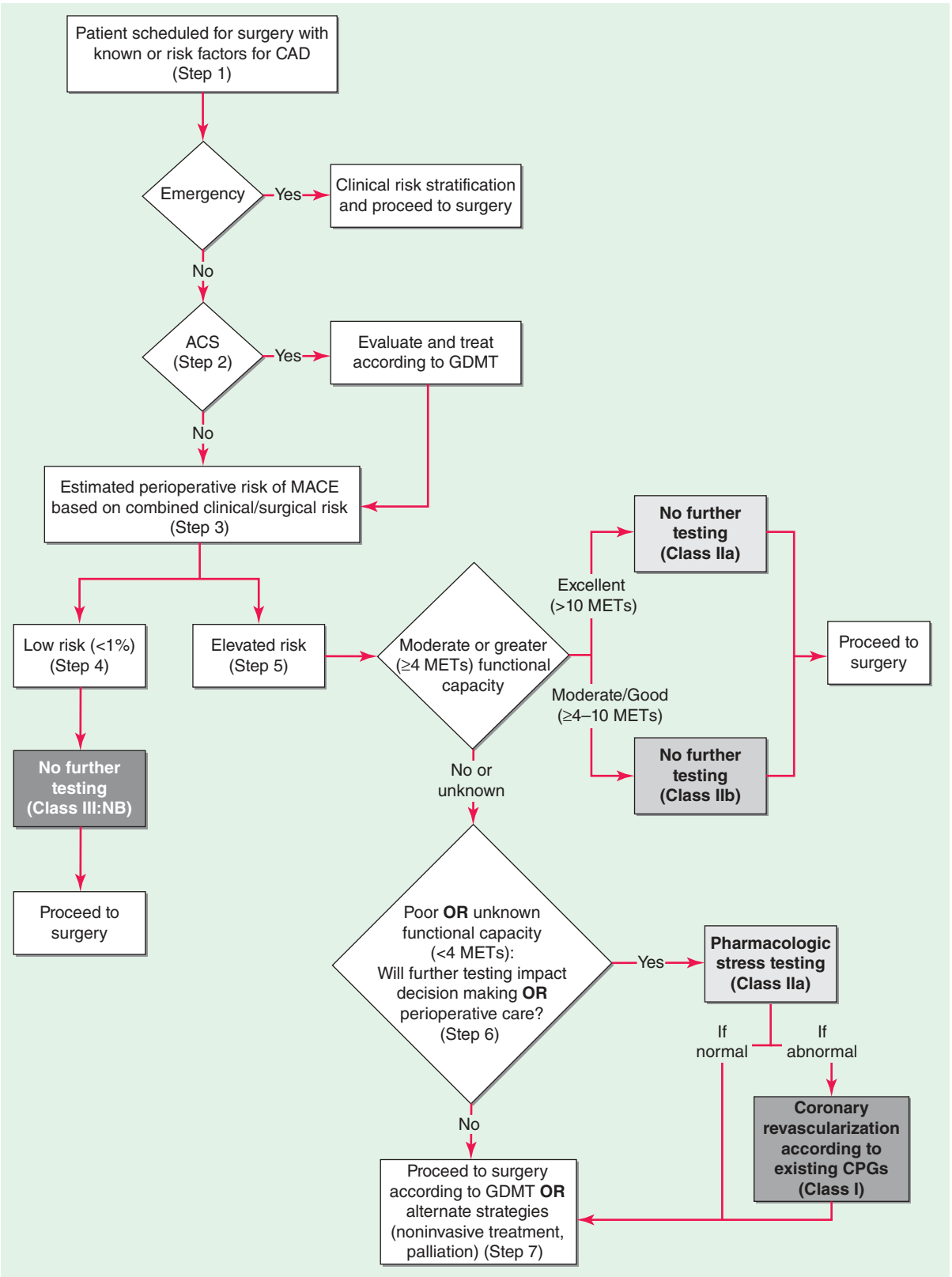


FIGURE 467-1 Composite algorithm for cardiac risk assessment and stratification in patients undergoing noncardiac surgery. Preoperative evaluation involves a stepwise clinical evaluation. Those individuals requiring emergency surgery should proceed without further risk stratification. Acute coronary syndrome (step 2) should be evaluated and treated, accordingly to goal-directed medical therapy. For patients awaiting non-emergent surgeries and without acute coronary syndrome, perioperative risk is a combination of clinical and surgical risk. Select procedures and surgeries (e.g., select endoscopic procedures) are associated with low perioperative (<1%) risk and no further clinical testing is generally necessary. For those procedures associated with elevated risk, an assessment of functional capacity informs the decision for further testing. Those individuals with moderate or greater functional capacity do not require further testing and should proceed to surgery. Individuals with poor or unknown functional capacity may require pharmacologic stress testing if it would change decision-making or perioperative care. *(From LA Fleisher et al: Circulation 2014;130:e278-e333, with permission.)*

TABLE 467-2 Clinical Markers Included in the Revised Cardiac Risk Index

High-Risk Surgical Procedures
Vascular surgery (except carotid endarterectomy)
Major intraperitoneal or intrathoracic procedures
Ischemic Heart Disease
History of myocardial infarction
Current angina considered to be ischemic
Requirement for sublingual nitroglycerin
Positive exercise test
Pathological Q-waves on ECG
History of PCI and/or CABG with current angina considered to be ischemic
Congestive Heart Failure
Left ventricular failure by physical examination
History of paroxysmal nocturnal dyspnea
History of pulmonary edema
S_3 gallop on cardiac auscultation
Bilateral rales on pulmonary auscultation
Pulmonary edema on chest x-ray
Cerebrovascular Disease
History of transient ischemic attack
History of cerebrovascular accident
Diabetes Mellitus
Treatment with insulin
Chronic Renal Insufficiency
Serum creatinine >2 mg/dL

Abbreviations: CABG, coronary artery bypass grafting; ECG, electrocardiogram; PCI, percutaneous coronary interventions.

Source: Adapted from TH Lee et al: Circulation 100:1043, 1999.

TABLE 467-3 Assessment of Cardiac Risk by Functional Status

Risk	Higher	• Has difficulty with adult activities of daily living
		• Cannot walk four blocks or up two flights of stairs or does not meet a MET level of 4
		• Is inactive but has no limitations
		• Is active: easily does vigorous tasks
	Lower	• Performs regular vigorous exercises

Source: From LA Fleisher et al: Circulation 116:1971, 2007.

B-ADRENERGIC ANTAGONISTS The use of perioperative beta blockade should be based on a thorough assessment of a patient's perioperative clinical and surgery-specific cardiac risk (e.g., as with the RCRI). The paradigm for beta blockade in the perioperative period has shifted in recent years owing, firstly, to the publication of the PeriOperative Ischemic Evaluation (POISE) trial demonstrating that, while perioperative beta blockade reduces the perioperative risk for MI, this is at the expense of increased death and stroke. Regarding POISE, this trial has been criticized for the use of an excessive dose of beta blocker in the perioperative period and one that may not be reflective of clinical practice, nor one that was titrated in the days or weeks preceding the procedure or surgery. Secondly, research misconduct has discredited the Dutch Echocardiographic Cardiac Risk Evaluation Applying Stress Echocardiography (DECREASE) family of studies, which previously contributed to the bedrock of data supporting the use of perioperative beta blockade but have now been retracted.

Current guidelines emphasize the following key points:

1. Continuation of beta blockade in patients undergoing surgery and who have been receiving such therapy chronically.
2. Avoidance of beta-blocker withdrawal or initiation on the day of surgery.
3. Consideration of initiation of beta-blocker therapy perioperatively (ideally far enough in advance to assess safety and tolerability) in very select high-risk patients, namely, those with intermediate- or high-risk ischemia or three more RCRI risk factors.

HMG-COA REDUCTASE INHIBITORS (STATINS) A number of prospective and retrospective studies support the perioperative prophylactic use of statins for reduction of cardiac complications in patients with established atherosclerosis. For patients undergoing noncardiac surgery and currently taking statins, statin therapy *should be continued* to reduce perioperative cardiac risk. Initiation of statin therapy is *reasonable* for patients undergoing vascular surgery independent of clinical risk. Perioperative initiation of statin therapy should be considered in patients undergoing elevated risk procedures if there is an indication for such therapy separate from the surgery and according to clinical practice guidelines.

ANGIOTENSIN-CONVERTING ENZYME (ACE) INHIBITORS It is important to maintain continuity of therapy with ACE inhibitors (when such therapy is used for the treatment of heart failure or hypertension).

ORAL ANTIPLATELET AGENTS The 4- to 6-week period following implantation of an intracoronary stent (bare metal or drug eluting) constitutes the period of time of greatest risk for the development of stent thrombosis. If possible, noncardiac surgery should be avoided in this vulnerable period. The duration of DAPT thereafter is dictated by the circumstances in which PCI was performed and whether the indication was stable ischemic heart disease or acute coronary syndrome. For the former among patients treated with a drug eluting stent, dual anti-platelet therapy should be given for at least 6 months. For the latter, dual anti-platelet therapy should be given for at least 12 months. However, DAPT may be interrupted to allow for noncardiac surgery 30 days after BMS and 6 months after DES, respectively. If $P2Y_{12}$ inhibitor therapy (clopidogrel, prasugrel, or ticagrelor) is interrupted or discontinued in patients who have received intracoronary stents, aspirin should be continued perioperatively (save select circumstances where the risk of bleeding may be catastrophic as in neurosurgical or spinal procedures) and the $P2Y_{12}$ receptor inhibitor should be restarted as soon as possible post-operatively. Decisions surrounding antiplatelet management in the perioperative setting among patients who have received intracoronary stents are complex and should involve multidisciplinary decision-making.

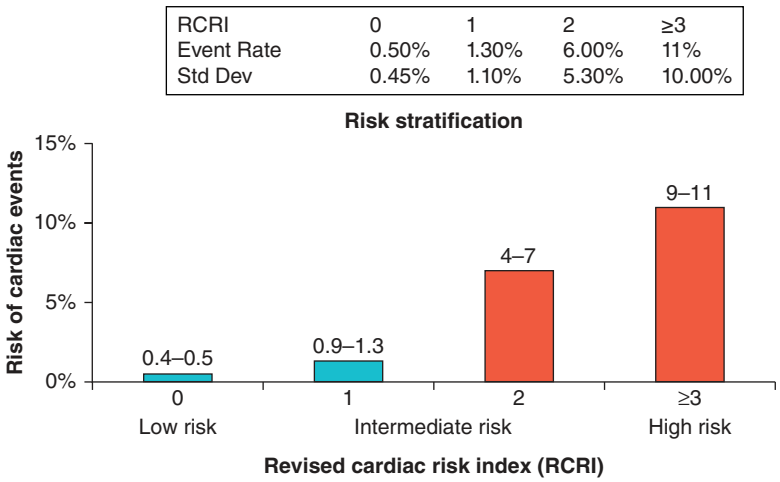


FIGURE 467-2 Risk stratification based on the revised cardiac risk index; derivation and prospective validation of a simple index for prediction of cardiac risk in patients undergoing major noncardiac surgery. Cardiac events include myocardial infarction, pulmonary edema, ventricular fibrillation, cardiac asystole, and complete heart block. *(Adapted from TH Lee et al: Circulation 100:1043, 1999.)*

α_2 AGONISTS Based on the results of POISE-2 (a large multicenter, international, blinded randomized clinical trial of aspirin and clonidine), α_2 agonists for prevention of cardiac events are not recommended in patients who are undergoing noncardiac surgery. In this trial, clonidine increased the rate

TABLE 467-4 Gradation of Mortality Risk of Common Noncardiac Surgical Procedures

Higher	• Emergent major operations, especially in the elderly • Aortic and other noncarotid major vascular surgery (endovascular and nonendovascular) • Prolonged surgery associated with large fluid shift and/or blood loss
Intermediate	• Major thoracic surgery • Major abdominal surgery • Carotid endarterectomy surgery • Head/neck surgery • Orthopedic surgery • Prostate surgery
Lower	• Eye, skin, and superficial surgery • Endoscopic procedures

Source: From LA Fleisher et al: Circulation 116:1971, 2007, with permission.

of nonfatal cardiac arrest and clinically important hypotension, while reducing the rate of death or nonfatal MI.

CALCIUM CHANNEL BLOCKERS Evidence is lacking to support the use of calcium channel blockers as a prophylactic strategy to decrease perioperative risk in major noncardiac surgery.

ANESTHETICS Mortality risk is low with safe delivery of modern anesthesia, especially among low-risk patients undergoing low-risk surgery (Table 467-4). Inhaled anesthetics have predictable circulatory and respiratory effects: all decrease arterial pressure in a dose-dependent manner by reducing sympathetic tone and causing systemic vasodilation, myocardial depression, and decreased cardiac output. Inhaled anesthetics also cause respiratory depression, with diminished responses to both hypercapnia and hypoxemia, in a dose-dependent manner; in addition, these agents have a variable effect on heart rate. Prolonged residual neuromuscular blockade also increases the risk of postoperative pulmonary complications due to reduction in functional residual lung capacity, loss of diaphragmatic and intercostal muscle function, atelectasis, and arterial hypoxemia from ventilation-perfusion mismatch.

Several meta-analyses have shown that rates of pneumonia and respiratory failure are lower among patients receiving neuroaxial anesthesia (epidural or spinal) rather than general anesthesia. However, there were no significant differences in cardiac events between the two approaches. Evidence from a meta-analysis of randomized controlled trials supports postoperative epidural analgesia for >24 h for the purpose of pain relief. However, the risk of epidural hematoma in the setting of systemic anticoagulation for venous thromboembolism prophylaxis (see below) and postoperative epidural catheterization must be considered.

PREOPERATIVE PULMONARY RISK ASSESSMENT

Perioperative pulmonary complications occur frequently and lead to significant morbidity and mortality. Clinical practice guidelines recommend the following:

1. All patients undergoing noncardiac surgery should be assessed for risk of pulmonary complications (Table 467-5).
2. Patients undergoing emergency or prolonged (3–4 h) surgery; aortic aneurysm repair; vascular surgery; major abdominal, thoracic, neurologic, head, or neck surgery; and general anesthesia should be considered to be at elevated risk for postoperative pulmonary complications.
3. Patients at higher risk of pulmonary complications should undergo incentive spirometry, deep-breathing exercises, cough encouragement, postural drainage, percussion and vibration, suctioning and ambulation, intermittent positive-pressure breathing, continuous positive airway pressure, and selective use of a nasogastric tube for postoperative nausea, vomiting, or symptomatic abdominal distention to reduce postoperative risk. Multiple pulmonary risk indices are available to estimate the postoperative risk of respiratory failure, pneumonia, and other pulmonary complications; among these is the ARISCAT risk index, which accounts for the following seven risk factors: age, low preoperative oxygen saturation, respiratory infection within the preceding month, upper abdominal or thoracic surgery, surgery lasting >2 h, and emergency surgery (Table 467-6).
4. Preoperative spirometry and chest radiography should not be used routinely for predicting risk of postoperative pulmonary complications but may be appropriate for patients with chronic obstructive pulmonary disease or asthma.
5. Spirometry is of value before lung resection in determining candidacy for coronary artery bypass; however, it does not provide a spirometric threshold for extrathoracic surgery below which the risks of surgery are unacceptable.

TABLE 467-5 Predisposing Risk Factors for Pulmonary Complications

1. Upper respiratory tract infection: cough, dyspnea
2. Age >60 years
3. Chronic obstructive pulmonary disease
4. Cigarette use
5. American Society of Anesthesiologists Class ≥2
6. Functional dependence
7. Congestive heart failure
8. Serum albumin <3.5 g/dL
9. Obstructive sleep apnea
10. Impaired sensorium (confusion, delirium, or mental status changes)
11. Abnormal findings on chest examination
12. Alcohol use
13. Weight loss
14. Spirometry threshold before lung resection
    a. FEV_1 <2 L
    b. MVV <50% of predicted
    c. PEF <100 L or 50% predicted value
    d. PCO_2 ≥45 mmHg
    e. PO_2 ≤50 mmHg

Abbreviations: FEV_1, forced expiratory volume in 1 s; MVV, maximal voluntary ventilation; PEF, peak expiratory flow rate; PCO_2, partial pressure of carbon dioxide; PO_2, partial pressure of oxygen.

Source: A Qaseem et al: Ann Intern Med 144:575, 2006. Modified from GW Smetana et al: Ann Intern Med 144:581, 2006, and from DN Mohr et al: Postgrad Med 100:247, 1996.

TABLE 467-6 Risk Modification to Reduce Perioperative Pulmonary Complications

Preoperatively
• Cessation of smoking for at least 8 weeks before and until at least 10 days after surgery
• Training in proper lung expansion techniques
• Inhalation bronchodilator and/or steroid therapy, when indicated
• Control of infection and secretion, when indicated
• Weight reduction, when appropriate
Intraoperatively
• Limited duration of anesthesia
• Avoidance of long-acting neuromuscular blocking drugs, when indicated
• Prevention of aspiration and maintenance of optimal bronchodilation
Postoperatively
• Optimization of inspiratory capacity maneuvers, with attention to:
◦ Mobilization of secretions
◦ Early ambulation
◦ Encouragement of coughing
◦ Selective use of a nasogastric tube
◦ Adequate pain control without excessive narcotics

Source: From VA Lawrence et al: Ann Intern Med 144:596, 2006, and WF Dunn, PD Scanlon: Mayo Clin Proc 68:371, 1993.

6. Pulmonary artery catheterization, administration of total parenteral nutrition (as opposed to no supplementation), or total enteral nutrition have no consistent benefit in reducing postoperative pulmonary complications.

PERIOPERATIVE MANAGEMENT AND PROPHYLAXIS

DIABETES MELLITUS

(See also Chaps. 396–398) Many patients with diabetes mellitus have significant symptomatic or asymptomatic CAD and may have silent myocardial ischemia due to autonomic dysfunction. Intensive (versus lenient) glycemic control in the perioperative period is generally not associated with improved outcomes, and may increase the risk of hypoglycemia. Practice guidelines advocate a target glucose range from 100 to 180 mg/dL in the perioperative period. Oral hypoglycemic agonists should not be given on the morning of surgery. Perioperative hyperglycemia should be treated with IV infusion of short-acting insulin or SC sliding-scale insulin. Patients whose diabetes is diet controlled may proceed to surgery with close postoperative monitoring.

INFECTIVE ENDOCARDITIS

(See also Chap. 123) Prophylactic antibiotics should be administered to the following patients before dental procedures that involve manipulation of gingival tissue, manipulation of the periapical region of teeth, or perforation of the oral mucosa: those with prosthetic cardiac valves (including transcatheter prosthetic valves); prosthetic material used in valve repair (annuloplasty ring or artificial chord); previous infective endocarditis; cardiac transplant recipients with valvular regurgitation from a structurally abnormal valve; and unrepaired cyanotic congenital heart disease or repaired congenital heart disease, with residual shunts or valvular regurgitation at the site of adjacent to the site of a prosthetic patch or prosthetic device.

VENOUS THROMBOEMBOLISM

(See also Chap. 273) Perioperative prophylaxis of venous thromboembolism should follow established guidelines of the American College of Chest Physicians. Aspirin is not supported as a single agent for thromboprophylaxis. Low-dose unfractionated heparin (≤5000 units SC bid), low-molecular weight heparin (e.g., enoxaparin, 30 mg bid or 40 mg qd), or a pentasaccharide (fondaparinux, 2.5 mg qd) is appropriate for patients at moderate risk; unfractionated heparin (5000 units SC tid) is appropriate for patients at high risk. Graduated compression stockings and pneumatic compression devices are useful supplements to anticoagulant therapy or in patients at excessive bleeding risk.

FURTHER READING

Fleisher LA et al: 2014 ACC/AHA Guideline on perioperative cardiovascular evaluation and management of patients undergoing noncardiac surgery. Circulation 130:e278, 2014.

Levine GN et al: 2016 ACC/AHA guideline focused update on duration of dual antiplatelet therapy in patients with coronary artery disease. A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol 68:1082, 2016.

Nishimura RA et al: 2017 AHA/ACC focused update of the 2014 AHA/ACC guideline for the management of patients with valvular heart disease. A report of the American College of Cardiology/American Heart Association Task Force on clinical practice guidelines. Circulation 135:1, 2017.

Smetana GW et al: American College of Physicians. Preoperative pulmonary risk stratification for noncardiothoracic surgery: Systematic review for the American College of Physicians. Ann Intern Med 144:581, 2006.

468 Behavioral Economics and Health

Kevin G. Volpp, George Loewenstein, David A. Asch

Tobacco use, physical inactivity, unhealthy diet, excessive alcohol use, and other individual behaviors are estimated to underlie 40% of premature mortality in the United States. Approximately 75% of the $3 trillion currently spent on health care in the United States is attributable to cancer, heart disease, type 2 diabetes, and obesity, and each of these conditions is strongly influenced by behavior. Nearly one-half of patients prescribed medications to lower their cholesterol within 1 year following myocardial infarction stop taking these drugs—even when they are provided free of charge. Despite great advances in the science and technology of health care, a large gap separates theoretically achievable goals in health and health care from what individuals and populations actually reach. Human behavior is both a major contributor to health problems and a barrier to the successful implementation of solutions to address them.

Recognizing that there are many reasons people do not take action to improve their health, behavior change experts have focused efforts on strategies targeted at individual behavior, introducing incentives for weight loss and smoking cessation, environmental strategies such as mandated food labeling, and combination approaches. For example, Section 2705 of the Affordable Care Act (ACA) allows employers to provide incentives of up to 50% of total premiums based on outcomes such as reduced body mass index, lowered blood pressure or cholesterol, and smoking cessation; this legislation puts as much as $300 billion worth of employee health incentives in play annually.

Many of these approaches have built-in limitations, largely because they have been designed around the pervasive view that people always act to improve their self-interest. Existing policy solutions presuppose that health care decisions are rationally based economic transactions and that rational people will dispassionately assess the net present value of the costs and benefits of alternative paths and pursue the best path forward. These approaches are normatively appealing but seem better suited to support the health of people who behave as economists assume they do, and perhaps less effective when exposed to such realities of human behavior as limited attention, overconfidence, and problems of self-control. It is not just the magnitude of incentives that matters, but also other critical features such as the specific nature of rewards, feedback frequency, saliency, and framing. Public health programs, including those involving financial incentives, are more likely to achieve their goals if designed based not on how perfectly rational people *ought* to make health decisions but rather on how humans *actually* make them. The field of behavioral economics, which uses insights from psychology to identify ways that human decision-making often falls short of the ideal, and which offers a more realistic account of human behavior, provides a natural framework for such efforts, including a range of insights relevant to understanding how people respond to financial and nonfinancial (e.g., health) incentives.

CONCEPTS OF CLASSICAL ECONOMICS AND HOW THEY DIFFER FROM BEHAVIORAL ECONOMICS

Classical economics posits that individuals are rational utility maximizers, meaning that they are able to dispassionately identify alternative decisions, calculate the probabilities of and utility/disutility for each potential outcome, and then, through a process of backward induction, implement the decision that has the highest net present value. When it comes to health behaviors, classical economic theory would assume that if people are obese they must have decided that the costs of obesity are outweighed by the benefits of the behaviors that lead to it and that if people smoke they must have decided that the pleasures of doing so outweigh the costs. This assumption of rational utility maximization has two major consequences for public policy, both of which are applicable in the health domain.

The first is that conventional economic thinking radically limits the range of situations in which it makes sense to intervene at a policy level. Under the assumptions of conventional economics, regulatory interventions such as targeted taxes and subsidies are considered appropriate only in situations characterized by externalities (costs that an individual's actions impose on others such as second-hand cigarette smoke), the presence of "market failures" (e.g., monopolies), or the presence of certain types of information asymmetries. The second is that standard economics offers a relatively restricted array of policy tools, including taxes, subsidies, and mandates regarding information provision, and makes unrealistic assumptions about how each of these approaches will influence the behavior of individuals.

Behavioral economics builds on conventional economics by enriching its conception of individual behavior, offering enhancements on both of these dimensions. First, it broadens the range of situations in which policy interventions make sense by introducing the notion of "internalities." While externalities are costs (or benefits) individual behaviors impose on others, internalities are the costs individuals impose on themselves—typically their future selves. While interventions to address smoking and obesity can be justified on externality grounds (e.g., the health care costs are borne by individuals other than the immediately affected individual), they can also be justified on grounds of internalities—e.g., people often irrationally discount the delayed consequences of their behavior. In that sense people might want to be protected not just from others, but from the decisions of their prior selves. Policy intervention can be further justified by the ubiquitous exploitation of individual weaknesses by businesses. Commercial enterprises may take advantage of such vulnerabilities for individual profit rather than customer health, such as with the formulation of tempting but unhealthy processed foods or the pricing of meal "deals" that do not account for the health consequences of economically predatory offers. This behavior is found across industries: Credit card companies and automobile manufacturers lure new customers with "$0 down" and fleeting but appealing teaser rates of "0% interest," playing on the common propensity to focus on the present rather than on the future. Banks earn revenue by charging high fees for minor mistakes such as account overdrafts or breaches of minimum balance rules hiding the description of such fees in small print and complicated jargon. States market lottery tickets that return $0.45 on the dollar and promote these games in ways that ignore more realistic expectations using one-sided messages such as "you can't win if you don't play" rather than, for example, the equally accurate message, "you can't lose if you don't play."

Second, behavioral economics substantially broadens the range of potential policy interventions far beyond those offered by traditional economics. Behavioral economics has become best known for the concepts of "libertarian paternalism" and "asymmetric paternalism." In contrast to "heavy-handed" paternalism, asymmetric paternalism attempts to protect people without limiting freedom of choice. That is, it is asymmetric because it seeks to help individuals who are prone to making irrational decisions without restricting the freedom of choice of others making informed, deliberate decisions. For example, arranging the presentation of food in a cafeteria line so that the healthy foods appear first is likely to increase the amount of healthy food chosen, without depriving those who want the unhealthy foods of the opportunity to purchase them. People who believe that individuals behave optimally should not object to asymmetric paternalism because it does not limit freedom, and those who acknowledge the limits of human rationality should endorse such measures. Frequently, these measures are called "nudges."

A nudge has been defined as "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid." The most prominent, and to date successful, application of nudges has been the use of defaults to increase enrollment in defined contribution retirement savings plans, and secondarily the use of automatic escalation to encourage higher savings rates. These ideas and research findings have had a major impact on retirement savings policies worldwide, including the Pension Protection Act of 2006 in the United States. Building on this success story in savings, and bolstered by the establishment of so-called "nudge units" worldwide, the nudge agenda has positioned behavioral economics at the center of public policy.

The applicability of behavioral economics to policy, including health-related policies, however, goes well beyond nudges. Many health-relevant insights from behavioral economics do not fit the definition of a nudge. For example, there have been incentive programs aimed at changing health behaviors for maximum cost-effectiveness, improvements in the delivery of health-related information, such as nutrition labels, and new designs for physician incentives or health insurance. A common theme of this work harnesses natural tendencies toward predictable decision errors by redirecting those tendencies to help people achieve longer-term health goals or other socially valuable purposes, much the way some martial arts redirect an adversary's strength against him.

Many of the same messages, incentives, and choice structures used so effectively to lure people into self-destructive health behaviors can be redirected to attract them to healthier choices that improve their long-term health and well-being. Some features of human decision-making, such as our propensity to experience regret when we make a bad choice, and our aversion to putting ourselves into such situations (a phenomenon known as "regret aversion" and arguably not really an example of an error) are additionally important features of human psychology that can be exploited in the service of improving health behaviors and outcomes.

There is a common misconception that if you deploy financial incentives in order to promote behavior change, then you are engaged in behavioral economics. But that kind of activity is not behavioral economics; it is simply economics. Indeed, a large number of everyday transactions, such as being paid to go to work, or getting a fine for parking in the wrong place, reflect traditional economic incentives to encourage or discourage certain behaviors. A central lesson from the field of behavioral economics is that how incentives are delivered can matter more than their objective magnitude. There are ways of delivering large incentives that make them ineffective in changing behavior, and there are ways that can greatly magnify the effectiveness of comparatively small incentives. This observation is a source of optimism, implying that with careful design we can leverage relatively small investments to improve public health.

USING BEHAVIORAL ECONOMICS TO PROMOTE SELF-BENEFICIAL HEALTH BEHAVIORS

Behavioral economics builds on neoclassical economics, which has at its core *expected utility theory*. Expected utility theory both presumes to describe how people make decisions and offers a prescription for how such decisions should be made. While utility maximization is a powerful normative model of how we ought to behave, it turns out to be a poor descriptive model of how real people actually behave. Efforts to alter human behavior that rely on this incomplete model often fall short.

Over the past several decades, behavioral economics has described various ways in which people's decisions differ from standard economic models (Table 468-1). While economists mapped out concepts of "bounded rationality" in the 1950s and identified limitations in the dominant expected utility model of decision making under risk, the publication of Prospect Theory by Kahneman and Tversky is widely credited with being seminal in the development of behavioral economics. Prospect theory provides an overarching conceptual framework for describing observations about human behaviors that could not be explained by expected utility theory.

Table 468-1 Traditional vs Behavioral Economics

TRADITIONAL ECONOMICS	BEHAVIORAL ECONOMICS
Core theory: Expected utility maximization	Core theory: Prospect theory
Assumes perfect rationality	Recognizes that people make decision errors
Starting point independent	Assessment depends on your starting point
Framing doesn't matter	Framing affects assessment even when utilities are the same
Stable preferences	Time-inconsistent preferences
People discount the future at constant rates	People discount the near future to a greater degree and have time inconsistent discounting
Intervene only when my actions adversely affect others (negative externalities)	Consider interventions when people will harm their future selves (internalities)
Regulations and policies generally geared to protecting people from the actions of others	Regulations and policies often geared to protecting people from themselves

Although these deviations from expected utility theory can be seen as psychological foibles and errors, the real value of this work has come less from identifying these errors, and more from recognizing that they occur in predictable ways. It is the predictability of these errors that allows the design of strategies to overcome them (Table 468-2). Indeed, it is the predictability and reliability of a number of key features of human decision-making identified by behavioral economics, including but not limited to those connected to prospect theory, allows them to be used in order to promote self-advantageous health-related behaviors.

Loss Aversion and Framing Effects Key tenets of prospect theory include (1) how people feel about a set of possible outcomes depends on their starting point; this is the notion of reference-dependence in which decision-makers evaluate outcomes as gains or losses depending on their starting point; (2) people dislike losses much more than they like equivalent-sized gains, a phenomenon known as "loss aversion"; and (3) there is diminished sensitivity to both gains and losses (Fig. 468-1).

One example of how we can apply the insights of loss aversion can be seen in designing provider payment systems. Financial incentives to physicians to achieve quality goals might be presented—"framed" in the language of behavioral economics—as either rewards (in which physicians receive a bonus if, for example, they increase their colorectal cancer screening rates or increase the percentage of their patients who reach a glycosylated hemoglobin target) or penalties (in which physicians would fail to receive an expected payment if they do not

TABLE 468-2 Key Decision Errors and Suggestions for Addressing Them

Present-biased preferences	Provide feedback and rewards quickly
Nonlinear probability weighting	Motivate people with probabilistic rewards (lotteries)
Overoptimism and loss aversion	Get people to precommit and put money at risk as an effective motivational tool
Peanuts effect	Deliver rewards in bundles, avoiding many small rewards
Narrow bracketing	Frame rewards in terms of effort per day rather than per month or year
Regret aversion	Help people anticipate the regret of poor choices
Defaults/status quo bias	Change the architecture or environment of choice to shift the path of least resistance to favor healthy decisions
Rational world bias	Move beyond the assumption that simply providing information will lead to desired behaviors

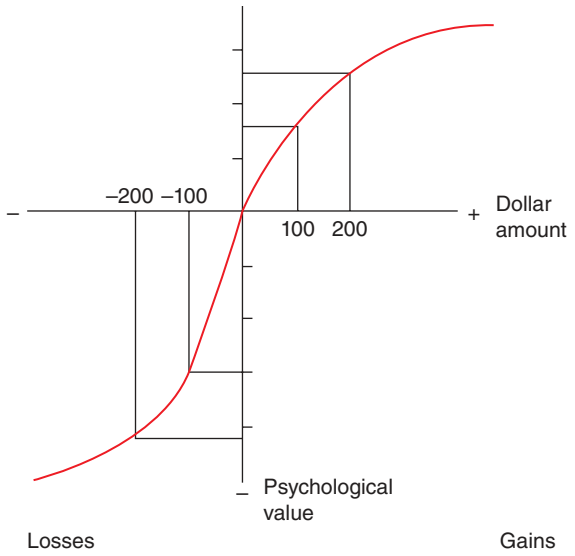


FIGURE 468-1 Prospect theory. (*From D Kahneman: Thinking, Fast and Slow. New York, NY: Farrar, Straus and Giroux, 2011.*)

meet these targets). Classical economists would consider equivalently sized rewards and penalties as identically motivating because each reflects a structure in which one receives $X for a certain outcome (Fig. 468-1). However, loss aversion reminds us that the disutility of losing money is much greater than the utility of gaining the same amount of money. A number of studies have shown that people have a "loss aversion ratio" in a range of 1.5 to 2.5. That means that a potential penalty of $1000 for failing to meet a quality target ought to be about as potent a motivator as a potential reward of $1500 to $2500 for meeting the same target. That multiplier is nonsensical from the standpoint of classical economics, but as an empirically verifiable descriptor of human behavior, it can be exploited in the designs of programs for clinicians or patients to improve health.

In a now classic example in public health, Tversky and Kahneman presented research participants with the following problem:

> *Imagine that the United States is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the program are as follows:*
> *If program A is adopted, 200 people will be saved*
> *If program B is adopted, there is a one-third probability that 600 people will be saved and a two-thirds probability that no one will be saved.*

Given these choices, a substantial majority (about 70%) choose Option A. People preferred to remove uncertainty in favor of the sure bet of saving 200 lives.

A randomly selected second group was presented with an alternative frame:

> *If program A′ is adopted, 400 people will die.*
> *If program B′ is adopted, there is a one-third probability that no one will die and a two-thirds probability that 600 people will die.*

About 70% of participants in this group choose option B′. However, in a population of 600 people, 200 people surviving (option A) is equivalent to 400 people dying (option A′) and options B and B′ are similarly different statements of the same outcome. When outcomes are viewed as gains, as they are in the choice between A and B, decision-makers tend to avoid gambles and are risk averse, choosing option A. When choosing between two bad outcomes, as in the choice between A′ and B′, decision-makers tend to take a gamble (be risk seeking), choosing option B′. In this case people would rather take a chance that everyone could be saved rather than consign themselves to a choice that involved a large number of deaths with 100% certainty.

All of us—including physicians—are highly susceptible to how information is framed. In a set of experiments patients, students, and physicians were presented cases of lung cancer that could be treated either by surgery or radiation therapy. Among all three groups, the choice of surgery was more popular when its outcomes were framed in terms of the probability of survival (e.g., a 68% chance of living for more than 1 year) rather than in terms of the probability of death (e.g., a 32% chance of dying by the end of 1 year). We can say that such sensitivity to framing is irrational since a 68% chance of survival is logically equivalent to a 32% chance of dying—but these irrational decisions fall into predictable patterns of behavior, and that predictability can be used to influence those decisions. This means that clinicians have enormous opportunities to lead patients toward particular decisions by framing the outcomes of those decisions in specific ways, even as they remain truthful. Such an understanding could lead to the view that clinicians should be careful to balance their framing (for example, by following up the statement of 68% chance of survival with a statement like "That means there is a 32% chance of death") in order to provide information in a way that is likely to lead to a particular choice. Alternatively, it could lead to the view that clinicians should deliberately frame outcomes in certain ways in order to lead patients to particular choices—a much more paternalistic stance. Patients often rely on trusted clinicians to help them make the best decisions and, in some settings, that reliance may justify using the principles of behavioral economics strategically even if the same actions in other settings might be seen as paternalistic, anti-libertarian, or coercive.

Loss Aversion and Overoptimism The power of loss aversion can be most effectively leveraged when combined with a well-documented decision error: overoptimism, or unrealistically high expectations about future outcomes. Overoptimism is especially pronounced in the context of people's predictions about their own likelihood of exerting self-control, sometimes referred to as the "false hope syndrome." Although in some contexts overoptimism seems to be beneficial, it can also result in suboptimal patterns of behavior. For example, people prefer paying a flat rate for gym memberships even though they would spend less if they were to pay on a per-visit basis, in part because they overestimate their future gym attendance.

As described earlier, loss aversion reflects the tendency to put greater weight on losses than on equivalently sized gains. It can produce a variety of undesired behaviors, from excessive risk aversion to the tendency for people to hold on to losing investments, such as houses or stocks for too long.

Loss aversion can, in theory, be deployed to advance social goals by framing reward outcomes in terms of losses by "fronting" a sum of money that gets lost if goals are not met, rather than providing equivalent gains for meeting goals (the economic but not psychological equivalent). Yet, despite the greater potency of losses, program administrators are often reluctant to use loss framing, perhaps because such programs can seem more punitive than organizations may wish to appear. However, it is possible to take advantage of loss aversion by designing programs in which people voluntarily put their own money at risk in the service of achieving health-behavior goals that they themselves desire. Many people enter into commercial weight loss programs and pay for a full year up front because they are overly optimistic about their chances of success. Once deposited, however, such optimism can become a self-fulfilling prophecy as loss aversion provides extra motivation to meet goals.

The combination of overoptimism and loss aversion has been used to help people lose weight by giving them the opportunity to participate in deposit contracts, in which they could deposit $0.01–$3.00 per day of their own money, with a 100% match. Participants reported their weight daily and received the sum of the deposit and the matching funds each day they were on track to meeting their monthly weight loss targets, but they forfeited both if they were not on track. The deposit contract leveraged participants' overly optimistic self-predictions of how much weight they would lose as well as loss aversion once deposits were made. In this 16-week study, average weight loss was 14.0 pounds in the deposit group compared with 3.9 pounds in the control group. This work was extended in a 32-week study in which weight loss was sustained for the duration of the intervention (8.7 vs 2.2 pounds in the control group). Although these results are promising,

to be effective as a health strategy this approach needs to achieve high ongoing participation rates to sustain its population effect. Deposit contract approaches are powerful motivators of behavior change, but they are not always popular even among those who initially opt to try them, posing a challenge to long-term success.

Peanuts Effects The prospect theory value function assumes diminishing marginal utility, which means that small gains and losses are disproportionately motivating relative to larger ones (e.g., two $500 rewards would be more potent than one $1000 reward). However, this may be too simple for small rewards (e.g., two $5 rewards may be less motivating than one $10 reward). This "peanuts effect" may be part of the reason why charities and retailers often describe costs in terms of "pennies a day." This observation challenges the expected efficacy of programs that emphasize efforts to repeatedly achieve small changes such as in weight loss programs. It is easy for a patient to rationalize that no single cigarette causes lung cancer or that no single trip to the gym prevents heart disease. If self-destructive patterns of behavior, such as cigarette smoking, weight gain, or cell phone use while driving, are seen as individual instances rather than parts of a composite whole, it is easier to understand how they can be seen as acceptable. The pleasure of smoking a cigarette or eating a dessert and the convenience and engagement of conducting business or socializing in otherwise "dead" travel time are immediate and tangible, but the marginal costs—increased risks of developing lung cancer, being overweight, or having a car accident—seem inappreciably small. Across a lifetime or a population, however, the cumulative costs and/or probabilities are not small at all.

The tendency to underweight small events can also be used to people's advantage, such as by inducing people to put away small sums for retirement savings in short cycle lengths or to make small periodic investments in their health via medication adherence. In each of these efforts, we need to think asymmetrically. In our incentive programs we should provide frequent (often daily) feedback on rewards because of present-biased preferences, but if we are *delivering* financial rewards, we want to aggregate them so that the rewards appear substantial enough to warrant attention. To use the peanuts effect to advantage, one might alert people to their rewards daily but then deliver them monthly to create larger aggregate payments.

PRESENT-BIASED PREFERENCES

Another central observation from behavioral economics is the concept of hyperbolic discounting, or "present-bias." It is standard in conventional economics to assume that people discount the future; for example, $1000 today is worth more than $1000 a year from now, since money can be invested and earn interest. However, people tend to discount outcomes that are close in time more steeply than outcomes that are farther off in time; the degree of time discounting is disproportionately greater for short time delays than for long ones, in contrast to the assumption of standard economic models.

The medical implications of present bias are profound. For example, most people would desperately like to avoid a stroke and many patients with hypertension have an understanding that taking their antihypertensive medications is one of the best ways to avoid a stroke in the future. While the classical economist would see daily adherence to antihypertensive treatment as "a good investment" to avoid a future stroke, the stroke that is avoided is far in the future and uncertain; moreover, the stroke that is prevented is never noticed. In contrast, even the relatively small effort required to stay on antihypertensive medication is immediate, continuous, and comes without any immediate compensatory benefit. To the extent that patients overly discount the future harms of a later stroke, they will be less motivated to invest today in their own medication adherence. To the classical economist, these patients are behaving irrationally because they are failing to make investments that, if they did the calculations, would clearly make them more likely to be better off; this reasoning is parallel to how classical economists consider people who undersave for retirement (the vast majority of Americans) to be irrational. To the behavioral economist, these errors are targets for therapy, in the way that we can see genetic mutations or defects in chemical pathways as therapeutic targets in the management of illness.

Present bias is somewhat subtler than simply steep discounting of the future. In fact, it reflects *two* behavioral tendencies: (1) the tendency to overweight immediate costs and benefits relative to those occurring in the future, as just discussed, and (2) the tendency to take a more evenhanded approach to *future* costs and benefits. People are much more willing to begin dieting *tomorrow*, because the overweighting of immediate costs deters us from the immediate deprivation of dieting, and the more balanced perspective on delayed deprivation makes us willing to impose these costs on ourselves in the future.

Although present-biased preferences typically promote unhealthy behaviors, policy makers can use them for beneficial effects. The motivational impact of benefits and costs, such as rewards for good behavior and punishments for bad behavior, can be increased substantially if they are made immediate. These consequences should coincide as closely as possible with the timing of the behaviors they are meant to encourage or deter. Funds for this could be provided by employers or insurers for whom this might be a cost-effective way to improve worker health and productivity.

Such programs have been shown to have dramatic effects in the area of drug addiction. This success is particularly striking, because many individuals with drug addiction already face major adverse consequences, such as loss of livelihood and disenfranchisement from their families; yet these costs are often insufficient to motivate abstinence. Similarly, small incentives offered on proof of abstinence have succeeded in tripling smoking-cessation rates where the far larger (but delayed) incentives in terms of improved health have failed. Small, daily, lottery-based incentives have significantly increased medication adherence and weight loss in part because they bring immediate rewards (money, excitement) to a situation in which the benefits of avoiding ill health are typically distant and uncertain.

Thus, rather than requiring individuals to make decisions based on consideration of their long-term best interests, it might be useful to change short-term incentives so that beneficial actions are easier and more attractive to choose. Some school districts have begun to use this approach by removing various products such as soda and candy from vending machines so that the cost of obtaining them now includes a walk off campus while healthier food and beverage options are immediately and readily available. In addition, people's willingness to commit to future changes can be leveraged by giving them choices between health-benefiting and health-harming behaviors well before they actually have to act on them. An example of this is scheduling gym visits and laboratory tests to monitor cholesterol ahead of time and having patients voluntarily accept financial penalties for last-minute cancellations in advance.

Nonlinear Probability Weighting In discussing prospect theory, we have so far discussed the concepts of reference-dependence, loss aversion, and diminishing sensitivity, collectively the properties that define prospect theory's value function (Fig. 468-1). However, prospect theory also encompasses a second important dimension: the way that people deal with—and weight—probabilities. In contrast to the standard expected utility model, which assumes that people weight outcomes according to their raw probabilities, prospect theory assumes that people overweight small probabilities but are insensitive to differences in probabilities—for example, between a 0.001 and a 0.00001 chance of winning a prize, even though the probabilities differ by several orders of magnitude—except where they provide a transition to certainty. Such overweighting of small probabilities is partly responsible for the enormous attraction of lottery tickets; yet, like present-biased preferences, this overweighting can be used to advantage in public health interventions.

Following these cognitive pathways, lottery-based reward systems have been introduced in programs aimed at motivating diverse health behaviors (described more fully below). These interventions exploit overweighting of small probabilities and also play on other psychological insights. Because people tend to be motivated by both the experience of past rewards and the prospect of future rewards, these

lottery-based systems provide frequent small payoffs and infrequent large payoffs. This approach has been demonstrated effective in a variety of areas, including helping people lose weight (52.6% of people achieve 16-week weight loss goals compared with about 10.5% in a control group) and reducing medication nonadherence (from about 23% to about 3%). However, results across studies and contexts have been inconsistent, suggesting that the active ingredients of these lottery-based incentives have not been fully elucidated.

Regret Aversion People dislike regretting decisions they have made, often voicing laments such as "If only I had …." Moreover, people are sufficiently far-sighted to anticipate possible future regret and seek to make decisions today that reduce that risk in the future. The avoidance of anticipated regret is a useful exception to present-biased preferences.

Regret aversion helps to explain the success of the Dutch postal code lottery, in which winning postal codes are selected and those living within the selected areas who purchased tickets receive prizes. Those who did not purchase tickets learn that they would have won had they done so. Individuals see their neighbors winning large prizes that they do not share, and their desire to avoid future regret drives subsequent lottery participation.

Anticipated regret has been shown to affect a variety of preventive behaviors, such as the significant increase in vaccination use among people who experienced illness after failing to get vaccinated. Lottery-based incentive programs where eligibility is conditioned on adherence (for example, you aren't eligible to play the lottery unless you had taken your medication or checked your blood pressure the previous day) can incorporate regret aversion into their design by notifying both winners and losers. Those whose number comes up but are ineligible for a reward because of nonadherence are told they would have won had they only taken their medication, checked their blood pressure, or done whatever it was the lottery program was designed to promote. Because people hate the feeling of regret, they are more likely to engage in behaviors to avoid that feeling. Indeed, the advertising campaigns of some traditional lottery systems take advantage of regret aversion. Many people play the same favorite number when the buy lottery tickets. It easy to imagine the disappointment you would feel if you missed buying a ticket one week and that was the week your favorite number came through. Advertising slogans like "don't let your number win without you" keep people in the game in order to escape that feeling of regret. The same techniques used to promote the marketing of lottery tickets can be deployed toward health promotion.

Defaults Although most of the interventions playing on behavioral economics that we have discussed so far do not qualify as "nudges," nudges remain one of the most powerful ways to influence choice. Traditional economic thinking is silent on the power of the default option—the path that is "selected" when no selection is made. However, the default bias, or status quo bias, reflects our tendency to take "the path of least resistance"—to continue doing what we have been doing, or to do what comes automatically, even when superior alternatives exist. Defaults have been blamed for a wide range of suboptimal outcomes, from the failure of employees to put aside retirement funds in companies whose default contribution rate is zero, to suboptimal allocation between investment alternatives, to excessive consumption of French fries and large sodas as part of "supersized" meals at fast-food franchises. In western European countries that have "opt-in" policies for organ donation—that is, the default is nonparticipation (as in the United States), donation rates tend to be in the range of 10%. In contrast, in countries with "opt-out" policies, in which citizens are automatically enrolled as organ donors unless they actively choose not to be, organ donation rates are often close to 99%. Defaults have been shown to increase the rate at which patients with terminal lung diseases choose comfort-oriented plans of care, and they could be used more widely to encourage the choice of beneficial health options that would vary based on the clinical context.

If used tactically, a choice architect (the person who makes decisions on how choices are presented to the end user) can utilize defaults such as changing scheduled automatic prescription refills from 30 to 90 days (or longer) for patients requiring lifelong chronic-disease therapy or changing the default option in fast-food restaurants in combination meals from large sodas to small sodas or water or french fries into carrots unless you ask for french fries to help propel people toward self-beneficial behaviors. A default toward generic prescription embedded in a health system electronic physician order system moved generic prescribing to nearly 100% in a context in which previous attempts at physician education had shown no effect at all.

Such approaches cost nothing since a default has to be set one way or another, preserve freedom of choice, and could change behavior substantially. There is little question we miss out on many opportunities to nudge people toward healthier lifestyles because we haven't considered the default option as a tactical choice to be actively incorporated into an overall health-promoting strategy.

The Rational-World Bias Perhaps the most consequential decision error affecting health-related behaviors is the tendency of public health officials and private-sector benefit designers to make policy decisions based on the assumption that people's choices are deliberative and rational. This in turn leads to assumptions that information provision is all that is needed for optimal decision-making, and that when financial incentives are offered, the amount is all that really matters.

One significant manifestation of the rational-world bias is the complexity of health insurance plans. Health insurance is complicated for many reasons, including insurance companies' desire to shield themselves from costs and, possibly, to hide or "shroud" information that might be perceived as unfavorable. The result is benefit designs with layers of rules about different forms of cost sharing for different kinds of services that vary in and out of networks. A major rationale for this complexity is to include benefit-design elements that incentivize patients to engage in specific cost-minimizing and health-maximizing behaviors across a wide range of activities. For these built-in incentives to exert their desired influence, however, patients need to understand the incentives they are facing, and there is considerable evidence that they do not. This complexity of insurance benefit design, rather than creating perfectly tailored incentives, is part of insurance design's undoing—because an incentive that can't be understood can't be effective. Most consumers lack an understanding of the most basic insurance concepts, such as deductibles, copays, and coinsurance and, when given a simplified version of a traditional insurance plan, are unable to compute the costs they would incur for basic services.

Giving a nod to the benefits of simplicity, the Affordable Care Act in the United States requires insurance plans to be presented in a standardized document that describes plan features such as premiums, deductibles, and coinsurance. However, describing something that is inherently complex in simple terms raises the risk of glossing over important subtleties. A more productive approach would be to provide consumers with a truly simplified insurance product. For example, insurers might create a plan designed with copayments only, since copayments are most easily understood by consumers: copayments are analogous to paying a price for a good or service when shopping, whereas coinsurance and deductibles make it difficult for people to estimate how much their care would cost. It is difficult, if not impossible, for a consumer to estimate 10% coinsurance on a hospital stay or an emergency room visit since they have no idea of the full amount on which the 10% will be calculated. For medical markets and patient cost-sharing elements embedded in plan design to affect decision-making *ex ante*, consumers need some way of accurately gauging not only benefits and risks of treatments but also the costs. This should occur before receiving a diagnostic test or treatment as opposed to what typically happens now: a medical bill is received after the fact.

Another manifestation of the rational-world bias is the more insidious and pervasive belief that if only people knew more about how to advance their health they would do so. This error leads to making education the centerpiece of many health interventions even though public health experts have long recognized that knowledge alone

rarely translates to health-enhancing behavior. For example, nearly everyone knows the health hazards of smoking. While raising and maintaining awareness of the dangers of smoking is an important goal, the financial budgets of the Centers for Disease Control and Prevention, or the time budgets of clinicians, might be more efficiently allocated toward efforts to change behavior that don't presume that the deficit is a lack of knowledge. Indeed, it has been argued that smokers tend to *overestimate* the hazards of tobacco, just as many women tend to *overestimate* their risks of developing breast cancer. In these cases, better or more accurate education might lead people to *lower* their estimates of risk and might, if they were perfectly rational, lead them to reduce health-promoting behaviors. In cases where the deficit is not in knowledge but in behavior, reliance on education as a primary avenue for reducing health risks may divert efforts and resources from other activities that might be much more effective.

APPLICATIONS

Weight Loss Efforts to combat obesity using incentives started in the 1970s. This early work was motivated by the observation that participants who deposited money and other valuables with a therapist and signed contracts in which return of their valuables was contingent on progress towards pre-specified goals lost tremendous amounts of weight. Even though participants received no training in weight loss or maintenance strategies, participants lost an average of 32 pounds. This small initial study lacked control groups and long-term follow-up but provided an important proof of concept.

In the first systematic study of deposit or pre-commitment contracts, participants responding to an advertisement for a weight loss program were informed that study participation required a deposit of $200 (1974 dollars), which would be fully refunded contingent on satisfactory weight loss. After an 11-session, 10-week program, those who received incentives for losing weight or for limiting calories lost significantly more weight than those who received incentives for attending sessions. Of the participants in the former two groups, 70% lost more than 15 pounds. The major limitation to this approach was that only 15% of the prospective participants who responded to the initial newspaper advertisement ended up enrolling, suggesting that deposit contracts that require participants to commit substantial funds up front are very effective for people who agree to participate, but that this requirement likely deters a substantial portion of high-risk participants from entering and its effectiveness may just as much reflect selection of those highly motivated to lose weight.

A subsequent study tested the effects on weight loss of deposit contracts of $30, $150, or $300, with the deposits returned based on either individual or group weight loss over 15 weeks. Participants in the intervention group could win $1, $5, or $10 per pound lost up to a maximum cumulative weight loss of 2 pounds per week (either individual or group average). Mean weight loss was large in all 3 groups but did not differ significantly based on contract size. However, the proportion who reached the goal of 30 pounds weight loss was significantly higher in the larger dollar groups.

Because it is increasingly difficult to shed pounds as weight is lost, the investigators also tested whether a deposit contract with increasing payments ($5, $10, $20, $40, $75) for each 5-pound increment of weight loss would be more effective than offering $30 for each 5-pound increment of weight loss. Participants in both conditions were also offered a maintenance program requiring a $100 deposit, returned in $25 increments for attendance at follow-up visits every 3 months. The increasing contract resulted in qualitatively larger weight loss during the weight loss phase, but the maintenance program did not prevent weight re-gain, likely because the magnitude of the deposit contract for maintenance was small, and feedback was infrequent (only every 3 months).

Paying participants for weight loss using direct payments was less effective than deposit contracts. In a randomized trial, cash payments up to $25 per week for making 100% of proportional progress toward goal, $12.50 for 50% of goal, and $2.50 for not gaining weight did not result in greater weight loss in the payment group than among control subjects.

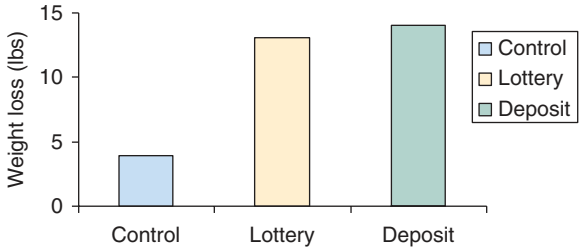


FIGURE 468-2 Weight loss in incentives versus control. *(Figure created using data from KG Volpp et al: Financial incentive-based approaches for weight loss: A randomized trial. JAMA 300:2631, 2008.)*

Studies that have shown no effects on either initial weight loss or maintenance typically have used incentives of small magnitude or were targeted at behaviors, like attendance at weight loss programs that, by themselves, do not ensure weight loss. In recent years, weight loss incentives have become a common feature in programs used by employers and health plans and a variety of start-ups like *stickk.com* and *dietbet.com* use deposit contracts as a way of trying to help people lose weight.

Newer approaches have provided proof of concept that daily lottery-type incentives and precommitment contract incentives promote initial weight loss over 16 weeks (lottery = 13.1 lbs; $p = .014$ for lottery vs control; deposit contract = 14.0 lbs; $p = .003$ vs control; **Fig. 468-2**). However, participants regained most of the weight they had lost over the following 3 months. Longer trials (8 months) found no difference in effectiveness of deposit contracts for continuous 8-month weight loss versus 6-month weight loss with 2 months of maintenance, but both were successful in achieving a mean weight loss of about 10 pounds. Subsequent tests included an employer-based study showing greater effectiveness of team competitive versus individual incentives, although there was a confound in that participants in the teams had the potential to win more money than those in the individual arms.

While both of these studies suggest that financial incentives promote weight loss, many employers use premium adjustments (increases or decreases to health insurance payments) as their standard approach to using financial incentives for health promotion. The effectiveness of premium-based financial incentives to promote weight loss has been assessed in a workplace wellness program and with a goal of losing 5% of initial weight over the next year. Participants were randomly assigned to a control group with no other intervention or one of three financial incentive groups. Two intervention groups were offered a premium reduction of $550 if they achieved their weight loss goal by 1 year. The "delayed" group would receive the premium adjustment in the following year, spread across each pay period. The "immediate" group would have their premiums adjusted as soon as the weight loss goal was met. A third intervention group was offered a daily lottery with about a 1 in 5 chance of winning $10 and a 1 in 100 chance of winning $100. To be eligible to win each day, participants had to weigh in and be on track to lose 5% of their initial weight loss by 6 months, with maintenance for the subsequent 6 months. After 1 year in the program, none of the intervention approaches showed a significant degree of weight loss. The control group gained <0.1kg.

The relative ineffectiveness of premium-based financial incentives is not surprising given their design. Premium adjustments are logistically appealing because the infrastructure to do so is already in place, but the evidence for their effectiveness in on the negative side of unknown and theory argues against it. Such incentives are typically hidden in paychecks that are directly deposited in bank accounts and may go unnoticed by the individual. While a $550 incentive seems like a large amount, it is only $20 in each biweekly paycheck. Typically, these incentives are administered on an all-or-nothing basis contingent on meeting a specific threshold such as a body mass index of ≤25 kg/m^2, meaning that those who are close to that target may be motivated by a goal within reach, but those who are farther away (and have the most weight to lose and the most health to gain) may be less motivated

(perhaps even demotivated) by a goal that doesn't seem attainable. The best evidence and theory now suggests that the standard approach of using premium-based incentives is not that effective and that employers should consider alternative delivery channels for financial incentives to promote health.

A common problem employers and health plans struggle with is getting high levels of engagement among employees/health plan members. Weight Watchers, the largest commercial weight loss program in the United States, worked with a team of academic investigators to conduct a randomized controlled trial involving more than 23,000 participants testing the impact of offering employer subsidization of Weight Watchers membership fees of 50%, 80%, 100% or 50% that could turn into 100% conditional on attending at least 3 Weight Watchers classes a month. Higher subsidies led to higher program enrollment ($p<.0001$). Enrollment differed significantly by subsidy level ($p<.0001$). The 100% subsidy produced the highest enrollment (7.7%), significantly higher than each of the lower subsidies (vs 80% subsidy: 6.2%, $p = .002$; vs 50% subsidy: 3.9%, $p<.0001$; vs hybrid: 3.7%, $p<.0001$). Enrollment in the 80% subsidy group was significantly higher than both lower subsidy groups (vs 50% subsidy: 3.9%, $p<.0001$; vs hybrid: 3.7%, $p<.0001$). Among enrollees, there were no differences among the four groups in attendance or weight loss. In all groups overall weight loss was modest, with a mean weight loss of 2.6 pounds (95% confidence interval [CI], 5.7-lb loss to 0.3-lb gain) in the 100% subsidy arm; 1.5 pounds (95% CI, 5.6-lb loss to 1.3-lb gain) in the 80% subsidy arm; 3.8 pounds (95% CI, 7.9-lb loss to 0.4-lb gain) in the 50% subsidy arm, and 4.0 pounds (95% CI, 8.1-lb loss to 0.1-lb gain) in the hybrid subsidy arm. In all arms, attendance rates drop steadily over time, suggesting that while the hypothesized tradeoff between higher subsidies and lower ongoing engagement in the program did not exist, ongoing participation is a challenge across the board and subsidies that lower the cost of participating are insufficient to achieve high levels of ongoing engagement **(Fig. 468-3)**.

In another study, 281 overweight and obese adults were randomly assigned to a control group or one of three incentive groups for a 13-week physical activity program. All participants were given a goal of achieving 7000 steps a day, tracked automatically using their smart phones. Participants in each of the three incentive arms were offered the same magnitude incentive, $1.40 per day, and were told that accumulated earnings would be sent via a check at the end of each month. However, the incentive in each group was framed differently. In the standard gain incentive group, participants were told that they could earn $1.40 each day they achieved the goal. In the regret lottery incentive group, participants were in a daily regret lottery in which they had an 18 in 100 chance of winning $5 and a one in 100 chance of winning $50, which averages to $1.40 per day. In the loss framing incentive group, participants were told at the beginning of each month that $42 had been placed in a virtual account and that they would lose $1.40 each day the goal was not achieved. During the 13-week intervention, participants in the control, standard gain, and regret gain arms achieved their daily step goal about 30%, 35% and 36% of the time, respectively, but those gain-framed incentive arms were not significantly different from control. However, in the loss framing group, participants achieved the goal 45% of the time, a 50% relative increase, which was significantly greater than the control arm ($p = .001$). This study demonstrated how loss framing can be used to motivate behavior change. It is also one of the first studies to create a loss frame without requiring participants to put their own money at risk using a deposit contract. This is important because fewer people are willing to engage in deposit contract-based incentives than reward-based incentives.

Experience with behavioral economics and weight loss provides several general lessons that are transferable to other health applications. While early studies largely emphasized research in the size of financial incentives, later studies have revealed that the design of the incentive strategy is at least as important. Moreover, designs can vary considerably based on the timing of incentives that can be immediate or delayed, or frequent or one time; the setting of targets that can be achievable, aspirational, or demotivating; the certainty of incentives that can be fixed or probabilistic; the channel for incentives that can either be delivered separately or bundled through payrolls; or the framing of incentives as gains or losses. Classical economists would see only the size of the incentive as an available lever for motivation, but behavioral economists face a much larger set of considerations in designing and testing effective therapies.

Medication Adherence Numerous studies have shown that at least a third of patients fail to adhere to medication regimens. One approach to improving medication adherence is to change some of the underlying defaults by, for example, using 90-day prescriptions for chronic illness medications as opposed to 30-day prescriptions. While we are unaware of empirical evidence supporting longer prescription cycles, it seems logical that adherence rates would be higher over a year if people had to get refills three times as opposed to 11 times—the latter provides more opportunities to forget, experience delays, or fall off the wagon. Automatic refills through prescription mail order might similarly prevent some people from inadvertently falling off the wagon

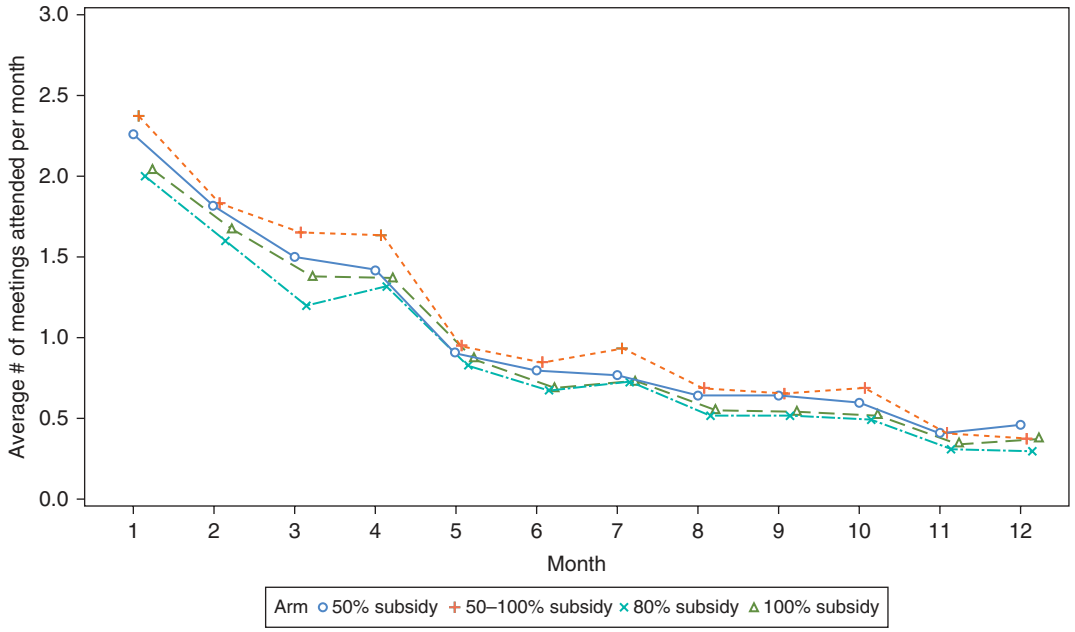


FIGURE 468-3 Attendance rates at Weight Watchers over time in different subsidy groups. *(From LK John et al: "The effect of cost sharing on an employee weight loss program": A randomized trial. Am J Health Promot 32:170, 2018.)*

of medication adherence, giving them one less thing to think about, or fail to think about.

Ninety-day prescriptions or automatic refills could be set up as the default and patients or their providers could opt out if desired. Of course, opt-out defaults are not always possible. A large pharmacy benefits manager wanted to encourage automatic refills for patients on long-term medication but could not have members opt out of such a system because of the potential that those who missed the implications of the opt-out would be surprised or angry about finding credit card charges for automatic refills. In some non-health settings, marketers have used a process sometimes called "active choice" to force explicit consideration of alternatives. Often when ordering airline tickets online the sale cannot proceed unless you affirmatively accept or reject the offer to buy travel insurance. In those settings, the requirement is often seen as a nuisance, particularly if the offer is perceived as a way to cross sell an otherwise unwanted insurance product.

But the same principles can be applied in more health promoting and prosocial settings. Switching from an opt-in system to embedding a choice (yes or no) within the prescription refill process and highlighting the advantages in terms of convenience ("we can send your refills to you automatically or you can get your refills manually if you prefer") resulted in more than twice as many patients choosing to be in the automatic-refill program.

A feature of many health insurance plans is that they require patients to pay some costs out-of-pocket, and hence discourage the use of a number of high-value elements of care, such as the treatment of hypertension or the use of statins by patients with diabetes—care that is widely seen as worth its cost. Support for the use of health insurance deductibles that require patients to have "skin in the game" for their health expenditures derives from insurance theory as well as the seminal RAND health insurance experiment, which demonstrated that these deductibles help overcome moral hazard and reduce the consumption of health care services. Copayments, deductibles, and other out-of-pocket costs make consumers more cost conscious and so aim to make them better shoppers for health care services. Indeed, the rise of high deductible health plans largely aims to increase patients' skin in the game, to make them more value-conscious shoppers.

However, while deductibles and copayments make sense as a way to reduce overutilization of some lower-value health care services, deductibles and copayments make considerably less sense when patients receive medications to manage their hypertension, diabetes, or hyperlipidemia. Given that deductibles and copayments are designed to reduce utilization, why would we ever want to apply them to antihypertensives or statins or insulin given the high health value of these drugs? Why put any barriers between patients and these drugs? Indeed, as the RAND health insurance study showed high-deductible health plans are as likely to discourage the use of high-value services as the use of low-value services. Because patients lack knowledge about what tests or services are of high or low value, and do not have information about the relationship between price and quality, such plans discourage spending on *all* tests and services, including those of high value.

Value-based insurance design—which involves discounting, or making free, services that are deemed to be high in value—is an attempt to sharpen the blunt incentives inherent in deductibles and copayments. Value-based insurance design was inspired by research that showed the use of higher copayments significantly reduced the use of services such as prescriptions but ultimately raised costs, because lower rates of medication nonadherence led to higher rates of emergency department visits and adverse outcomes. Extrapolating from these results, it was natural to conclude that lowering cost sharing for high-value activities, such as taking medications for chronic conditions, would increase adherence and potentially thereby reduce costs. The Affordable Care Act incorporates a kind of value-based insurance design in its requirement that preventive services be offered to patients at no charge.

Unfortunately, value-based insurance design has not delivered on the hope that it would both save money and improve health. From the perspective of the purchaser (for example, the employer or insurer) the economic impact of value-based insurance design depends on whether it can make enough people adherent who were previously nonadherent—and on the health and cost consequences of that improved adherence—to offset the loss of the copayments from those who were already adherent. Although some experimental tests of value-based insurance design have found that copayment reductions increase adherence, those effects have typically been small, in the range of 3–6 percentage points. Even among patients who had recent heart attacks and were given their cardiovascular medications for free, average adherence was only about 45%, just a few percentage points higher than that seen with regular copayments. One reason for these disappointing results is the "dog that didn't bark" problem. People who are nonadherent don't notice that their copays have been reduced because they aren't using (and thus aren't paying for) the service.

Indeed, one of the valuable lessons learned from efforts to introduce value-based insurance design has been a reminder of the asymmetry of the forces that surround patient engagement. Based on conventional economic thought, it might seem reasonable to assume that decreasing copayments would create effects equal and opposite to those of increasing copayments. However, behavioral economic research reveals that framing matters and that losses (in this case, higher copayments intended to reduce use) loom larger in patients' minds than gains (lowered copayments). Furthermore, people who would be deterred by higher copayments are different from people who might become adherent with lower copayments, because the first group consists of those who take their medications while the second group consists of those who do not. Behavioral economic thinking, therefore, helps to explain what has been observed: Increases in copayments have larger effects in reducing adherence than decreases in copayments have in raising adherence. In general, copayment increases lead to far smaller decreases in medication adherence than copayment decreases lead to increases in medication adherence.

Value-based insurance design is an appealing idea. But its benefits could be increased through the application of ideas from behavioral economics, such as simple changes in reward delivery to increase salience (e.g., retaining the copay, but sending a rebate) and communications from insurers to patients so that even those who are nonadherent are aware of the benefit. Better designs might also reflect that most medication adherence happens at least daily, and so reinforcements to that behavior probably need to occur more frequently than the 30- or 90-day cycles coinciding with prescription refills.

A series of studies have used daily lottery-based financial incentives to improve medication adherence. Early work tested the impact of a lottery on medication adherence among patients on warfarin. Participants were eligible for the lottery daily if they correctly took their warfarin the day before. In the first study where the lottery had an expected value (EV) of $5 per day, the proportion of incorrect pill taking was 2.3% (97.7% adherence), compared with a historic mean of 22% incorrect pill taking in this clinic population. In the second study (EV of $3 per day), the overall mean adherence was 98.4% (1.6% days nonadherent), similar to the $5/day study. In a two-arm randomized trial of lotteries for warfarin adherence, the *a priori* subgroup with baseline international normalized ratios (INRs) below the therapeutic range showed no change relative control, but in the a priori subgroup with out-of-range INRs there was a significant reduction in out-of-range INR in the lottery arm vs the control arm (adjusted odds ratio, 0.39; 95% CI, 0.25–0.62). This study highlighted the importance of targeting nonadherent patients in terms of interventions and provided evidence from a randomized, controlled trial (RCT) that lottery incentives can be effective in achieving improved clinical outcomes.

A four-arm NHLBI-funded RCT tested the impact of daily lotteries and daily reminders in a 2 × 2 factorial design on warfarin adherence to address the question of the degree to which a daily lottery is effective due to the fact it also constitutes a daily reminder. However, this study found that while participants in the reminder group had the lowest percentage of time out of target INR range, with an adjusted odds of an out-of-range INR 36% lower than among those in the control group (95% CI, 7–55), the only group with significant improvement in incorrect adherence was the lottery group (incorrect adherence: 12.1% compared with 23.7% in the control group; difference of 7.4%; 95% CI,

–14 to –0.3). There was no relationship between changes in adherence and anticoagulation control in the lottery group, highlighting that participants may appear to change their behavior without perhaps taking the medication, highlighting the importance of serologic or biometric confirmation when possible.

The 5000 Hours Problem A major challenge is determining the optimal method to reach patients and reinforce their behavior each day if we want to significantly improve medication adherence. Even patients with chronic illnesses may spend only a few hours a year with a doctor or nurse, but they spend about 5000 waking hours a year doing just about everything else. Those 5000 hours are when they live their lives and make choices about what to eat and whether to exercise, smoke, take their medications, or visit the doctor.

Although what people do in those hours almost certainly affects their health outcomes, the hours are typically ignored by the U.S. health care system, in part because current approaches to U.S. health care financing support health care during visits to the doctor, not between them, and because "hovering over" people during the hours between visits is personnel-intensive, often requiring nurses or other clinicians to call or visit patients or to staff telemedicine programs. Hovering also requires a fair amount of the very kind of engagement in their own health and health care that is so often missing in the patients these interventions aim to reach. As a result, many of the most promising efforts in telemedicine and home health care have been disappointing.

If some form of hovering is required to engage people who are otherwise hard to engage during the 5000 hours, it almost certainly has to become substantially more automated—both because providers must reduce the need for expensive personnel and because patients have in many cases already revealed limits to their willingness to exert themselves to improve their health. Nevertheless, there is reason for optimism based on the increasing use of cell phones and other wireless devices that make it technologically easier to embed reminders and other forms of touch into patients' lives. Indeed, one key lesson from behavioral economics is that rather than trying to change people's behavior patterns to promote health, it is better to restructure their environment and circumstances so that their existing behavior patterns are more likely to lead to better outcomes. Those efforts require a substantial amount of hovering over patients. Cell phones and other wireless devices don't necessarily change behavior on their own, but because they are already part of many patients' everyday lives they allow previously private behaviors to be witnessed and at times acted upon.

Indeed, an error in early approaches to technology and health behavior overgeneralized from the technologies that support the "quantified self" movement. Apps and wearables that track your diet, physical activity, and biometrics were largely designed for people passionate about measuring themselves. Such individuals don't need a lot of encouragement to wear devices or enter data. The same approaches are far less likely to be useful for patients with difficult-to-manage chronic illness. Many of the internal and external challenges that make their chronic illness hard to manage also make such monitoring hard to manage. A patient who is nonadherent to medication is likely also to be nonadherent to using a new electronic device, but devices like cell phones that are already in use, or other devices that require much less active involvement (like wireless pill bottles) offer more conceptual appeal.

Patients with poorly controlled diseases typically exhibit multiple risk behaviors: poorly controlled diabetes, for example, can be exacerbated by poor diet, lack of exercise, obesity, and medication nonadherence. In testing daily use rates of wireless glucometers and blood pressure cuffs within a population of patients with poorly controlled blood sugars within a health system population, patients who were randomized to be asked to use their wireless glucometer or blood pressure cuff daily used these devices only 50% of the time by the end of 3 months, whereas patients randomized to receive daily lotteries conditional on device use used their devices more than 80% of the time and achieved better glycemic control.

FUTURE PERSPECTIVES

Human health derives from the interaction of basic biologic processes, environmental exposures, social structures, and behavior. The field of behavioral economics has greatly contributed to our understanding of behavior and has made significant contributions to the science of public policy. Given that understanding, we have the opportunity to replace older policies based on unrealistic normative models of rational decision-making with newer policies reflecting our most up-to-date understanding of how humans actually make decisions. Individual and population health outcomes would be very different if people were able to weigh the present and future costs of their actions carefully and dispassionately and had the necessary information and self-control to implement behavioral plans and overcome decision errors that contribute to unhealthy behaviors. Because few people can meet any of those challenges, let alone all, we should not structure our behavior change interventions and public health policies around such models of behavior.

There is broad potential for using understanding of human motivation rooted in behavioral economics to improve private and public approaches to health behavior. One major question is whether developed economies will continue to invest the majority of their health care dollars in treatment (typically about 97% of health care dollars) rather than shifting it toward innovations that seek to keep people healthy. This will be particularly important in settings where treatment options are limited and highly effective methods of prevention exist. For example, it has been estimated that a combination of low-cost cardiovascular drugs could reduce cardiovascular events by 62–88% with perfect adherence, revealing that reducing atherosclerotic cardiovascular disease risk is largely a *behavioral* challenge, given that adherence to medications remain low despite effective pharmacologic solutions. Shifts in health care financing away from fee-for-service toward various forms of payment that require health delivery systems to take on financial risk for populations of patients may drive greater interest in addressing these behavioral and social determinants of health. Research expenditures, which in many developed countries are also roughly allocated about 97% to new treatments and about 3% to prevention, similarly could be shifted to focus more on testing of innovative approaches to improve population health.

The same errors that misdirect patients and providers also misdirect policy makers. In part because of present bias, preventive services often are covered by insurance only if they show a positive return on investment, and yet treatments of existing disease are not held to the same standard. An employer wondering whether to introduce a smoking cessation program for employees wonders about the return on that investment in terms of reduced illness and its cost. The same employer might never question the return on investment of treating lung cancer in the same employee pool, despite what would almost certainly be a negative return on investment. These asymmetries are so embedded in policy making as to be nearly invisible or at least unchallengeable. In fact, the cost of treatments is not even allowed to be considered in Medicare coverage decisions. This prohibition naturally leads to overinvestment in treatments of low value and underinvestment in prevention. The same standard for assessing the impact of health programs, with the goal of achieving the most health possible with the available resources, should be used for both preventive and therapeutic services.

Despite the promise of behavioral economics in structuring policy solutions to social goals, plenty of existing policies that have nothing to do with behavioral economics are effective. For example, raising taxes on cigarettes and other unhealthy goods where it is in the public interest to consume less is a powerful policy tool derived from classical economics. Indeed, tobacco taxes represent one of the most effective ways to curb the use of tobacco and its initiation among youth. Behavioral economics can help make such policies more effective but should not be seen as a substitute for them.

For private-sector entities, the implications of choosing defaults wisely are recognized by many organizations that aim to shift the "path of least resistance" towards healthier choices. Setting up defaults in benefit program design to favor health plans that provide better coverage of preventive services, changing the environment in workplaces to make it easier to take the stairs, and serving more healthful food in cafeterias represent approaches to gently lead people toward individual and population goals.

While medical research continues to generate new tests, interventions, and drugs successfully targeting conditions recently seen as intractable, even the most effective drugs will not work if physicians fail to prescribe them and if patients fail to take them. Although the dominant forms of investigation in medicine seek cellular or molecular therapeutic targets to modify disease, behavioral sciences have revealed cognitive pathways that operate nearly as predictably as the genetic code. The opportunity for behavioral economics to improve health and health care delivery derives from its recognition of these behavioral pathways and the growing empirical evidence about how to best make use of them.

FURTHER READING

Asch DA et al: Automated hovering in health care—watching over the 5000 hours. N Engl J Med 367:1, 2012.

Asch DA et al: Asymmetric thinking about return on investment. N Engl J Med 374:606, 2016.

Connolly T, Butler DU: Regret in economic and psychological theories of choice. J Behav Decis Mak 19:148, 2006.

John LK et al: The effect of cost sharing on an employee weight loss program: A randomized trial. Am J Health Promot 32:170, 2018.

Kahneman D: *Thinking, Fast and Slow*. New York, Farrar, Straus and Giroux, 2011.

Kahneman D, Tversky A: Prospect theory: An analysis of decision under risk. Econometrica 47:263, 1979.

Loewenstein G et al: Asymmetric paternalism to improve health behaviors. JAMA 298:2415, 2007.

Loewenstein G et al: Behavioral economics holds potential to deliver better results for patients, insurers, and employers. Health Aff (Millwood) 32:1244, 2013.

Patel MS et al: Generic medication prescription rates after health system-wide redesign of default options within the electronic health record. JAMA Intern Med 176:847, 2016.

Thaler RH, Sunstein CR: *Nudge: Improving Decisions About Health, Wealth and Happiness*. New York, Penguin Books, 2008.

Volpp KG et al: Financial incentive-based approaches for weight loss: A randomized trial. JAMA 300:2631, 2008.

Volpp KG et al: Assessing value in health care programs. JAMA 307:2153, 2012.

469 Complementary, Alternative, and Integrative Health Approaches

Josephine P. Briggs

The search for health includes many beliefs and practices that are outside conventional medicine. Physicians are important sources for information and guidance about health matters, but our patients also rely on a wide range of other sources including family and friends, cultural traditions, alternative practitioners, and increasingly the Internet, popular media, and advertising. An important step in patient-centered care is understanding what patients are doing to manage their health. This understanding is important to harness potential benefits and to help patients avoid harm.

DEFINITIONS AND SCOPE

Complementary health approaches include a broad range of practices, interventions, and natural products, which are not typically part of conventional medical care, or which may have origins outside of usual Western practice. *Complementary* approaches are defined as those used together with conventional therapies, distinguishing them from *alternative* practices, those used as a substitute for standard care. Complementary practices can roughly be divided into two major groups—mind and body practices, and natural products. Mind and body practices and disciplines are usually administered by or taught to others by a clinician, trained practitioner, or teacher, and include acupuncture, massage, meditation, and hypnosis. Natural products include a diverse group of orally or topically administered substances such as botanical products, unconventional diets, dietary supplements, herbal medicines, homeopathic remedies, probiotics, and others. Brief definitions for some of the common complementary and alternative health practices are provided in **Table 469-1**. Although some complementary health practices are recommended or provided by a physician or a complementary health care provider such as a chiropractor, acupuncturist, or naturopathic practitioner, many of these practices are undertaken as "self-care." Although some are reimbursed, most are paid for out of pocket.

In the last decade or so, the terms *integrative health care* and *integrative medicine* have entered the dialogue. The term *integrative health* care emphasizes a holistic, patient-focused approach to health care and wellness. Most integrative health care is team-based, often bringing conventional and complementary approaches together with self-care in a coordinated way. Physicians advocating this approach generally included selected complementary health practices in the care they offer patients, and many have established practice settings that include complementary health practitioners. Integration of select complementary approaches is common in Veterans Administration and Department of Defense facilities, particularly as part of management of pain and post-traumatic stress disorder. A 2007 national survey conducted by the Centers for Disease Control and Prevention's National Center for Health Statistics found that 42% of hospices had integrated complementary health practices into the care they provide. Although the integrative approach appears to be attractive to many patients, the heavy use of dietary supplements and the weaknesses in the evidence base for a number of the interventions offered in integrative practices continue to attract substantial concern and controversy.

Until a decade ago or so, *complementary and alternative medicine* could be defined as practices that are neither taught in medical schools nor reimbursed, but this definition is no longer workable, since medical students increasingly seek and receive some instruction about complementary health practices, and some practices are reimbursed by third-party payers.

By its nature, the demarcation between mainstream medicine and complementary health practices is porous, varying from culture to culture and over time. Traditional Chinese medicine and the Indian practice of Ayurvedic medicine were once the dominant health teachings in those cultures. Certain health practices that arose as challenges to the mainstream have been integrated gradually into conventional care. Examples include the teachings of Fernand Lamaze that led to the widespread use of relaxation techniques during childbirth, the promotion of lactation counseling by the La Leche League, and the teaching of Cicely Saunders and Elizabeth Kübler-Ross that established the hospice movement.

The late nineteenth century saw the development of a number of healing philosophies by care providers who were critical of the medicine of the time. Of these, naturopathy and homeopathy, which arose in Germany, and chiropractic and osteopathy, which developed in the United States, have endured. Osteopathic medicine is currently thoroughly integrated into conventional medicine, although the American Medical Association (AMA) labeled it a cult as late as 1960. The other three traditions have remained largely separate from mainstream medicine, although chiropractic care is increasingly available in some conventional care settings.

PATTERNS OF USE

The first large survey of use of these practices was performed by David Eisenberg and associates in 1993. It surprised the medical community by showing that >30% of Americans use complementary approaches. Many studies since that time have extended those conclusions. The National Health Interview Survey (NHIS), a large, national household survey of health practices conducted by the National Center for Health Statistics, a component of the Centers for Disease Control and

TABLE 469-1 Terminology of Complementary Health Approaches	
Mind and Body Practices	
Acupuncture and acupressure	A family of procedures involving stimulation of defined anatomic points, a component of the major Asian medical traditions; most common application involves the insertion and manipulation of thin metallic needles
Alexander technique	A movement therapy that uses guidance and education to improve posture, movement, and efficient use of muscles for improvement of overall body functioning
Guided imagery	The use of relaxation techniques followed by the visualization of images, usually calm and peaceful in nature, to invoke specific images to alter neurologic function or physiologic states
Hypnosis	The induction of an altered state of consciousness characterized by increased responsiveness to suggestion
Massage	Manual therapies that manipulate muscle and connective tissues to promote muscle relaxation, healing, and sense of well-being
Meditation	A group of practices, largely based in Eastern spiritual traditions, intended to focus or control attention and obtain greater awareness of the present moment, or mindfulness
Reflexology	Manual stimulation of points on hands or feet that are believed to affect organ function
Rolfing/structural integration	A manual therapy that attempts to realign the body by deep tissue manipulation of fascia
Spinal manipulation	A range of manual techniques, employed by chiropractors and osteopaths, for adjustments of the spine to affect neuromuscular function and other health outcomes
Tai chi	A mind and body practice originating in China that involves slow, gentle movements and sometimes is described as "moving meditation"
Therapeutic touch	Secular version of the laying on of hands, described as "healing meditation"
Yoga	An exercise practice, originally East Indian, that combines breathing exercises, physical postures, and meditation
Traditional Medical Systems	
Ayurvedic medicine	The major East Indian traditional medicine system; treatment includes meditation, diet, exercise, herbs, and elimination regimens (using emetics and diarrheals)
Curanderismo	A spiritual healing tradition common in Latin American communities that uses ritual cleansing, herbs, and incantations
Native American medicine	Diverse traditional systems that incorporate chanting, shaman healing ceremonies, herbs, laying on of hands, and smudging (ritual cleansing with smoke from sacred plants)
Siddha medicine	An East Indian medical system (prevalent among Tamil-speaking people)
Tibetan medicine	A medical system that uses diagnosis by pulse and urine examination; therapies include herbs, diet, and massage
Traditional Chinese medicine	A medical system that uses acupuncture, herbal mixtures, massage, exercise, and diet
Unani medicine	An East Indian medical system, derived from Persian medicine, practiced primarily in the Muslim community; also called "hikmat"
"Modern" Medical Systems	
Anthroposophic medicine	A spiritually based system of medicine that incorporates herbs, homeopathy, diet, and a movement therapy called eurythmy
Chiropractic	Chiropractic care involves the adjustment of the spine and joints to alleviate pain and improve general health; primarily used to treat back problems, musculoskeletal complaints, and headaches
Homeopathy	A medical system with origins in Germany that is based on a core belief in the theory of "like cures like"—compounds that produce certain syndromes, if administered in very diluted solutions, will be curative
Naturopathy	A clinical discipline that emphasizes a holistic approach to the patient, herbal medications, diet, and exercise; practitioners have degrees as doctors of naturopathy
Osteopathy	A clinical discipline, now incorporated into mainstream medicine, that historically emphasized spinal manipulative techniques to relieve pain, restore function, and promote overall health

Prevention, has addressed the use of complementary health practices in 2002, 2007, and 2012. The NHIS survey uses methods that create a nationally representative sample and has a sample size large enough to permit valid estimates about some subgroups. Information was obtained from 31,000 adults in 2002; 23,300 adults and 9400 children in 2007; and 34,500 adults and 10,200 children in 2012. In all three surveys, approximately one-third of adults report using some form of complementary therapy or health practice. In the 2012 survey, 32.2% of adults and 11.6% of children had used one or more modalities. These surveys yield the estimate that nonvitamin, nonmineral dietary supplements are used by ~18% of adults and 5% of children. The most prevalent mind and body practices are yoga, chiropractic or osteopathic manipulation, meditation, and therapeutic massage. Americans are willing to pay for these services; the estimated out-of-pocket expenditure for complementary health practices in 2012 was $30.2 billion ($28.3 billion for adults and $1.9 billion for children), representing 1.1% of total health expenditures and 9.2% of out-of-pocket costs.

The appeal of unproven complementary health approaches continues to perplex many physicians. Many factors contribute to these choices. In NHIS surveys, the most common reason cited for use of complementary health approaches is the management of pain. In one analysis of data from the 2012 NHIS survey, >40% of people with a musculoskeletal pain disorder used a complementary health approach. This was significantly higher than use by people without a musculoskeletal pain disorder (24%). Some patients seek out complementary health practitioners because they offer optimism or greater personal attention. For others, alternative approaches reflect a "self-help" approach to health and wellness or satisfy a search for "natural" or less invasive alternatives, since dietary supplements labelled as natural are believed, often incorrectly, to be inherently healthy.

PRACTITIONER-BASED DISCIPLINES

Licensure and Accreditation At present, six fields of complementary health practice—osteopathic manipulation, chiropractic, acupuncture and traditional Chinese medicine, therapeutic massage, naturopathy, and homeopathy—are subject to some form of educational accreditation and state licensure. Accreditation of educational programs is the responsibility of professional organizations or commissions under federal oversight by the Department of Education. Licensure, in contrast, is strictly a state matter, generally determined by state legislatures. Legal recognition establishes public access to therapies even when there is no scientific consensus about their clinical value.

Osteopathic Manipulative Therapy Founded in 1892 by the physician Andrew Taylor Still, osteopathic medicine was originally based on the belief that manipulation of soft tissue and bone can

correct a wide range of diseases of the musculoskeletal and other organ systems. Over the ensuing century, the osteopathic profession has welcomed increasing integration with conventional medicine. Today, the postgraduate training, practice, credentialing, and licensure of osteopathic physicians are virtually indistinguishable from those of allopathic physicians. Osteopathic medical schools, however, include training in manual therapies, particularly spinal manipulation.

Chiropractic The practice of chiropractic care, founded by David Palmer in 1895, is the most widespread practitioner-based complementary health practice in the United States. Chiropractic practice emphasizes manual therapies for treatment of musculoskeletal complaints, although the scope of practice varies widely, and in some rural areas, chiropractors may serve a primary care role, due in part to the lack of other providers. According to the NHIS, ~8% of American adults and 3% of children receive chiropractic manipulation in a given year.

Since the mid-1970s, chiropractors have been licensed in all 50 states and reimbursed by Medicare. Chiropractic educational standards mandate 2 years of undergraduate training, 4 years of training at an accredited school of chiropractic, and in most states, successful completion of a standardized board examination. Postgraduate training is not required. The U.S. Department of Labor estimates that there are 45,000 licensed chiropractors (2015 figure). There is substantial geographic variation, with greater numbers of practitioners and greater use in the midwest, particularly in rural areas, and lower use in the southeast.

Historically, the relationship between the medical and chiropractic professions has been strained. Extending through the 1970s, the AMA set forth standards prohibiting physicians consulting or entering into professional relationships with chiropractors, but in 1987, after a decade of complex litigation, the U.S. District Court found the AMA in violation of antitrust laws. An uneasy truce has followed, with continued physician skepticism, but also evidence for robust patient demand and satisfaction.

The role of both osteopathic and chiropractic spinal manipulative therapies (SMTs) in management of low-back pain has been the subject of a number of carefully performed trials and many systematic reviews. Conclusions are not consistent, but the most recent 2017 guidelines from the American College of Physicians on noninvasive treatments for acute, subacute, and chronic low-back pain conclude that spinal manipulation has a small effect on improving function and pain compared with control—either a sham manipulation or an inert treatment. Although evidence for spinal manipulation for chronic low-back pain is graded as *low-quality*, the recommendation for consideration of nonpharmacologic treatment including spinal manipulation is graded as a *strong recommendation*, reflecting increasing concern with the impact of chronic opioid use for low-back pain.

The evidence of benefit of spinal manipulation for neck pain is not as extensive, and continued concern that cervical manipulation may occasionally precipitate vascular injury clouds a contentious debate.

Naturopathy Naturopathy is a discipline that emerged in central Europe in the nineteenth century as part of the Natural Cure movement and was introduced to the United States in the early twentieth century by Benjamin Lust. Nineteen states and the District of Columbia currently license naturopathic physicians, with considerable variation in the scope of practice. The naturopathic profession is actively seeking licensure in other states. There are estimated to be ~5000 licensed naturopathic physicians in the United States. There is also a robust naturopathy presence in Canada. Conventional and unconventional diagnostic tests and medications are prescribed, with an emphasis on relatively low doses of drugs, herbal medicines, healthy diet, and exercise. While there is some support for success of naturopathic practitioners in motivating healthy behaviors, concern exists about the heavy promotion of dietary supplements, most with little rigorous evidence.

Homeopathy Homeopathy was widespread in the United States in the late nineteenth and early twentieth centuries and continues to be a common alternative practice in many European countries, but estimates from the NHIS suggest that <2.2% of Americans visit a homeopathic practitioner in any given year. In the United States, licensure as a homeopathic physician is only possible in three states (Arizona, Connecticut, and Nevada) where it is restricted to licensed physicians. The number of practitioners is uncertain, however, because some states include homeopathy within the scope of practice of other fields, including chiropractic and naturopathy, and some practitioners may self-identify as homeopathic practitioners. As discussed below, the regulatory framework for homeopathic remedies differs from other dietary supplements. Homeopathic remedies are widely available and commonly recommended by naturopathic physicians, chiropractors, and other licensed and unlicensed practitioners.

Therapeutic Massage The field of therapeutic massage is growing rapidly, as use by the public is increasing. According to U.S. Department of Labor statistics, there are ~168,000 licensed massage therapists employed in the United States, and by 2024, this number is projected to grow by 22%. Forty-five states and the District of Columbia currently have laws regulating massage therapy; however, there is little consistency, and in some states, regulation is by town ordinance. States that do provide licensure for massage therapists typically require a minimum of 500 h of training at an accredited institution, as well as meeting specific continuing education requirements and carrying malpractice insurance. Massage training programs generally are approved by a state board, but some may also be accredited by an independent agency, such as the Commission on Massage Therapy Accreditation (COMTA). The development of regulatory standards for therapeutic massage has not yet caught up with the evolution of the field or the high demand. Many techniques used are also employed by physical therapists.

Acupuncture and Traditional Chinese Medicine A venerable component of traditional Chinese medicine, with a history of use that extends at least 2000 years, acupuncture became better known in the United States in 1971, when *New York Times* reporter James Reston wrote about how doctors in China used needles to ease his pain after surgery. More than 3 million adults in the United States use acupuncture, according to NHIS data. In a number of European countries, acupuncture is performed primarily by physicians. In the United States, the training and licensure processes for physicians and nonphysicians differ. Currently, acupuncture is licensed in 47 states and the District of Columbia, with licensure standards varying within the scope of practice of each state. Licensure for nonphysicians generally requires 3 years of accredited training and the successful completion of a standardized examination. The main accrediting organization is the Accreditation Commission for Acupuncture and Oriental Medicine. Acupuncture is included in doctor of medicine (MD) and doctor of osteopathic medicine (DO) licensure in 35 states, with 15 states requiring additional training for physicians performing acupuncture.

MIND AND BODY INTERVENTIONS

Mind and body practices and disciplines consist of physical procedures or exercises, manual therapies, or mental techniques that are administered or taught to others by a clinician, trained practitioner, or teacher. Examples include acupuncture, massage therapy, meditation, relaxation techniques, spinal manipulation, and yoga. These approaches are being used more frequently in mainstream health care facilities for both patients and health care providers. Mind and body practices such as meditation and yoga are not licensed in any state, and training in those practices is not subject to national accreditation.

Painful conditions are the most common reasons why American adults use complementary health approaches. About 40 million American adults experience severe pain in any given year, and they spend >$14 billion out-of-pocket on complementary approaches to manage such painful conditions as back pain, neck pain, and arthritis. Chronic pain management is often refractory to conventional medical approaches, and standard pharmacologic approaches have substantial drawbacks. Health care guidelines of the American Pain Society, the American College of Physicians, and other professional organizations

recognize the value of certain complementary approaches as alternatives or adjuncts to pharmacologic management.

The evidence base for the effectiveness of these modalities is still relatively incomplete, but a few rigorous examples where there is promise of usefulness and safety include acupuncture and tai chi for osteoarthritis of the knee pain; massage for neck pain; tai chi for fibromyalgia pain; relaxation techniques for headaches and migraine; and acupuncture, massage, yoga, and spinal manipulation for chronic back pain. In addition, new research is shedding light on the effects of meditation and acupuncture on central mechanisms of pain processing and perception and regulation of emotion and attention. Although many unanswered questions remain about these effects, findings are pointing to scientifically plausible mechanisms by which these modalities might yield benefit.

DIETARY SUPPLEMENTS

Regulation The Dietary Supplements Health and Education Act (DSHEA), passed in 1994, gives authority to the U.S. Food and Drug Administration (FDA) to regulate dietary supplements, but with expectations that differ in many respects from the regulation of drugs or food additives. Purveyors of dietary supplements cannot claim that they prevent or treat any disease. They can, however, claim that they maintain "normal structure and function" of body systems. For example, a product cannot claim to treat arthritis, but it can claim to maintain "normal joint health." Homeopathic products predate FDA drug regulations and are sold with no requirement that they be proved effective. Although homeopathic products are widely believed to be safe because they are highly dilute, one product, a nasal spray called Zicam, was withdrawn from the market when it was found to produce anosmia, probably because of a significant zinc content. In January 2017, the FDA warned consumers about homeopathic teething tablets containing belladonna that pose a serious risk to infants and children.

Regulation of advertising and marketing claims is the purview of the Federal Trade Commission (FTC). The FTC does take legal action against promoters or websites that advertise or sell dietary supplements with false or deceptive statements. Misleading marketing of homeopathic products, and indeed other complementary health products and practices, contributes to the very significant risk that individuals will use them instead of effective conventional modalities.

Inherent Toxicity Although the public may believe that "natural" equates with "safe," it is abundantly clear that natural products can be toxic. Misidentification of medicinal mushrooms has led to liver failure. Contamination of tryptophan supplements caused the eosinophilia-myalgia syndrome. Herbal products containing particular species of *Aristolochia* were associated with genitourinary malignancies and interstitial nephritis. In 2013, dietary supplements containing 1,3-dimethylamylamine (DMAA), often touted as a "natural" stimulant, led to cardiovascular problems, including heart attacks. Among the most controversial dietary supplements is *Ephedra sinica*, or ma huang, a product used in traditional Chinese medicine for short-term treatment of asthma and bronchial congestion. The scientific basis for these indications was revealed when ephedra was shown to contain the ephedrine alkaloids, especially ephedrine and pseudoephedrine. With the promulgation of the DSHEA regulations, supplements containing ephedra and herbs rich in caffeine sold widely in the U.S. marketplace because of their claims to promote weight loss and enhance athletic performance. Reports of severe and fatal adverse events associated with use of ephedra-containing products led to an evidence-based review of the data surrounding them, and in 2004, the FDA banned their sale in the United States.

Another major current concern with dietary supplements is adulteration with pharmacologic active compounds. Multi-ingredient products marketed for weight loss, body building, "sexual health," and athletic performance are of particular concern. Recent FDA recalls have involved contamination with steroids, diuretics, stimulants, and phosphodiesterase type 5 inhibitors.

TABLE 469-2 Resources for Dietary Supplement–Drug Interactions

Medscape	
http://www.medscape.com/druginfo/druginterchecker?cid=med	
This website is maintained by WebMD and includes a free drug interaction checker tool that provides information on interactions between two or more drugs, herbals, and/or dietary supplements.	
Natural Medicines Comprehensive Database	
http://naturaldatabase.therapeuticresearch.com	
This website provides an interactive natural product–drug interaction checker tool that identifies interactions between drugs and natural products, including herbals and dietary supplements. This service is available by subscription. A PDA version is available.	

Abbreviation: PDA, personal digital assistant.

Herb–Drug Interactions A number of herbal products have potential impact on the metabolism of drugs. This effect was illustrated most compellingly with the demonstration in 2000 that consumption of St. John's wort interferes with the bioavailability of the HIV protease inhibitor indinavir. Later studies showed its similar interference with metabolism of topoisomerase inhibitors such as irinotecan, with cyclosporine, and with many other drugs. The breadth of interference stems from the ability of hyperforin in St. John's wort to upregulate expression of the pregnane X receptor, a promiscuous nuclear regulatory factor that promotes the expression of many hepatic oxidative, conjugative, and efflux enzymes involved in drug and food metabolism.

Because of the large number of compounds that alter drug metabolism and the large number of agents some patients are taking, identification of all potential interactions can be a daunting task. Several useful Web resources are available as information sources **(Table 469-2)**. Clearly, attention to this problem is particularly important with drugs with a narrow therapeutic index, such as anticoagulants, antiseizure medications, antibiotics, immunosuppressants, and cancer chemotherapeutic agents.

PATIENT AND PROVIDER RESOURCES

Physicians regularly face difficult challenges in providing patients with advice and education about complementary practices. Of particular concern to all physicians are practices of uncertain safety and practices that raise inappropriate hopes. Cancer therapies, antiaging regimens, weight-loss programs, sexual function, and athletic performance are frequently targeted for excessive claims and irresponsible marketing. A number of Internet resources provide critical tools for patient education **(Table 469-3)**. Because many complementary health products and practices are used as self-care and because many patients research these approaches extensively on the Internet, directing patients to responsible websites can often be very helpful.

The scientific evidence regarding complementary therapies is fragmentary and incomplete. Nonetheless, in some areas, particularly pain management, it is increasingly possible to perform the kind of rigorous systematic reviews of complementary health approaches that are the cornerstone of evidence-based medicine. A particularly valuable resource in this respect is the Cochrane Collaboration, which has performed >300 systematic reviews of complementary health practices. Practitioners will find this a valuable source to answer patient questions. Practice guidelines, particularly for pain management, are also available from several professional organizations. Links to these resources are provided in Table 469-3.

SUMMARY

The use of complementary, alternative, and integrative health approaches reflects an active interest in improved health. An array of unproven modalities will always be used by our patients. While some of these choices need to be actively discouraged, many are in fact innocuous and can be accommodated. Some may be genuinely helpful, particularly in the management of troublesome symptoms. An important step in patient-centered care is understanding patients' beliefs and expectations and what they are doing to manage their health, including the use of complementary health approaches, and using those insights to help guide health-seeking practices in a constructive way.

TABLE 469-3 Internet Resources on Complementary Health Approaches

The Cochrane Collaboration Complementary Medicine Reviews
This website offers rigorous systematic reviews of mainstream and complementary health interventions using standardized methods. It includes >300 reviews of complementary health practices. Complete reviews require institutional or individual subscription, but summaries are available to the public.
http://www.cochrane.org/evidence
MedlinePlus All Herbs and Supplements, A–Z List
MedlinePlus Complementary and Alternative Medicine
NLM FAQ: Dietary Supplements, Complementary, or Alternative Medicines
These National Library of Medicine (NLM) Web pages provide an A–Z database of science-based information on herbal and dietary supplements; basic facts about complementary health practices; and federal government sources on information about using natural products, dietary supplements, medicinal plants, and other complementary health modalities.
http://www.nlm.nih.gov/medlineplus/druginfo/herb_All.html
http://www.cochrane.org/evidence
http://www.nlm.nih.gov/medlineplus/dietarysupplements.html
National Institutes of Health National Center for Complementary and Integrative Health (NCCIH)
This National Institutes of Health NCCIH website contains information for consumers and health care providers on many aspects of complementary health products and practices. Downloadable information sheets include short summaries of complementary health approaches, uses and risks of herbal therapies, and advice on wise use of dietary supplements.
http://www.nccih.nih.gov
Resources for Health Care Providers: *http://www.nccih.nih.gov/health/providers*
NCCIH Clinical Digest e-Newsletter: *http://www.nccih.nih.gov/health/providers/digest*
Continuing medical education lectures: *http://www.nccih.nih.gov/training/videolectures*

Acknowledgment

The late Dr. Stephen Straus contributed this chapter in prior editions, and some material from his chapter has been retained here.

FURTHER READING

Barnes PM et al: Complementary and alternative medicine use among adults and children: United States, 2007. Natl Health Stat Rep 12:1, 2008.

Eisenberg DM et al: Trends in alternative medicine use in the United States, 1990–1997: Results of a follow-up national survey. JAMA 280:1569, 1998.

MacPherson H et al: Closing the evidence gap in integrative medicine. BMJ 339:b3335, 2009.

Manek NJ et al: What rheumatologists in the United States think of complementary and alternative medicine: Results of a national survey. BMC Complement Altern Med 10:5, 2010.

Nahin RL et al: Evidence-based evaluation of complementary health approaches for pain management in the United States. Mayo Clin Proc 91:1292, 2016.

Nahin RL et al: Expenditures on complementary health approaches: United States, 2012. Natl Health Stat Rep 95:1, 2016.

Paige NM et al: Association of spinal manipulative therapy with clinical benefit and harm for acute low back pain: Systematic review and meta-analysis. JAMA 317:1451, 2017.

Qaseem A et al: Noninvasive treatments for acute, subacute, and chronic low back pain: A clinical practice guideline from the American College of Physicians. Ann Intern Med 166:514, 2017.

470 Telomere Disease

Rodrigo T. Calado, Neal S. Young

DEFINITION

In telomere diseases (also called *telomeropathies* or *telomere spectrum disorders*), organ dysfunction is caused by loss of the ends of chromosomes, a process termed *accelerated telomere attrition*. Inadequate repair or insufficient protection of telomeres and their resulting erosion induces cell death, deficient cell proliferation, and chromosome instability; affected tissues show defective organ regeneration, fibrosis or replacement by fat, and a proclivity for cancer. A variety of regenerative disorders affecting especially the bone marrow, lungs, liver, and skin share telomere dysfunction and loss as their common molecular mechanism. It is important to note that telomeres appear to shorten with time based on cross-sectional data of average telomere length in groups of people at different ages. However, limited data exist about telomere shortening longitudinally in individual people. Despite shortening of telomeres over time, normal aging is not associated with the development of disease from short telomeres. In normal aging, sufficient stem cell number and function are maintained to sustain vital processes. Even a patient who receives a limited number of hematopoietic cells from an adult donor is capable of maintaining normal hematopoiesis for many years, at least in part related to normal telomerase function and telomere repair. When symptoms develop as a consequence of short telomeres, a disease process is at work.

DISEASE MECHANISM

Telomeres, the physical termini of linear chromosomes, are repeated hexanucleotide sequences physically associated with specific proteins. Telomeres function to protect the chromosome ends against recognition as damaged or infectious DNA by the DNA repair machinery (Fig. 470-1). During mitosis, the DNA polymerase employs an RNA oligonucleotide with a 3′ hydroxyl group to prime replication. The primer dissociates as the DNA polymerase advances along the template strand, and a gap is left at the ends of linear DNA molecules: the newly synthesized DNA strand is necessarily shorter than the original template—the "end-replication problem." Chromosome erosion is thus inevitable with mitotic cell division, but the noncoding telomeric long, repetitive structure buffers loss of genetic information. In human cells, telomeres are composed of hundreds to thousands of TTAGGG tandem repeats in the leading and CCCTAA in the lagging DNA strand. At birth, telomeres are relatively long but they inexorably shorten with chronological aging (Fig. 470-1). In an individual cell, critically short telomere length triggers the p53 pathway, usually leading to proliferative arrest, senescence, and apoptosis. Telomere loss is the molecular basis for the "Hayflick phenomenon," the limit to cell division and thus to cell proliferation in tissue culture. If a cell overrides proliferation arrest, extremely short telomeres may engage the DNA damage repair machinery, and chromosome end-to-end fusions, chromosome breaks, aneuploidy, and chromosome instability may occur. In addition to the telomere repeated sequences, a group of specialized proteins, collectively termed *shelterins*, directly bind to or indirectly associate with telomeres, assisting in the organization of the telomere tertiary structure and inhibiting activity of DNA damage response proteins (Fig. 470-1).

To escape telomere attrition, cells with high proliferative demand, including embryonic and adult stem cells, lymphocytes, and the majority of cancer cells have at least two mechanisms to preserve telomere length: recombination and the capability to synthesize telomeric repeats. GTTAGG hexanucleotides are added to the 5′ end of the leading DNA strand by telomerase, a reverse transcriptase enzyme (TERT), and TERC, its RNA template, (Fig. 470-1). The telomerase holoenzyme complex is composed of two copies of TERT, TERC and dyskerin, and associated proteins. TERC binds to TERT and serves as the RNA template for its function as a reverse transcriptase. Dyskerin, encoded by

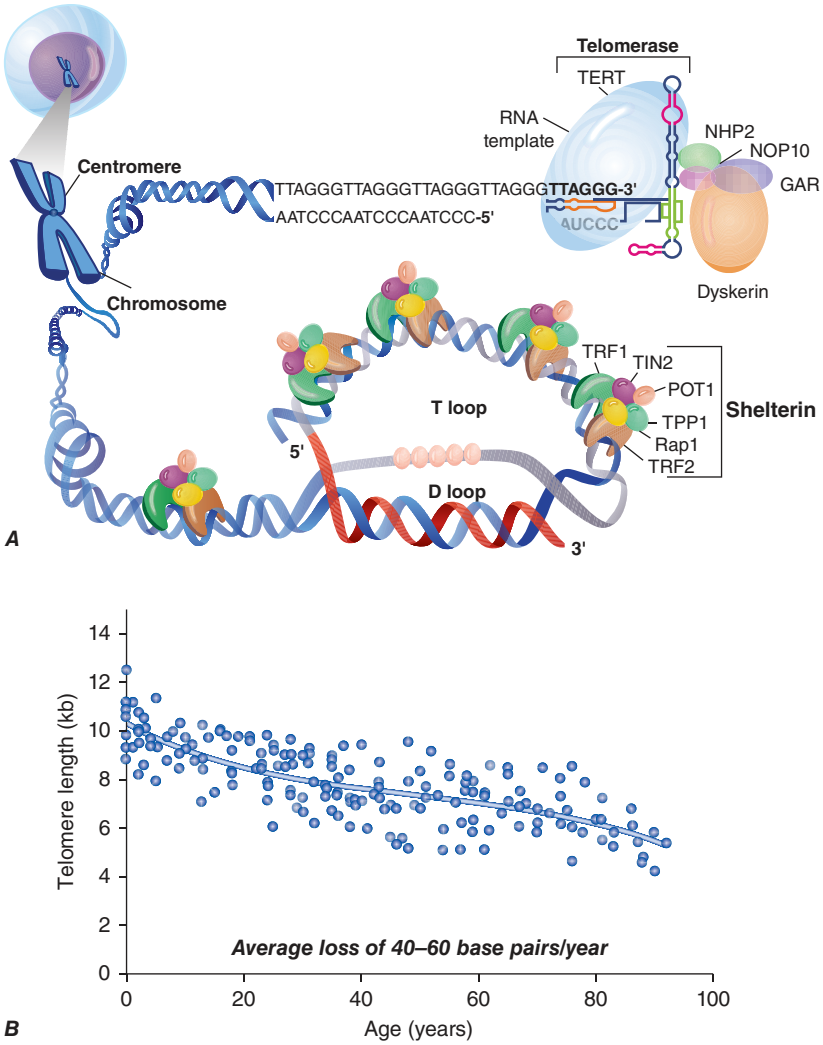


FIGURE 470-1 Telomeres and telomerase. ***A.*** Telomeres are ribonucleoprotein structures located at the termini of linear chromosomes inside the cell nucleus composed of hundreds of tandem hexameric DNA repeats. A group of proteins bind directly or indirectly to telomere sequences in order to provide protection to the structure and are collectively termed *shelterin* or *telosome* (TRF1, TRF2, TIN2, POT1, TPP1, and RAP1). As the 3' end of the leading strand forms a single-stranded overhang, it folds back and invades the telomeric double helix, forming a lariat termed *T loop*. The telomerase complex is composed of the enzyme telomerase reverse transcriptase (TERT), its RNA component (TERC), the protein dyskerin, and associated proteins (NHP2, NOP10, and GAR). This enzymatic complex elongates telomeres by adding GTTAGG hexameric repeats to the 3' end of the telomeric leading strand, using a sequence in TERC as the template. ***B.*** The average telomere length in human leukocytes varies: it is longer at birth (10–11 kilobases) and progressively shortens with aging (6–7 kilobases at age 90 years) at an average loss of 40–60 base pairs/year. However, there is significant variability in telomere length in each given age.

DKC1, stabilizes the complex, and TCAB1, encoded by the *WRAP53* gene, aids telomerase trafficking to the Cajal bodies, nuclear structures for ribonucleoprotein processing where telomerase associates with telomeres for elongation. Telomerase expression is tightly regulated. MYC and sex hormones stimulate *TERT* transcription, whereas in mature cells the *TERT* gene is heavily repressed. In addition, shelterin proteins can also regulate telomerase function and processivity, modulating its catalytic activity on telomeres. Other proteins also are important for telomere length maintenance. RTEL1, an essential DNA helicase, dismantles t-loops and resolves g-quadruplexes, ensuring adequate telomere elongation.

Pathologic accelerated telomere attrition has a genetic origin. Germ-line loss-of-function mutations in genes involved in telomere maintenance and function impair telomere length repair, thus increasing the rate of telomere erosion in highly proliferative cells. Telomeres reach critically short lengths faster than with normal aging; the consequences are limited cell proliferation and impaired tissue regeneration. Some organs appear to be particularly susceptible to telomere erosion. Billions of blood cells are produced daily (Chap. 92), and telomere attrition curtails cell proliferation, producing a hypocellular marrow and often low blood counts. The liver also is an organ with high proliferative capacity, and telomere dysfunction impairs hepatic regeneration after injury, with a variety of pathologic consequences. The lung alveolar epithelium is in contact with exogenous toxins that stimulate regeneration, and telomere loss may hamper these physiologic responses. However, it remains unclear why other regenerative tissues, like the intestine, are relatively unaffected by telomere dysfunction, or the mechanism by which telomere loss provokes a fibrotic response in the lungs (pulmonary fibrosis), an adipose response in the marrow (aplastic anemia), and both in the liver (hepatic steatosis and cirrhosis).

When telomeres are critically short, the DNA damage response machinery may be recruited, mistaking telomeres for damaged or infectious DNA and forcing inappropriate repair. Activation of this pathway may cause chromosome instability due to end-to-end fusion of chromosomes or translocations; these alterations generate unstable, potentially malignant clones. That telomere dysfunction increases the risk of cancer development has been demonstrated in murine models of telomerase deficiency, and patients with telomere diseases are especially prone to develop acute myeloid leukemia and head and neck squamous cell carcinomas.

GENETICS

The pattern of inheritance is variable: X-linked, autosomal recessive, and autosomal dominant. At least 13 genes are implicated in the etiology of telomeropathies, which may be grouped in three categories (Table 470-1).

TABLE 470-1 Genetic Variants in 13 Genes Involved in Telomere Maintenance, Inheritance Pattern, and Phenotype

GENE	DYSKERATOSIS CONGENITA	APLASTIC ANEMIA	PULMONARY FIBROSIS	CIRRHOSIS	MDS/LEUKEMIA
Telomerase					
DKC1	X-L				
TERT	AD/AR	AD/AR	AD	AD	AD/AR
TERC	AD/AR	AD	AD	AD	AD
NOP10	AR				
NHP2	AR				
WRAP53	AR				
Shelterin					
TINF2	AD	AD	AD		
TERF2		AD			
ACD	AD				
Others					
RTEL1	AR	AD/AR	AD		AD
CTC1	AR	AR			
PARN			AD		
USB1	AD				

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; X-L, X-linked.

CLINICAL MANIFESTATIONS

Presentation of telomere disease in the clinic is highly variable—in the tissues affected, in the severity of organ dysfunction, and in patterns of disease within families and between families with similar mutations. In a same family, one individual may be severely affected while close relatives carrying the same mutation are asymptomatic and have normal laboratory results. Asymptomatic carriers, however, may display subclinical organ dysfunction, which may be detected by directed or specialized testing (reduced forced vital capacity on pulmonary function test, hypocellular bone marrow at biopsy, hepatic steatosis on ultrasound). In addition, relatives sharing the same mutation and short telomeres may have different organs affected: for instance, while one individual may have aplastic anemia, another within the pedigree may suffer from pulmonary fibrosis. Environmental factors (smoking, alcohol consumption, viral infection) appear to cooperate with organ damage in these patients and contribute to disease heterogeneity among individuals.

Disease anticipation, in which clinical phenotype manifests at an earlier age in successive generations, is observed in some families with telomeropathies, due to the inheritance of short telomeres in sperm and oocytes.

The diagnosis usually is suggested by personal and family history and can be confirmed by simple measurement of leukocyte telomeres and next-generation sequencing of the most prominent genes encoding for telomere repair enzyme complex and shelterin components.

Dyskeratosis Congenita Dyskeratosis congenita is the classic telomere disease, a mainly pediatric syndrome diagnosed in the first two decades of life. Affected children have at least two features of the mucocutaneous triad of ungual dystrophy, reticular skin pigmentation, and oral leukoplakia (Fig. 470-2). In severe cases, affected newborns have cerebellar hypoplasia (Hoyeraal-Hreidarsson syndrome) or exudative retinopathy (Revesz syndrome) (Fig. 470-3). Telomeres are usually very short, below the first percentile expected for age (Fig. 470-4). Most patients with dyskeratosis congenita develop bone marrow failure, often requiring transfusions and bone marrow transplant. Pulmonary fibrosis appears in as many as 20% of cases and liver disease in 10%, not infrequently after bone marrow transplant. Other tissues and organs also may be affected (Fig. 470-3). The genetic defects most commonly observed in dyskeratosis congenita patients are in *DKC1*, *TINF2*, *TERT*, and *TERC* genes (Table 470-1).

Bone Marrow Failure Aplastic anemia (Chap. 98) is the most common major clinical manifestation of dyskeratosis congenita. However, young or older patients carrying a telomere defect, without typical physical stigmata of dyskeratosis, also may also develop marrow failure. Mutations are monoallelic (one mutated allele and one wild-type allele), and *TERT*, *TERC*, and *RTEL1* are the genes usually affected. Telomere loss in these cases is often less intense than in classic dyskeratosis congenita. As a result of inadequate telomerase function, the stem cell pool is limited in size and in its ability to regenerate. There is insufficient production of erythrocytes, platelets, and granulocytes with anemia, thrombocytopenia, and leukopenia of peripheral blood and relatively low marrow cellularity (Fig. 470-5). Patients can present with apparently sudden onset of severe aplastic anemia, but more common is a long history of macrocytic mild to moderate anemia and/or thrombocytopenia, with preservation of leukocyte numbers. A comprehensive personal and family history is important, querying especially for blood abnormalities, and also for lung and liver disease; early hair graying, while not specific to telomeropathies, strongly suggests telomere disease in the appropriate context.

When treated, immunosuppressive therapy is generally ineffective in these patients, and they may be more susceptible to pulmonary or hepatic complications after hematopoietic stem cell transplant.

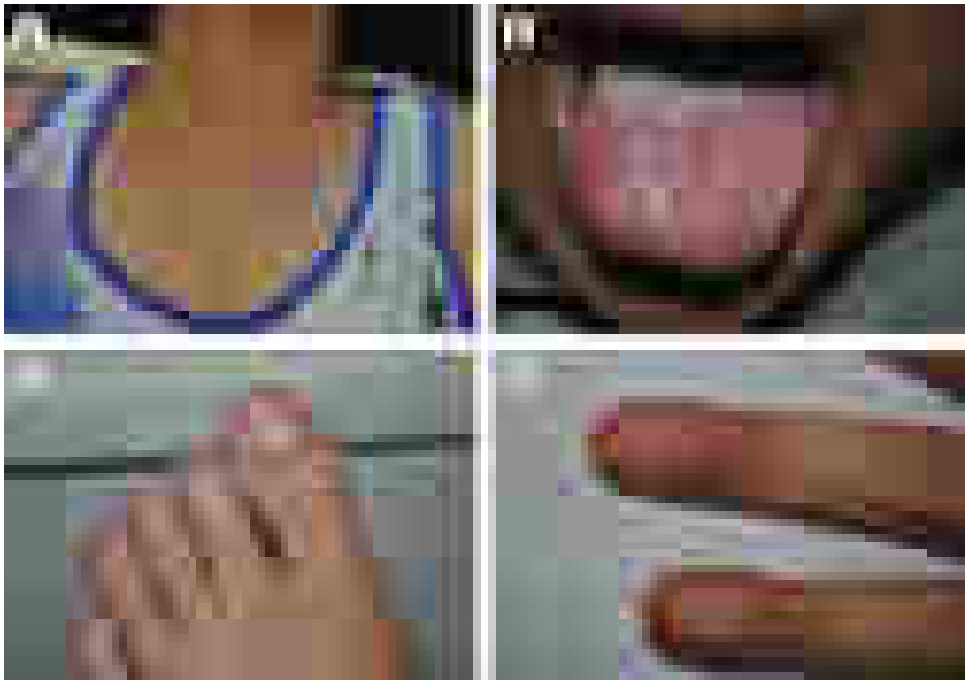


FIGURE 470-2 Skin manifestations of dyskeratosis congenita. The pediatric syndrome dyskeratosis congenita characterizes by the mucocutaneous triad of ***(A)*** reticular skin pigmentation, ***(B)*** oral leukoplakia, and ***(C, D)*** nail dystrophy.

FIGURE 470-3 Clinical consequences of telomere diseases. Telomere dysfunction affects a variety of organs: cerebellum, eyes, lungs, liver, skin, gastrointestinal tract, and the bone marrow.

Myeloid Neoplasms Some patients diagnosed with myelodysplastic syndrome (Chap. 98) or acute myeloid leukemia (Chap. 100) have a family history of bone marrow failure or of other myeloid neoplasms. One of the genetic causes for myeloid neoplasia predisposition is a telomere defect, and these disorders are now classified together by the World Health Organization as "myeloid neoplasms associated with telomere biology disorders." Telomere length measurement may be confounded by the presence of circulating immature cells, which may have very short telomeres, precluding appropriate test interpretation.

Pulmonary Fibrosis Pulmonary fibrosis appears in about 20% of children with dyskeratosis congenita, and approximately 10–15% of patients with idiopathic pulmonary fibrosis (Chap. 287) or familial pulmonary fibrosis have an etiologic telomerase gene mutation. Most pulmonary fibrosis patients, regardless of mutation status, have short telomeres for their age but not as short as in dyskeratosis congenita.

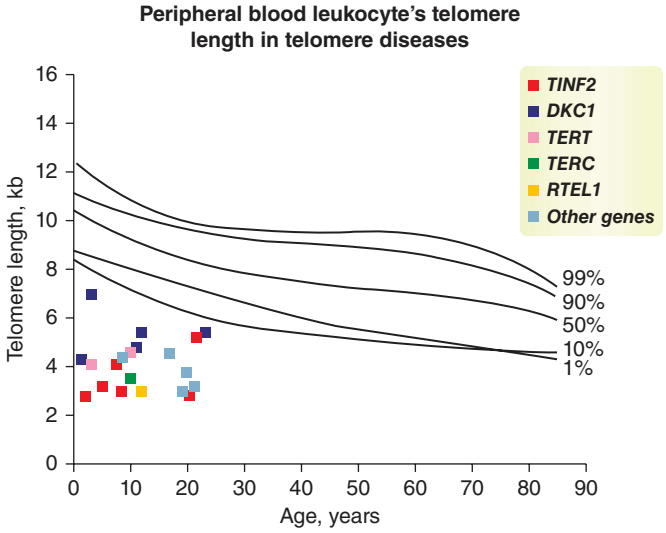


FIGURE 470-4 Telomere length measurement in dyskeratosis congenita. Telomeres shorten with aging, and solid curves represent the percentiles for age in healthy subjects. Telomeres are considered "short" when below the tenth percentile and very short when below the first percentile. In patients with dyskeratosis congenita, telomeres are usually below the first percentile, regardless of the gene lesion.

How telomere erosion causes pulmonary fibrosis is unknown, but it may prevent adequate proliferation and regeneration of pneumocytes type II. Idiopathic pulmonary fibrosis due to a telomere disease usually manifests after the fourth decade of life, with a restrictive pattern on pulmonary function testing associated with decreased diffusion capacity for carbon monoxide (DLCO) and diffuse interstitial lesion on high-resolution CT scan (Fig. 470-5). Histopathology of biopsied lung most commonly shows usual interstitial pneumonia. The pulmonary clinical presentation of patients with telomere disease is indistinguishable from idiopathic pulmonary fibrosis, except that those with an underlying telomere defect may have cryptic hepatic cirrhosis, macrocytosis, cytopenias, or a family history of lung, liver, or bone marrow disease. Pulmonary arteriovenous malformation leading to right-to-left shunting is observed in patients with pulmonary fibrosis due to telomere disease. Patients with idiopathic pulmonary fibrosis or familial pulmonary fibrosis should have their leukocyte telomere length measured and, if telomeres are short, undergo screening for mutations in telomere-associated genes and telomeres; however, telomere length may be normal in some cases despite the presence of pathogenic mutations. *TERT*, *TERC*, *RTEL1*, and *PARN* are the most commonly mutated genes.

Liver Disease Genetic telomere defects may cause hepatic cirrhosis (Chap. 337), nodular regenerative hyperplasia of the liver, nonalcoholic fatty liver disease (Chap. 336), and hepatocellular carcinoma (Chap. 78), and hepatocytes of most patients with cirrhosis have very short telomeres. Eroded telomeres limit hepatocyte proliferation, especially upon chronic injury. Additionally, hepatocytes with short telomeres display abnormal metabolic pattern and defective mitochondrial function. Abnormal liver pathology may be uncovered incidentally during the evaluation of telomeropathy patients with aplastic anemia or pulmonary fibrosis, but cirrhosis also may be the sole or most prominent clinical presentation of a telomere defect. A minority of individuals with cirrhosis associated with virus B or C infection or alcohol-associated liver disease may carry a telomere-associated gene mutation. Liver histopathology is variable, but cirrhosis is usually associated with inflammation and inflammatory cells (Fig. 470-5), increased iron deposit, positivity for CD34 in sinusoid endothelial cells, and widening of hepatocyte plates. Defective telomere maintenance may increase susceptibility of the liver to environmental challenges, such as viruses and toxins, increasing the risk for developing severe hepatic disease in mutation carriers.

TELOMERE LENGTH MEASUREMENT

Length of telomeres can be accurately measured in peripheral blood leukocytes by commercial laboratories. Of several methods available, flow-FISH and quantitative real-time polymerase chain reaction (qPCR) are the most widely utilized. Both methods have advantages and limitations and require high-quality samples, usually fresh or freshly processed, as cell death and DNA degradation impact the accuracy of testing. Results are usually expressed as leukocyte telomere length in kilobases. However, the interpretation of length must account for patient age, due to physiologic telomere loss. A normal range for telomeres is available for each year of age, longest at birth and shortening at 40–60 base pairs per year (Fig. 470-1). For each age bracket, the percentile curves are calculated and a given patient's test result is interpreted in the context of normal age variation: telomeres below the tenth percentiles for age are defined as "short" and telomeres below the first percentile are considered "very short" (Fig. 470-4).

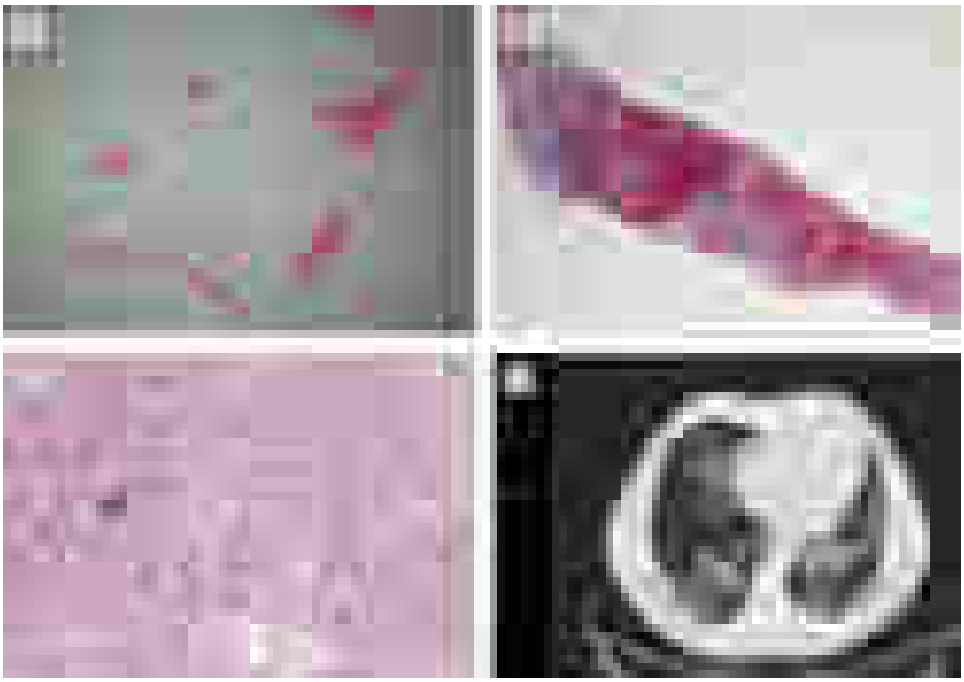


FIGURE 470-5 Pathologic manifestations of telomere diseases. ***A.*** In the bone marrow, telomere erosion predisposes to aplastic anemia, characterized by an empty hematopoietic marrow replaced by fat (hematoxylin and eosin). ***B.*** In the liver, telomere attrition predisposes to cirrhosis (hematoxylin and eosin). ***C.*** Telomere shortening may also result in nodular regenerative hyperplasia of the liver (reticulin stain). ***D.*** In the lungs, telomere dysfunction predisposes to pulmonary fibrosis mainly in the subpleural regions, which may be detected by high-resolution CT scan.

Telomeres may also be short in groups of patients with some chronic conditions, such as cardiovascular disease or diabetes. However, in these settings telomere erosion is not assumed to be etiologic, but rather a consequence of chronic inflammation; telomere testing is not known to have clinical utility and is not recommended. Likewise, telomere length tests have no known clinical utility in the assessment of aging and longevity or as a basis for therapeutic interventions absent a diagnosis of telomere disease.

Flow-FISH uses a fluorescent-labeled nucleotide probe specific for telomere repeats in order to estimate telomere content in an individual cell by flow cytometry. It has the advantage of determining telomere length in individual cells and in leukocyte subpopulations (neutrophils, lymphocytes, monocytes); lymphocyte telomere shortening is more specific for telomere diseases than in other cells. However, flow-FISH requires intact cells for analysis, not always available, and neutrophils are susceptible to damage during processing, freezing, and thawing.

Quantitative PCR (qPCR) utilizes telomere-binding modified primers to measure telomere content in comparison to a housekeeping gene in the whole leukocyte population and thus does not require intact cells. qPCR provides an estimate of the average telomere length of a given sample without determining telomere length in individual cells. Good DNA quality is essential for adequate qPCR testing and automation or semi-automation required for clinical purposes, as variability in conditions among batches may result in inter-assay variation.

GENETIC TESTING

When a patient with a suspected telomeropathy has short or very short telomeres, genetic screening for mutations in genes involved in telomere maintenance and biology is warranted (Table 470-1). Genetic testing should be restricted to patients with suspected telomere disorders. Sanger sequencing has been used for screening purposes but has been replaced by next-generation sequencing, and for telomerase complex and shelterin genes commercial panels are available. Mutations may be bi-allelic (especially in dyskeratosis congenita), but in most cases of aplastic anemia, myelodysplastic syndrome, acute myeloid leukemia, idiopathic pulmonary fibrosis, and hepatic cirrhosis, only one gene is mutated. Thus, it is crucial to appropriately interpret genetic screening results, as several rare singleton mutations of unknown significance have been identified in large cohorts of healthy individuals. In silico analysis, mutation location, and functional studies are utilized before declaring a mutation pathogenic.

Adequate genetic counseling is necessary after genetic screening, as the inheritance pattern may be autosomal dominant, mutation penetrance is highly variable, and phenotypes may be diverse even within a pedigree. Potential family stem cell donors must be screened before transplantation to ensure that they do not have mutations.

TREATMENT

Telomere Disease

Patients with severe aplastic anemia due to telomere disease may undergo allogeneic hematopoietic stem cell transplant when a suitable donor is available. Treatment-related mortality may be increased due to pulmonary and hepatic complications. Lung transplant for pulmonary fibrosis is feasible but often not performed due to coexisting cytopenias, especially thrombocytopenia, and other comorbidities. Similarly, there is no specific treatment for the liver in telomere disease; liver transplant has been attempted in rare cases. Telomeropathy patients should be advised to avoid toxins (metal dust, busulfan, amiodarone), ionizing radiation, cigarette smoke, and alcohol as possibly harmful.

Sex hormones may mitigate telomere attrition and even elongate telomere length. In case reports and a small research trial, danazol improved blood counts in marrow failure patients, and it may even slow progression of lung and liver disease in patients with telomeropathies.

FURTHER READING

Blackburn EH et al: Human telomere biology: A contributory and interactive factor in aging, disease risks, and protection. Science 350:1193, 2015.

Calado RT, Young NS: Telomere diseases. N Engl J Med 361:2353, 2009.

Collins J, Dokal I: Inherited bone marrow failure syndromes. Hematology 20:433, 2015.

Devine MS, Garcia CK: Genetic interstitial lung disease. Clin Chest Med 33:95, 2012.

Holohan B et al: Telomeropathies: An emerging spectrum disorder. J Cell Biol 205:289, 2014.
Newton CA et al: Telomere-related lung fibrosis is diagnostically heterogeneous but uniformly progressive. Eur Respir J 48:1710, 2016.
Savage SA, Bertuch AA: The genetics and clinical manifestations of telomere biology disorders. Genet Med 12:753, 2010.
Townsley DM et al: Bone marrow failure and the telomeropathies. Blood 124:2775, 2014.
Townsley DM et al: Danazol treatment for telomere diseases. N Engl J Med 374:1922, 2016.

471 The Role of Epigenetics in Disease and Treatment

Brian C. Capell, Shelley L. Berger

The term *epigenetics* was first coined by Waddington in 1942, as he sought to explain how changes in phenotype could occur throughout development independent of any changes to genotype. Appending the prefix *epi-* (Greek, meaning "over, outside of, around") to genetics aptly describes the numerous mechanisms by which gene expression and phenotypes are influenced, independent of any changes to the underlying DNA sequence. Today, epigenetics occupies one of the most exciting topics in biology and medicine, offering profound opportunities for discovery, as well as promise for the development of new therapies for disease. Interdisciplinary by nature, the field crosses virtually all areas of science and medicine: chemistry and genetics, development and differentiation, immunology, cancer, aging, and neuroscience.

The continuous introduction of ever more powerful techniques for interrogating the epigenome has led epigenetics to become one of the most innovative fields within the biomedical sciences. Given the vast expanse of the topic and limitations of space, in this chapter we provide a broad overview of the field and highlight key areas from across the landscape of biomedicine where epigenetics plays critical roles in disease, and importantly, where epigenetics-based therapies have demonstrated success in clinical medicine.

THE BIOCHEMICAL BASES OF EPIGENETICS

Fundamental to epigenetic regulation is the intricate organization of each cell's genome into chromatin **(Chap. 456)**. The fundamental unit of the packaging into chromatin is the nucleosome, consisting of 147 base pairs of DNA wrapped around an octamer of 8 histone proteins (two copies of each of the four core histone proteins: H2A, H2B, H3, and H4). The level of compaction of this chromatin structure determines the accessibility of the DNA strand to transcription factors, the DNA repair machinery, and other DNA-binding entities. Thus, compaction has a profound influence on gene expression levels and on local DNA mutation rates. Open regions of chromatin (euchromatin) tend to be transcriptionally active, whereas compacted chromatin (heterochromatin) tends to be transcriptionally repressed.

Histones include the four core histones, which are the most abundant and most frequently found throughout the genome, and variant histones of H2A, H2B, and H3. The structure of core and variant histones include amino- and carboxyl-terminal "tails," which are extended and unstructured, and highly conserved globular domains. The x-ray crystal structure of the nucleosome particle has illuminated the dynamic alterations of chromatin by an astounding range of regulatory mechanisms, summarized below.

The three main processes that regulate chromatin compaction, and thus access to the DNA template, include direct methylation modifications of the DNA strand itself, post-translational modifications of histones, and remodeling of nucleosomes to alter their location and composition with variant histones **(Fig. 471-1)**. The major modification of DNA is cytosine methylation of CpG dinucleotides (5-mC), associated with gene repression and catalyzed by the DNMT1, DNMT3A, and DNMT3B enzymes. DNMT3A and 3B catalyze the addition of methyl groups on unmethylated DNA de novo at CpG dinucleotides that are typically located throughout transcribed genes and in intergenic regions, but lacking at promoters, while DNMT1 is critical for the maintenance of the methylation state after DNA replication and after transcription during the S phase of the cell cycle. To further alter and reverse methylation, the TET enzymes (TET1–3) progressively oxidize 5-methylcytosine (5-mC) to 5-hydroxymethylcytosine (5-hmC), to 5-formylcytosine (5-fC) and 5-carboxylcytosine (5-caC), which are unable to be recognized by DNMT1 but can be removed by additional enzymes. Hence, these are mechanisms to passively lose 5-mC following DNA replication, or to actively remove 5-mC, both returning to unmethylated cytosine.

Histone post-translational modifications (hPTMs) are rich sources of diverse signaling to, and marking of, the chromatin template, including at least 60 different covalent chemical modifications on the histone N- and C-terminal tails and within the globular domains. The hPTMs are added (written) and removed (erased) by enzymes, and also serve as sequence- and PTM-specific binding surfaces for effector proteins and complexes (readers) to carry out a wide range of downstream actions including transcription, replication, and DNA repair and recombination. One key point is that the staggering numbers of writers, erasers, and readers provide unlimited potential for diagnostic and therapeutic pharmacologic discovery.

Throughout this chapter, we focus on histone methylation and acetylation, the most abundant and the most well-studied hPTMs (Fig. 471-1), although several others, such as serine/threonine/tyrosine phosphorylation, lysine ubiquitination, lysine SUMOylation, and lysine ADP-ribosylation also play important roles in epigenetic regulation. For instance, histone phosphorylation targets histone H2A at Ser139 (γH2A.X), which marks DNA double-strand breaks immediately following damage and is critical for the recruitment of the DNA repair machinery. Histone mono-ubiquitination functions similarly to other hPTMs, in signaling and marking the chromatin template, in particular serving to mark the initiation region or elongation of transcribed genes for future rounds of transcription, whereas histone SUMOylation plays a role in transcriptional repression. Poly-ubiquitination serves to tag proteins for degradation by the proteasome, and dysfunction in this system may play a role in the pathogenesis of neurodegenerative diseases such as Alzheimer's, Parkinson's, and Huntington's disease. ADP-ribosylation involves a class of enzymes, the poly-ADP-ribose polymerases (PARPs), which transfer ADP-ribose units from NAD+ to a variety of nuclear proteins. This PARylation alters the chromatin environment through the recruitment and modification of chromatin-associated proteins. In general, future studies of the wealth of types and functions of hPTMs will enhance our understanding of these chromatin-based mechanisms and processes and will illuminate new opportunities and targets for therapies.

In contrast, there is extensive understanding of histone lysine acetyltransferases (KATs) and methyltransferases (KMTs). KATs, previously known as HATs, were among the first identified histone modification enzymes. They attach acetyl groups on the lysine residues of histone tails and other proteins, resulting in both a novel side chain (acetyl-lysine) and an increase in negative charge (from positive charged lysine to neutral acetyl-lysine). This alteration results in loosening of chromatin structure to become more permissive to the binding of transcription factors, and it also creates a novel binding surface for the association of reader proteins. Acetylation on core histones, such as lysine 9 on histone H3 (H3K9ac) or lysine 27 (H3K27ac), is typically associated with transcriptional activation. Acetylation is very dynamic and can be rapidly removed by histone deacetylases (HDACs), of which there are multiple classes, including HDACs and sirtuins (NAD-dependent deacetylases), which return the lysine to unmodified ground state.

Methylation of histone tails by KMTs provides more nuanced regulation, in that particular methylated lysines are associated with transcriptional activation (e.g., H3K4me3, H3K36me3, H3K79me3), transcriptional repression (e.g., H3K27me3), or DNA repeat and

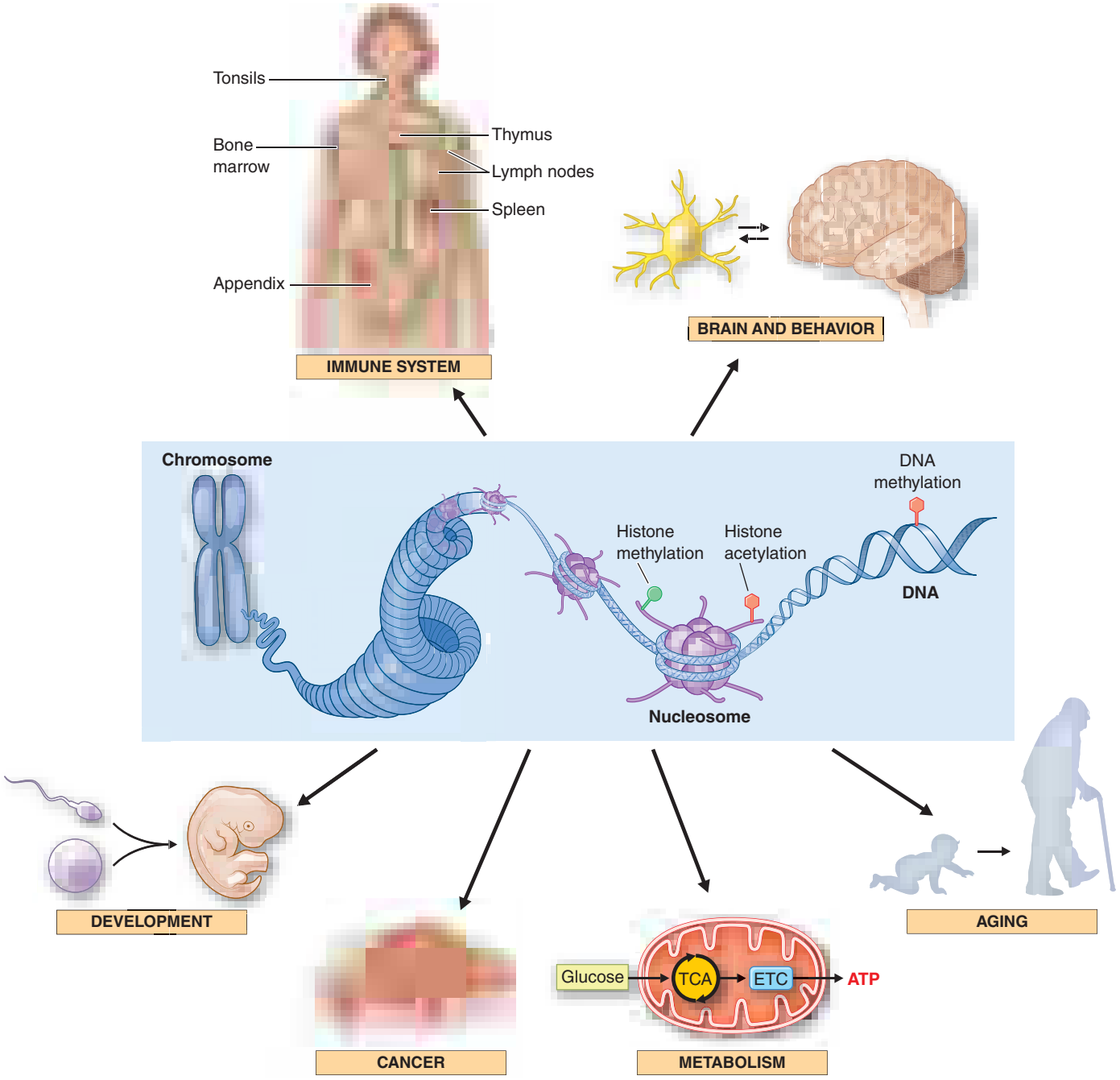


FIGURE 471-1 Epigenetic pathways influence multiple physiologic and disease pathways. As depicted in the center of the illustration, epigenetics refers to the chemical modifications of DNA and histones, which influence chromatin structure, gene expression, and susceptibility to mutations. These molecular pathways, in turn, play important roles in development, cancer, metabolism, aging, neural function, and behavior, and in the immune system.

centromeric silencing (e.g., H3K9me3). The output is strictly determined by effector protein binding, as methylation of lysine does not alter side chain electrostatic charge. Lysine methylation is a more stable chemical modification than is acetylation, and, while demethylases are identified for some of the specific methylated sites (H3K4, H3K9, H3K36, H3K27), it is provocative that H3K79 and H4K20 demethylases have not yet been discovered.

Frequently coordinating with histone modification enzymes are nucleosome remodeling enzymes, which use the energy derived from the hydrolysis of ATP to reposition and remove nucleosomes along the DNA template, and to exchange core histones and variant histones (including variants that are located at the transcriptional initiation sites [H2AZ] and over the transcribed genes [H3.3]). These complexes activate or repress transcription. The SWI/SNF family creates nucleosome-free regions for transcriptional activation, the ISWI family evenly spaces nucleosomes to repress transcription, and the INO80 family exchanges H2A with H2AZ at transcription start sites to poise transcriptional activation. Other remodeling complexes play key roles in the DNA damage response and apoptosis, among additional genomic processes.

Because multiple enzymes redundantly write, erase, and recognize many of the hPTMs, there is great complexity and the potential for fine-tuning of gene regulation. While extensive knowledge gaps remain to fully explicate mechanisms of chromatin regulation, epigenetics has become a fully established discipline within biomedical research. In the coming years, it is likely that the basic understanding of these processes will be harnessed for further betterment of human health.

EPIGENETICS IN DEVELOPMENT AND DIFFERENTIATION

Epigenetic processes are critical to organismal development and to cellular differentiation and reprogramming of cell fate (Fig. 471-1).

Transcription factors establish the epigenomic landscape that enables and stabilizes cell type–specific gene expression while simultaneously ensuring stable repression of alternative cell fates. This results in chromatin profiles that display remarkable cell-type specificity in differentiated cells, particularly at the key regulatory nodes of gene enhancers. In fact, epigenome profiling of the chromatin landscape in tumors of unknown cell origin can provide a better index of the origin tissue than does sequencing the tumor itself to catalog genetic mutations.

The cell type–specific epigenetic program is first derived from the template of embryonic stem cells, where numerous genes required for differentiation exist in a "bivalent" state, marked by both the activating histone modification, H3K4me3, and the repressive modification, H3K27me3. From this state, the genes are thought to be "poised" and ready either for activation or for repression, depending on their subsequent cell fate. Critical genes directing toward a specific cell fate will be turned on, whereas genes leading toward alternative fates will be repressed. Once differentiated, an epigenetic barrier will prevent the cells from returning to the stem cell state. For example, heterochromatin in the form of H3K9me3 can serve as a barrier to cellular reprogramming when attempting to create induced pluripotent stem cells, and inhibiting the enzymes that catalyze H3K9me3, such as SUV39H1, can enhance reprogramming efficiency.

DNA methylation contributes to the specification of cell fate and to other developmental pathways. DNA methylation alterations are involved in critical processes ranging from sex chromosome dosage compensation to coordinating expression of imprinted genes. Disruption of this latter process can lead to imprinting disorders including Prader-Willi syndrome, Angelman syndrome, and Beckwith-Wiedemann syndrome.

Beyond embryonic development, epigenetics can provide the necessary coordination and balance between adult stem cell renewal and differentiation. This epigenetic control is critical, as impaired self-renewal can lead to stem cell exhaustion and premature aging, while excessive self-renewal may promote cancer. Key epigenetic regulators tend to play conserved roles across diverse tissue types. For instance, BMI1, a component of the polycomb repressive complex 1 (PRC1), is required for stem cell proliferation and self-renewal, and its ablation leads to stem cell depletion in hematopoietic, epidermal, muscle, intestinal, and mammary stem cells. Similarly, the DNA methyltransferase DNMT1 also is required for stem cell self-renewal in hematopoietic, epidermal, and mammary stem cells. HDACs 1 and 2 possess some overlapping functions and are required for normal epidermal differentiation. Likewise, a loss of these enzymes in hematopoietic stem cells can lead to failure of differentiation and severe anemia. These factors represent repressive chromatin regulation, leading to the concept that restraining specific transcription pathways related to differentiation are crucial to maintaining undifferentiated stem cell pools.

The epigenetic regulation of the tumor suppressor p16 (*CDKN2A*) locus during differentiation provides a prime example of this finely tuned system. For example, as mentioned above, DNMT1 is necessary for self-renewal in human epidermal stem cells. Levels of DNMT1 are high in the basal undifferentiated layer of the epidermis, decreasing progressively with epidermal stratification, leading to derepression of the tumor suppressors p16 and p15, thereby promoting cell cycle arrest and full differentiation. BMI1 displays a similar phenotype in both hematopoietic and epidermal stem cells, repressing key genes that promote differentiation, such as p16 and p19ARF. Consistently, a loss of BMI1 leads to premature differentiation and defective self-renewal. In addition to the repression provided by DNMT1 and BMI1, the p16 locus is highly decorated with the repressive H3K27me3 catalyzed by EZH2 in epidermal stem cells. Then, during epidermal differentiation, H3K27me3 is removed by the KDM6B (JMJD3) histone demethylase. Loss of this control over programmed p16 expression occurs in epithelial cancers, such as squamous cell carcinoma (SCC), where EZH2 is overexpressed and KDM6B expression is lost. In the breast, progesterone can lead to increased levels of EZH2 to promote mammary epithelial cell proliferation. However, excessive EZH2 expression occurs in breast cancers, exemplifying how epigenetics can integrate environmental signals and have a profound influence on the fine balance between stem cell maintenance and overt carcinogenesis. Indeed, loss of key chromatin regulation that promotes cell differentiation, and gain of activities that promote stemness, is a recurrent theme in cancer.

Chromatin modifying enzymes also play a major role in influencing cell type specificity. High levels of EZH2 that modify H3K27me3 promote adipogenesis while simultaneously inhibiting osteogenesis. In contrast, the H3K27me3 demethylases, KDM6A (UTX) and KDM6B (JMJD3), derepress those same genes, driving stem cells toward osteogenesis. Through interactions with tissue-specific master regulators, epigenetic modifiers also shape cell type specificity. In the epidermis, p63, the p53 family member that is a master regulator of the epidermal compartment, interacts with several chromatin regulators including HDAC1 and HDAC2, SATB1, and BRG1 to orchestrate epidermal differentiation. Similarly, the gene-activating H3K4 histone methyltransferases, MLL3 (KMT2C) and MLL4 (KMT2D), are required for adipogenesis by forming a complex with the transcriptional activator ASC2 and the transcription factor PPARγ to induce adipogenic genes.

EPIGENETICS OF METABOLISM

One of the fascinating aspects of epigenetics is that it represents a mechanism for direct connection between the environment and gene expression. Numerous studies in the field of metabolism have identified a complex interplay between diet, metabolism, and the epigenome (Fig. 471-1). Seminal findings in *Drosophila* and mice have shown that changes in diet, particularly the paternal diet, and other environmental factors, can influence the metabolism of offspring, ultimately promoting obesity in later generations. Epidemiologic studies in humans have supported these results, as the nutritional status of grandparents has been correlated with phenotypic effects in grandchildren. In fact, diet can directly affect the levels and activity of chromatin modifiers. For instance, HDAC9 is elevated by high-fat diets, and inhibition of HDAC9 can be protective against the deleterious effects of high-fat diets in mice, including weight gain, decreased glucose tolerance, and increased insulin resistance.

These connections are driven by metabolites that constitute key cofactors for post-translational histone modifications such as acetylation and methylation. These include acetyl-CoA for histone acetylation and S-adenosylmethionine (SAM) for histone and DNA methylation. Levels of acetyl-CoA, in comparison to all measured metabolites, are indeed the best predictor of histone acetylation levels. Consistent with this, increased acetyl-CoA correlates with rising levels of total histone acetylation, including at the promoters of growth-associated genes. This increase in nuclear acetylation is associated with cell cycle progression and proliferation, and it can have clinically relevant downstream effects. For example, high levels of acetyl-CoA can delay stem cell differentiation and suppress autophagy. The oncogenes MYC and AKT can both hijack metabolic networks to enhance nutrient uptake by cancer cells, thus promoting acetyl-CoA production and resulting in both the initiation and progression of tumorigenesis.

Methylation is also altered by metabolism. Dietary factors are estimated to explain 30% of the variation in human serum methionine concentration, and this can alter histone methylation. For example, dietary methionine availability and intracellular production of SAM affects the levels of histone H3K4me3 associated with transcriptional activation. Furthermore, these fluctuations can have critical physiologic consequences: DNA methylation levels in rectal mucosa and colonic polyps are increased by higher levels of dietary folate, and a diet low in methyl donors reduces the formation of gastrointestinal cancers in mice predisposed to these tumors. Methionine metabolism and the availability of SAM regulates stem cell differentiation and contributes to carcinogenesis. For instance, cancers that display hypermethylation, including those with *IDH* mutations, are associated with poorly differentiated gene expression profiles. In contrast, loss of the *MTAP* gene, which is part of the 9p21 locus containing p16, and one of the most frequent events in human cancer, disrupts normal methionine metabolism. While this lowers methylation levels, interestingly, it can also sensitize cancer cells to inhibitors of the PRMT5 methyltransferase,

therefore opening a new therapeutic opportunity. These observations illustrate how connections between epigenetics and metabolism can generate unanticipated advances in medicine.

CANCER EPIGENETICS

Cancer is now understood to be a mixed genetic and epigenetic disease, as epigenetic dysregulation is pervasive in human cancers (Fig. 471-1). Beyond simple activation of oncogenes or reduced expression of tumor suppressors, epigenetic mechanisms can contribute to chemotherapy resistance and to failure of antitumor immunity. Accordingly, the development of drugs targeting epigenetic pathways is one of the most active areas of clinical and pharmaceutical development, with several compounds already approved for human use and shown to be effective in a variety of cancers. Epigenetic perturbations in cancer largely affect chromatin-regulating enzymes, which represent robust targets for development of novel small-molecule inhibitors, especially as compared with canonical oncogenic transcription factors (e.g., MYC) and tumor suppressors (e.g., p53).

Epigenetics can contribute to carcinogenesis in a variety of ways. First, on a global scale, chromatin organization is the single most influential factor in determining local mutation rate across the genome. Analysis of abundant tumor sequencing data has demonstrated that heterochromatic regions of the genome contain a higher frequency of mutations compared with more open euchromatic regions. This difference is due to the improved accessibility of the DNA repair machinery to less compact, more open regions of chromatin.

The first discovery of an epigenetic mutation was found in 1998 when the chromatin remodeler *SMARCB1* was shown to drive the formation of malignant rhabdoid tumors. Extensive sequencing of human tumors from the majority of cancer types has been performed by The Cancer Genome Atlas (TCGA) consortium, and, remarkably, 25–30% of identified cancer driver mutations occur in chromatin regulatory proteins. Similar to *SMARCB1*, numerous other chromatin modifiers (e.g., methyltransferases *MLL3* and *MLL4*, and acetyltransferases *EP300* and *CBP*) and nucleosome remodeling enzymes and associated complex components (e.g., *SMARCA4*, *SMARCA2*, *ARID1A*) are heavily mutated and inactivated in many cancers. The majority of these mutations are loss-of-function mutations, and, indeed, enzymes like MLL4 and demethylase KDM6A possess tumor-suppressive activity. In contrast, the H3K27me3 histone methyltransferase EZH2 is an oncogene, and accordingly, it is overexpressed in many advanced-stage or metastatic solid tumors such as breast cancer, prostate cancer, and melanoma. Mechanistically, EZH2 represses the p16 tumor suppressor and other cell cycle genes required for cell cycle exit via H3K27me3 deposition. Consistent with a broad growth regulatory role, EZH2 inhibitors are therapeutically successful for a number of cancers in preclinical models and are being actively studied for B cell lymphoma, melanoma, and other solid tumors.

Recently, provocative evidence has emerged for a direct tumorigenic role of histones based on the discovery of causative mutations, such as histone H3 mutations identified in pediatric high-grade gliomas. Specifically, the majority of these mutations are in the H3 variant H3.3, where lysine 27 is replaced by methionine (K27M). Similarly, over 90% of chondroblastomas replace lysine 36 with methionine (K36M) in histone H3.3. These effects appear to be dominant negative because (1) in H3.3 these are heterozygous mutations, and (2) the mutations also occur in the canonical H3, which exists in approximately 30 orthologous genes in the human genome. Thus, a minority of H3/H3.3 mutant protein leads to global defects in the associated histone modifications (K27 or K36 methylation), possibly via irreversible inhibition of the cognate enzymes by the mutant histones. These histone mutations promote resistance to apoptosis and failure of normal differentiation in a number of pediatric cancers.

Beyond mutations, genetic translocations involving chromatin modifiers also implicate chromatin pathways as direct drivers in cancer. MLL1, the H3K4 histone methyltransferase, is a frequent translocation partner occurring in adult and pediatric acute myeloid leukemia (AML), and in approximately 80% of infant acute lymphoid leukemia (ALL) cases. MLL1 can fuse with more than 70 translocation partners, and these mutant proteins prevent normal hematopoietic differentiation. Consistent with a causative role of MLL1 in these gene fusions, drugs inhibiting the catalytic activity of MLL1 are effective in preclinical models of AML.

Given the abundance of epigenetic abnormalities in cancer combined with the inherent reversibility of epigenetic changes, extensive efforts are underway to develop epigenetic drugs. The first epigenetic therapeutic involved the use of DNA methylation inhibitors (DNMTi) to reactivate tumor suppressor genes. Interestingly, the mechanism of traditional chemotherapeutics, such as azacitidine and decitabine, is to inhibit DNMT1, thereby promoting global hypomethylation; these are currently in clinical use for myelodysplastic syndrome (MDS) and AML. In a second broad mechanism, loss of acetylation occurs in many cancers, and thus HDAC inhibitors (HDACi) are under intensive development. HDACi are effective and approved for treatment in cutaneous T cell lymphoma and multiple myeloma. Bromodomain (BRD)-containing proteins bind to lysine acetylated target proteins, including histones, and rationally designed BET inhibitors (BETi) block their binding. BETi reduce the amplified expression of oncogenes such as *MYC* in hematologic cancers. Current studies are focused on optimizing combinatorial therapies of these pan-DNMTi and pan-HDAC inhibitors with conventional chemotherapies and immunotherapies, particularly given the ability of epigenetic therapeutics to promote re-expression of tumor antigens and interferon (IFN)-mediated antitumor immunity.

There are several hundred chromatin enzymes and binding proteins in the human genome, and the current focus is on the identification of more specific inhibitors. Indeed, targeted inhibitors of numerous mutated chromatin regulators have been developed, with more than 30 compounds currently in various stages of development and preclinical trials. Some notable examples showing early clinical success include EZH2 inhibitors for lymphomas, sarcomas, and melanoma, IDH inhibitors for AML and gliomas carrying mutant *IDH1* or *IDH2* genes, LSD1 inhibitors for AML and small cell lung cancer, and DOT1L and MLL1 inhibitors for leukemias with activated MLL1.

EPIGENETICS OF AGING

Like many diseases of aging, human aging itself results from the complex interplay between genes and the environment. Evidence that the epigenome may be the key link between these processes derives from observations that numerous environmental stimuli and stressors—ranging from diet and exercise to hormones and circadian rhythms—contribute to both aging and epigenetic alterations (Fig. 471-1). Thus, a lifetime of exposures progressively disrupts the chromatin landscape. These age-dependent changes in chromatin organization increase the susceptibility of the genome to mutations and also reduce transcriptional fidelity. Further, provocative findings in model systems demonstrate that stress-induced epigenetic changes can be transmitted over several generations and can even affect the lifespan of later generations. Among these global epigenetic alterations, there is dysregulation of histone modifications and a general loss of histone proteins with aging across taxa. Amazingly, experimental increases in histone levels, particularly histones H3 and H4, but not H2A or H2B, can reverse these age-related changes in mammalian cells and yeast.

Thus, the sum of current evidence suggests a model of aging via a general increase in activating epigenetic modifications along with a loss of repressive modifications. Together these changes create a state of transcriptional instability and "noise" that inhibit accurate transcription. Cells from patients with Hutchinson-Gilford progeria syndrome (HGPS), the most severe form of human premature aging, display reduced levels of both H3K9me3 and H3K27me3 repressive chromatin. In another premature aging disease, Werner syndrome, DNA damage induces global loss of H3K9me3 and H3K27me3 due to the inherent absence of the Werner syndrome ATP-dependent DNA helicase, which is critical for DNA repair. Such heterochromatin loss is not limited to premature aging conditions, as aged cells derived from healthy older humans display age-dependent loss of H3K9me3 leading to aberrant expression of normally repressed transposable elements. Activation

of these mobile elements correlates with neurodegenerative disorders and may also promote other aging-related phenotypes such as cancer. Human fibroblasts undergoing cellular senescence (exit from cell cycle due to replicative or other stress) undergo destabilization of compact heterochromatin adjacent to the nuclear periphery, in so-called lamin-associated-domains (LADs). At LADs, in addition to a reduction of repressive histone modifications as discussed above, there are broad new regions of the euchromatic histone modification H3K4me3.

In addition to age-associated alterations of histone modifications, direct manipulation of chromatin modifying enzymes that control these marks affects the balance between heterochromatic and euchromatic regions, and it alters the lifespan of model organisms. Inhibiting the H3K27me3 histone demethylase UTX (KDM6A) results in increased repressive H3K27me3 and extended lifespan in *Caenorhabditis elegans*. Consistent with this, genetic reduction of enzymes (*ash-2*, *set-2*, *wdr-5*) that add the activating H3K4me3 histone modification also extends lifespan in *C. elegans*. The consequences of these genetic manipulations nicely correspond to the observed changes in histone modifications as described above. Beyond histone modifying enzymes, dysregulation of the levels or function of chromatin remodelers can also affect lifespan in model organisms. This dysregulation occurs in humans as well, as in the nucleosome remodeling deacetylase complex (NuRD), which is reduced in HGPS fibroblasts and in aged healthy donors.

In addition to age-related changes in histone methylation, histone acetylation also contributes to aging phenotypes. Dysregulation of histone acetyltransferases (HATs) and HDACs is associated with reduced longevity across model organisms. Further, sirtuin deacetylases (class III NAD^+-dependent HDACs) promote healthspan and lifespan across species as key mediators of pro-longevity effects of caloric restriction. Indeed, loss of Sirt6 results in premature aging in mice. As discussed previously, metabolism and acetylation are intricately linked, and the sirtuins, via NAD^+ levels, and other HDACs may play key roles connecting the environment, gene expression, and physiologic output. For instance, exercise in humans reduces activity of HDACs 4 and 5, leading to increased H3K36ac in skeletal muscle, which likely promotes beneficial gene expression.

Epigenetic alterations with aging are not limited to histone modifications and extend to DNA methylation. Consistent with the histone patterns, DNA methylation data support the model described above—that is, general decompaction of the epigenome with aging. Specifically, levels of 5-methylcytosine (5-mC) are reduced in senescent humans cells, and global DNA hypomethylation occurs across the human genome with aging. Concurrent with this overall hypomethylated state, there are local regions of hypermethylation focused near CpGs at gene promoters, particularly at genes that maintain cellular differentiation and cell identity. This epigenetic disruption during aging thus leads to profound changes in transcription. For example, in HSCs, DNA hypermethylation blocks proper binding of transcription factors, resulting in dysregulation of normal gene expression with aging. Importantly, these patterns are not merely stochastic alterations in response to environmental stressors throughout aging. Indeed, the methylation status of a defined number of CpG sites is a highly accurate predictor of chronologic age in human tissues. This work reveals that DNA methylation status with aging outperforms previous standard biomarkers of aging, such as p16 expression levels and telomere length, and will be highly valuable in the near future to gauge effects of treatment aiming to ameliorate diseases of aging

EPIGENETICS OF THE BRAIN AND BEHAVIOR

Brain disorders are among the greatest clinical challenges to understand and to treat. Most neurologic and psychiatric disorders result from complex dysregulation of numerous genes and pathways. In this interplay between underlying genetic predisposition and external environmental factors, aberrant epigenetic regulation is increasingly recognized as a potentially key modulator (Fig. 471-1).

More directly, however, several progressive neurodevelopmental disorders are caused by germline mutations in chromatin regulators. Mutations in methyl CpG binding protein 2 (*MECP2*), a protein important for binding to methylated DNA and contributing to gene repression, is the major cause of Rett syndrome. MeCP2 loss leads to overactive gene transcription in neurons and impaired presynaptic excitatory functions. Similarly, Kabuki syndrome, another progressive neurodevelopmental disorder, is caused by germline mutations in either the H3K4me1 histone methyltransferase, *MLL4* (*KMT2D*), or the H3K27me3 demethylase, *UTX* (*KDM6A*). This disorder may derive from dysregulation of transcriptional enhancers, a major class of gene regulatory elements, as both MLL4 and UTX play a key role in activation of enhancers. Finally, the acetyltransferase CBP (*CREBBP*) also is important for gene enhancer function, and when mutated can lead to Rubinstein-Taybi syndrome, a cause of intellectual disability.

Beyond germline mutations, altered methylation dynamics can drive disorders of neural development and of neurodegeneration. Fragile X syndrome, characterized by learning disabilities and cognitive impairment, is caused by mutations in the *FMR1* or *FMR2* genes, or by hypermethylation of the transcriptional promoters regulating *FMR1* or *FMR2*. Similarly, Prader-Willi syndrome and Angelman syndrome, neurodevelopmental conditions caused by abnormal imprinting of the paternal or maternal chromosomal region (15q11-13), respectively, are frequently caused by aberrant DNA methylation. Further, DNA hypomethylation is implicated in some neurodegenerative conditions. For instance, in Parkinson's disease, several genes involved in pathogenesis are hypomethylated due to DNMT1 depletion, including the α-synuclein gene (*SCNA*). In Alzheimer's disease (AD), DNA hypomethylation occurs at promoters of key pathogenic genes such as amyloid precursor protein (*APP*). Indeed, APP promoter methylation is responsive to environmental factors, including aging, a major risk factor for AD. Likewise, presenilin-1 (*PSEN1*) is implicated in AD and displays altered DNA methylation in response to variations in metabolic stimuli. Studies of Huntington's disease (HD) have demonstrated DNA hypomethylation and decreased histone acetylation, in part due to altered function of the acetyl transferase CBP, leading to transcriptional dysregulation. Together these observations underscore epigenetic regulation as a crucial feature of neurodegeneration.

Additional gene regulatory proteins in the nervous system interact with and are regulated by chromatin modifiers. REST (repressor element 1–silencing transcription factor) is important in neuronal homeostasis through its ability to recruit chromatin regulatory enzymes, such as histone deacetylases and histone methyltransferases, and via its control over gene expression. REST levels increase with aging and serve a protective function in neurons against age-associated stressors and loss of cognitive function associated with AD. Similar to REST, brain-derived neurotrophic factor (BDNF), another important mediator of neural development and homeostasis, is implicated in a variety of neurologic and psychiatric disorders including depression, schizophrenia, bipolar disorder, and autism. Knockdown of BDNF in the dentate gyrus leads to depression-like behavior in mouse models, and BDNF mediates effects of antidepressant therapies. Chromatin pathways, including DNA methylation/MeCP2 and H3K27me3, play a key role in BDNF regulation as observed in brains from patients with schizophrenia.

Finally, addiction medicine is another frontier where epigenetics holds great promise to reveal connections between environmental exposure and phenotypes. Although still in its early stages in terms of mechanistic understanding, emerging evidence demonstrates disruption of epigenetic homeostasis as a consequence of addictive substances ranging from alcohol to cocaine. For example, the acetylation of regulatory elements in the *FOSB* gene by the histone acetyltransferase CBP is associated with behavioral effects of cocaine. Ethanol also induces histone acetylation and a decompacted chromatin structure.

EPIGENETIC INFLUENCES ON INFECTION, IMMUNITY, AND INFLAMMATION

Alterations in gene expression patterns are important determinants of immune-mediated disease, and in turn, epigenetics regulates immunity and inflammation (Fig. 471-1). Treatment with immune-stimulating agents such as lipopolysaccharide (LPS) and TNF-α activate expression of numerous inflammatory genes within hours, with precise gene pathways and activation kinetics determined by the cellular epigenetic

state. HATs and HDACs are critical components of this response, coordinating with pro-inflammatory transcription factors such as AP-1 and NF-κB, to activate (HATs) and to repress (HDACs). For example, corticosteroids recruit HDAC2 to promoters of NF-κB-stimulated inflammatory genes to prevent activation during asthma treatment.

Type 1 IFN responses are exceptional examples of regulatory complexity governed by epigenetic control. In an unstimulated state, the H3K9 methyltransferases G9a (*EHMT2*) and EHMT1 suppress expression of IFN and IFN-induced genes. Upon induction of IFN-stimulated genes, STAT transcription factors recruit chromatin remodeling complexes, such as BAF (*SMARCA4*), and recruit HATs including p300, CBP, and GCN5 (*KAT2A*). In turn, chromatin remodeling and acetylation recruit chromatin binding proteins including the bromodomain protein, BRD4, which promotes transcriptional elongation and full activation.

Major regulators of adaptive immunity pathways are similarly epigenetically regulated. CD4+ and CD8+ T cells undergo extensive changes in histone modification profiles during differentiation to distinct subsets of effector T cells. For example, genes associated with effector T cell functions in CD8+ memory T cells (such as *PRDM1*, *KLRG1*, *IFNG*) display enrichment of H3K4me3 and low levels of H3K27me3 compared with those genes in naïve T cells. DNA methylation also plays an important regulatory role and may contribute to disease. For example, CD4+ T cells from individuals with rheumatoid arthritis (RA), systemic scleroderma, and latent autoimmune diabetes in adults display hypermethylation of the *FOXP3* gene, which activates regulatory T cells that dampen immune responses. In addition, hypermethylation of the *CTLA4* locus occurs in regulatory T cells from RA patients, impairing their immunosuppressive abilities.

Although mechanistic studies remain limited, there are numerous examples of epigenetic-based therapies associated with extensive effects on the immune system, underscoring the potential hope for eventual treatment of immune-related conditions. For example, the DNA methylation inhibitors azacitidine and decitabine have immunosuppressive effects possibly mediated by enhanced expression of *FOXP3*, which generally suppresses immune responses. HDAC inhibitors upregulate and downregulate immune genes, and they inhibit cytokine production in macrophages from patients with RA. Further, the HDAC inhibitors vorinostat and panobinostat inhibit primary B cell responses and antibody production in vitro and in vivo. Given these broad effects, it is not surprising that the HDAC inhibitor trichostatin A (TSA) has efficacy in various model systems for treatment of RA, systemic lupus erythematosus (SLE), asthma, acute kidney injury, sepsis-induced lung and cardiac damage, and acute pancreatitis. Similarly, BET inhibitors also display broad effects. They block antigen presentation and T and B cell activation, and thus have beneficial protective effects in a variety of inflammatory settings including autoimmunity, sepsis, atherosclerosis, psoriasis, periodontitis, and arthritis. Beyond these "broad-spectrum" epigenetic inhibitors, a specific inhibitor of the H3K27me3 demethylases KDM6A and KDM6B, GSK-J4, has anti-inflammatory activity, presumably by preventing loss of H3K27me3 repression over inflammatory genes.

CONCLUSIONS

Due to the enormity and complexity of the chromatin and epigenetics fields, and their reach into all areas of biology and medicine, it is not possible to cover such a broad scope in a single chapter. Thus, here we provide a concise snapshot highlighting key areas of development in medicine. We hope to have conveyed the tremendous excitement and promise that pervades the discipline. Indeed, given the exponential growth in uncovering the interface between the epigenome and epigenetic therapies with the environment and disease, there is little doubt that the coming years will bring important additions to this field.

FURTHER READING

Allis CD, Jenuwein T: The molecular hallmarks of epigenetic control. Nat Rev Genet 17:487, 2016.

Avgustinova A, Benitah SA: Epigenetic control of adult stem cell function. Nat Rev Mol Cell Biol 17:643, 2016.

Benayoun BA et al: Epigenetic regulation of ageing: Linking environmental inputs to genomic stability. Nat Rev Mol Cell Biol 16:593, 2015.

Brien GL et al: Exploiting the epigenome to control cancer-promoting gene-expression programs. Cancer Cell 29:464, 2016.

Jones PA et al: Targeting the cancer epigenome for therapy. Nat Rev Genet 17:630, 2016.

Kinnaird A et al: Metabolic control of epigenetics in cancer. Nat Rev Cancer 16:694, 2016.

Sen P et al: Epigenetic mechanisms of longevity and aging. Cell 166:822, 2016.

Tough DF et al: Epigenetic drug discovery: Breaking through the immune barrier. Nat Rev Drug Discov 15:835, 2016.

Tsankova N et al: Epigenetic regulation in psychiatric disorders. Nat Rev Neurosci 8:355, 2007.

Yao B et al: Epigenetic mechanisms in neurogenesis. Nat Rev Neurosci 17:537, 2016.

472 Mitochondrial DNA and Heritable Traits and Diseases

Karl L. Skorecki, Bruce H. Cohen

Mitochondria are cytoplasmic organelles whose major function is to generate ATP by the process of oxidative phosphorylation under aerobic conditions. This process is mediated by the respiratory electron transport chain (ETC) multiprotein enzyme complexes I–V and the two electron carriers, coenzyme Q_{10} (CoQ_{10}) and cytochrome c. Other cellular processes to which mitochondria make a major contribution include apoptosis (programmed cell death) and additional cell type-specific functions **(Table 472-1)**. The efficiency of the mitochondrial ETC in ATP production is a major determinant of overall body energy balance and thermogenesis. In addition, mitochondria are the predominant source of reactive oxygen species (ROS), whose rate of production also relates to the coupling of ATP production to oxygen consumption. Given the centrality of oxidative phosphorylation to the normal activities of almost all cells, it is not surprising that mitochondrial dysfunction can affect almost any organ system **(Fig. 472-1)**. Until recently, it was thought that disruption of energy production was the source of the pathophysiology in those with mitochondrial dysfunction, but recent evidence suggests that free-radical production and the redox state of the mitochondria may play a role as well. Thus, physicians in many disciplines might encounter patients with mitochondrial diseases and should be aware of their existence and characteristics.

The integrated activity of estimated 1500 gene products is required for normal mitochondrial biogenesis, function, and integrity. Aside from the 37 genes that comprise the mitochondrial DNA (mtDNA)

TABLE 472-1 Functions of Mitochondria

All Cells and Tissues
Oxidative phosphorylation
Free radical production
Calcium homeostasis
Apoptosis (programmed cell death)
Tissue- or Cell-Specific
Cholesterol metabolism
Amino and organic acid metabolism
Fatty acid beta oxidation
Sex steroid synthesis
Heme synthesis
Hepatic ammonia detoxification
Neurotransmitter metabolism

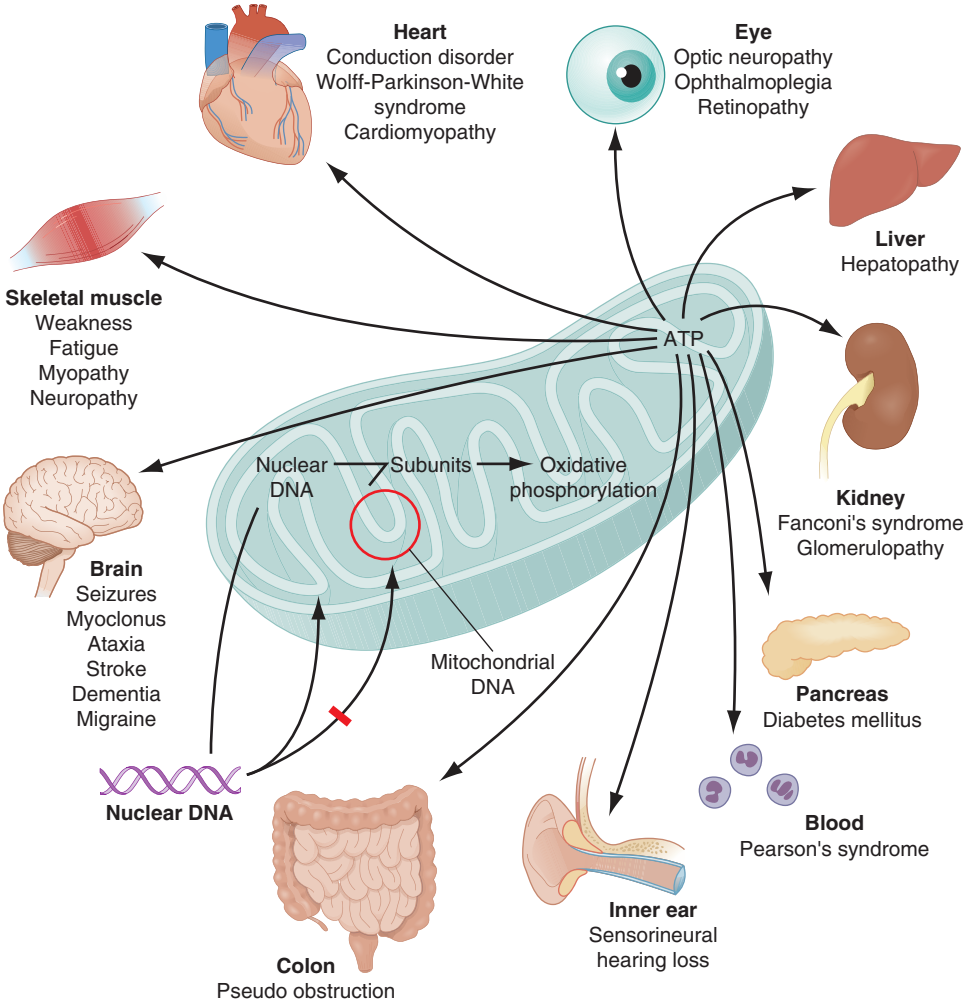


FIGURE 472-1 Dual genetic control and multiple organ system manifestations of mitochondrial disease. *(Reproduced with permission from DR Johns: Mitochondrial DNA and disease. N Engl J Med 333:638, 1995.)*

molecule, the remaining 1400+ gene products are encoded by nuclear genes (referred to as nDNA) and thus follow the rules and patterns of nuclear genomic inheritance **(Chap. 456)**. These nuclear-encoded proteins are synthesized in the cell cytoplasm and imported to their location of activity within the mitochondria through a complex biochemical process. This process includes unfolding of the nuclear-encoded protein, attachment to a chaperone protein that shuttles it through a specific channel to a specific mitochondrial location, and detachment from the chaperone followed by assembly with other mtDNA- and nDNA-encoded proteins. In addition, the mitochondria contain their own small genome consisting of numerous copies (polyploidy) per mitochondrion of a circular, double-strand mitochondrial DNA (mtDNA) molecule comprising 16,569 nucleotides. This mtDNA sequence (also known as the "mitogenome") might represent the remnants of endosymbiotic prokaryotes from which mitochondria are thought to have originated. The mtDNA sequence contains a total of 37 genes, of which 13 encode mitochondrial protein components of the ETC **(Fig. 472-2)**. The remaining 22 tRNA- and 2 rRNA-encoding genes are mitochondria-specific and dedicated to the process of translating the 13 mtDNA-encoded proteins. The mtDNA itself replicates constantly, independent of cell division, and requires its own unique polymerase, referred to as polymerase gamma (polγ), which is encoded by the nuclear protein *POLG*, disorders of which are discussed in **Chaps. 457 and 441**. However, mutations in *POLG* can disrupt the endonuclease function of polγ, resulting in somatic mutations in the mtDNA that endure with future replication. Unless this mutation occurs and repopulates in an oocyte, it is not heritable. This dual nuclear and mitochondrial genetic control of mitochondrial function results in unique and diagnostically challenging patterns of inheritance. The current chapter focuses on heritable traits and diseases related to the mtDNA component of the dual genetic control of mitochondrial function. The reader is referred to **Chaps. 456 and 441** for consideration of mitochondrial disease originating from mutations in the nuclear genome. The latter include: (1) disorders due to mutations in nuclear genes directly encoding structural components or assembly factors of the oxidative phosphorylation complexes, (2) disorders due to mutations in nuclear genes encoding proteins indirectly related to oxidative phosphorylation, (3) mtDNA depletion syndromes (MDSs) characterized by a reduction of mtDNA copy number in affected tissues without mutations or rearrangements in the mtDNA, and (4) disorders due to mutations in nuclear genes that disrupt normal mitochondrial dynamics (biosynthesis, mitophagy, fission, and fusion).

The classic physical structure of the mitochondria is that of a thread-like organelle, which under fixed conditions, such as observed with immunohistochemical stains or electron microscopy, has a submarine-shape and measures about 1 μm in length. However, in the living state, mitochondrial shape is highly variable based on the cell type, and manifests a complex and ever-changing syncytial form, with continuous appearance and disappearance of budding structures (representing mitochondrial fission) and reorganization of separate mitochondria (representing mitochondrial fusion). Although we often think of the *mitochondrial number* in an individual cell, in fact the more accurate concept in a living cell is probably mitochondrial volume.

Although the presence of mitochondria have been known for >150 years, the first knowledge of their respiratory function was proposed about 100 years ago, and the initial description of an illness linked to mitochondrial dysfunction was only made in 1962. The presence of mtDNA was noted in the 1960s and it was not until 1988 when the first mutations in the mtDNA causing human illness were described. These included the demonstration of a large-scale mtDNA deletion causing Kearn-Sayre Syndrome (KSS) and the discovery of a point mutation in *ND4*, an mtDNA-encoded complex I gene causing Leber's Hereditary Optic Neuropathy (LHON). Following these two discoveries, >400 pathogenic mtDNA mutations or deletions have been reported to cause human disease.

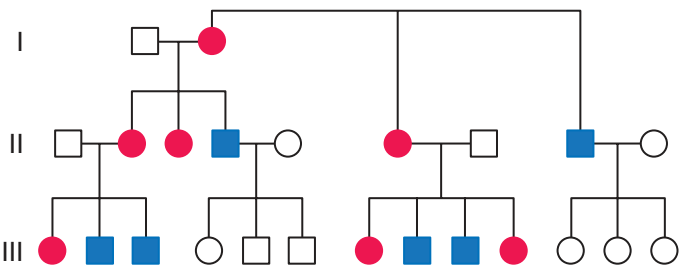


FIGURE 472-2 Maternal inheritance of mitochondrial DNA (mtDNA) disorders and heritable traits. Affected women (*filled circles*) transmit the trait to their children. Affected men (*filled squares*) do not transmit the trait to any of their offspring.

MITOCHONDRIAL DNA STRUCTURE AND FUNCTION

As a result of its circular structure and extranuclear location, the replication and transcription mechanisms of mtDNA differ from the corresponding mechanisms in the nuclear genome, whose nucleosomal packaging and structure are more complex. Specifically, mitochondria have their own transcription system, and the mtDNA itself replicates independently of cellular replication. Because each cell contains many copies of mtDNA, and because the number of mitochondria can vary during the lifetime of each cell, mtDNA copy number is not directly coordinated with the cell cycle. Thus, vast differences in mtDNA copy number are observed between different cell types and tissues and during the lifetime of a cell. Another important feature of the mtDNA replication process is a reduced stringency of proofreading and replication error correction, leading to a greater degree of sequence variation compared to the nuclear genome. Some of these sequence variants are silent polymorphisms that do not have the potential for a phenotypic or pathogenic effect, whereas others may be considered pathogenic mutations. There are some mutations that may be considered ecogenetic, as they typically remain silent, meaning they do not cause disease, unless an external event occurs. One classic example is seen in a common (1:800) mutation in the mitochondrial 12S rRNA gene, m.A1555G., which is associated with hearing loss but is rapidly exacerbated by exposure to normal dosages of an aminoglycoside.

With respect to transcription, initiation can occur on both strands and proceeds through the production of an intronless polycistronic precursor RNA, which is then processed to produce the 13 individual mRNA and 24 individual tRNA and rRNA products. The 37 mtDNA genes comprise fully 93% of the 16,569 nucleotides of the mtDNA in what is known as the *coding region*. The *control region*, which is contained in the D-loop, consists of ~1.1 kilobases (kb) of noncoding DNA, which is thought to have an important role in replication and transcription initiation.

MATERNAL INHERITANCE AND LACK OF RECOMBINATION

In contrast to homologous pair recombination that takes place in the nucleus, mtDNA molecules do not undergo recombination, such that mutational events represent the only source of mtDNA genetic diversification. Moreover, it is only the maternal DNA that is transmitted to the offspring. The fertilized oocyte degrades mtDNA carried from the sperm in a complex process involving the ubiquitin proteasome system and autophagy which takes place on the inner membrane of the oocyte. Thus, although mothers transmit their mtDNA to both their sons and daughters, only the daughters are able to transmit the inherited mtDNA to future generations. Accordingly, mtDNA sequence variation and associated phenotypic traits and diseases are inherited exclusively along maternal lines, meaning both sons and daughters have equal chances of having symptomatic disease, with the only significant exception being LHON as described below.

The phenotypic expression, including age of onset and the exact pattern of organ dysfunction, of a pathogenic mtDNA mutation may vary greatly, even within families. Because of this complex relationship between mtDNA mutations and disease expression, sometimes it is difficult to recognize the maternal pattern of inheritance at the clinical or pedigree level. However, evidence of paternal transmission can almost certainly exclude an mtDNA genetic origin of phenotypic variation or disease; conversely, a disease affecting both sexes without evidence of paternal transmission strongly suggests a heritable mtDNA disorder (Fig. 472-2).

MULTIPLE COPY NUMBER (POLYPLOIDY), HIGH MUTATION RATE, HETEROPLASMY, AND MITOTIC SEGREGATION

Each aerobic cell in the body has multiple mitochondria, often numbering many hundreds or more in cells with extensive energy production requirements. Furthermore, the number of copies of mtDNA within each mitochondrion varies from several to hundreds; this is true of both somatic as well as germ cells, including oocytes in females. In the case of somatic cells, this means that the impact of most newly acquired somatic mtDNA mutations is likely to be very small in terms of total cellular or organ system function; however, because of the many fold higher mutation rate during mtDNA replication, numerous different mutations may accumulate with aging of the organism. It has been proposed that the total cumulative burden of acquired somatic mtDNA mutations with age may result in an overall perturbation of mitochondrial function, contributing to age-related reduction in the efficiency of oxidative phosphorylation and increased production of damaging ROS. Because mtDNA (and nDNA) mutations may result in electron leak within the ETC, the ROS damage may rise above the normal baseline in some with specific mutations, resulting in increased susceptibility to somatic mtDNA damage and disease expression. The accumulation of such acquired somatic mtDNA mutations with aging may contribute to age-related diseases, such as metabolic syndrome and diabetes, cancer, and neurodegenerative and cardiovascular disease in any given individual. However, somatic mutations are not carried forward to the next generation, and the hereditary impact of mtDNA mutagenesis requires separate consideration of events in the female germline.

The multiple mtDNA copy number within each cell, including the maternal germ cells, results in the phenomenon of heteroplasmy, in contrast to much greater uniformity (homoplasy) of somatic nuclear DNA sequence. Heteroplasmy for a given mtDNA sequence variant or mutation arises as a result of the coexistence within a cell, tissue, or individual of mtDNA molecules bearing more than one version of the sequence variant **(Fig. 472-3)**. The importance of the heteroplasmy phenomena to the understanding of mtDNA-related mitochondrial diseases is critical. The coexistence of mutant and nonmutant (wild-type) mtDNA and the variation of the mutant load, which can be thought of as the percentage of mutant mtDNA molecules within a specific cell, tissue, organ or organism, contributes to the expression of a phenotype among individuals from the same maternal sibship. At the level of the oocyte, the percentage of mtDNA molecules bearing each version of the polymorphic sequence variant or mutation depends on

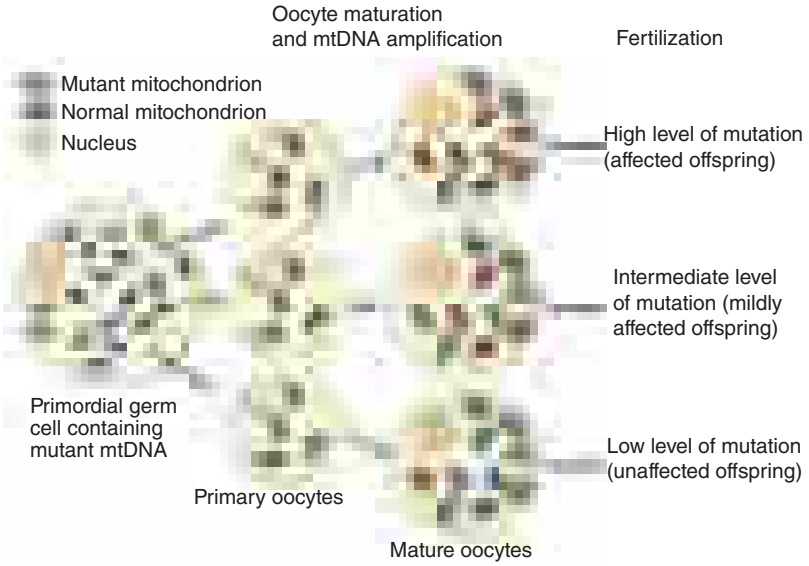


FIGURE 472-3 Heteroplasmy and the mitochondrial genetic bottleneck. During the production of primary oocytes, a selected number of mitochondrial DNA (mtDNA) molecules are transferred into each oocyte. Oocyte maturation is associated with the rapid replication of this mtDNA population. This restriction-amplification event can lead to a random shift of mtDNA mutational load between generations and is responsible for the variable levels of mutated mtDNA observed in affected offspring from mothers with pathogenic mtDNA mutations. Mitochondria that contain mutated mtDNA are shown in *red*, and those with normal mtDNA are shown in *green*. *(Reproduced with permission from R Taylor, D Turnbull: Mitochondrial DNA mutations in human disease. Nat Rev Genetics 6:389, 2005.)*

stochastic events related to partitioning of mtDNA molecules during the process of oogenesis itself. Thus, oocytes differ from each other in the degree of heteroplasmy for that sequence variant or mutation. In turn, the heteroplasmic state is carried forward to the zygote and to the organism as a whole, to varying degrees, depending on mitotic segregation of mtDNA molecules during organ system development and maintenance. For this reason, in vitro fertilization, followed by preimplantation genetic diagnosis (PGD), is not as predictive of the genetic health of the offspring in the case of mtDNA mutations as in the case of the nuclear genome. Similarly, the impact of somatic mtDNA mutations acquired during development and subsequently also shows an enormous spectrum of variability. In general, a higher mutant load will result in a more severe, and earlier phenotypic presentation. However, measuring heteroplamy in one tissue (lymphocytes from blood or urine sediment containing kidney and bladder epithelial cells for example) may not represent the percentage of mutant heteroplasmy in the tissue or organs most affected, such as the cardiac atrioventricular node or brain. Furthermore, the threshold of mutant heteroplasmy which results in clinical illness may vary depending on the specific mutation.

Mitotic segregation refers to the unequal distribution of wild-type and mutant versions of mtDNA molecules during all cell divisions that occur during prenatal development and subsequently throughout the lifetime of an individual. The phenotypic effect or disease impact will be a function not only of the inherent disruptive effect (pathogenicity) on the mtDNA-encoded gene (coding region mutations) or integrity of the mtDNA molecule (control region mutations), but also of its distribution among the multiple copies of mtDNA in the various mitochondria, cells, and tissues of the affected individual. Thus, one consequence can be the generation of a bottleneck due to the marked decline in given sets of mtDNA variants, pathogenic and nonpathogenic, consequent to such mitotic segregation. It is postulated that the main effects of this bottleneck occur between the primordial germ cell state and the primary oocyte stage of development. Heterogeneity arises from differences in the degree of heteroplasmy among oocytes of the transmitting female, together with subsequent, probably random, mitotic segregation of the pathogenic mutation during tissue and organ development, and throughout the lifetime of the individual offspring. The actual expression of disease might then depend on a threshold percentage of mitochondria whose function is disrupted by mtDNA mutations. This in turn confounds hereditary transmission patterns and hence genetic diagnosis of pathogenic heteroplasmic mutations. Generally, if the proportion of mutant mtDNA is <60%, the individual is unlikely to be affected, whereas proportions exceeding 90% cause clinical disease. One notable exception is LHON, in which these mutations are present either in 100% mutant homoplasmy, which causes the disease expression, or 100% wild-type homoplasmy. It is not understood why this specific phenotype and the several known mtDNA alleles that result in LHON behave in this manner.

HOMOPLASMIC VARIANTS AND HUMAN mtDNA PHYLOGENY

In contrast to classic mtDNA diseases, most of which begin in childhood and are the result of heteroplasmic mutations as noted above, during the course of human evolution, certain mtDNA sequence variants have drifted to a state of homoplasmy, wherein all of the mtDNA molecules in the organism contain the new sequence variant. This arises due to a "bottleneck" effect followed by genetic drift during the very process of oogenesis itself (Fig. 472-3). In other words, during certain stages of oogenesis, the mtDNA copy number becomes so substantially reduced that the particular mtDNA species bearing the novel or derived sequence variant may become the increasingly predominant, and eventually exclusive, version of the mtDNA for that particular nucleotide site. All of the offspring of a woman bearing an mtDNA sequence variant or mutation that has become homoplasmic will also be homoplasmic for that variant and will transmit the sequence variant forward in subsequent generations.

Considerations of reproductive fitness limit the evolutionary or population emergence of pathogenic homoplasmic mutations that are lethal or cause severe disease in infancy or childhood. Thus, with a number of notable exceptions (e.g., mtDNA mutations causing LHON; see below), most homoplasmic mutations are considered to be neutral markers of human evolution, which are useful and interesting in the population genetics analysis of shared maternal ancestry but which have little significance in human phenotypic variation or disease predisposition.

More importantly is the understanding that this accumulation of homoplastic mutations occurs at a genetic locus that is transmitted only through the female germline, and that lacks recombination. In turn, this enables reconstruction of the sequential topology and radiating phylogeny of mutations accumulated through the course of human evolution since the time of the most recent common mtDNA ancestor of all contemporary mtDNA sequences, some 200,000 years ago. The term *haplogroup* is usually used to define major branching points in the human mtDNA phylogeny, nested one within the other, which often demonstrate striking continental geographic ancestral partitioning. At the level of the complete mtDNA sequence, the term haplotype is usually used to describe the sum of mutations observed for a given mtDNA sequence and as compared to a reference sequence, such that all haplotypes falling within a given haplogroup share the total sum of mutations that have accumulated since the most recent common ancestor and the bifurcation point they mark. The remaining observed variants are private to each haplotype. Consequentially, human mtDNA sequence is an almost perfect molecular prototype for a nonrecombining locus, and its variation has been extensively used in phylogenetic studies. Moreover, the mtDNA mutation rate is considerably higher than the rate observed for the nuclear genome, especially in the control region, which contains the displacement, or D-loop, in turn comprising two adjacent hypervariable regions (HVR-I and HVR-II). Together with the absence of recombination, this amplifies drift to high frequencies of novel haplotypes. As a result, mtDNA haplotypes are more highly partitioned across geographically defined populations than sequence variants in other parts of the genome. Despite extensive research, it has not been well established that such haplotype-based partitioning has a significant influence on human health conditions. However, mtDNA-based phylogenetic analysis can be used both as a quality assurance tool and as a filter in distinguishing neutral mtDNA variants comprising human mtDNA phylogeny from potentially deleterious mutations.

MITOCHONDRIAL DNA DISEASE

The true prevalence of mtDNA disease is difficult to estimate because of the phenotypic heterogeneity that occurs as a function of heteroplasmy, the challenge of detecting and assessing heteroplasmy in different affected tissues, and the other unique features of mtDNA function and inheritance described above. It is estimated that at least 1 in 200 healthy humans harbors a pathogenic mtDNA mutation with the potential to causes disease, but that heteroplasmic germline pathogenic mtDNA mutations actually affect up to ~1 in 5000 individuals.

The true disease burden relating to mtDNA sequence variation will only be known when the following capabilities become available: (1) ability to distinguish a completely neutral sequence variant from a true phenotype-modifying or pathogenic mutation, (2) accurate assessment of heteroplasmy that can be determined with fidelity, and (3) a systems biology approach **(Chap. 476)** to determine the network of epistatic interactions of mtDNA sequence variations with mutations in the nuclear genome.

OVERVIEW OF CLINICAL AND PATHOLOGIC FEATURES OF HUMAN mtDNA DISEASE

Given the vital roles of mitochondria in all nucleated cells, it is not surprising that mtDNA mutations can affect numerous tissues with pleiotropic effects. More than 200 different disease-causing, mostly heteroplasmic mtDNA mutations have been described affecting ETC function. **Figure 472-4** provides a partial mtDNA map of some of the better characterized of these disorders. A number of clinical clues can increase the index of suspicion for a heteroplasmic mtDNA mutation as an etiology of a heritable trait or disease, including (1) familial clustering with absence of paternal transmission; (2) adherence to one of the

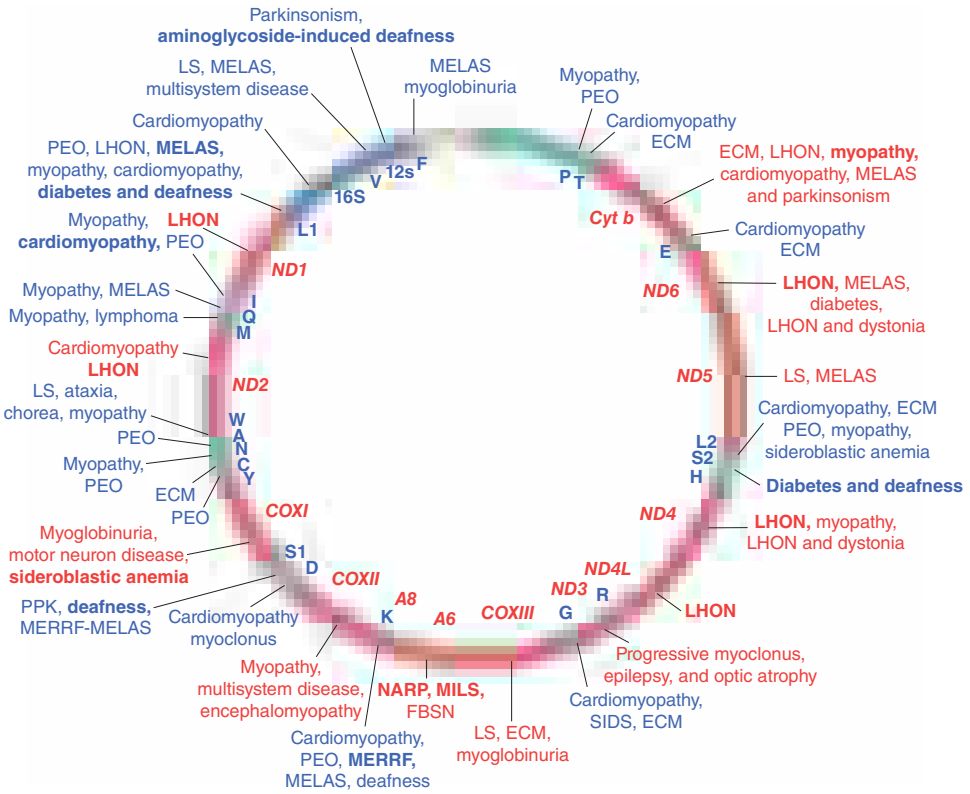


FIGURE 472-4 Mutations in the human mitochondrial genome known to cause disease. Disorders that are frequently or prominently associated with mutations in a particular gene are shown in *boldface*. Diseases due to mutations that impair mitochondrial protein synthesis are shown in *blue*. Diseases due to mutations in protein-coding genes are shown in *red*. ECM, encephalomyopathy; FBSN, familial bilateral striatal necrosis; LHON, Leber's hereditary optic neuropathy; LS, Leigh syndrome; MELAS, mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes; MERRF, myoclonic epilepsy with ragged red fibers; MILS, maternally inherited Leigh syndrome; NARP, neuropathy, ataxia, and retinitis pigmentosa; PEO, progressive external ophthalmoplegia; PPK, palmoplantar keratoderma; and SIDS, sudden infant death syndrome. *(Reproduced with permission from S DiMauro, E Schon: Mitochondrial respiratory-chain diseases. N Engl J Med 348:2656, 2003.)*

classic syndromes (see below) or paradigmatic combinations of disease phenotypes involving several organ systems that normally do not fit together within a single nuclear genomic mutation category; (3) a complex of laboratory and pathologic abnormalities that reflect disruption in cellular energetics (e.g., lactic acidosis and neurodegenerative and myodegenerative symptoms with the finding of ragged red fibers, reflecting the accumulation of abnormal mitochondria under the muscle sarcolemmal membrane); or (4) a mosaic pattern reflecting a heteroplasmic state. There are no truly sensitive and specific biomarkers of disease, and the presence of a historically quintessential finding of ragged red fibers can be seen in numerous muscle disorders, so laboratory tests must always be interpreted in the context of their limitations and should not be used to define the disease.

Heteroplasmy can sometimes be elegantly demonstrated at the tissue level using histochemical staining for enzymes in the oxidative phosphorylation pathway, with a mosaic pattern indicating heterogeneity of the genotype for the coding region for the mtDNA-encoded enzyme. Complex II, CoQ, and cytochrome c are exclusively encoded by nuclear DNA. In contrast, complexes I, III, IV, and V contain at least some subunits encoded by mtDNA. Just 3 of the 13 subunits of the ETC complex IV enzyme, cytochrome c oxidase (COX), are encoded by mtDNA, and, therefore, this enzyme has the lowest threshold for dysfunction when a threshold level of mutated mtDNA is reached. Histochemical staining for COX activity in tissues of patients affected with heteroplasmic inherited mtDNA mutations (or with the somatic accumulation of mtDNA mutations, see below) can show a mosaic pattern of reduced histochemical staining in comparison with histochemical staining for the complex II enzyme, succinate dehydrogenase (SDH) **(Fig. 472-5)**. Heteroplasmy can also be detected at the genetic level through direct Sanger-type mtDNA genotyping under special conditions, although clinically significant low levels of heteroplasmy can escape detection in genomic samples extracted from whole blood using conventional genotyping and sequencing techniques.

Next-generation sequencing (NGS) has dramatically improved the clinical genetic diagnostic evaluation of mitochondrial diseases at the level of both the nuclear genome and mtDNA. In the context of the larger nuclear genome, the ability of NGS techniques to dramatically increase the speed at which DNA can be sequenced at a fraction of the cost of conventional Sanger-type sequencing technology is particularly beneficial. Low sequencing costs and short turnaround time expedite "first-tier" screening of panels of hundreds of previously known or suspected mitochondrial disease genes or screening for the entire exome or genome in an attempt to identify novel genes and mutations affecting different patients or families. In the context of the mtDNA, NGS approaches now provide rapid and reliable detection of heteroplasmy in different affected tissues. Although Sanger sequencing allows for complete coverage of the mtDNA, it is limited by the lack of deep coverage and low sensitivity for heteroplasmy detection when levels are <50%. In contrast, NGS technology is an excellent tool for rapidly and accurately obtaining a patient's predominant mtDNA sequence and also lower frequency heteroplasmic variants and can typically detect mutant heteroplasmy <10%. Lower levels are often only clinically relevant if in the setting of a striking difference in heteroplasmy in different tissues. This capability to detect heteroplasmy at levels that assist in pointing to an mtDNA-based disease process emanates from deep coverage of the genome through multiple independent sequence reads. Accordingly, recent studies making use of NGS techniques have demonstrated sequence accuracy equivalent to Sanger-type sequencing, but also have uncovered heretofore unappreciated heteroplasmy rates ranging between 10 and 50% and detection of single-nucleotide heteroplasmy down to levels of <10%.

Clinically, the most striking overall characteristic of mitochondrial genetic disease is the phenotypic heterogeneity associated with mtDNA mutations. This extends to intrafamilial phenotypic heterogeneity for the same mtDNA pathogenic mutation and, conversely, to the overlap of phenotypic disease manifestations with distinct mutations. Thus, although fairly consistent and well-defined "classic" syndromes have been attributed to specific mutations, frequently "nonclassical" combinations of disease phenotypes ranging from isolated myopathy to extensive multisystem disease are often encountered, rendering genotype-phenotype correlation challenging. In both classical and nonclassical mtDNA disorders, there is often a clustering of some combination of abnormalities affecting the neurologic system (including optic nerve atrophy, pigment retinopathy, and sensorineural hearing loss), cardiac and skeletal muscle (including extraocular muscles), and endocrine and metabolic systems (including diabetes mellitus). Additional organ systems that may be affected include the hematopoietic, renal, hepatic, and gastrointestinal systems, although these are more frequently involved in infants and children. Disease-causing mtDNA coding region mutations can affect either one of the 13 protein encoding genes or one of the 24 protein synthetic genes. Clinical manifestations

FIGURE 472-5 Cytochrome c oxidase (COX) deficiency in mitochondrial DNA (mtDNA)-associated disease. Transverse tissue sections that have been stained for COX and succinate dehydrogenase (SDH) activities sequentially, with COX-positive cells shown in *brown* and COX-deficient cells shown in *blue*. ***A.*** Skeletal muscle from a patient with a heteroplasmic mitochondrial tRNA point mutation. The section shows a typical "mosaic" pattern of COX activity, with many muscle fibers harboring levels of mutated mtDNA that are above the crucial threshold to produce a functional enzyme complex. ***B.*** Cardiac tissue (left ventricle) from a patient with a homoplasmic tRNA mutation that causes hypertrophic cardiomyopathy, which demonstrates an absence of COX in most cells. ***C.*** A section of cerebellum from a patient with mtDNA rearrangement that highlights the presence of COX-deficient neurons. ***D, E.*** Tissues that show COX deficiency due to clonal expansion of somatic mtDNA mutations within single cells—a phenomenon that is seen in both postmitotic cells (***D***; extraocular muscles) and rapidly dividing cells (***E***; colonic crypt) in aging humans. *(Reproduced with permission from R Taylor, D Turnbull: Mitochondrial DNA mutations in human disease. Nat Rev Genetics 6:389, 2005.)*

do not readily distinguish these two categories, although lactic acidosis and specific muscle pathologic findings (e.g., ragged-red and ragged-blue fibers, described immunohistochemical staining, paracrystalline inclusions on ultrastructure) tend to be more prominent in the latter. In all cases, either defective ATP production due to disturbances in the ETC or enhanced generation of ROS has been invoked as the mediating biochemical mechanism between mtDNA mutation and disease manifestation.

mtDNA DISEASE PRESENTATIONS

The clinical presentation of adult patients with mtDNA disease can be divided into three categories: (1) clinical features suggestive of mitochondrial disease (Table 472-2), but not a well-defined classic syndrome; (2) classic mtDNA syndromes; and (3) clinical presentation confined to one organ system (e.g., isolated sensorineural deafness, cardiomyopathy, or diabetes mellitus).

Table 472-3 provides a summary of eight illustrative classic mtDNA syndromes or disorders that affect adult patients and highlights some of the most interesting features of mtDNA disease in terms of molecular pathogenesis, inheritance, and clinical presentation. The first five of these syndromes result from heritable point mutations in either protein-encoding or protein synthetic mtDNA genes; the other three result from rearrangements or deletions that usually do not involve the germline.

TABLE 472-2 Common Features of mtDNA-Associated Diseases in Adults

Neurologic: stroke, epilepsy, migraine headache, peripheral neuropathy, ataxia, dystonia, myoclonus, cranial neuropathy (optic atrophy, sensorineural deafness, dysphagia, dysphasia)
Skeletal myopathy: ophthalmoplegia, exercise intolerance, myalgia, weakness
Cardiac: conduction block, cardiomyopathy
Respiratory: hypoventilation, aspiration pneumonitis
Endocrine: diabetes mellitus, premature ovarian failure, hypothyroidism, hypoparathyroidism
Ophthalmologic: cataracts, pigment retinopathy, neurologic and myopathic (optic atrophy, ophthalmoplegia)

LHON is a common cause of maternally inherited visual failure. LHON typically presents during young adulthood with subacute painless loss of vision in one eye, with symptoms developing in the other eye 6–12 weeks after the initial onset. In some instances, cerebellar ataxia, peripheral neuropathy, and cardiac conduction defects are observed. In >95% of cases, LHON is due to one of the three homoplasmic point mutations of mtDNA that affect genes encoding different subunits of complex I of the mitochondrial ETC; however, not all individuals who inherit a primary LHON mtDNA mutation develop optic neuropathy, and males are four to five times more likely than females to be affected, indicating that additional environmental (e.g., tobacco exposure) or additional independent genetic factors are important in the etiology of the disorder. Both the nuclear and mitochondrial genomic backgrounds modify disease penetrance. Indeed, a region of the X chromosome containing a high-risk haplotype for LHON has been identified, supporting the formulation that nuclear genes act as modifiers and affording an explanation for the male prevalence of LHON. This haplotype can be used in predictive genomic testing and prenatal screening for this disease. In contrast to the other classic mtDNA disorders, it is of interest that patients with this syndrome are often homoplasmic for the disease-causing mutation. The somewhat later onset in young adulthood and modifying effect of protective background nuclear genomic haplotypes may have enabled homoplasmic pathogenic mutations to have escaped evolutionary censoring.

Mitochondrial encephalomyopathy, lactic acidosis, and stroke-like episodes (MELAS) is a multisystem disorder with a typical onset between 2 and 10 years of age. Following normal early psychomotor development, the most common initial symptoms are seizures, recurrent headaches, anorexia, and recurrent vomiting. Exercise intolerance or proximal limb weakness can be the initial manifestation, followed by generalized tonic–clonic seizures. Short stature is common. Seizures are often associated with stroke-like episodes of transient hemiparesis or cortical blindness that may produce recurrent encephalopathy with impaired consciousness. It is often not possible to determine if the encephalopathy is due to refractory seizures or should be attributed to an independent effect. The cumulative residual effects of the stroke-like episodes gradually impair motor abilities, vision, and cognition, often by adolescence or young adulthood. Sensorineural hearing loss adds to the progressive decline of these individuals. A plethora of less common symptoms have been described including myoclonus, ataxia, episodic coma, optic atrophy, cardiomyopathy, pigmentary retinopathy, ophthalmoplegia, diabetes mellitus, hirsutism, gastrointestinal dysmotility, and nephropathy. The typical age of death ranges from 10 to 35 years, but some individuals live into their sixth decade. Intercurrent infections or intestinal obstructions are often the terminal events. It is not atypical for some family members to have much less severe, or later onset illness, presumably because of a lessor mutation load, and "MELAS" is not used as a diagnosis for these restricted phenotypes. This creates somewhat of a disconnect between the genotype for MELAS (most commonly the m.3243A>G mutation), and a diverse phenotype, which includes the syndrome MELAS, as well as a syndrome of high-frequency hearing loss and diabetes with onset later in life, as well as many other phenotypes between these two extreme syndromes. Certain other mtDNA mutations can also cause such patterns

TABLE 472-3 Mitochondrial Diseases Due to mtDNA Point Mutations and Large-Scale Rearrangements

DISEASE	PHENOTYPE	MOST FREQUENT mtDNA MUTATIONS	HOMOPLASMIC (USUALLY)	MATERNAL
NARP, Leigh disease	Loss of central vision leading to blindness in young adult life	m.1778G>A, m.14484T>C, m.3460G>A	Heteroplasmic	Maternal
MELAS	*M*itochondrial encephalomyopathy, *l*actic*a*cidosis, and stroke-like episodes; may manifest only as diabetes mellitus	Point mutation in $tRNA^{leu}$	Heteroplasmic	Maternal
MERRF	*M*yoclonic epilepsy, *r*agged *r*ed *f*ibers in muscle, ataxia, increased CSF protein, sensorineural deafness, dementia	Point mutation in $tRNA^{lys}$	Heteroplasmic	Maternal
Deafness	Progressive sensorineural deafness, often induced by aminoglycoside antibiotics	m.1555A>G mutation in 12S rRNA	Homoplasmic	Maternal
	Nonsyndromic sensorineural deafness	m.7445A>G mutation in 12S rRNA	Homoplasmic	Maternal
Chronic progressive external ophthalmoplegia (PEO)	Late-onset bilateral ptosis and ophthalmoplegia, proximal muscle weakness, and exercise intolerance	Single deletions or duplications	Heteroplasmic	Mostly sporadic, somatic mutations
Pearson syndrome	Pancreatic insufficiency, pancytopenia, lactic acidosis	Large deletion	Heteroplasmic	Sporadic, somatic mutations
Kearns-Sayre syndrome (KSS)	External ophthalmoplegia, heart block, retinal pigmentation, ataxia	The 5-kb "common deletion"	Heteroplasmic	Sporadic, somatic mutations

Abbreviations: CSF, cerebrospinal fluid; NARP, neuropathy, ataxia, and retinitis pigmentosa.

of diverse phenotypic expression. Laboratory investigation commonly demonstrates elevated lactate concentrations at rest with excessive increase after moderate exercise. Brain imaging during stroke-like episodes shows areas of involvement on T2- or FLAIR sequences, with decreased signal on perfusion-weighted sequences, which typically involve the posterior cerebrum and not conforming to the distribution of major arteries. These MRI abnormalities may be temporary or evolve to subsequent atrophy **(Fig. 472-6)**. Electrocardiography (ECG) may show evidence of cardiomyopathy, preexcitation, or incomplete heart block. Electromyography and nerve conduction studies are consistent with a myopathic process, without or with coexisting axonal and sensory neuropathic findings. Muscle biopsy typically shows ragged red fibers with the modified Gomori trichrome stain or "ragged blue fibers" with the SDH histochemical stain, resulting from the hyperintense reaction. The diagnosis of MELAS is based on a combination of clinical findings and molecular genetic testing. Mutations in the mtDNA gene *MT-TL1* encoding $tRNA^{leu}$ are causative. The most common mutation, present in ~80% of individuals with typical clinical findings, is an A-to-G transition at nucleotide 3243 (m.3243A>G). Mutations can usually be detected in mtDNA from leukocytes in individuals with typical MELAS; however, the occurrence of heteroplasmy can result in varying tissue distribution of mutated mtDNA. In the absence of specific treatment, various manifestations of MELAS are treated according to standard modalities for prevention, surveillance, and treatment. Recent developments in therapy are described below.

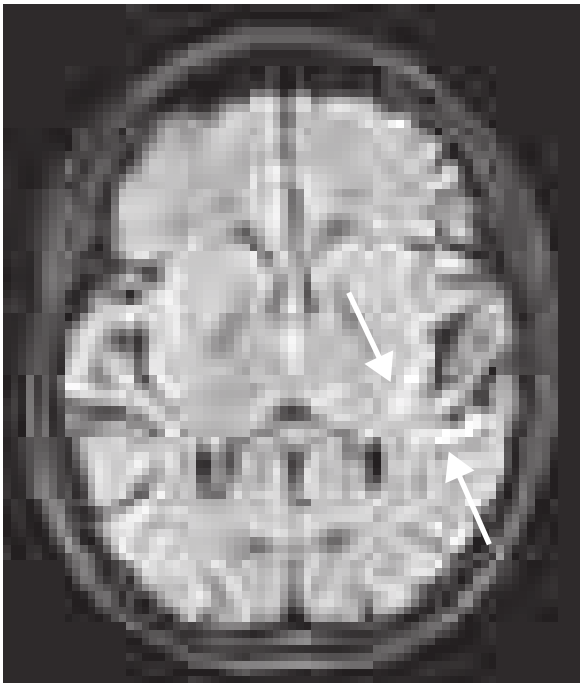


FIGURE 472-6 A 15-year-old girl with MELAS due to m.A3243G ($tRNA^{Leu(UUR)}$), 85% mutant heteroplasmy, presenting at age 5 with focal motor seizures, ataxia and short stature, with episodes of acute language and motor dysfunction and progress cognitive impairment. The FLAIR MRI shows increased signal intensity (white arrows) in the left temporal-parietal region in addition to global mild volume loss (increased extra-axial CSF spaces).

Myoclonus epilepsy with ragged red fiber (MERRF) is a multisystem disorder characterized by myoclonus, seizures, ataxia, and myopathy with ragged red fibers. Hearing loss, exercise intolerance, neuropathy, and short stature are often present. Cerebrospinal fluid (CSF) analysis reveals an elevated protein content. Almost all MERRF patients have a mutation in the mtDNA $tRNA^{lys}$ gene, and the m.8344A>G mutation in the mtDNA gene encoding the lysine amino acid tRNA is responsible for 80–90% of MERRF cases.

Neuropathy, ataxia, and retinitis pigmentosa (NARP) is characterized by moderate diffuse cerebral and cerebellar atrophy and symmetric lesions of the basal ganglia on magnetic resonance imaging (MRI: **Figs. 472-7 and 472-8**). A heteroplasmic m.8993T>G mutation in the ATPase 6 subunit gene has been identified as causative. Ragged red fibers are not observed in muscle biopsy. When >95% of mtDNA molecules are mutant, a more severe clinical, neuroradiologic, and neuropathologic picture (Leigh syndrome) emerges. Not uncommonly, an infant is diagnosed with Leigh syndrome due to the m.8993T>G mutation and not until several years later will the mother present with symptoms of NARP; a situation that highlights the concept of a higher threshold for lower levels of tissue heteroplasmy.

Point mutations in the mtDNA gene encoding the 12S rRNA (m.A1555G) result in heritable nonsyndromic hearing loss. One such mutation causes heritable ototoxic susceptibility to aminoglycoside antibiotics, which opens a pathway for a simple pharmacogenetic test in the appropriate clinical settings. This is an example of an eco-genetic disorder in that most people with this mutation do not develop any symptoms until exposed to an external agent.

KSS, sporadic progressive external ophthalmoplegia (PEO), and Pearson syndrome are three disease phenotypes caused by large-scale mtDNA rearrangements including partial deletions or partial duplication. The majority of single large-scale rearrangements of mtDNA are thought to result from clonal amplification of a single sporadic mutational event, occurring in the maternal oocyte or during early embryonic development. The typical mtDNA deletion is specifically at nucleotide 4977, accounting for most KSS and PEO of mtDNA deletion origin. Because germline involvement is rare, most cases are sporadic

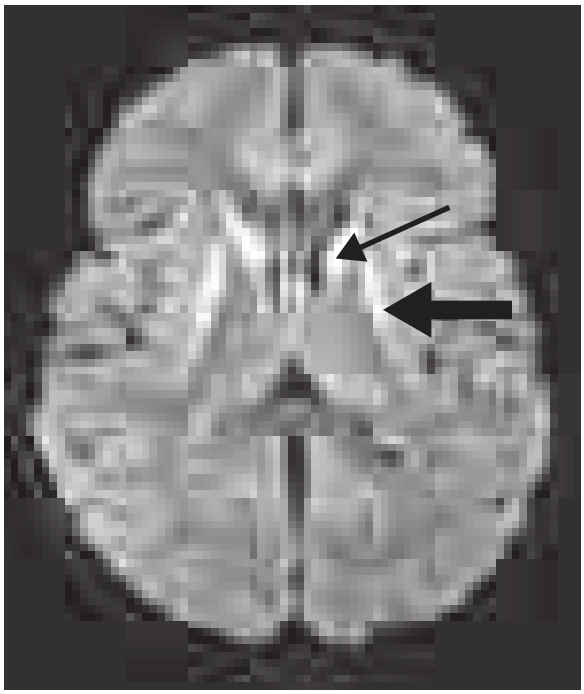


FIGURE 472-7 A 9-year-old girl with Leigh syndrome due to m.T8993G (ATPase subunit 6), 99% heteroplasmy, presenting at age 14 months with a motor delay and underwent an MRI at 24 months, at which time she had just begun to walk. She has moderate cognitive impairment, arm chorea, and distal leg dystonia. The FLAIR MRI shows symmetric bilateral increased signal in the caudate nuclei (*thin arrow*) and putamen (*thick arrow*); only left-sided lesions indicated with arrows.

rather than inherited. KSS is characterized by the triad of onset before age 20, chronic PEO, and pigmentary retinopathy. Cerebellar syndrome, heart block, increased CSF protein content, diabetes mellitus, and short stature are also part of the syndrome. Single deletions/duplication

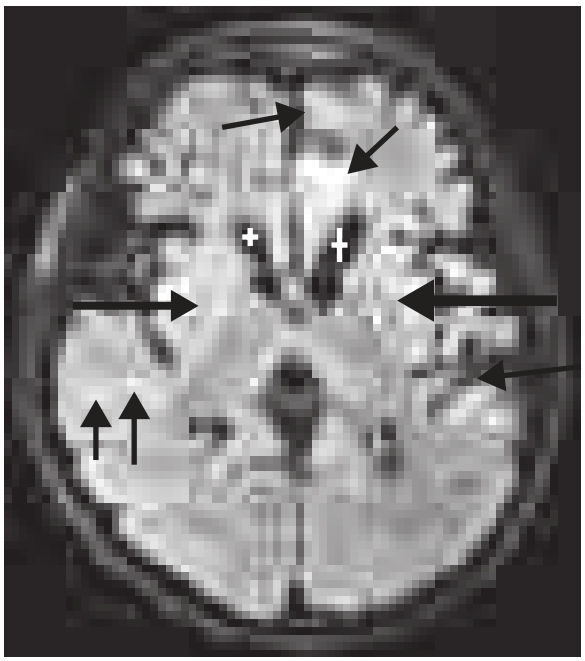


FIGURE 472-8 A 12-year-old boy with Leigh syndrome due to m.T10191C (ND3 gene, complex I), heteroplasmy percentage not determined, presenting with infantile spasms at 8 months of life. He responded well to Adrenocorticotropic Hormone (ACTH) and his MRI and development were normal until 30 months when he developed dystonia and progressive medically intractable epilepsy. The FLAIR MRI at 6 years of life shows global atrophy with large extraaxial CSF spaces, increased signal intensity in the cortex (*thin arrows*), necrotic bilaterally symmetric lesions in the putamina, and enlarged lateral ventricles due to loss of bilateral caudate nuclei volume (*stars*).

can also result in milder phenotypes such as PEO, characterized by late-onset PEO, proximal myopathy, and exercise intolerance. In both KSS and PEO, diabetes mellitus and hearing loss are frequent accompaniments. Pearson syndrome is also characterized by infantile onset of a sideroblastic anemia accompanied by lactic acidosis and failure to thrive caused in part by exocrine pancreatic insufficiency. If the child survives, the manifestations appear phenotypically similar to that of severe KSS with myopathy, PEO, encephalopathy, and cardiomyopathy. Pearson syndrome is generally caused by large-scale sporadic deletion of several mtDNA genes that differ from the common deletion seen in KSS. Typically, the deletion size is larger in Pearson syndrome than in KSS or PEO, but this is not always the case.

Two important dilemmas in classic mtDNA disease have benefited from recent important research insights. The first relates to the greater involvement of neuronal, muscular, renal, hepatic, and pancreatic manifestations in mtDNA disease in these syndromes. This observation has appropriately been mostly attributed to the high energy utilization of the involved tissues and organ systems and, hence, greater dependency on mitochondrial ETC integrity and health. However, because mutations are stochastic events, mitochondrial mutations should occur in any organ during embryogenesis and development. Recently, additional explanations have been suggested based on studies of the common m.3243A>G transition. The proportion of this mutation in peripheral blood cells was shown to decrease exponentially with age. A selective process acting at the stem cell level with a strong bias against the mutated form would have its greatest effect to reduce the mutant mtDNA only in highly proliferating cells, such as those derived from the hematopoietic system. Tissues and organs with lower cell turnover, such as those involved with mtDNA mutations, would not benefit from this effect and, thus, would be the most affected.

The other dilemma arises from the observation that only a subset of mtDNA mutations accounts for the majority of the familial mtDNA diseases. The random occurrence of mutations in the mtDNA sequence should yield a more uniform distribution of disease-causing mutations. However, recent studies using the introduction of one severe and one mild point mutation into the female germline of experimental animals demonstrated selective elimination during oogenesis of the severe mutation and selective retention of the milder mutation, with the emergence of mitochondrial disease in offspring after multiple generations. Thus, oogenesis itself can act as an "evolutionary" filter for mtDNA disease.

THE INVESTIGATION OF SUSPECTED mtDNA DISEASE

The clinical presentations of classic syndromes, groupings of disease manifestations in multiple organ systems, or unexplained isolated presentations of one of the disease features of a classic mtDNA syndrome should prompt a systematic clinical investigation as outlined in Fig. 472-9. Indeed, mitochondrial disease should be considered in the differential diagnosis of any progressive multisystem disorder. Despite the centrality of disruptive oxidative phosphorylation, an elevated blood lactate level is neither specific nor sensitive, because there are many causes of blood lactic acidosis, and many patients with mtDNA defects presenting in adulthood have normal blood lactate. An elevated CSF lactate is a more specific test for mitochondrial disease if there is central nervous system involvement. The serum creatine kinase may be elevated but is often normal, even in the presence of a proximal myopathy. Recently, testing for elevated levels of Growth Differentiating Factor 15 (GDF15) has shown a high degree of sensitivity and specificity in those with a mitochondrial myopathy, but it is not known yet if the degree of elevation for an individual patient reflects the severity of the illness or is in any way a marker of disease activity. Urinary organic and amino acids may also be abnormal, reflecting metabolic and kidney proximal tubule dysfunction. Every patient with seizures, episodes of confusion or atypical behavioral changes, or cognitive decline should have an electroencephalogram. A brain computed tomography (CT) scan may show calcified basal ganglia or bilateral hypodense regions with cortical atrophy. MRI is indicated in patients with brainstem signs or stroke-like episodes.

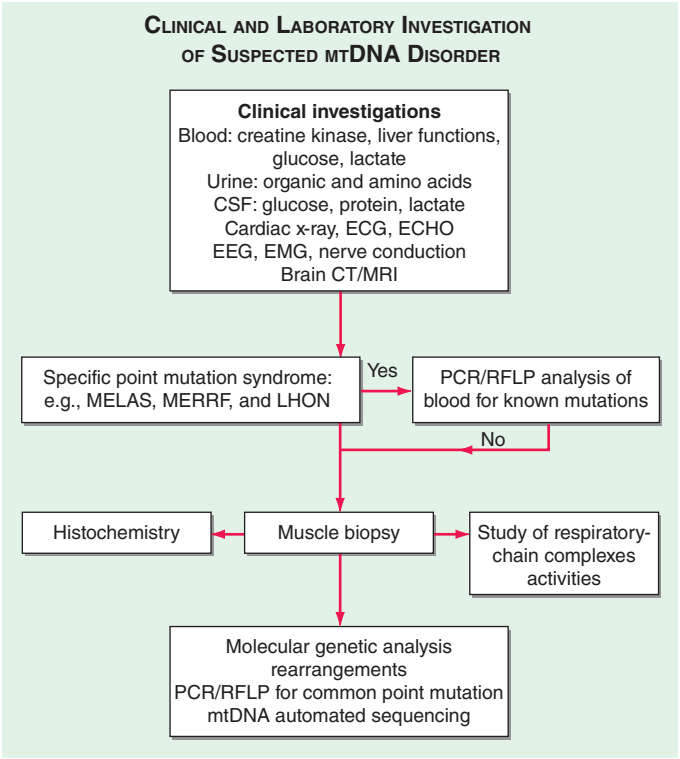


FIGURE 472-9 Clinical and laboratory investigation of a suspected mitochondrial DNA (mtDNA) disorder. CSF, cerebrospinal fluid; CT, computed tomography; ECG, electrocardiogram; ECHO, echocardiography; EEG, electroencephalogram; EMG, electromyogram; LHON, Leber's hereditary optic neuropathy; MELAS, mitochondrial encephalomyopathy, lactic acidosis, and stoke-like episodes; MERFF, myoclonic epilepsy with ragged red fibers; MRI, magnetic resonance imaging; PCR, polymerase chain reaction; RFLP, restriction fragment length polymorphism.

For some mitochondrial diseases, it is possible to obtain an accurate diagnosis with a simple molecular genetic screen. For examples, 95% of patients with LHON harbor one of the three mtDNA point mutations (m.11778A>G, m.A3460A>G, or m.14484T>C). These patients have very high levels of mutated mtDNA in peripheral blood cells, and, therefore, it is appropriate to send a blood sample for molecular genetic analysis by polymerase chain reaction (PCR) or restriction fragment length polymorphism (RFLP). The same is true for most MERRF patients who harbor a point mutation in the lysine tRNA gene at position 8344. In contrast, patients with the m.3243A>G MELAS mutation often have low levels of mutated mtDNA in blood. If clinical suspicion is strong enough to warrant peripheral blood testing, then patients with a negative result should be repeated in a saliva sample, or investigated further by performing a skeletal muscle biopsy.

Muscle biopsy histochemical analysis had been the historical cornerstone for investigation of patients with suspected mitochondrial disease. Histochemical analysis may show subsarcolemmal accumulation of mitochondria with the appearance of ragged red fibers, especially in those with mtDNA mutations affecting the tRNA and rRNA genes. Electron microscopy might show abnormal mitochondria with paracrystalline inclusions. Muscle histochemistry may show COX-deficient fibers, which indicate mitochondrial dysfunction (Fig. 472-5). Respiratory chain complex assays may also show reduced enzyme function. If enzymatic or polarographic data are used to aid in the confirmation of diagnosis, a standard method of analysis should be employed. Either of these two abnormalities, within the exact context of established peer-reviewed criteria may confirm the presence of a mitochondrial disease, to be followed by an in-depth molecular genetic analysis. In some centers, genetic testing may immediately follow establishment of a clinical phenotype and screening biochemical labs, which may obviate the need for invasive testing.

Recent evidence has provided important insights into the importance of nuclear-mtDNA genomic cross-talk and has provided a descriptive framework for classifying and understanding disorders that emanate from perturbations in this cross-talk. Although not strictly considered as mtDNA genetic disorders, manifestations do overlap those highlighted above **(Fig. 472-10)**.

IMPACT OF HOMOPLASMIC SEQUENCE VARIATION ON HERITABLE TRAITS AND DISEASE

The relationship among the degree of heteroplasmy, tissue distribution of the mutant mtDNA, and disease phenotype simplifies inference of a clear causative relationship between heteroplasmic mutation and disease. With the exception of certain mutations (e.g., those causing most cases of LHON), drift to homoplasmy of such mutations would be precluded normally by the severity of impaired oxidative phosphorylation and the consequent reduction in reproductive fitness. Therefore, sequence variants that have reached homoplasmy should be neutral in terms of human evolution and, hence, useful only for tracing human evolution, demography, and migration, as described above. One important exception is in the case of one or more of the homoplasmic population-level variants, which designate the mtDNA haplogroup J, and the interaction with the mtDNA mutations causing LHON. Reduced disease predilection suggests that one or more of the ancient sequence variants designating mtDNA haplogroup J appear to attenuate predisposition to degenerative disease, in the face of other risk factors. Whether or not additional epistatic interactions between population-level mtDNA haplotypes and common health conditions will be found remains to be determined. If such influences do exist, then they are more likely to be relevant to health conditions in the post-reproductive age groups, wherein evolutionary filters would not have had the opportunity to censor deleterious effects and interactions and wherein the effects of oxidative stress may play a role. Although much has been written about the possible associations of population-level

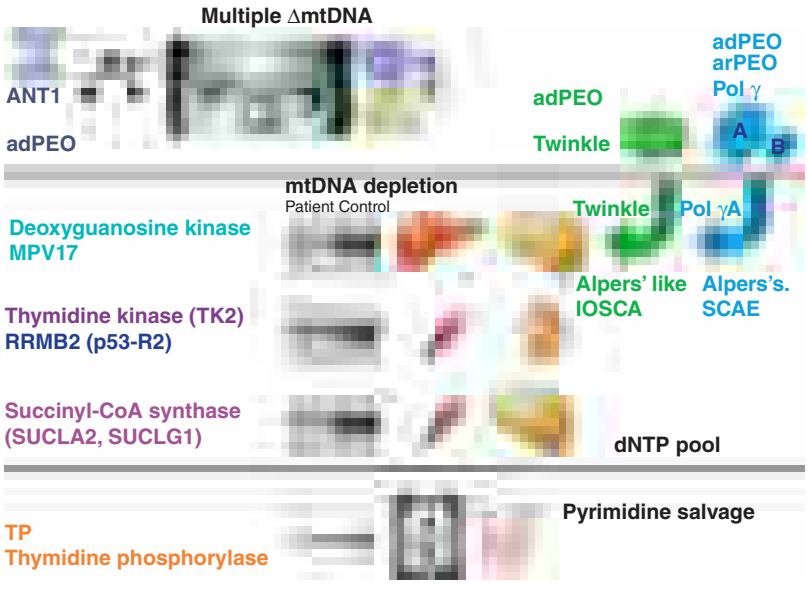


FIGURE 472-10 Disorders associated with perturbations in nuclear-mitochondrial genomic cross-talk. Clinical features and genes associated with multiple mitochondrial DNA (mtDNA) deletions, mtDNA depletion, and mitochondrial neurogastrointestinal encephalomyopathy syndromes. ANT, adenine nucleotide translocators; adPEO, autosomal dominant progressive external ophthalmoplegia; arPEO, autosomal recessive progressive external ophthalmoplegia; IOSCA, infantile-onset spinocerebellar ataxia; SCAE, spinocerebellar ataxia and epilepsy. *(Reproduced with permission from A Spinazzola, M Zeviani: Disorders from perturbations of nuclear-mitochondrial intergenomic cross-talk. J Intern Med 265:174, 2009.)*

common mtDNA variants and human health and disease phenotypes or adaptation to different environmental influences (e.g., climate), a word of caution is in order.

Many studies that purport to show such associations with phenotypes such as longevity, athletic performance, and metabolic and neurodegenerative disease are limited by small sample sizes, possible genotyping inaccuracies, and the possibility of population stratification or ethnic ancestry bias. Because mtDNA haplogroups are so prominently partitioned along phylogeographic lines, it is difficult to exclude the possibility that a haplogroup for which an association has been found is simply a marker for differences in populations with a societal or environmental difference or with different allele frequencies at other genomic loci, which are actually causally related to the heritable trait or disease of interest. The difficulty in generating cellular or animal models to test the functional influence of homoplasmic sequence variants (as a result of mtDNA polyploidy) further compounds the challenge. The most likely formulation is that the risk conferred by different mtDNA haplogroup-defining homoplasmic mutations for common diseases depends on the concomitant nuclear genomic background, together with environmental influences. Progress in minimizing potentially misleading associations in mtDNA heritable trait and disease studies should include ensuring adequate sample size taken from a large sample recruitment base, using carefully matched controls and population structure determination, and performing analysis that takes into account epistatic interactions with other genomic loci and environmental factors.

IMPACT OF ACQUIRED SOMATIC mtDNA MUTATION ON HUMAN HEALTH AND DISEASE

Studies on aging humans and animals have shown a potentially important correlation of age with the accumulation of heterogeneous mtDNA mutations, especially in those organ systems that undergo the most prominent age-related degenerative tissue phenotype. Sequencing of PCR-amplified single mtDNA molecules has demonstrated an average of two to three point mutations per molecule in elderly subjects when compared with younger ones. Point mutations observed include those responsible for known heritable heteroplasmic mtDNA disorders, such as the m.3344A>G and m.3243A>G mutations responsible for the MERRF and MELAS syndromes, respectively. However, the cumulative burden of these acquired somatic point mutations with age was observed to remain well below the threshold expected for phenotypic expression (<2%). Point mutations at other sites not normally involved in inherited mtDNA disorders have also been shown to accumulate to much higher levels in some tissues of elderly individuals, with the description of tissue-specific "hot spots" for acquired somatic mtDNA point mutations. Along the same lines, an age-associated and tissue-specific accumulation of mtDNA deletions has been observed, including deletions involved in known heritable mtDNA disorders, as well as others. The accumulation of functional mtDNA deletions in a given tissue is expected to be associated with mitochondrial dysfunction, as reflected in an age-associated patchy and reduced COX activity on histochemical staining, especially in skeletal and cardiac muscle and brain. A particularly well-studied and potentially important example is the accumulation of mtDNA deletions and COX deficiency observed in neurons of the substantia nigra in Parkinson's disease patients.

The progressive accumulation of ROS has been proposed as the key factor connecting mtDNA mutations with aging and age-related disease pathogenesis **(Fig. 472-11)**. As noted above, ROS are a by-product of oxidative phosphorylation and are removed by detoxifying antioxidants into less harmful moieties; however, exaggerated production of ROS or impaired removal results in their accumulation. One of the main targets for ROS-mediated injury is DNA, and mtDNA is particularly vulnerable because of its proximity to the origin of free radical production, the lack of protective histones, and less efficient injury repair systems compared with nuclear DNA. In turn, accumulation of mtDNA mutations results in inefficient oxidative phosphorylation, with the potential for excessive production of ROS, generating a "vicious cycle" of cumulative mtDNA damage. Indeed, measurement

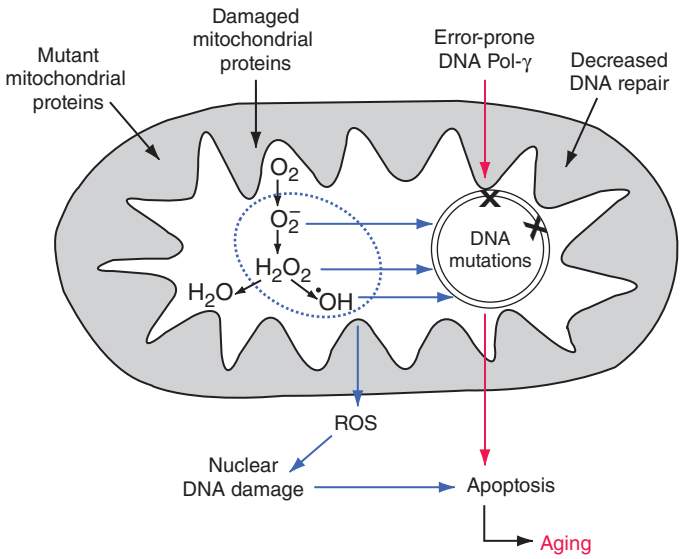


FIGURE 472-11 Multiple pathways of mitochondrial DNA (mtDNA) damage and aging. Multiple factors may impinge on the integrity of mitochondria that lead to loss of cell function, apoptosis, and aging. The classic pathway is indicated with *blue arrows*; the generation of reactive oxygen species (ROS; superoxide anion, hydrogen peroxide, and hydroxyl radicals), as a by-product of mitochondrial oxidative phosphorylation, results in damage to mitochondrial macromolecules, including the mtDNA, with the latter leading to deleterious mutations. When these factors damage the mitochondrial energy-generating apparatus beyond a functional threshold, proteins are released from the mitochondria that activate the caspase pathway, leading to apoptosis, cell death, and aging. *(Reproduced with permission from L Loeb et al: The mitochondrial theory of aging and its relationship to reactive oxygen species damage and somatic mtDNA mutations. Proc Natl Acad Sci USA 102:18769, 2005.)*

of the oxidative stress biomarker 8-hydroxy-2-deoxyguanosine has been used to measure age-dependent increases in mtDNA oxidative damage at a rate exceeding that of nuclear DNA. It should be noted that mtDNA mutations can potentially occur in postmitotic cells as well, because mtDNA replication is not synchronized with the cell cycle. Two other proposed links between mtDNA mutation and aging, besides ROS-mediated tissue injury, are the perturbations in efficiency of oxidative phosphorylation with disturbed cellular aerobic function and perturbations in apoptotic pathways, whose execution steps involve mitochondrial activity.

Genetic intervention studies in animal models have sought to clarify the potential causative relationship between acquired somatic mtDNA mutation and the aging phenotype, and the role of ROS in particular. Replication of the mitochondrial genome is mediated by the activity of the nuclear-encoded *POLG*. A transgenic homozygous mouse knock-in mutation of this gene renders the polymerase enzyme deficient in proofreading and results in a threefold to fivefold increase in mtDNA mutation rate. Such mice develop a premature aging phenotype, which includes subcutaneous lipoatrophy, alopecia, kyphonia, and weight loss with premature death. Although the finding of increased mtDNA mutation and mitochondrial dysfunction with age has been solidly established, the causative role and specific contribution of mitochondrial ROS to aging and age-related disease in humans has yet to be proved. Similarly, although many tumors display higher levels of heterogeneous mtDNA mutations, a causal relationship to tumorigenesis has not been proved.

Besides the age-dependent acquired accumulation in somatic cells of heterogeneous point mutations and deletions, a quite different effect of nonheritable and acquired mtDNA mutation has been described affecting tissue stem cells. In particular, disease phenotypes attributed to acquired mtDNA mutation have been observed in sporadic and apparently nonfamilial cases involving a single individual or even tissue, usually skeletal muscle. The presentation consists of decreased exercise tolerance and myalgias, sometimes progressing to rhabdomyolysis. As in the case of the sporadic, heteroplasmic, large-scale deletion, classic syndromes of chronic PEO, Pearson syndrome, and KSS, the

absence of a maternal inheritance pattern, together with the finding of limited tissue distribution, suggest a molecular pathogenic mechanism emanating from mutations arising de novo in muscle stem cells after germline differentiation (somatic mutations that are not sporadic and occur in tissue-specific stem cells during fetal development or in the postnatal maintenance or postinjury repair stage). Such mutations would be expected to be propagated only within the progeny of that stem cell and affect a particular tissue within a given individual, without evidence of heritability.

PROSPECTS FOR CLINICAL MANAGEMENT OF mtDNA DISEASE

TREATMENT OF mtDNA DISORDERS

No specific curative treatment for mtDNA disorders is currently available; therefore, the management of mitochondrial disease is largely supportive. Management issues may include early diagnosis and treatment of epilepsy, gastrointestinal dysfunction, diabetes mellitus, cardiac pacing, ptosis correction, and intraocular lens replacement for cataracts. Less specific interventions in the case of other disorders involve combined treatment strategies including dietary intervention and removal of toxic metabolites. Cofactors and vitamin supplements are widely used in the treatment of diseases of mitochondrial oxidative phosphorylation, although there is little evidence, apart from anecdotal reports, to support their use. This includes administration of artificial electron acceptors, including vitamin K_3, vitamin C, and ubiquinone (coenzyme Q10); administration of cofactors (coenzymes) including riboflavin, carnitine, and creatine; and use of oxygen radical scavengers, such as vitamin E, copper, selenium, ubiquinone, and idebenone. Drugs that could interfere with mitochondrial function, such as the anesthetic agent propofol, barbiturates, and high doses of valproate, should be avoided. The use of valproate in patients with pathogenic mutations in *POLG* and possibly other mutations affecting mtDNA stability and replication are especially contraindicated. Supplementation with the nitric oxide synthase substrate, L-arginine, and more recently L-citrulline has been advocated as a vasodilator treatment during stroke-like episodes as well as for chronic management in patients with MELAS. Open label studies demonstrate that levoarginine and levocitrulline may be helpful in reducing the stroke-like symptoms in MELAS, but may have serious side effects. As CSF folate deficiency has been reported in some cases of mitochondrial disease, this can be treated with folinic acid.

The physician should also be familiar with environmental interactions, such as the strong and consistent association between visual loss in LHON and smoking or ethanol consumption. A clinical penetrance of 93% was found in men who smoked. Asymptomatic carriers of an LHON mtDNA mutation should, therefore, be strongly advised not to smoke and to moderate their alcohol intake. Although not a cure, these interventions might stave off the devastating clinical manifestations of the LHON mutation. Another example is strict avoidance of aminoglycosides in the familial syndrome of ototoxic susceptibility to aminoglycosides in the presence of the mtDNA m.1555A>G mutation of the 12SrRNA encoding gene.

Clinical trials using novel agents are currently being conducted and analyzed. These include α-tocotrienol (EPI-743, BioElectron Technology Corporation), cysteamine bitartrate (RP-103, Horizon Pharma), omaveloxolone (RTA-408, Reata Pharma), and elamipretide (Stealth Biotheraputics). In an open label study of α-tocotrienol used to treat 10 children with Leigh syndrome, there were improvements in the primary endpoints including the Newcastle Pediatric Mitochondrial Diseases Scale, the Gross Motor Function Measure, and the PedsQL Neuromuscular Module. Ongoing studies continue for α-tocotrienol in children, and both omaveloxolone and elamipritide in adults with primary mitochondrial myopathy.

GENETIC COUNSELING, PRENATAL DIAGNOSIS, AND PGD IN mtDNA DISORDERS

The provision of accurate genetic counseling and reproductive options to families with mtDNA mutations is challenging due to the unique genetic features of mtDNA inheritance that distinguish it from Mendelian genetics. mtDNA defects are transmitted by maternal inheritance. mtDNA de novo mutations are often large deletions, affect one family member, and usually represent no significant risk to other members of the family. In contrast, mtDNA point mutations or duplications can be transmitted down the maternal line. Accordingly, the father of an affected individual has no risk of harboring the disease-causing mutation, and a male cannot transmit the mtDNA mutation to his offspring. In contrast, the mother of an affected individual usually harbors the same mutation but might be completely asymptomatic. This wide phenotypic variability is primarily related to the phenomena of heteroplasmy and the mutation load carried by different members of the same family. Consequently, a symptomatic or asymptomatic female harboring a disease-causing mutation in a heteroplasmic state will transmit to her offspring variable amounts of the mutant mtDNA molecules. The offspring will be symptomatic or asymptomatic primarily according to the mutant load transmitted via the oocyte and, to some extent, subsequent mitotic segregation during development. Interactions with the mtDNA haplotype background or nuclear human genome (as in the case of LHON) serve as an additional important determinant of disease penetrance. Because the severity of the disease phenotype associated with the heteroplasmic mutation load is a function of the stochastic differential segregation and copy number of mutant mtDNA during the oogenesis bottleneck and, subsequently, following tissue and organ development in the offspring, it is rarely predictable with any degree of accuracy. For this reason, prenatal diagnosis (PND) and PGD techniques that have evolved into integral and well-accepted standards of practice are severely hampered in the case of mtDNA-related diseases.

The value of PND and PGD is limited, partly due to the absence of data on the rules that govern the segregation of wild-type and mutant mtDNA species (heteroplasmy) among tissue in the developing embryo. Three factors are required to ensure the reliability of PND and PGD: (1) a close correlation between the mutant load and the disease severity, (2) a uniform distribution of mutant load among tissues, and (3) no major change in mutant load with time. These criteria are suggested to be fulfilled for the NARP m.8993T>G mutation but do not seem to apply to other mtDNA disorders. In fact, the level of mutant mtDNA in a chorionic villous or amniotic fluid sample may be very different from the level in the fetus, and it would be difficult to deduce whether the mutational load in the prenatal samples provides clinically useful information regarding the postnatal and adult state.

PREVENTION OF MITOCHONDRIAL DISEASE INHERITANCE BY ASSISTED REPRODUCTIVE TECHNOLOGIES

Because the treatment options for patients with mitochondrial disease are rather limited, with no current U.S. Food and Drug Administration (FDA) approved therapies for established mitochondrial DNA disease, preventive interventions that eliminate the likelihood of transmission of affected mtDNA into offspring are desirable. The poor reliability of prenatal and preimplantation approaches in predicting mitochondrial DNA disease has resulted in the search for alternative preventive approaches. The common purpose underlying various emerging approaches is to reduce mutant heteroplasmy levels to a level below a pathogenic threshold. This is based on the observed relationship between heteroplasmy and disease inheritance patterns, which indicates that even a small increase in copy number of nonmutant mtDNA molecules in the fertilized egg can exceed the threshold required to ameliorate serious clinical disease. Use of gene editing, with Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) technology for example, to shift the heteroplasmy load in affected tissues will require future development of corrective gene delivery techniques. Likewise, induced pluripotent cell technology has not yet met with widespread success in the preclinical research setting. This has prompted the application of Mitochondrial Replacement Therapy (MRT) approaches **(Fig. 472-12)**. Both of these approaches substitute in vitro the entire oocyte or zygote complement of mitochondria, together with their mtDNA from the carrier mother, with the

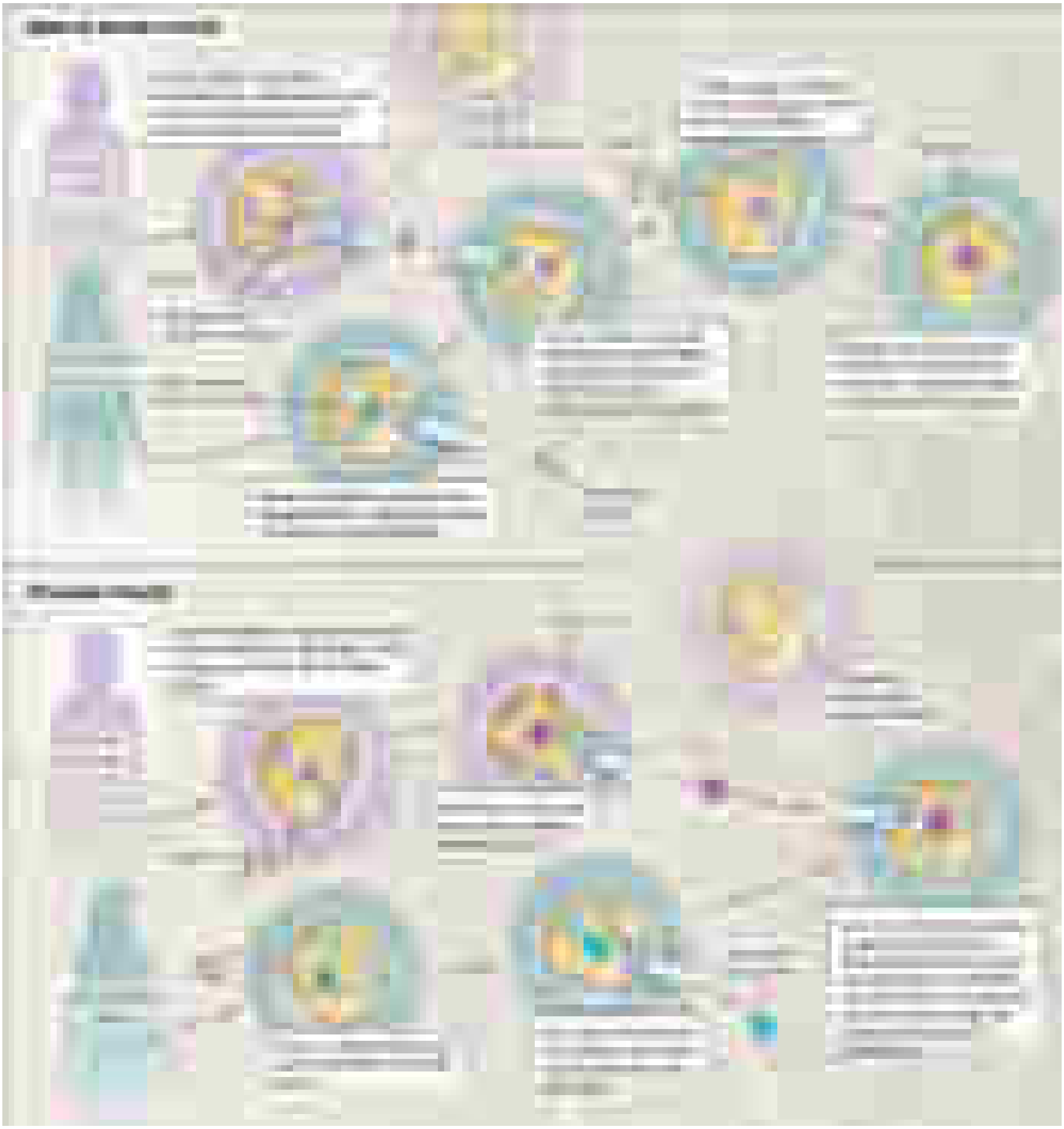


FIGURE 472-12 Mitochondrial replacement techniques—maternal spindle transfer and pronuclear transfer. In both procedures, some mutant mtDNA, estimated at 1–2%, might be carried over together with the spindle or pronucleus, but the levels are low enough to avoid disease risk. IVF denotes in vitro fertilization. *(Used with permission from MJ Falk et al: Mitochondrial Replacement Techniques–Implications for the Clinical Community. N Engl J Med 374: 1103, 2016.)*

unaffected complement of mitochondria and their unaffected mtDNA from a donor woman. This can be accomplished either by removing and transferring the carrier mother's spindle with her nuclear DNA into the unfertilized oocyte of the donor, or alternatively by transferring the pronucleus from the fertilized oocyte of the carrier mother to the unfertilized donor oocyte from which the pronucleus has been removed. Both of these approaches provide a "bulk" substitution and hence do not target the specific mtDNA mutation, and are potentially applicable to a wide variety of mtDNA disorders. This is a form of germline genetic therapy, and therefore projects onto future generations in the case of a female offspring. Accordingly, ethical and regulatory bodies have appropriately weighed in on the societal implications of such approaches, and have been tentatively supportive of human clinical investigation for situations of preventing great suffering and when the clinical need is clear and unambiguous, subject to specified conditions and principles and subject to ethical scrutiny. Several such studies have been initiated, and careful examination and follow up are needed to determine developmental and longer-term health and fertility of children who had undergone genetic manipulation at the earliest stages of human development, and whose genomes comprise separate maternal origins of nuclear and mtDNA genomes. It has been recommended that such studies be limited to male offspring, who cannot then transmit the donor mtDNA to future generations, until such time as the health, ethical, and societal issues are well understood, and live up to the exciting promise of reducing the burden of clinical mtDNA disease in the future.

FURTHER READING

ALSTON CL et al: The genetics and pathology of mitochondrial disease. J Pathol 241:236, 2017.

CAMP K et al: Nutritional interventions in primary mitochondrial disorders: Developing an evidence base. Mol Genet Metab 119:187, 2016.

DIMAURO S: Mitochondrial encephalomyopathies—Fifty years on: The Robert Wartenberg Lecture. Neurology 81:281, 2013.

DIMAURO S: et al: The clinical maze of mitochondrial neurology. Nat Rev Neurol 9:429, 2013.

HAAS RH, et al: Mitochondrial disease: A practical approach for primary care physicians. Pediatrics 129:1326, 2007.

KOOPMAN WJ et al: Monogenic mitochondrial disorders. N Engl J Med 366:1132, 2012.

PARIKH S et al: Diagnosis and management of mitochondrial disease: A consensus statement from the Mitochondrial Medicine Society. Genet Med 17:689, 2015.

SANETO RP et al (eds): *Mitochondrial Case Studies, Underlying Mechanisms and Diagnosis*. London, Academic Press/Elsevier, 2016.

473 Applications of Stem Cell Biology in Clinical Medicine

John A. Kessler

Damage to an organ initiates a series of events that lead to the reconstruction of the damaged tissue, including proliferation, differentiation, and migration of various cell types; release of cytokines and chemokines; and remodeling of the extracellular matrix. Endogenous stem and progenitor cells are among the cell populations that are involved in these injury responses. In normal steady-state conditions, an equilibrium is maintained in which endogenous stem cells intrinsic to the tissue replenish dying cells. After tissue injury, stem cells in organs such as the liver and skin have a remarkable ability to regenerate the organ, whereas other stem cell populations, such as those in the heart and brain, have a much more limited capability for self-repair. In rare circumstances, circulating stem cells may contribute to regenerative responses by migrating into a tissue and differentiating into organ-specific cell types. The goal of stem cell therapies is to promote cell replacement in organs that are damaged beyond their ability to self-repair.

GENERAL STRATEGIES FOR STEM CELL REPLACEMENT

At least three different therapeutic concepts for cell replacement can be envisaged **(Fig. 473-1)**. One therapeutic approach involves direct administration of stem cells. The cells may be injected directly into the damaged organ, where they can differentiate into the desired cell type. Alternatively, stem cells may be injected systemically since they have the capacity to home in on damaged tissues by following gradients of cytokines and chemokines released by the diseased organ. A second approach involves transplantation of differentiated cells derived from stem cells. For example, pancreatic islet cells can be generated from stem cells before transplantation into diabetic patients, and cardiomyocytes can be generated to heart disease. A third approach involves stimulation of endogenous stem cells to facilitate repair. This goal might be accomplished by administration of appropriate growth factors and drugs that amplify the number of endogenous stem/progenitor cells and/or direct them to differentiate into the desired cell types. Therapeutic stimulation of precursor cells is already a clinical reality in the hematopoietic system, where factors such as erythropoietin, granulocyte colony-stimulating factor, and granulocyte-macrophage colony-stimulating factor are used to increase production of specific blood elements. In addition to these strategies for cell replacement, a number of other approaches could involve stem cells for ex vivo or in situ generation of tissues, a process termed tissue engineering. Stem cells are excellent candidates as vehicles for cellular gene therapy **(Chap. 458)**, and they also potentially can be used to modify immune responses. Finally, transplanted stem cells may exert paracrine effects to promote repair of damaged tissues without differentiating to replace lost cells.

Stem cell transplantation is not a new concept but rather is already part of established medical practice. Hematopoietic stem cells (HSCs) **(Chap. 92)** are responsible for the long-term repopulation of all blood elements in recipients of bone marrow transplants, and hematopoietic stem cell transplantation is the gold standard against which other stem cell transplantation therapies will be measured. Transplantation of differentiated cells is also a clinical reality, and donated organs and tissues are often used to replace damaged tissues. However, the need for transplantable tissues and organs far outweighs the available supply, and organ transplantation has limited potential for some tissues, such as the brain. Stem cells offer the possibility of a renewable source of replacement cells for virtually all organs.

SOURCES OF STEM CELLS FOR TISSUE REPAIR

A variety of different types of stem cells could be used in regenerative strategies, including embryonic stem (ES) cells, induced pluripotent stem (iPS) cells, umbilical-cord blood stem cells (USCs), organ-specific

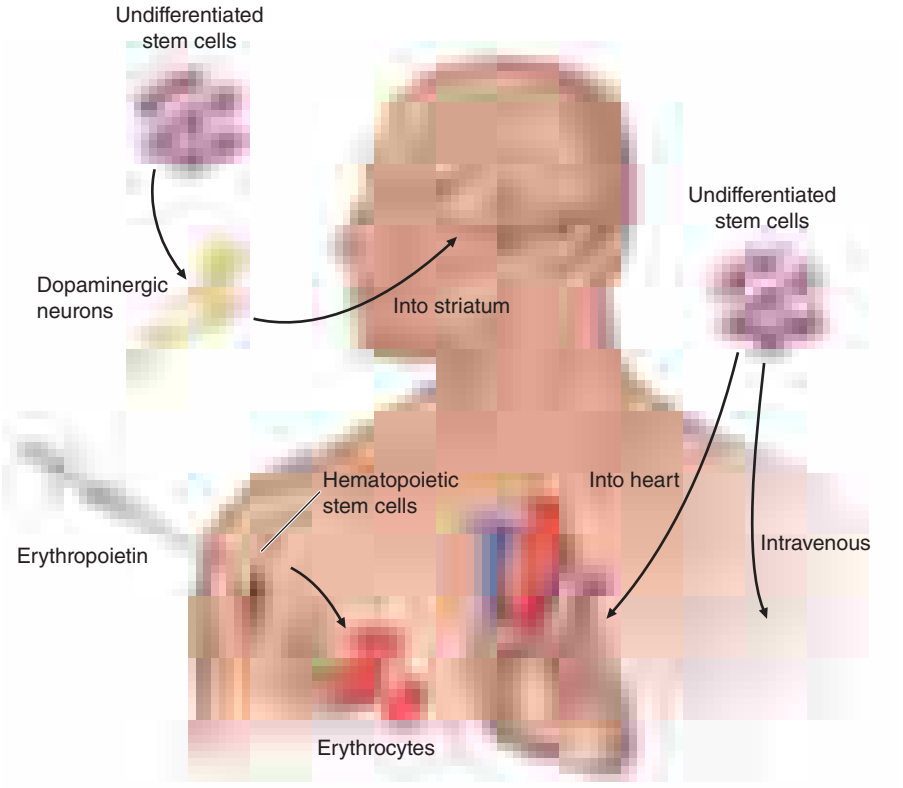


FIGURE 473-1 Strategies for transplantation of stem cells. 1. Undifferentiated or partially differentiated stem cells may be injected directly into the target organ or intravenously. 2. Stem cells may be differentiated ex vivo before injection into the target organ. 3. Growth factors or other drugs may be injected to stimulate endogenous stem cell populations.

somatic stem cells (e.g., neural stem cells for treatment of the brain), and somatic stem cells that generate cell types specific for the target organ rather than the donor organ (e.g., bone marrow mesenchymal stem cells [MSCs] or CD34+ HSCs for cardiac repair). Although each cell type has potential advantages and disadvantages, there are a number of generic challenges associated with developing any of these cell types into a useful and reliable clinical tool.

Embryonic Stem Cells ES cells have the potential to generate all of the cell types in the body; thus, in theory, there are no restrictions on the organs that could be regenerated. ES cells can self-renew endlessly, so that a single cell line with carefully characterized traits potentially could generate almost limitless numbers of cells. In the absence of moral or ethical constraints (see "Ethical Issues," below), unused human blastocysts from fertility clinics could be used to derive new ES cell lines that are matched immunologically with potential transplant recipients. Alternatively, somatic cell nuclear transfer ("therapeutic cloning") could be used to create ES cell lines that are genetically identical to those of the patient, although this endeavor has been technically refractory for human cells. However, human ES cells are difficult to culture and grow slowly. Techniques for differentiating them into specific cell types are just beginning to be developed. Cells tend to develop abnormal karyotypes and other abnormalities with increased time in culture, and ES cells have the potential to form teratomas if all cells are not committed to the desired cell types before transplantation. Further, human ES cells are ethically controversial and, on these grounds, their use would be unacceptable to some patients and physicians despite their therapeutic potential. Nevertheless, there have been limited clinical trials of ES-derived cells in a number of disorders, including macular degeneration, myopia, heart failure, diabetes, and spinal cord injury (SCI).

Induced Pluripotent Stem Cells The field of stem cell biology was transformed by the discovery that adult somatic cells can be converted ("reprogrammed") into pluripotent cells through the overexpression of four transcription factors normally expressed in pluripotent cells. These iPS cells share most properties with ES cells, although there are distinct differences in gene expression between ES and iPS cells. The initial use of viruses to insert the transcription factors into somatic cells made the resulting cells unsuitable for clinical use. However, a number of strategies have since been developed to circumvent this problem, including the insertion of modified mRNAs, proteins, or microRNAs rather than cDNAs and treatment with small molecules; the use of non-integrating viruses such as Sendai virus; the insertion of transposons with the programming factors, followed by their subsequent removal; and the use of floxed viral constructs, followed by treatment with Cre recombinase to excise those constructs. iPS cells derived from patients with different disorders are currently being used extensively for disease modeling and drug discovery. However, the safety of iPS cells for use in regenerative strategies in humans remains to be demonstrated. The first clinical trial in macular degeneration was suspended after treatment of one patient because of discovery of a mutation in cells derived for the second patient, but it is scheduled to resume. Potential advantages of iPS cells are that somatic cells from patients would generate pluripotent cells genetically identical to those of the patient, and that these cells are not subject to the same ethical constraints as ES cells. It is not clear whether the differences in gene expression between ES and iPS cells will have any impact on their potential clinical utility, and studies of both cell types will be needed to resolve this issue.

Umbilical-Cord Stem Cells USCs are widely available. These cells appear to be associated with less graft-versus-host disease than are some other cell types, such as marrow stem cells. They have less human leukocyte antigen restriction than adult marrow stem cells and are less likely to be contaminated with herpesvirus. However, it is unclear how many different cell types can be generated from USCs, and methods for differentiating these cells into non-hematopoietic phenotypes are largely lacking. Nevertheless, there are ongoing clinical trials of these cells in dozens of disorders, including cirrhosis, cardiomyopathies, multiple sclerosis, burns, stroke, autism, and critical limb ischemia.

Organ-Specific Multipotent Stem Cells Organ-specific multipotent stem cells have the advantage of already being somewhat specialized so that the inducement of desired cell types may be easier. Cells potentially could be obtained from the patient and amplified in cell culture, circumventing the problems associated with immune rejection. Stem cells are relatively easy to harvest from some tissues, such as bone marrow and blood, but are difficult to harvest from other tissues, such as heart and brain. Moreover, these populations of cells are more limited in potentiality than are pluripotent ES or iPS cells, and they may be difficult to obtain in large quantities from many organs. Therefore, substantial efforts have been devoted to developing techniques for using more easily obtainable stem cell populations, such as bone marrow MSCs, CD34+ HSCs, cardiac mesenchymal cells, and adipose-derived stem cells (ASCs), for use in regenerative strategies. Tissue culture evidence suggests that these stem cell populations may be able to generate differentiated cell types unrelated to their organ source (including myocytes, chondrocytes, tendon cells, osteoblasts, cardiomyocytes, adipocytes, hepatocytes, and possibly neurons) in a process known as *transdifferentiation*. However, it is still unclear whether these stem cells are capable of generating differentiated cell types that integrate into organs, survive, and function after transplantation in vivo. A number of early studies of MSCs transplanted into heart, liver, and other organs suggested that the cells had differentiated into organ-specific cell types with beneficial effects in animal models of disease. Unfortunately, subsequent studies revealed that the stem cells had simply fused with cells resident in the organs and that the observed beneficial effects were due to paracrine release of trophic and anti-inflammatory cytokines. Further studies will be necessary to determine whether transdifferentiation of MSCs, ASCs, or other stem cell populations occurs at a high enough frequency to make these cells useful for stem cell replacement therapy. Despite the remaining issues, clinical trials of MSCs, autologous HSCs, USCs, and ASCs are being performed in many disorders, including ischemic cardiac disease, cardiomyopathy, diabetes, stroke, cirrhosis, amyotrophic lateral sclerosis (ALS), and muscular dystrophy. Another approach has been to derive organ-specific stem cells by parthenogenesis, and a clinical trial of neural stem cells derived this way has begun for Parkinson's disease.

Regardless of the source of the stem cells used in regenerative strategies, a number of generic problems must be overcome for the development of successful clinical applications. Methods must be devised to reliably generate large numbers of specific cell types, to minimize the risk of tumor formation or proliferation of inappropriate cell types, to ensure the viability and function of the engrafted cells, to overcome immune rejection when autografts are not used, and to facilitate revascularization of regenerated tissue. Each organ system also will pose tissue-specific problems for stem cell therapies.

DISEASE-SPECIFIC APPLICATIONS OF STEM CELLS

Ischemic Heart Disease and Cardiomyocyte Regeneration Because of the high prevalence of ischemic heart disease, extensive efforts have been devoted to the development of strategies for stem cell replacement of cardiomyocytes. Historically, the adult heart has been viewed as a terminally differentiated organ without the capacity for regeneration. However, recent studies have demonstrated that the heart has the capacity for low levels of cardiomyocyte regeneration (Chap. 232). This regeneration appears to be accomplished by cardiac stem cells resident in the heart and possibly also by cells originating in the bone marrow. The heart might be an ideal source of stem cells for therapeutic use, but techniques for isolating, characterizing, and amplifying large numbers of these cells have not yet been perfected. For successful myocardial repair, stem cell therapy must deliver cells either systemically or locally, and the cells must survive, engraft, and differentiate into functional cardiomyocytes that couple mechanically and electrically with the recipient myocardium. The optimal method for cell delivery is not clear, and various experimental and clinical studies have successfully employed intramyocardial, transendocardial, intravenous, intracoronary, and retrograde coronary venous injections. In experimental myocardial infarction, functional improvements have

been achieved after transplantation of a variety of different cell types, including ES cells, HSCs, MSCs, USCs, and ASCs. Early studies suggested that each of these cell types might have the potential to engraft and generate cardiomyocytes. However, most investigators have found that the generation of new cardiomyocytes by these cells is at best a rare event and that graft survival over long periods is poor. The preponderance of evidence suggests that the observed beneficial effects of most experimental therapies were not derived from direct stem cell generation of cardiomyocytes but rather from indirect effects of the stem cells on resident cells. It is not clear whether these effects reflect the release of soluble trophic factors, the induction of angiogenesis, the release of anti-inflammatory cytokines, or another mechanism. A wide variety of cell delivery methods, cell types, and cell doses have been used in a progressively enlarging series of clinical trials, but the fate of the cells and the mechanisms by which they alter cardiac function are still open questions. In aggregate, however, these studies have shown a small but measurable improvement in cardiac function and, in some cases, reduction in infarct size. Further, transplantation of bone marrow-derived stem cells improved outcome for patients in heart failure. In short, the available evidence suggests that the beneficial clinical impact reflects an indirect effect of the transplanted cells rather than cell replacement. However, genuine cell replacement may become possible as new protocols are being developed for generating cardiomyocytes from pluripotent and multipotent stem cells.

Diabetes Successes with islet cell and pancreas transplantation have provided proof of concept for cell-based therapies for type 1 diabetes. However, the demand for donor pancreases far exceeds the number available, and maintenance of long-term graft survival remains a problem. The search for a renewable source of stem cells capable of regenerating pancreatic islets has therefore been intensive. Pancreatic beta cell turnover occurs even in the normal pancreas, although the source of the new beta cells remains controversial. This persistent turnover suggests that, in principle, it should be possible to develop strategies for reconstituting the beta cell population in diabetics. Attempts to devise techniques for promoting endogenous regenerative processes by using combinations of growth factors, drugs, and gene therapy have failed thus far, but this remains a potentially viable approach. A number of different cell types are candidates for use in stem cell replacement strategies, including iPS cells, ES cells, hepatic progenitor cells, pancreatic ductal progenitor cells, and MSCs. Successful therapy will depend on the development of a source of cells that can be amplified to produce large numbers of progeny with the ability to synthesize, store, and release insulin when it is required, primarily in response to changes in the ambient level of glucose. The proliferative capacity of the replacement cells must be tightly regulated to avoid excessive expansion of beta cell numbers and the consequent development of hyperinsulinemia/hypoglycemia; moreover, the cells must withstand immune rejection. Although ES and iPS cells can be differentiated into cells that produce insulin, these cells have a lower content of insulin and a higher rate of apoptosis than pancreatic beta cells, and generally lack the capacity to fully normalize blood glucose levels in diabetic animals. However, clinical trials of encapsulated ES cell-derived pancreatic progenitor cells are currently in progress.

During embryogenesis, the pancreas, liver, and gastrointestinal tract are all derived from the anterior endoderm, and transdifferentiation of pancreas to liver and vice versa has been observed in a number of pathologic conditions. There is also substantial evidence that multipotential stem cells reside within gastric glands and intestinal crypts. These observations suggest that hepatic, pancreatic, and/or gastrointestinal precursor cells may be reasonable candidates for cell-based therapy for diabetes, although it is unclear whether insulin-producing cells derived from pancreatic stem cells or liver progenitors can be expanded in vitro to clinically useful numbers. MSCs and neural stem cells both reportedly have the capacity to generate insulin-producing cells, but there is no convincing evidence that either cell type will be clinically useful. Clinical trials of MSCs, USCs, HSCs, and ASCs in both type 1 and type 2 diabetes are ongoing.

Nervous System Substantial progress has been made in the development of methodologies for generating neural cells from different stem cell populations. Human ES or iPS cells can be induced to generate cells with the properties of neural stem cells, and these cells in turn give rise to neurons, oligodendroglia, and astrocytes. Reasonably large numbers of these cells can be transplanted into the rodent brain with formation of appropriate cell types and no tumor formation. Multipotent stem cells present in the adult brain also can be easily amplified in number and used to generate all the major neural cell types, but the need for invasive procedures to obtain autologous cells is a major limitation. Fetal neural stem cells derived from miscarriages or abortions are an alternative but raise ethical concerns. Nevertheless, clinical trials of fetal neural stem cells have commenced in ALS, stroke, and several other disorders. Transdifferentiation of MSCs and ASCs into neural stem cells, and vice versa, has been reported by numerous investigators, and clinical trials of such cells have begun for a number of neurologic diseases. Clinical trials of a conditionally immortalized human cell line and of USCs in stroke are also in progress. Because of the incapacitating nature of neural disorders and the limited endogenous repair capacity of the nervous system, clinical trials of stem cells in neurologic disorders have been particularly numerous, including trials in SCI, multiple sclerosis, epilepsy, Alzheimer's disease, ALS, acute and chronic stroke, numerous genetic disorders, traumatic brain injury, Parkinson's disease, and others. In diseases such as ALS, possible benefits are more likely to be due to indirect trophic effects than to neuron replacement. In Parkinson's disease, the major motor features of the disorder result from the loss of a single cell population: dopaminergic neurons within the substantia nigra; this circumstance suggests that cell replacement should be relatively straightforward. However, two clinical trials of fetal nigral transplantation failed to meet their primary endpoint and were complicated by the development of dyskinesia. Transplantation of stem cell–derived dopamine-producing cells offers a number of potential advantages over the fetal transplants, including the ability of stem cells to migrate and disperse within tissue, the potential for engineering regulatable release of dopamine, and the ability to engineer cells to produce factors that will enhance cell survival. Nevertheless, the experience with fetal transplants points out the difficulties that may be encountered.

At least some of the neurologic dysfunction after SCI reflects demyelination, and both ES cells and MSCs can facilitate remyelination after experimental SCI. Clinical trials of MSCs in this disorder have commenced in a number of countries, and SCI was the first disorder targeted for the clinical use of ES cells. The first trial of ES cell derived oligodendroglial progenitor cells in SCI was terminated early for nonmedical reasons, but another trial has commenced. At present, no population of transplanted stem cells has been shown to have the capacity to generate neurons that extend axons over long distances to form synaptic connections (as would be necessary for replacement of upper motor neurons in ALS, stroke, or other disorders). For many injuries, including SCI, the balance between scar formation and tissue repair/regeneration may prove to be an important consideration. For example, it may ultimately prove necessary to limit scar formation so that axons can reestablish connections.

Liver Liver transplantation is currently the only successful treatment for end-stage liver diseases, but the shortage of liver grafts limits its application. Clinical trials of hepatocyte transplantation demonstrate its potential as a substitute for organ transplantation, but this approach is limited by the paucity of available cells. Potential sources of stem cells for regenerative strategies include endogenous liver stem cells (such as oval cells), ES cells, MSCs, and USCs. Although a series of studies in humans as well as animals suggested that transplanted MSCs and HSCs can generate hepatocytes, fusion of the transplanted cells with endogenous liver cells, giving the erroneous appearance of new hepatocytes, appears to be the underlying event in most circumstances. The available evidence suggests that transplanted HSCs and MSCs can generate hepatocyte-like cells in the liver only at a very low frequency, but there are beneficial consequences presumably related to indirect paracrine effects. ES cells can be differentiated into hepatocytes

and transplanted in animal models of liver failure without the formation of teratomas. Clinical trials are in progress in cirrhosis with numerous cell types, including MSCs, USCs, HSCs, and ASCs.

Other Organ Systems and the Future The use of stem cells in regenerative strategies has been studied for many other organ systems and cell types, including skin, eye, cartilage, bone, kidney, lung, endometrium, vascular endothelium, smooth muscle, and striated muscle, and clinical trials in these and other organs are ongoing. In fact, the potential for stem cell regeneration of damaged organs and tissues is virtually limitless. However, numerous obstacles must be overcome before stem cell therapies can become a widespread clinical reality. Only HSCs have been adequately characterized by surface markers so that they can be unambiguously identified, a prerequisite for reliable clinical applications. The pathways for differentiating stem cells into specific cellular phenotypes are largely unknown, and the ability to control the migration of transplanted cells or predict the response of the cells to the environment of diseased organs is presently limited. Some strategies may employ the coadministration of scaffolding, artificial extracellular matrix, and/or growth factors to orchestrate differentiation of stem cells and their organization into appropriate constituents of the organ. There is currently no way to image stem cells in vivo after transplantation into humans, and it will be necessary to develop techniques to do so. Fortunately, stem cells can be engineered before transplantation to contain a contrast agent that may make in vivo imaging feasible. The potential for tumor formation and the problems associated with immune rejection are impediments, and it will also be necessary to develop techniques for ensuring vascularization of regenerated tissues. There already are many strategies for supporting cell replacement, including coadministration of vasoactive endothelial growth factor to foster vascularization of the transplant. Some strategies also include the genetic engineering of stem cells with an inducible suicide gene so that the cells can be easily eradicated in the event of tumor formation or another complication. The potential for stem cell therapies to revolutionize medical care is extraordinary, and disorders such as myocardial infarction, diabetes, and Parkinson's disease, among many others, are potentially curable by such therapies. However, stem cell–based therapies are still at a very early stage of development, and perfection of techniques for clinical transplantation of predictable, well-characterized cells is going to be a difficult and lengthy undertaking.

ETHICAL ISSUES

Stem cell therapies raise ethical and socially contentious issues that must be addressed in parallel with the scientific and medical opportunities. Society has great diversity with respect to religious beliefs, concepts of individual rights, tolerance for uncertainty and risk, and boundaries for how scientific interventions should be used to alter the outcome of disease. In the United States, the federal government has authorized research using existing human ES cell lines but still restricts the use of federal funds for developing new human ES cell lines. Ongoing studies of existing lines have indicated that they develop abnormalities with time in culture and that they may be contaminated with mouse proteins. These findings highlight the need to develop new human ES cell lines. The development of iPS cell technology may lessen the need for deriving new ES cell lines, but it is still not clear whether the differences in gene expression by ES and iPS cells are important for potential clinical use.

In considering ethical issues associated with the use of stem cells, it is helpful to draw from experience with other scientific advances, such as organ transplantation, recombinant DNA technology, implantation of mechanical devices, neuroscience and cognitive research, in vitro fertilization, and prenatal genetic testing. These and other precedents have pointed to the importance of understanding and testing fundamental biology in the laboratory setting and in animal models before applying new techniques in carefully controlled clinical trials. When these trials occur, they must include full informed consent and careful oversight by external review groups.

Ultimately, there will be medical interventions that are scientifically feasible but ethically or socially unacceptable to some members of a society. Stem cell research raises fundamentally difficult questions about the definition of human life, and it has raised deep fears about the ability to balance issues of justice and safety with the needs of critically ill patients. Health care providers and experts with backgrounds in ethics, law, and sociology must help guard against the premature or inappropriate application of stem cell therapies and the inappropriate involvement of vulnerable population groups. However, these therapies offer important new strategies for the treatment of otherwise irreversible disorders. An open dialogue among the scientific community, physicians, patients and their advocates, lawmakers, and the lay population is critically important to raise and address important ethical issues and balance the benefits and risks associated with stem cell transfer.

FURTHER READING

Benthuysen JR, Carrano AC, Sander M: Advances in β cell replacement and regeneration strategies for treating diabetes. J Clin Invest 126:3651, 2016.

Goodell MA, Rando TA: Stem cells and healthy aging. Science 350:1199, 2015.

Guidelines for Stem Cell Research and Clinical Translation International Society for Stem Cell Research May 2016. Available from *http://www.isscr.org/docs/default-source/guidelines/isscr-guidelines-for-stem-cell-research-and-clinical-translation.pdf?sfvrsn=2*. Accessed June 26, 2017.

Nguyen PK, Rhee J-W, Wu JC: Adult Stem Cell Therapy and Heart Failure, 2000 to 2016: A Systematic Review. JAMA Cardiol 1:831, 2016.

Passier R, Orlova V, Mummery C: Complex tissue and disease modeling using hiPSCs. Cell Stem Cell 18:309, 2016.

Rowe RG et al: Engineering hematopoietic stem cells: Lessons from development. Cell Stem Cell. 18:707, 2016.

Steinbeck JA, Studer L: Moving stem cells to the clinic: Potential and limitations for brain repair. Neuron 86:187, 2015.

Tapia N, Schöler HR: Molecular obstacles to clinical translation of iPSCs. Cell Stem Cell 19:298, 2016.

Trounson A, DeWitt ND: Pluripotent stem cells progressing to the clinic. Nat Rev Mol Cell Biol 17:194, 2016.

474 Microbial Genomics and Infectious Disease

Roby P. Bhattacharyya, Yonatan H. Grad, Deborah T. Hung

Just as microscopy opened up the worlds of microbiology by providing a tool with which to visualize microorganisms, technological advances in genomics are now providing microbiologists with powerful new methods to characterize the genetic map that underlies all microbes with unprecedented resolution, thereby illuminating their complex and dynamic interactions with each other, the environment, and human health. The field of infectious disease genomics encompasses a vast frontier of active research that is beginning to transform the clinical practice of infectious diseases. While genetics has long played a key role in elucidating the process of infection and impacting clinical infectious diseases, the ability to extend our thinking and our approaches beyond the study of single genes to an examination of the sequence, structure, and function of entire genomes is allowing us to identify new possibilities for research and opportunities to change clinical practice. From the development of diagnostics with unprecedented sensitivity, specificity, and speed to the design of novel public health interventions, technical and statistical genomic innovations are reshaping our understanding of the influence of the microbial world on human health and providing us with new tools to diagnose, track, and combat infection. In this chapter, we explore the application of genomics methods to

TABLE 474-1 Glossary of Selected Terms in Genomics

TERM	DEFINITION
Contig	A DNA sequence representing a continuous fragment of a genome, assembled from overlapping sequences; relevant for de novo assembly of sequence data that do not align to previously sequenced genomes
Genome	The entire set of heritable genetic material within an organism
Horizontal gene transfer	The transfer of genes between organisms through mechanisms other than by clonal descent, such as through transformation, conjugation, or transduction
Metagenomics	Analysis of genetic material from multiple species directly from primary samples without requiring prior culture steps
Microarray	A collection of DNA oligonucleotides ("oligos") spatially arranged on a solid surface and used to detect or quantify sequences in a sample of interest that are complementary (and therefore bind) to one or more of the arrayed oligos
Microbial genome-wide association study (GWAS)	An analytic framework to test statistical associations between microbial genotypes and phenotypes of interest, such as antibiotic resistance and virulence
Mobile genetic elements	DNA elements that can move within a genome and can be transferred between genomes through horizontal gene transfer (e.g., plasmids, bacteriophages, and transposons)
Multilocus sequence typing	A method for typing organisms based on DNA sequence fragments from a prespecified set of genes
Next-generation sequencing	High-throughput sequencing using a parallelized sequencing process that produces millions of sequences concurrently, far beyond the capacity of prior dye-terminator methods
Nucleic acid amplification test (NAAT)	A biochemical assay that evaluates for the presence of a particular string of nucleic acids through amplification by one of several methods, including polymerase and ligase chain reactions
Polymerase chain reaction (PCR)	A type of NAAT used to amplify a specific region of DNA by means of specific oligonucleotide primers and a DNA polymerase
Single-nucleotide polymorphism (SNP)	Point mutations, the number of which in different microbial isolates is a measure of their genetic distance from one another
Transcriptome	The catalog of the full set of messenger RNA (mRNA) transcripts from a cell or organism, which are typically measured by microarray or by next-generation sequencing of complementary DNA (cDNA) via a process called RNA-Seq
Whole-genome sequencing (WGS)	A process that determines the full DNA sequence of an organism's genome; has been greatly facilitated by next-generation sequencing technology

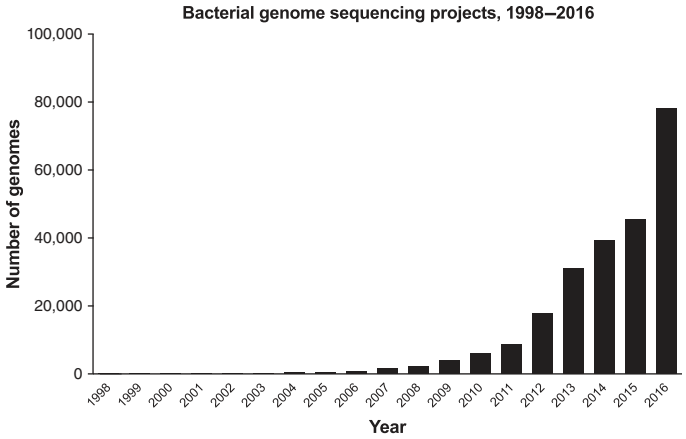


FIGURE 474-1 Bacterial genome sequencing projects submitted to the Genomes Online Database, a manually curated repository for genome and metagenome sequencing data, from 1998 to 2016. *(Data compiled from https://gold.jgi.doe.gov/statistics, accessed July 17, 2017. See S Mukherjee et al: Genomes OnLine Database (GOLD) v.6: Data updates and feature enhancement. Nucleic Acids Res 45:D446, 2017.)*

microbial pathogens and the infections they cause. We discuss innovations that are driving the development of diagnostic approaches as well as the discovery of new pathogens, providing insight into novel therapeutic approaches and paradigms, and advancing methods in infectious disease epidemiology and the study of pathogen evolution that can inform infection control measures, public health responses to outbreaks, and vaccine development. We draw on examples in current practice and from the recent scientific literature as signposts that point toward ways in which the insights from pathogen genomics may influence infectious diseases in the short and long terms. Table 474-1 provides definitions for a selection of important terms used in genomics.

MICROBIAL DIAGNOSTICS

The basic goals of a clinical microbiology laboratory are to establish the presence of a pathogen in a clinical sample, to identify the pathogen, and, when possible, to provide other information that can help guide clinical management and even affect prognosis, such as antibiotic susceptibility profiles or the presence of virulence factors. To date, clinical microbiology laboratories have largely approached these goals phenotypically by growth-based assays and biochemical testing. Bacteria, for instance, are algorithmically grouped into species by their characteristic microscopic appearance, nutrient requirements for growth, and ability to catalyze certain reactions. Antibiotic susceptibility is determined in most cases by assessing bacterial growth in the presence of antibiotic.

With the sequencing revolution paving the way to easy access of complete pathogen genomes (Fig. 474-1), we are now able to more systematically understand the genetic basis for these observable phenotypes. Compared with traditional growth-based methods for bacterial diagnostics that dominate the clinical microbiology laboratory, nucleic acid–based diagnostics promise improved speed, sensitivity, specificity, and breadth of information. Bridging clinical and research laboratories, adaptations of genomic technologies have begun to deliver on this promise (Table 474-2).

HISTORICAL LIMITATIONS AND PROGRESS THROUGH GENETIC APPROACHES

The molecular diagnostics revolution in the clinical microbiology laboratory is well under way, born of necessity in the effort to identify microbes that are refractory to traditional culture methods. Historically, diagnosis of many so-called unculturable pathogens has relied largely on serology and antigen detection. However, these methods provide only limited clinical information because of their suboptimal sensitivity and specificity as well as the long delays that diminish their utility for real-time patient management. Newer tests to detect pathogens, based on nucleic acid content, have already offered improvements in the select cases in which they have been applied.

Unlike direct pathogen detection, serologic diagnosis—measurement of the host's response to pathogen exposure—can typically be made only in retrospect, requiring both acute- and convalescent-phase serum samples. For chronic infections, distinguishing active from latent infection or identifying repeat exposure from serology alone can be difficult or impossible, depending on the syndrome. In addition, serologic diagnosis is variably sensitive, depending on the organism and the patient's immune status. For instance, tuberculosis is notoriously difficult to identify by serologic methods; tuberculin skin testing using purified protein derivative (PPD) is especially insensitive in active disease and possibly cross-reactive with vaccines or other mycobacteria. Even the newer interferon γ release assays (IGRAs), which measure cytokine release from T lymphocytes in response to *Mycobacterium tuberculosis*–specific antigens in vitro, have limited sensitivity in immunodeficient hosts. Neither PPD testing nor IGRAs can distinguish latent from active infection. Serologic Lyme disease diagnostics suffer similar limitations: in patients from endemic regions, the presence of IgG antibodies to *Borrelia burgdorferi* may reflect prior exposure rather than active disease, while IgM antibodies are imperfectly sensitive and specific (50% and 80%, respectively, in early disease). The complicated nature of these tests, particularly in view of the nonspecific symptoms that may accompany Lyme disease, has had substantial implications for public

TABLE 474-2 Selected Clinical Applications of Infectious Disease Genomics

APPLICATION	TECHNOLOGY	NOTES/EXAMPLES
Organism Identification		
Viral detection	PCR	Identification of HIV, HBV, HCV, respiratory viruses including influenza, and others for diagnosis and response to therapy
TB detection	PCR	Amplification of the *rpoB* gene for species-specific identification of *Mycobacterium tuberculosis*
Pathogen detection	PCR, NAAT	Multiplexed identification of dozens of viruses, bacteria, yeasts, and parasites from a variety of clinical specimens
Bacterial detection	16S ribosomal gene sequencing	Targeted amplification and sequencing of regions of the 16S rRNA gene for identification of suspected bacterial infections undiagnosed by conventional methods
Pathogen Discovery		
Bacterial pathogens	Sequencing, metagenomic assembly	Unbiased "shotgun" sequencing of isolated nucleic acid from patient samples to identify associated pathogens; proofs-of-concept: new *Bradyrhizobium* species associated with cord colitis; *Escherichia coli* 0104:H4 from 2011 diarrheal outbreak in Germany; *Leptospira* species from one patient's cerebrospinal fluid; research use only at this time
Viral pathogens	Microarray, sequencing	Hybridization of clinical samples to microarrays from phylogenetically diverse known viruses identified the SARS coronavirus and others. Direct sequencing has identified West Nile virus and the MERS coronavirus, among others. Use is primarily in research.
Antibiotic Resistance		
MRSA detection	PCR	Detection of the *mecA* gene, the genotypic cause of methicillin resistance in *Staphylococcus aureus*
VRE detection	PCR	Detection of the *vanA* or *vanB* gene, the main genotypic causes of vancomycin resistance in *Enterococcus*
MDR-TB detection	PCR, NAAT	Detection of polymorphisms in the *rpoB* gene from *M. tuberculosis*, which account for 95% of rifampin resistance. Other probes available for *inhA* and *katG* genes can detect up to 85% of isoniazid resistance.
Carbapenemase detection	PCR	Detection of genes encoding one of several types of enzymes (KPC, NDM, OXA-48, IMP-1, VIM) that hydrolyze carbapenems, accounting for much but not all carbapenemase resistance in Enterobacteriaceae
HIV resistance detection	Targeted sequencing	Targeted sequencing of specific genes with known resistance-conferring mutations; now the standard of care prior to initial therapy in the United States and Europe
Epidemiology		
Outbreak and epidemic tracking	Sequencing	Application to tracking outbreaks and epidemics on local and international scales, including spread of carbapenemase-producing *Klebsiella*, *S. aureus*, *M. tuberculosis*, *E. coli*, *Vibrio cholerae*, Ebola virus, Zika virus, and influenza virus
Evolution and spread of pathogens	Sequencing	Sequencing collections of pathogens to shed light on pathogen dissemination, virulence factors, and antibiotic resistance determinants; innumerable examples, including *V. cholerae*, influenza virus, Ebola virus, and Zika virus

Abbreviations: HBV, hepatitis B virus; HCV, hepatitis C virus; MDR, multidrug-resistant; MERS, Middle East respiratory syndrome; MRSA, methicillin-resistant *S. aureus*; NAAT, nucleic acid amplification test; PCR, polymerase chain reaction; SARS, severe acute respiratory syndrome; TB, tuberculosis; VRE, vancomycin-resistant enterococci.

perception of Lyme disease and antibiotic misuse in endemic areas. Similarly, syphilis, a chronic infection caused by *Treponema pallidum*, is notoriously difficult to stage by serology alone, requiring multiple different nontreponemal and treponemal tests (e.g., rapid protein reagin and fluorescent treponemal antibody, respectively) in conjunction with clinical suspicion. Complementing serology, antigen detection can improve sensitivity and specificity in select cases, but has been validated only for a limited set of infections. Typically, structural elements of pathogens are detected, including components of viral envelopes (e.g., hepatitis B surface antigen, HIV p24 antigen), cell surface markers in certain bacteria (e.g., *Streptococcus pneumoniae*, *Legionella pneumophila* serotype 1) or fungi (e.g., *Cryptococcus*, *Histoplasma*), and less specific fungal cell-wall components such as galactomannan and β-glucan (e.g., *Aspergillus* and other dimorphic fungi).

Given the impracticality of culture and the lack of sensitivity or sufficient clinical information afforded by serologic and antigenic methods, the push toward nucleic acid–based diagnostics originated in pursuit of viruses and fastidious bacteria, becoming part of the standard of care for select organisms in U.S. hospitals. Such tests, including polymerase chain reaction (PCR) and other nucleic acid amplification tests (NAATs), are now widely used for many viral infections, both chronic (e.g., HIV infection, hepatitis C) and acute (e.g., influenza). NAATs provide essential information about both the initial diagnosis and the response to therapy and in some cases genotypically predict drug resistance. Indeed, progression from antigen detection to PCR transformed our understanding of the natural course of HIV infection, with profound implications for treatment **(Fig. 474-2*A*)**. In the early years of the AIDS pandemic, p24 antigenemia was detected in acute HIV infection but then disappeared for years before emerging again with progression to AIDS **(Fig. 474-2*B*)**. Without a marker demonstrating viremia, the role of treatment during HIV infection prior to the development of clinical AIDS was uncertain, and assessing treatment efficacy was challenging. With the emergence of PCR as a progressively more sensitive test (now able to detect as few as 20 copies of virus per milliliter of blood), viremia was recognized as a near-universal feature of HIV infection. Given the challenges of phenotypic assays, genotypic antiviral resistance testing was also adopted early for HIV and is now the standard of care before the initiation of therapy in developed countries. These developments have been transformative in guiding therapy in early disease and, together with the development of less toxic therapies, have helped to shape policy that is moving toward ever-earlier introduction of antiretroviral therapy in HIV infection.

As they are for viral testing, nucleic acid–based tests have become the diagnostic tests of choice for fastidious bacteria, including the common sexually transmitted bacterial pathogens *Neisseria gonorrhoeae* and *Chlamydia trachomatis* as well as the tick-borne *Ehrlichia chaffeensis* and *Anaplasma phagocytophilum*. More recently, nucleic acid amplification–based detection has offered improved sensitivity for diagnosis of the important nosocomial pathogen *Clostridium difficile*, and NAATs have provided clinically relevant information on the presence of cytotoxins A and B as well as molecular markers of hypervirulence, such as the North American pulsotype 1 (NAP1) strain that is enriched in severe illness. The importance of genomics in selecting loci for diagnostic assays and in monitoring test sensitivity was highlighted by the emergence in

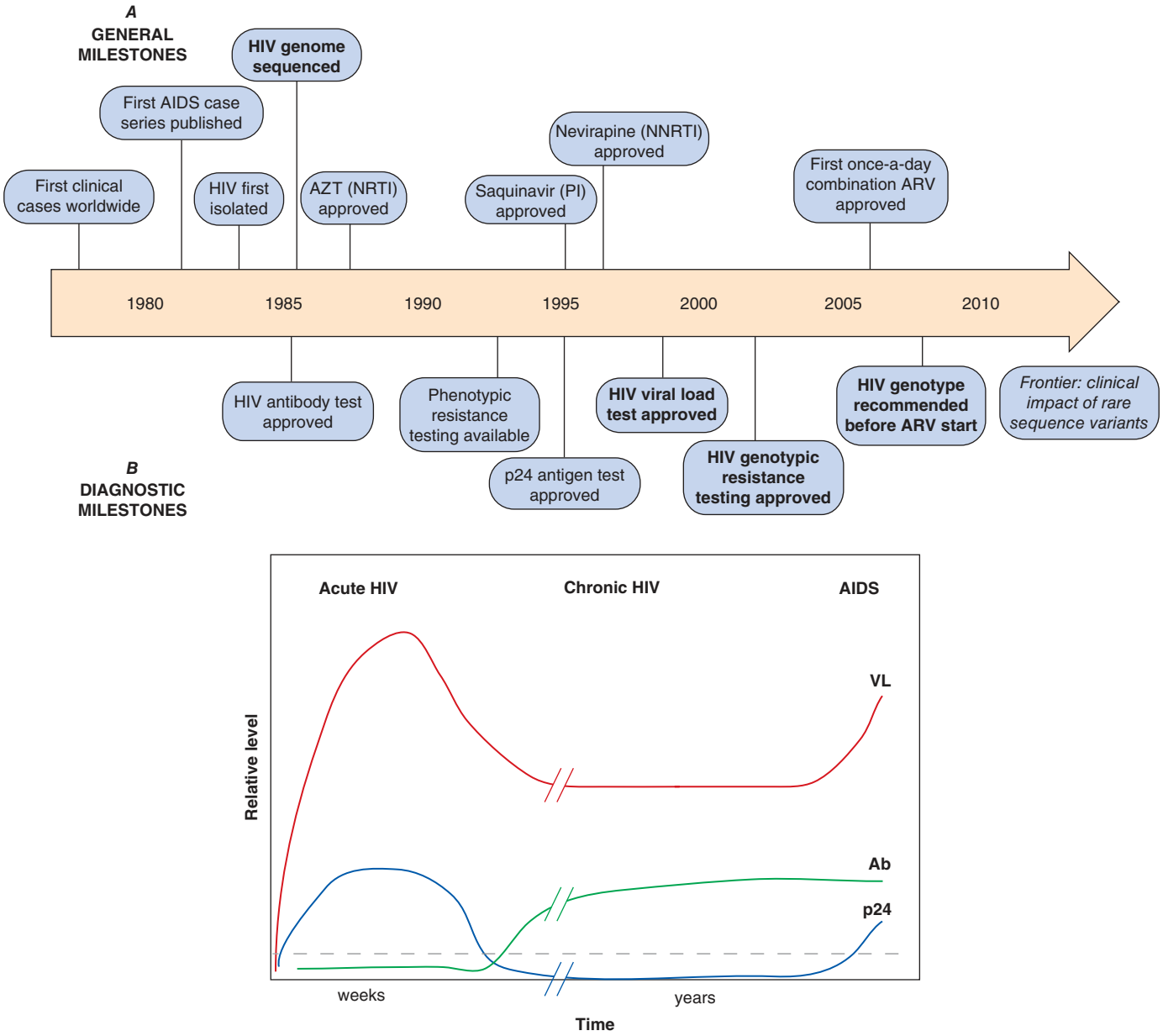


FIGURE 474-2 ***A.*** Timeline of select milestones in HIV management. Genomic advances are shown in bold type. The approvals and recommendations indicated apply to the United States. ARV, antiretroviral; AZT, zidovudine; NRTI, nucleoside reverse transcriptase (RT) inhibitor; NNRTI, nonnucleoside RT inhibitor; PI, protease inhibitor. ***B.*** Viral dynamics in the natural history of HIV infection. Three diagnostic markers are shown: HIV antibody (Ab), p24 antigen (p24), and viral load (VL). *Dashed gray line* represents limit of detection. *(Adapted from data in HH Fiebig et al: Dynamics of HIV viremia and antibody seroconversion in plasma donors: Implications for diagnosis and staging of primary HIV infection. AIDS 17:1871, 2003.)*

Sweden of a newly recognized variant of *C. trachomatis* with a deletion that includes the gene targeted by a set of commercial NAATs. By evading detection through this deletion (and thus avoiding treatment), this strain came to be highly prevalent in some areas of Sweden. While nucleic acid–based tests remain the diagnostic approach of choice for fastidious bacteria, this example serves as a reminder of the need for careful development and ongoing monitoring of molecular diagnostics.

In contrast, for typical bacterial pathogens for which culture methods are well established, growth-based assays followed by biochemical tests still dominate in the clinical laboratory. Informed by decades of clinical microbiology, these tests have served clinicians well, yet the limitations of growth-based tests—in particular, the delays associated with waiting for growth—have left opportunities for improvements. Driven by this need, mass spectrometry–based assays are already being adopted for highly accurate organism identification within a few hours of a positive blood culture. Looking ahead, molecular diagnostics, greatly informed by the vast quantity of microbial genome sequences generated in recent years, offers a way forward. First, sequencing studies can readily identify key genes (or noncoding nucleic acids) that can be developed into targets for clinical assays using PCR or hybridization assay platforms. Second, sequencing itself may eventually become cheap and rapid enough to be performed routinely on clinical specimens, with consequent unbiased detection of pathogens.

One of the biggest drivers for the implementation of novel molecular technologies in the diagnosis of infectious diseases is the desire for more rapid—or even real-time—pathogen identification, ideally with antibiotic susceptibility information on those microbes for which resistance to the current anti-infective armamentarium is of concern. Such real-time tests have the potential to transform infectious disease management, impacting antibiotic stewardship in the outpatient setting, mortality risk in the critically ill (i.e., patients in whom early administration of effective antibiotics is the most significant factor in decreasing mortality risk), hospital admission, and length of hospital stay; the extent of this impact will depend on the economic forces that will help define the breadth of their deployment. On the public health level, such tests will likely play a role in improving antibiotic stewardship,

thereby influencing the rise of antibiotic resistance and enabling surveillance of outbreaks by local, national, and international networks. In the United States and the United Kingdom, for example, public health agencies have shifted from pulsed-field gel electrophoresis to genome sequencing to track food-borne pathogens and identify outbreaks; in addition, these countries are rapidly expanding the routine use of genomics in identifying and characterizing other pathogens, from mycobacteria (both *M. tuberculosis* and nontuberculous mycobacteria) to *N. gonorrhoeae*. Further, international efforts to track the spread of viral diseases, including recent work on Ebola and Zika outbreaks and ongoing work on seasonal influenza, offer opportunities for improving interventions, surveillance, and prevention efforts, ranging from more accurate selection of the influenza virus strains to include in seasonal vaccine development to improved design of trials to evaluate novel vaccines and therapies.

Technological innovations are lowering several critical barriers to the widespread adoption of genomics and other molecular methods. Specifically, for clinical sequencing: (1) the cost and speed of sequencing and analysis methods continue to fall precipitously; (2) automation and miniaturization of the preparation of a sample for sequencing promise to reduce cost and minimize the expertise needed; and (3) direct sequencing technologies that eliminate the complex molecular biology required to prepare clinical samples for sequencing are improving in accuracy and robustness. Further barriers exist, including the need for standardized pipelines to process data and present clinicians with easily interpretable and readily actionable results. However, as these advances give rise to rapid, accurate diagnostic tests, the ultimate goal is to inform a clinician in real time whether antibiotics are indicated and, if so, which will be effective. Real-time diagnostics will allow more efficient deployment of our precious antibiotic arsenal, thus improving both societal and patient-specific outcomes in much the same way that a rapid, sensitive troponin assay has transformed bedside management of chest pain.

ORGANISM IDENTIFICATION

In order to adapt nucleic acid detection to diagnostic tests and thus to identify pathogens on a wide scale, sequences must be found that are conserved enough within a species to identify the diversity of strains that may be encountered in various clinical settings, but divergent enough to distinguish one species from another. Until recently, this problem has been solved for bacteria by targeting the element of a bacterial genome that is most highly conserved within a species, the 16S ribosomal RNA (rRNA) subunit. This method has now been used to confirm *Mycobacterium chimaera* infections in several patients after cardiothoracic surgery, leading ultimately to recognition of a widespread outbreak. At present, 16S PCR amplification from tissue specimens can be performed by specialty laboratories, though its sensitivity and clinical utility to date have remained somewhat limited, in part because of the rarity of pathogen nucleic acid in the sampled tissue, which necessitates reliable, sensitive nucleic acid amplification. As such barriers are reduced through technological advances and as the causes of culture-negative infection are clarified (perhaps in part through sequencing efforts), these tests may become both more accessible and more helpful.

With the wealth of sequencing data now available, other regions beyond 16S rRNA can be targeted for bacterial species identification. These other genomic loci can provide additional information about a clinical isolate that is relevant to patient management. For instance, detection of the presence, or potentially even the expression, of toxin genes such as *C. difficile* toxins A and B or Shiga toxin may provide clinicians with additional information that will help distinguish commensals or colonizing bacteria from pathogens and thus aid in prognostication and management as well as in diagnosis.

While amplification tests such as PCR exemplify one approach to nucleic acid detection, other approaches exist, including detection by hybridization. Although not currently used in the clinical realm, techniques for detection and identification of pathogens by hybridization to microarrays are being developed for other purposes. Of note, these different detection techniques require different degrees of conservation. Highly sensitive amplification methods require a high degree of sequence identity between PCR primer pairs and their short, specific target sequences; even a single base-pair mismatch (particularly near the 3′ end of the primer) may interfere with detection. In contrast, hybridization-based tests are more tolerant of mismatch and thus can be used to detect important regions that may be less precisely conserved within a species, thus potentially allowing detection of clinical isolates from a given species with greater diversity between isolates. Such assays take advantage of the predictable binding interactions of nucleic acids. The applicability of hybridization-based methods toward either DNA or RNA opens up the possibility of expression profiling, which can uncover phenotypic information from nucleic acid content.

Both PCR and hybridization methods target specific, known organisms. At the other extreme, as sequencing costs decline, metagenomic sequencing from patient samples is increasingly feasible. This shotgun sequencing approach is unbiased—i.e., is able to detect any microbial sequence, however divergent or unexpected. In one recent example, a clinical sample of cerebrospinal fluid from an immunocompromised patient with signs and symptoms of chronic meningitis was found through metagenomic sequencing and analysis to contain small amounts of *Leptospira* DNA. In light of this information, retrospective PCR testing confirmed the diagnosis of neuroleptospirosis, which had been missed prior to the sequencing result. The patient was treated with penicillin G and clinically recovered. Increasingly, efforts are under way to bring whole-genome sequencing to other clinical samples, including sputum and blood, in order to more readily identify pathogens. This new approach brings its own set of challenges, however, including the need to recognize pathogenic sequences against a background of expected host and commensal sequences, and to distinguish true pathogens from either colonizers or laboratory contaminants. The burgeoning field of microbiome research is driving technology development for sequencing and analyzing complex microbial communities. Lessons from this field will inform diagnostic efforts.

PATHOGEN DISCOVERY

In addition to clinical diagnostic applications, novel genomic technologies, including whole-genome sequencing, are being applied to clinical research specimens with a goal of identifying new pathogens in a variety of circumstances. The tremendous sensitivity and unbiased nature of sequencing is also ideal in searching clinical samples for unknown or unsuspected pathogens.

Causal inference in infectious diseases has progressed since the time of Koch, whose historical postulates provided a rigorous framework for attributing a disease to a microorganism. To modernize Koch's postulates: an organism, whether it can be cultured or not, should induce disease upon introduction into a healthy host if it is to be implicated as a causative pathogen. Current sequencing technologies are ideal for advancing this modern version of Koch's postulates because they can identify candidate causal pathogens with unprecedented sensitivity and in an unbiased way, unencumbered by limitations such as culturability. Yet, as direct sequencing on primary patient samples greatly expands our ability to recognize associations between microbes and disease states, critical thinking and experimentation will remain vital in establishing causality.

Virus discovery in particular has been greatly facilitated by new nucleic acid technology. These frontiers were first notably explored with high-density microarrays containing spatially arrayed sequences from a phylogenetically diverse collection of viruses. Despite bias toward those with homology to known viruses, novel viruses in clinical samples were successfully identified on the basis of their ability to hybridize to these prespecified sequences. This methodology famously contributed to identification of the coronavirus causing severe acute respiratory syndrome (SARS). Once discovered, the SARS coronavirus was rapidly sequenced: the full genome was assembled in April 2003, <6 months after recognition of the first case. This accomplishment illustrates the advancing power and speed of new diagnostic technologies.

With the advent of next-generation sequencing, unbiased pathogen discovery is now possible through a process known as *metagenomic assembly* **(Fig. 474-3)**. Sequences of random nucleotide fragments can

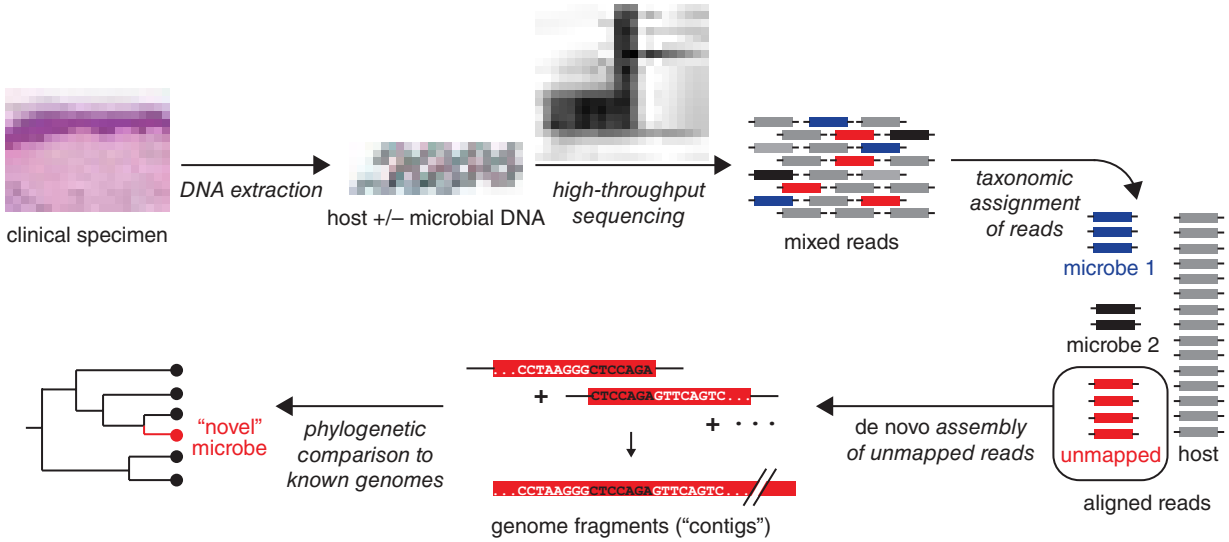


FIGURE 474-3 Workflow of metagenomic assembly for pathogen discovery. DNA is isolated from a specimen of interest (e.g., tissue, body fluid) containing a mixture of host DNA and nucleic acids from coexisting microbes, either commensal or pathogenic. All DNA (and RNA, if a reverse transcription step is added) is then sequenced, yielding a mixture of DNA sequence fragments ("reads") from the organisms present. Except for reads that do not align ("map") to any known sequence, these reads are aligned to existing reference genomes for the host or any known microbes. The unmapped reads are computationally assembled de novo into the largest contiguous stretches of DNA possible ("contigs"), representing fragments of previously unsequenced genomes. These genome fragments (contigs) are then mapped onto a phylogenetic tree based on their sequence. Some may represent known but as-yet-unsequenced organisms, while others will represent novel species. *(Figure prepared with valuable input from Dr. Ami S. Bhatt, personal communication.)*

be generated from clinical specimens with no a priori knowledge of pathogen identity through a process called *shotgun sequencing*. This collection of sequences can then be computationally aligned to host (i.e., human) sequences, with aligned sequences removed and remaining sequences compared with other known genomes to detect the presence of known microorganisms. Sequence fragments that remain unaligned suggest the presence of an additional organism that cannot be matched to a known, characterized genome; these reads can be assembled into contiguous nucleic acid stretches that can be compared with known sequences to construct the genome of a potentially novel organism. Assembled genomes (or parts of genomes) can then be compared with known genomes to infer the phylogeny of new organisms and identify related classes or traits. Thus, not only can this process identify unanticipated pathogens; it can even identify undiscovered organisms. Some early applications of sequencing on clinical samples have centered around the discovery of novel viruses, including such emerging pathogens as West Nile virus, SARS coronavirus, and the Middle East respiratory syndrome coronavirus (MERS-CoV) that has caused severe respiratory illnesses in healthy adults, as well as viral causes of myriad other conditions, from tropical hemorrhagic fevers to diarrhea in newborns.

As metagenomic sequencing and assembly techniques become more robust, this technology holds great promise for identifying microorganisms that are associated with clinical conditions of unknown etiology. Conventional methods already have unexpectedly linked numerous conditions with specific agents of infection—e.g., cervical and oropharyngeal cancers with human papillomavirus (HPV), Kaposi's sarcoma with human herpesvirus 8, and certain lymphomas with Epstein-Barr virus. Recently, Zika virus, first described in the 1940s, was found to be increasing in incidence as a cause of febrile syndromes, particularly in Central and South America. A concurrent increase in the incidence of microcephaly was noted that temporally and geographically matched the Zika epidemics. Zika was suspected to be neurotropic because of a previously recognized association with Guillain-Barré syndrome, but the strongest link between Zika virus and microcephaly came when the virus itself was detected by both quantitative reverse transcription PCR (RT-qPCR) and whole-genome sequencing in postmortem fetal brain tissue from microcephalic infants. An argument for causality was built on the foundation of epidemiologic evidence and direct viral detection, both of which were built on nucleic acid detection and genome sequencing. Sequencing techniques offer unprecedented sensitivity and specificity for identifying foreign nucleic acid sequences that may suggest other such pathogen-associated conditions—from malignancies to inflammatory conditions to unexplained fevers or other clinical syndromes—associated with organisms from viruses to bacteria to parasites. Caution is needed, though: in the absence of the ability to fulfill Koch's postulates, sequence-based identification of a microbe from patient specimens is not, on its own, sufficient to identify a novel pathogen. The increasing sensitivity of these methods warrants greater rigor and care in defining what is "noise" and what represents a pathogen.

As sequencing-based discovery expands, microbes may be found to be associated with conditions not classically thought of as infectious, such as the link between maternal Zika virus infection and fetal microcephaly. Studies of bowel flora in laboratory animals and even humans already suggest correlations between microbe composition and various aspects of metabolic and cardiovascular health. Improved methods for pathogen detection will continue to uncover unexpected correlations between microbes and disease states, but the mere presence of a microbe does not establish causality. Fortunately, once the relatively laborious and computationally intensive metagenomic sequencing and assembly efforts have identified a pathogen, further detection can more easily be undertaken with targeted methods such as PCR or hybridization, which may be more scalable and amenable to in situ confirmation. This capacity should facilitate the additional careful investigation that will be required to progress beyond correlation and to draw causal inference.

ANTIBIOTIC RESISTANCE

At present, antibiotic resistance in bacteria and fungi is conventionally determined by isolating a single colony from a cultured clinical specimen and testing its growth in the presence of drug. The requirement for multiple growth steps in these conventional assays has several consequences. First, only culturable pathogens can be readily processed. Second, this process requires considerable infrastructure to support the sterile environment needed for culture-based testing of diverse organisms. Finally, and perhaps most significantly, even the fastest-growing organisms require 1–2 days of processing for identification and 2–3 days for determination of susceptibilities. Some slow-growing organisms take even longer; for instance, weeks must pass before drug-resistant *M. tuberculosis* can be identified by growth phenotype. Given the clinical imperative in serious illness to begin effective therapy early, this inherent delay in susceptibility determination has obvious

implications for empirical antibiotic use: broad-spectrum antibiotics often must be chosen up front in situations where it is later shown that preferred narrower-spectrum drugs would have been effective or even that no antibiotics were appropriate (i.e., in viral infections). Even with this strategy, as resistant organisms become more common, the empirical choice can be incorrect, often with devastating consequences. Real-time identification of the infecting organism and information on its susceptibility profile would guide initial therapy and support judicious antibiotic use, ideally improving patient outcomes while aiding in the ever-escalating fight against antibiotic resistance by reserving the use of broad-spectrum agents for cases in which they are truly needed.

Molecular diagnostics and sequencing offer a way to accelerate detection of a pathogen's antibiotic susceptibility profile. If a genotype that confers resistance can be identified, this genotype can be targeted for molecular detection. In infectious disease, this approach has most convincingly come to fruition for HIV (Fig. 474-2*A*). (In a conceptually parallel application of genomic analysis, molecular detection of certain resistance determinants in cancers informs selection of targeted chemotherapy.) Extensive sequencing of HIV strains and correlations drawn between viral genotypes and phenotypic resistance have delineated the majority of mutations in key HIV genes, such as reverse transcriptase, protease, and integrase, that confer resistance to the antiretroviral agents that target these proteins. For instance, the single amino acid substitution K103N in the HIV reverse transcriptase gene predicts resistance to the first-line nonnucleoside reverse transcriptase inhibitor efavirenz, and its detection informs a clinician to choose a different agent. The effects of these common mutations on HIV susceptibility to various drugs—as well as on viral fitness—are curated in publicly available databases. Thus, genotypes are now routinely used to predict drug resistance in HIV, as phenotypic resistance assays are far more cumbersome than targeted sequencing. Indeed, the current recommendation in the United States is to sequence virus from a patient's blood before initiating antiretroviral therapy, which is then tailored to the predicted resistance phenotype. As new targeted therapies are introduced, this targeted sequencing–based approach to drug resistance will likely prove important in other viral infections, such as hepatitis C.

The challenge of predicting drug susceptibility from genotype is more daunting for bacteria than for HIV, yet considerable progress has been made toward sequencing-based determination of bacterial antibiotic susceptibility. Bacteria have far more complex genomes than viruses, with thousands of genes on their chromosomes (many of which can functionally interact in ways that escape a priori prediction) and the capacity to acquire many more through horizontal gene transfer of plasmids and mobile genetic elements within and between species. Thus, the task of comprehensively defining all possible genetic resistance mechanisms is orders of magnitude more complex in bacteria than in viruses, which typically have far more limited genomes. Despite these challenges, considerable progress has been made in recent years. In select cases where biological factors appear to have constrained the genotypic basis for resistance to a small, well-defined set of mutations, genotypic assays for antibiotic resistance are already being introduced into clinical practice. One important example is the detection of methicillin-resistant *Staphylococcus aureus* (MRSA). *S. aureus* is one of the most common and serious bacterial pathogens of humans, particularly in health care settings. Resistance to methicillin, the most effective class of antistaphylococcal antibiotics, has become very common, even in community-acquired strains. Vancomycin—the alternative drug to methicillin—is effective against MRSA but is measurably inferior to methicillin against methicillin-susceptible *S. aureus* (MSSA). Analysis of clinical MRSA isolates has demonstrated that the molecular basis for resistance to methicillin in essentially all cases stems from the expression of an alternative penicillin-binding protein (PBP2A) encoded by the gene *mecA*, which is found within a transferable genetic element called *mec*. This mobile cassette has spread rapidly through the *S. aureus* population via horizontal gene transfer and selection from widespread antibiotic use. Because methicillin resistance is essentially always due to the presence of the *mec* cassette, MRSA is particularly amenable to molecular detection. In recent years, a PCR test for the *mec* cassette, which saves hours to days compared with standard culture-based methods, has been approved by the U.S. Food and Drug Administration (FDA). Similar to MRSA, vancomycin-resistant enterococci (VRE) harbor one of a limited number of *van* genes found to be responsible for resistance to this important antibiotic, which occurs through alteration of the mechanism for cell wall cross-linking that vancomycin inhibits. Detection of one of these genes by PCR indicates resistance. More recently, identification of carbapenemase-encoding plasmids responsible for a significant fraction of carbapenem resistance (though not all instances) has led to multiplexed PCR assays to detect this important resistance element to this crucial antibiotic class. Finally, a PCR assay targeting the highly conserved RNA polymerase gene serves not only to detect *M. tuberculosis* directly in sputum samples but also to detect resistance to rifampin, since the determinants of resistance to this RNA polymerase inhibitor map almost exclusively to a short region of this gene. Since rifampin resistance is epidemiologically associated with, though not causal for, multidrug resistance, this assay identifies strains at high risk for multidrug resistance, enhancing its value.

Although identification and rapid detection of monogenic resistance determinants have improved, bacteria have tended to evolve multiple, diverse resistance mechanisms to most antibiotics; therefore, these tasks often require probing for and integration of multiple genetic lesions, targets, or mechanisms. For instance, at least five distinct modes of resistance to fluoroquinolones are known: reduced import, increased efflux, target site mutation, drug modification, and shielding of the target sites by expression of another protein. These mechanisms are typically present in combination in clinically resistant isolates; thus the problem of detecting genetic resistance is often a combinatorial one. In another clinically important example, while carbapenem resistance in Enterobacteriaceae is often explained by the presence of carbapenemases, resistance may also develop when other, less broad-spectrum β-lactamases are found in combination with porin mutations or efflux pumps. Thus, while multiplexed PCR assays for the most common carbapenemases (e.g., those encoded by the KPC, NDM, OXA-48, IMP-1, and VIM genes) have become a valuable tool for rapid identification of the subset of carbapenem-resistant Enterobacteriaceae in which resistance is caused by carbapenemases, their sensitivity is limited by their inability to detect other mechanisms of carbapenem resistance. To further complicate genetic prediction, changes in gene expression (which may be detectable through mutations in promoter regions or regulatory genes without coding mutations in known resistance determinants) and even gene copy number (which may occur without changes in primary sequence) of resistance determinants play critical roles in some cases of genetic resistance. Thus, while predicting resistance when determinants are found is rapidly becoming feasible, the more clinically relevant task of predicting *susceptibility* when no known resistance determinants are found remains more difficult.

To build on early successes with the goal of advancing beyond binary detection of monogenic resistance determinants, the ultimate frontier for genetic prediction of bacterial antibiotic resistance lies in more comprehensive prediction of a resistance phenotype from sequence information—a task similar to HIV resistance prediction. Yet there is no comprehensive compendium of genetic elements conferring resistance and their pairwise and higher-order interactions with each other and with the genetic background of bacterial pathogens. Nonviral genomes are much larger than viral ones, and their abundance and diversity are such that thousands of genetic differences often exist between clinical isolates of the same species, of which perhaps only one or a few may contribute to resistance. In addition, new mechanisms may emerge in the face of antibiotic deployment or with the release of new drugs, and genetic prediction of resistance will inevitably lag behind the emergence of unforeseen mechanisms. While confident prediction of bacterial antibiotic resistance from sequencing determinants may therefore seem daunting, the vast expansion of microbial sequencing capacity (Fig. 474-1), combined with analytic methods such as microbial genome-wide association studies and machine learning algorithms, offers powerful analytical approaches to this "needle in a haystack" problem and has permitted remarkable advances in the

predictive power of sequence determinants to date. Particularly in *M. tuberculosis*, where horizontal gene transfer is minimal and the pathogen is essentially restricted to human hosts, a remarkably wide array of phenotypic resistance can be explained by known genetic determinants. Even in more highly variable pathogens, with sequencing of sufficient numbers of susceptible and resistant pathogens, sequence-based prediction methods are improving in predictive accuracy, at least within the geographic region from which the test samples have been sequenced.

It is important to note that genotype-based analytical methods largely identify correlates, not necessarily surrogates or determinants, of resistance. In HIV diagnostics, surrogates (i.e., causal determinants of resistance) were found to be more reliable predictors than mere correlates in expanding sequencing-based resistance prediction to the general population. Without a mechanistic understanding of genetic resistance, a correlative relationship may be lineage-specific and less generalizable. Especially with multiple possible mechanisms of resistance to a given antibiotic and ongoing evolutionary pressure resulting in the development and acquisition of new modes of resistance, a genotypic approach to diagnosing antibiotic resistance is likely to remain challenging and to require ongoing vigilance in constantly correlating genotypic with more traditional phenotypic methods. An important corollary benefit of a genomic approach to resistance prediction, anchored in phenotypic validation, could be the systematic identification of outliers with unexplained resistance. These strains can form the basis for understanding newly emerging resistance mechanisms, which can in turn inform new drug development endeavors. Understanding resistance mechanisms may also help direct infection control efforts. For instance, the first identification of the *mcr-1* (mobilized colistin resistance) gene on a plasmid, together with other antibiotic resistance determinants, heightened concern about colistin-resistant Enterobacteriaceae identified first in China and later elsewhere because it implied rapid transmissibility of multidrug resistance. Early recognition of these potentially dangerous strains elucidated the immediate need for strict containment protocols.

In parallel with advancing sequencing technologies, progress in computational techniques, bioinformatics and statistics, and data storage as well as experimental confirmatory testing of hypotheses will be needed to advance toward the ambitious goal of a comprehensive compendium of global antibiotic resistance determinants. Open sharing and careful curation of new sequence information will be of paramount importance, as will iterative or even continuous comparison of predictions with ongoing phenotypic testing in order to assess performance and allow prediction algorithms to keep up with newly evolving or emerging resistance mechanisms.

We continuously observe the accumulation of new or unanticipated modes of resistance from ongoing evolutionary pressure caused by the widespread clinical use of antibiotics. Even with MRSA, perhaps the best-studied case of antibiotic resistance and a model of relative simplicity with a single known monogenic resistance determinant (*mecA*), a genotype-based approach to resistance detection proved imperfect. One limitation was a recall of the initial commercial genotypic resistance assay that was deployed for the identification of MRSA. A clinical isolate of *S. aureus* that emerged in Belgium expressed a variant of the *mec* cassette not detected by the assay's PCR primers. New primers were added to detect this new variant, and the assay was re-approved for use. This example illustrates the need for ongoing monitoring of any genotypic resistance assay. A second limitation is that a contradiction can occur between genotypic and phenotypic evidence for resistance. Up to 5% of MSSA strains have been reported to carry a copy of the *mecA* gene that is either nonfunctional or not expressed. Thus, the erroneous identification of these strains as MRSA by genotypic detection would lead to administration of the inferior antibiotic vancomycin rather than the preferred β-lactam therapy.

These examples illustrate one of the prime challenges of moving beyond growth-based assays: genotype is merely a proxy for the resistance phenotype that directly informs patient care. Alternative approaches currently under development attempt to circumvent the limitations of genotypic resistance testing by returning to phenotypic assays, albeit more rapid ones. One such approach is informed by genomic methods: transcriptional profiles serve as a rapid phenotypic signature for antibiotic response. Conceptually, since dying cells are transcriptionally distinct from cells fated to survive, susceptible bacteria enact different transcriptional profiles after antibiotic exposure than resistant ones, independent of the mechanism of resistance. These differences can be measured and, since transcription is one of the most rapid responses to cell stress (minutes to hours), can be used to determine whether cells are resistant or susceptible much more rapidly than is possible if growth in the presence of antibiotics is awaited (days). Like DNA, RNA can be readily detected through predictable rules governing base pairing via either amplification or hybridization-based methods. Changes in a carefully selected set of transcripts form an expression signature that can represent the total cellular response to antibiotic without requiring full characterization of the entire transcriptome. Preliminary proof-of-concept studies suggest that this approach may identify antibiotic susceptibility on the basis of transcriptional phenotype much more quickly than is possible with growth-based assays.

Because of its sensitivity in detecting even very rare nucleic acid fragments, sequencing provides an unprecedented depth of study into complex populations of cells and tissues. The strength of this depth and sensitivity applies not only to the detection of rare, novel pathogens in a sea of host signal, but also to the identification of heterogeneous pathogen subpopulations in a single host that may differ, for example, in drug resistance profiles or pathogenesis determinants. For instance, recent studies have highlighted the diversification of pathogens in chronic bacterial infections, such as *Pseudomonas* in the lungs of patients with cystic fibrosis or *M. tuberculosis* in disseminated infection, perhaps allowing for niche specialization within the host. Such diversification has long been recognized in chronic viral populations, as exemplified by HIV. Future studies will be needed to elucidate the clinical significance of these variable subpopulations, even as deep sequencing is now providing unprecedented levels of detail about majority and minority members of this population.

HOST-BASED DIAGNOSTICS

While pathogen-based diagnostics continue to be the mainstay for confirming infection, serologic testing and nonspecific biomarkers—such as erythrocyte sedimentation rate, C-reactive protein level, and even total white blood cell and neutrophil counts—have long been the basis of a strategy for measuring host responses to aid in the diagnosis of infection. Even recently identified host biomarkers of bacterial infection, such as procalcitonin, have fallen short in their versatility, with positive and negative predictive values that are thus far adequate for only a few narrow applications but inadequate for generalized clinical use. Here, too, the application of genomics is now being explored to improve upon this approach, given the previously described limitations of serologic testing and the lack of specificity of protein biomarkers identified to date. Rather than using antibody responses as a retrospective biomarker for infection, recent efforts have focused on transcriptomic analysis of the host response as a new direction with diagnostic implications for human disease. For instance, while pathogen-based diagnostic tests to distinguish active from latent tuberculosis infection have proven elusive, recent work shows that the transcriptional profile of circulating white blood cells exhibits a differential pattern of expression of nearly 400 transcripts that distinguish active from latent tuberculosis; this expression pattern is driven in part by changes in interferon-inducible genes in the myeloid lineage. In a validation cohort, this transcriptional signature was able to distinguish patients with active versus latent disease, to distinguish tuberculosis infection from other pulmonary inflammatory states or infections, and to track responses to treatment in as little as 2 weeks, with normalization of expression toward that of patients without active disease over 6 months of effective therapy. Such a test could play an important role not only in the management of patients but also as a marker of efficacy in clinical trials of new therapeutic agents. Similarly, considerable progress has been made toward identifying host transcriptional signatures in circulating blood cells that distinguish viral from bacterial causes

of upper respiratory infection, with better performance characteristics than current clinical parameters or available protein biomarkers. Additional host signatures have been reported that distinguish among bacterial infection, viral infection, and inflammatory states; identify Lyme disease; identify influenza; and even distinguish between gram-positive and gram-negative bacterial infections. In some cases, results have been extended to different host populations—including adults and children, and those with varying immune function—which obviously will be critical for generalizing such an approach. Thus, profiling of host transcriptional dynamics could augment the information obtained from studies of pathogens, both enhancing diagnosis and monitoring the progression of illness and the response to therapy.

In this era of genome-wide association studies and attempts to move toward personalized medicine, genomic approaches are also being applied to the identification of host genetic loci and factors that contribute to infection susceptibility. Such loci will have undergone strong selection among populations in which the disease is endemic. Through identification of the beneficial genetic alleles among individuals who survive in such settings, markers for susceptibility or resistance are being discovered; these markers can be translated to diagnostic tests to identify susceptible individuals in order to implement preventive or prophylactic interventions. Further, such studies may offer mechanistic insight into the pathogenesis of infection and inform new methods of therapeutic intervention. Such beneficial genetic associations were recognized long before the advent of genomics, as in the protective effects of the negative Duffy blood group or heterozygous hemoglobin abnormalities against *Plasmodium* infection. Genomic approaches allow more systematic and widespread application of this principle to identify not only people with increased susceptibility to prevalent diseases (e.g., HIV infection, tuberculosis, and cholera) but also host factors that contribute to and thus might predict the severity of disease.

THERAPEUTICS

Genomics has the potential to impact infectious disease therapeutics in two ways. By transforming the speed or type of diagnostic information that can be attained, it can influence therapeutic decision-making. Alternatively, by opening new avenues to a better understanding of pathogenesis, providing new ways to disrupt infection, and delineating new approaches to antibiotic discovery, it has the potential to facilitate the development of new therapeutic agents.

GENOMIC DIAGNOSTICS INFORMING THERAPEUTICS

Efforts at antibiotic discovery are declining, with few new agents in the pipeline and even fewer new drugs (in particular, few agents with new mechanisms of action) entering the market. This phenomenon is due in part to the lack of economic incentives for the private sector; however, it is also attributable in part to the enormous challenges involved in the discovery and development of antibiotics. Most recent efforts have focused on broad-spectrum antibiotics; the development of a chemical entity that works across an extremely diverse set of organisms (i.e., species more divergent from each other than a human is from an amoeba) is far more challenging than the development of an agent that is designed to target a single bacterial species. Nevertheless, the concept of narrow-spectrum antibiotics has heretofore been rejected because of the lack of early diagnostic information that would guide the selection of such agents. Thus, rapid diagnostics providing antibiotic susceptibility information that can guide antibiotic selection in real time has the potential to alter and simplify antibiotic strategies by allowing a paradigm shift away from broad-spectrum drugs and toward narrow-spectrum agents. Such a paradigm shift clearly would have additional implications for antibiotic resistance, helping to limit selective pressure applied to pathogens and commensal bacteria during therapy.

In yet another diagnostic paradigm with the potential to impact therapeutic interventions, genomics is opening new avenues to a better understanding not only of different host susceptibilities to infection but also of different host responses to therapy. For example, the role of glucocorticoids in tuberculous meningitis has long been debated. Recently, polymorphisms in the human genetic locus *LTA4H*, which encodes a leukotriene-modifying enzyme, were found to modulate the inflammatory response to tuberculosis. Patients with tuberculous meningitis who were homozygous for the proinflammatory *LTA4H* allele were most helped by adjunctive glucocorticoid treatment, while those who were homozygous for the anti-inflammatory allele were negatively affected by steroid treatment. Steroids have become part of the standard of care in tuberculous meningitis, but this study suggests that perhaps only a subset of patients benefit from this anti-inflammatory adjunct (while others may be harmed) and further suggests a genetic means of prospectively identifying this subset. Thus, genomic diagnostic tests may eventually approach the goal of personalized medicine, informing diagnosis, prognosis, and treatment decisions by revealing the pathogenic potential of the microbe and by detecting individualized host responses to both infection and therapy.

GENOMICS IN DRUG AND VACCINE DEVELOPMENT

Genomic technologies are dramatically changing research on host–pathogen interactions, with a goal of increasingly influencing the process of therapeutic discovery and development. Sequencing offers several possible avenues into antimicrobial therapeutic discovery. First, genome-scale molecular methods have paved the way for comprehensive identification of all essential genes encoded by a pathogen, thereby systematically identifying critical vulnerabilities within a pathogen that could be targeted therapeutically. Second, genome-scale methodologies offer rapid ways to address the mechanism of action of newly identified hits from compound screens. Whole-genome sequencing offers a rapid, unbiased way to detect mutations arising in resistant mutants during selection. Similarly, transcriptional profiling can provide insights into mechanisms of action of new candidate drugs. For instance, the transcriptional signature of cell wall disruptors (e.g., β-lactams) is distinct from that of DNA-damaging agents (e.g., fluoroquinolones) or protein synthesis inhibitors (e.g., aminoglycosides). Either approach can thus suggest a mechanism of action or flag compounds for prioritization because of a potentially novel activity. In an alternative genomic strategy for determining mechanisms of action, an RNA interference approach followed by targeted sequencing was used to identify genes required for antitrypanosomal drug efficacy. This approach provided new insights into the mechanism of action of drugs that have been in use for decades for human African trypanosomiasis. Third, sequencing can readily identify the most conserved regions of a pathogen's genomes and corresponding gene products; this information is invaluable in narrowing antigen candidates in vaccine development. These surface proteins can be expressed recombinantly and tested for the ability to elicit a serologic response and protective immunity. This process, termed *reverse vaccinology*, has proved particularly useful for pathogens that are difficult to culture or poorly immunogenic.

Genomics has been employed in both developing vaccines and defining their impact on microbial epidemiology and ecology. Examples include recent studies of influenza, malaria, *S. pneumoniae*, and HPV following vaccine introduction. Extensive sequencing of influenza viruses has been valuable in understanding the modest efficacy of seasonal influenza vaccination, and the combination of genomics and antigenic cartography is proving helpful in the selection of strains to include in subsequent influenza vaccines. The new RTS,S/AS01 malaria vaccine was analyzed by targeted sequencing of parasites from vaccinated and control populations during a phase 3 trial conducted at 11 sites in Africa; these analyses revealed reduced vaccine efficacy against parasites with amino acid mutations in the circumsporozoite protein targeted by the vaccine. Similarly, studies of the more established pneumococcal vaccine (the 7-valent polysaccharide conjugate vaccine, PCV-7) documented serotype replacement: strains targeted by the vaccine have dramatically decreased in prevalence following widespread vaccination campaigns. Given that specific serotypes of HPV (e.g., types 16 and 18) clearly are more strongly associated than others with carcinogenesis, HPV vaccines have capitalized on serotype replacement, targeting vaccine strains to specifically prevent infection with the more dangerous serotypes. Such a strategy, informed by pathogen genomics, aims to protect individuals and ideally to decrease the circulating burden of more virulent strains within society.

Large-scale gene content analysis from sequencing or expression profiling enables new research directions that provide novel insights into the interplay of pathogen and host during infection or colonization. One important goal of such research is to suggest new therapeutic approaches to disrupt this interaction in favor of the host. Indeed, one of the most immediate applications of next-generation sequencing technology has come from simply characterizing human pathogens and related commensal or environmental strains and then finding genomic correlates for pathogenicity. For instance, as *Escherichia coli* varies from a simple nonpathogenic, lab-adapted strain (K-12) to a Shiga toxin–producing enterohemorrhagic gastrointestinal pathogen (O157:H7), it displays up to a 25% difference in gene content, though it is classified as the same species. Similarly, some isolates of *Enterococcus*—a genus notorious for its increasing incidence of resistance to common antibiotics such as ampicillin, vancomycin, and aminoglycosides—also contain recently acquired genetic material comprising up to 25% of the genome on mobile genetic elements. This fact suggests that horizontal gene transfer plays an important role in the organisms' adaptation as nosocomial pathogens. On closer study, this genome expansion is associated with loss of regulatory elements called CRISPRs (*c*lustered, *r*egularly *i*nterspaced *s*hort *p*alindromic *r*epeats). Loss of CRISPR elements, which protect the bacterial genome from invasion by certain foreign genetic material, may thus facilitate the acquisition of antibiotic resistance–conferring genetic elements. While loss of this regulation appears to impose a competitive disadvantage in antibiotic-free environments, these drug-resistant strains thrive in the presence of even some of the best antienterococcal therapies. In addition to insights gained from genome sequencing, extension of unbiased whole-transcriptome sequencing (RNA-Seq) efforts to bacteria is beginning to identify unexpected regulatory, noncoding RNAs in many diverse species. While the functional implications of these new transcripts are as yet largely unknown, the presence of such features—conserved across many bacterial species—implies evolutionary importance and suggests areas for future study and possible new therapeutic avenues. Transcriptomic and proteomic profiling of pathogens under various conditions that mimic colonization or infection, including existence as biofilms or in polymicrobial communities, intracellular infection models, antibiotic exposure, and nutrient starvation, has begun to reveal novel biological features that may be targeted by the next generation of therapies. At the cutting edge of the host–pathogen interface, single-cell transcriptomic methodologies are rapidly increasing in feasibility and extent, revealing previously unknown heterogeneity in the potential outcomes of intracellular infection.

Thus, genomic studies are transforming our understanding of infection, offering evidence of virulence factors or toxins and providing insight into ongoing evolution of pathogenicity and drug resistance. One goal of such studies is to identify therapeutic agents that can disrupt the pathogenic process. There is currently much interest in the theoretical concept of antivirulence drugs that inhibit virulence factors rather than killing the pathogen outright as a means to intervene in infection. Further, with sequencing ever more accessible and efficient, ongoing large-scale studies have unprecedented statistical power to associate clinical outcomes with pathogen and host genotypes and thus to further reveal vulnerabilities in the infection process that can be targeted for disruption. Although this is just the beginning, such studies point to a tantalizing future in which the clinician is armed with genomic predictors of infection outcome and therapeutic response to guide clinical decision-making.

EPIDEMIOLOGY OF INFECTIOUS DISEASES

Epidemiologic studies of infectious diseases have several main goals: to identify and characterize outbreaks, to describe the pattern and dynamics of an infectious disease as it spreads through populations, and to identify interventions that can limit or reduce the burden of disease. One classic, paradigmatic example is John Snow's elucidation of the origin of the 1854 London cholera outbreak. Snow used careful geographic mapping of cases to determine that the likely source of the outbreak was contaminated water from the Broad Street pump, and, by removing the pump handle, he aborted the outbreak. Whereas that effort was undertaken without knowledge of the causative agent of cholera, advances in microbiology and genomics have expanded the purview of epidemiology to consider not just the disease but also the pathogen, its virulence factors, and the complex relationships between microbial and host populations.

Through use of genomic tools such as high-throughput sequencing, the diversity of a microbial population can be rapidly described with unprecedented resolution, with discrimination between isolates that have single-nucleotide differences across the entire genome and advancement beyond prior approaches that relied on phenotypes (such as antibiotic susceptibility profiles) or genetic markers (such as multilocus sequence typing). The development of statistical methods grounded in molecular genetics and evolutionary theory has established analytical approaches that translate descriptions of microbial population diversity and structure into descriptions of the origin and history of pathogen spread. By linking phylogenetic reconstruction with epidemiologic and demographic data, genomic epidemiology presents the opportunity to track transmission from person to person and across demographic and geographic boundaries, to infer transmission patterns of both pathogens and sequence elements that confer phenotypes of interest, and to estimate the transmission dynamics of outbreaks.

TRANSMISSION NETWORKS

Whole-genome sequencing of pathogen genomes can be used to infer transmission and identify point-source outbreaks. As reported in a seminal paper in 2010, a study of MRSA in a Thai hospital demonstrated the use of whole-genome sequencing in reconstructing the transmission of a pathogen from patient to patient by integrating the analysis of accumulation of mutations over time with the dates and hospital locations of the infected individuals. Since then, multiple instances of the use of whole-genome sequencing to define and motivate interventions aimed at interrupting transmission chains have been reported. In another MRSA outbreak in a special-care baby unit in Cambridge, United Kingdom, whole-genome sequencing extended the traditional infection control analysis, which relies on typing organisms by their antibiotic susceptibilities, to sequencing of isolates from clinical samples. This approach identified an otherwise unrecognized outbreak of a specific MRSA strain that was occurring against a background of the usual pattern of infection caused by a diverse circulating population of MRSA strains. The analysis showed evidence of transmission among mothers within the special-care baby unit and in the community and demonstrated the key role of MRSA carriage in a single health care provider in the persistence of the outbreak. In yet another example, in response to the observation of 18 cases of infection by carbapenemase-producing *Klebsiella pneumoniae* over 6 months at the National Institutes of Health Clinical Research Center, genome sequencing of the isolates was used to discriminate between the possibilities that these cases represented multiple, independent introductions into the health care system or a single introduction with subsequent transmission. On the basis of network and phylogenetic analysis of genomic and epidemiologic data, the authors reconstructed the likely relationships among the isolates from patient to patient, demonstrating that the spread of resistant *Klebsiella* infection was in fact due to nosocomial transmission of a single strain. Similar approaches have elucidated the extent to which presumed nosocomial *C. difficile*, VRE, and carbapenem-resistant Enterobacteriaceae represent within-hospital transmission rather than independent acquisitions. With these demonstrations of the potential contribution of genomics to hospital infection-control efforts, an important avenue of research seeks to develop statistical methods with which to ascertain when such tools are useful and their cost-effectiveness when compared with that of current nongenomic approaches.

The uncovering of unexpected transmission events by genomic epidemiology studies is motivating investigations into pathogen ecology and modes of transmission. For example, the rise in prevalence of infections with nontuberculous mycobacteria, including *Mycobacterium abscessus*, among patients with cystic fibrosis has led to speculation about the possible role of patient-to-patient transmission in the cystic fibrosis community; however, conventional typing

approaches have lacked the resolution to define pathogen population structure accurately, a critical component of inferring transmission. Past infection-control guidelines discounted the possibility of acquisition of nontuberculous mycobacteria in health care settings, as no strong evidence for such transmission had been described. In whole-genome sequencing studies of *M. abscessus* isolates from patients with cystic fibrosis, an analytical approach using genome sequencing, epidemiology, and Bayesian modeling revealed that, contrary to the prior belief that infections with *M. abscessus* are independently acquired, the majority of infections appear to be transmitted. Because there are often no clear epidemiologic links that place the infected patients in the same place at the same time, this finding highlights a need to explore preexisting notions of circumstances required for transmission, including the roles of fomites and aerosols, and a reconsideration of *M. abscessus* infection-control guidelines. In a clear example of the utility of whole-genome sequencing for revealing unexpected transmission networks, isolates of *M. chimaera* causing infections after cardiothoracic surgery in patients in different locations were all found to be closely related. These isolates differed from one another by at most 38 pairwise single-nucleotide polymorphisms out of >5 million bases; in contrast, they differed by >2900 single-nucleotide polymorphisms from the nonclonally related reference isolate. Although a hospital source was initially suspected when the first of these cases were identified, this whole-genome sequencing analysis strongly supported a single point-source for these geographically dispersed isolates. A subsequent investigation ultimately implicated *M. chimaera* contamination in the manufacturing chain of a temperature-control system used during cardiac bypass. Similar studies of other pathogens—particularly those that share human, other animal host, and environmental reservoirs—will continue to advance our understanding of the relative roles and prominence of sources of infection and the modes of spread through populations, thereby establishing evidence-based strategies for prevention and intervention.

As more studies aim to carefully define the origins and spread of infectious agents using the high-resolution lens of whole-genome sequencing, fundamental questions arise about the diversity of infecting and colonizing microbial populations. Traditional microbiologic methods include taking a single colony from a growth plate as representative of the population. However, the more diverse the colonizing or infecting pathogen population, the less representative these individual isolates are and the greater the possibility for introducing error into whole-genome sequencing–based methods while reconstructing transmission. Sequencing studies of multiple colonies of an *S. aureus* strain colonizing a single individual showed a "cloud" of diversity. What is the clinical significance of this diversity? What are the processes that generate and limit it? What amount of diversity is transmitted under different conditions and routes of transmission? How do the answers to these questions vary by infectious organism, type of infection, host, and response to treatment? More comprehensive descriptions of diversity, population dynamics, transmission bottlenecks, and the forces that shape and influence the growth and spread of microbial populations will be a critically important focus of future investigations.

ORIGINS AND DYNAMICS OF PATHOGEN SPREAD

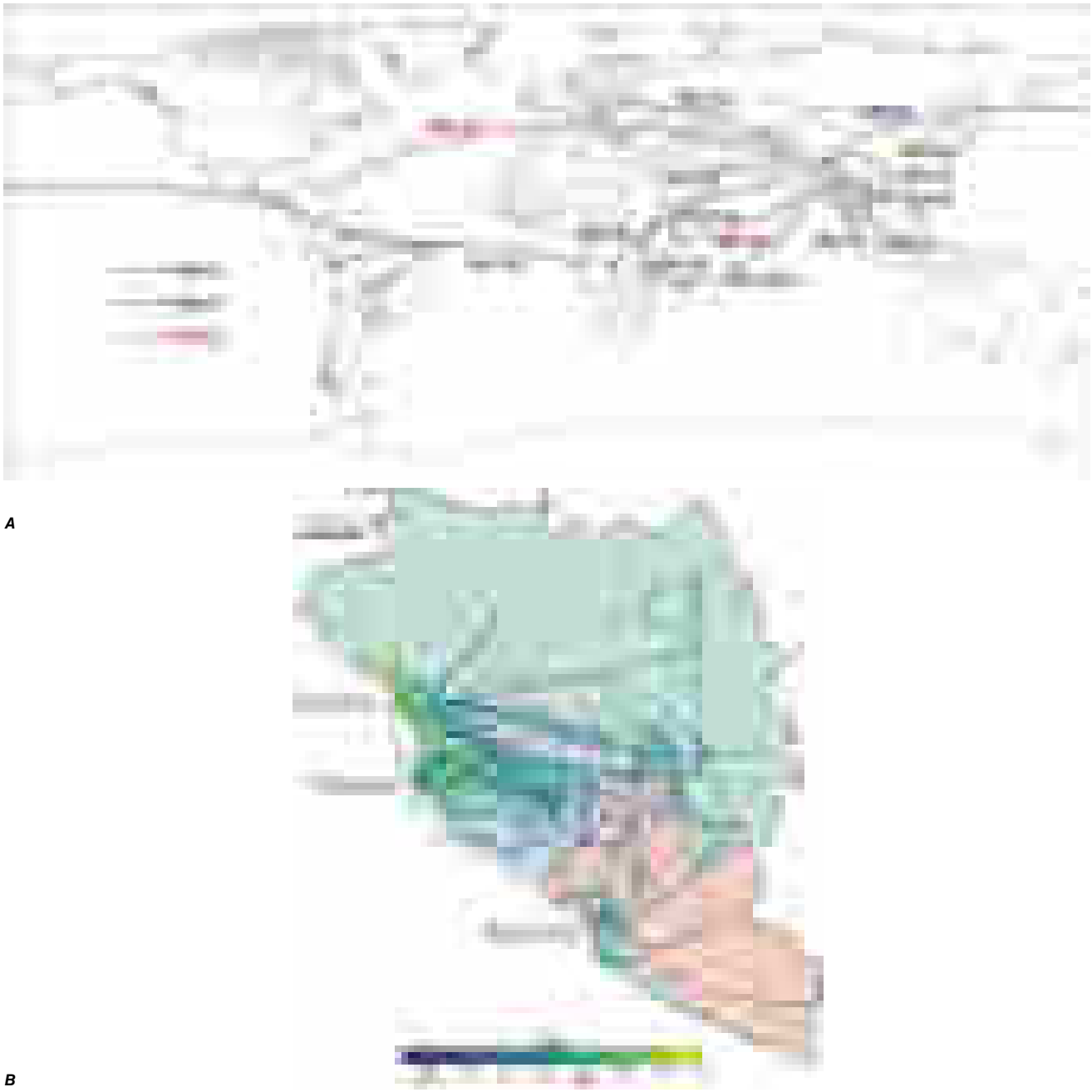
In addition to reconstructing the transmission chains of local outbreaks, genomics-based epidemiologic methods reveal broad-scale geographic and temporal spread of pathogens. Three recent examples include the origins of cholera in Haiti, the history of HIV-1 group M, and the spread of Ebola in West Africa. Cholera, a dehydrating diarrheal illness caused by infection with *Vibrio cholerae*, first spread worldwide from the Indian subcontinent in the 1800s and has since caused seven pandemics; the seventh pandemic has been ongoing since the 1960s. An investigation into the geographic patterns of cholera spread in the seventh pandemic used genome sequences from a global collection of 154 *V. cholerae* strains representing isolates from 1957 to 2010. This investigation revealed that the seventh pandemic has comprised at least three overlapping waves spreading out from the Indian subcontinent **(Fig. 474-4*A*)**. Further, analysis of the genome of an isolate of *V. cholerae* from the 2010 outbreak of cholera in Haiti showed it to be more closely related to isolates from South Asia than to isolates from neighboring Latin America, supporting the hypothesis that the outbreak was derived from *V. cholerae* introduced into Haiti by human travel (likely from Nepal) rather than by environmental or more geographically proximal sources. A subsequent study that dated the time to the most recent common ancestor of a population of *V. cholerae* isolates from Haiti provided further support for a single point-source introduction from Nepal. Application of similar methods that integrate pathogen genome sequences, mutation rates, geographic locations, and phylogenetic inference to HIV-1 group M dated the origin of the virus to the 1920s and the city of Kinshasa (then called Leopoldville), the capital of the Democratic Republic of the Congo (then called the Belgian Congo). This work established an understanding of how a boom in industry and a city with extensive railroad connections provide a scaffolding along which a virus can rapidly spread geographically.

More recently, genome sequencing has proven invaluable in understanding the geographic, demographic, climatic, and administrative factors that drove, sustained, and limited the 2013–2016 Ebola outbreak that ravaged West Africa **(Fig. 474-4*B*)** as well as the factors and patterns of transmission of Zika virus in the Americas. These efforts illustrate the remarkable promise of genome sequencing in improving outbreak response strategy by elucidating previously hidden paths of disease spread and details of the forces that shape epidemics. The combination of in-the-field sequencing with portable sequencing platforms, rapid data sharing, and rapid open analysis through sites such as nextstrain.org offers a paradigm by which real-time genomic epidemiology may contribute to "weather maps," enabling prediction of epidemic patterns and thus providing guidance for public health interventions to slow or control their spread.

Increasing numbers of investigations into the spread of many pathogens are contributing to a growing atlas of maps describing routes, patterns, and tempos of microbial diversification and dissemination, not just for agents of emerging infectious diseases but for common pathogens as well. Such studies will create a vast amount of data that can be used to investigate the diversity and microbiologic links within distinct niches and the patterns of spread from one niche to another. The increasingly broad adoption of genome sequencing by health care and public health institutions will ensure that the available catalog of genome sequences and associated epidemiologic data will grow very rapidly. For example, updating from the pulsed-field gel electrophoresis techniques that have been used to define strains of food-borne pathogens since the late 1980s, PulseNet—the U.S. Centers for Disease Control and Prevention network for monitoring these pathogens—is instituting routine genome sequencing. With higher-resolution description of microbial diversity and of the dynamics of that diversity over time and across epidemiologic and demographic boundaries and evolutionary niches, we will gain even greater insights into the relationships of transmission routes and patterns of historical spread.

EPIDEMIC POTENTIAL

Defining pathogen transmissibility is a critical step in the development of public health surveillance and intervention strategies as this information can help to predict the epidemic potential of an outbreak. Transmissibility can be estimated by a variety of methods, including inference from the growth rate of an epidemic and the generation time of an infection (the mean interval between infection of an index case and infection of the people infected by that index case). Genome sequencing and analysis of a well-sampled population provide another method by which to derive similar fundamental epidemiologic parameters. One key measure of transmissibility is the basic reproduction number, defined as the number of secondary infections generated from a single primary infectious case. When the basic reproduction number is >1, an outbreak has epidemic potential; when it is <1, the outbreak will become extinct. On the basis of sequences from influenza virus samples obtained from infected patients very early in the 2009 H1N1 influenza pandemic, the basic reproduction number was estimated through a population genomic analysis at 1.2; this result provided greater confidence to estimates derived by traditional epidemiologic



A

B

FIGURE 474-4 ***A.*** Transmission events inferred from phylogenetic reconstruction of 154 *Vibrio cholerae* isolates from the seventh cholera pandemic. Date ranges represent estimated time to the most recent common ancestor for strains transmitted from source to destination locations, based on a Bayesian model of the phylogeny. *(Reprinted with permission from A Mutreja et al: Evidence for several waves of global transmission in the seventh cholera pandemic. Nature 477:462, 2011.)* ***B.*** Inferred Ebola virus spread in West Africa (Liberia, *red*; Guinea, *green*; and Sierra Leone, *blue*) by phylogeographic methods using virus genome sequences, dates, and an evolutionary model. The lines reflect spread between population centroids of each administrative region, going from the thin end to the thick end and colored by a time scale. *(Reprinted with permission from G Dudas et al: Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature 544:309, 2017.)*

data, which ranged from 1.4 to 1.6. In addition, with the assumption of a molecular clock model, sequences of H1N1 samples together with information about when and where the samples were obtained have been used to estimate the date and location of the pandemic's origin, providing insight into disease origins and dynamics. Because the magnitude and intensity of the public health response are guided by the predicted size of an outbreak, the ability of genomic methods to cast light on a pathogen's origin and epidemic potential adds an important dimension to the contributions of these methods to infectious disease epidemiology.

PATHOGEN EVOLUTION

Beyond describing transmission and dynamics, pathogen genomics can provide insight into the evolution of pathogens and the interactions of selective pressures, the host, and pathogen populations, which can have implications for clinical decision-making and the development of vaccines and therapeutics. From a clinical perspective, this process is central to the acquisition of antibiotic resistance, the generation of increasing pathogenicity or new virulence traits, the evasion of host immunity and clearance (leading to chronic infection), and vaccine efficacy.

Microbial genomes evolve through a variety of mechanisms, including mutation, duplication, insertion, deletion, recombination, and horizontal gene transfer. Segmented viruses (e.g., influenza virus) can reassort gene segments within multiply infected cells. The pandemic 2009 H1N1 influenza A virus, for example, appears to have been generated through reassortment of several avian, swine, and human influenza strains. Such potential for the evolution of novel pandemic strains has precipitated concern about the possible evolution to transmissibility of virulent strains that have been associated with high mortality rates but have not yet exhibited efficient human infectivity.

Experiments with H5N1 avian influenza, for example, have defined five mutations that render it transmissible, at least in ferrets—the animal model system for human influenza. Studies that examine the genomes of pathogens collected longitudinally from individual infections have similarly demonstrated the evolution of bacteria as they adapt from colonization to invasion and to new host environments and new immune and therapeutic pressures.

The continuous antigenic evolution of seasonal influenza offers an example of how studies of pathogen evolution can impact surveillance and vaccine development. Frequent updates to the annual influenza vaccine are needed to ensure protection against the dominant strains. These updates are based on anticipating which viral populations from a pool of substantial locally and globally diverse circulating viruses will predominate in the upcoming season. Toward that end, sequencing-based studies of influenza virus dynamics have shed light on the global spread of influenza, providing concrete data on patterns of spread and helping to elucidate the origins, emergence, and circulation of novel strains. Through analysis of more than 1000 influenza A H3N2 virus isolates over the 2002–2007 influenza seasons, Southeast Asia was identified as the usual site from which diversity originates and spreads worldwide. Further studies of global isolate collections have shed further light on the diversity of circulating virus, showing that some strains persist and circulate outside of Asia for multiple seasons.

Not only do genomic epidemiology studies have the potential to help guide vaccine selection and development; they are also helping to track what happens to pathogens circulating in the population in response to vaccination. By describing pathogen evolution under the selective pressure of a vaccinated population, such studies can play a key role in surveillance and identification of virulence determinants and perhaps may even help to predict the future evolution of escape from vaccine protection. The seven-valent pneumococcal conjugate vaccine (PCV-7) targeted the seven serotypes of *S. pneumoniae* responsible for the majority of invasive disease at the time of its introduction in 2000; since then, PCV-7 has dramatically reduced the incidence of pneumococcal disease and mortality. However, sequencing of >600 Massachusetts pneumococcal isolates from 2001 to 2007 has shown that, in the pneumococcal population, previously rare nonvaccine serotypes are replacing vaccine serotypes and that some vaccine strains have persisted despite vaccination by recombining the vaccine-targeted capsule locus with a cassette of capsule genes from non-vaccine-targeted serotypes.

The large collections of pathogen genome sequences are driving development of tools to decipher the genetic basis for antibiotic resistance, virulence, and infection risk. Some pathogens have distinct types of clinical manifestations, the basis for which we are just beginning to unravel with the aid of genomics. For example, *Listeria* is a food-borne pathogen that can cause both central nervous system infections and maternal/neonatal infections. Although all *Listeria* isolates are treated the same from a public health perspective, variation in outcomes exists and appears to be linked to the strains' genomic background. Molecular analysis of a national reference laboratory's collections of well-characterized specimens, based on the fraction of immunocompetent people in which they caused disease, revealed that some clonal complexes of *Listeria* appear to be more virulent than others. Linking epidemiology and comparative genomics then enabled enumeration of putative virulence factors that contribute to the clinical phenotypes as well as identification and confirmation of a novel gene cluster that mediates central nervous system tropism. This approach illustrates progress toward a future in which we can link pathogen identification with risk, thereby informing resource use and allocation.

GLOBAL CONSIDERATIONS

While cutting-edge genomic technologies are largely implemented in the developed world, their application to infectious diseases perhaps offers the biggest potential impact in less developed regions where the burden of these infections is greatest. This globalization of genomic technology and its extensions has already begun in each of the areas of focus highlighted in this chapter; it has occurred both through the application of advanced technologies to samples collected in the developing world and through the adaptation and importation of technologies directly to the developing world for on-site implementation as they become more globally accessible.

Genomic characterization of the pathogens responsible for such important global illnesses such as tuberculosis, malaria, trypanosomiasis, and cholera has led to insights in diagnosis, treatment, and infection control. For instance, with the increasing burden of drug-resistant tuberculosis in the developing world, a molecular diagnostic test has been developed to detect rifampin-resistant tuberculosis. The genetic basis for rifampin resistance has been well defined by targeted sequencing: characteristic mutations in the molecular target of rifampin, RNA polymerase, account for the vast majority of instances of rifampin resistance. At least in areas that can afford to implement it, a rapid, automated PCR assay that can detect both *M. tuberculosis* and a rifampin-resistant allele of RNA polymerase directly in clinical samples has been implemented in parts of Africa and Asia, transforming the recognition and management of incident tuberculosis and multidrug resistance where they are most prevalent. Since rifampin resistance frequently accompanies resistance to other antibiotics, this test can suggest the presence of multidrug-resistant *M. tuberculosis* within hours instead of weeks, without the infrastructure required for culture.

High-resolution genomic tracking of the spread of epidemics—from cholera to Ebola to Zika—has yielded insights into which public health measures may prove most effective in controlling local epidemics. Many genomic tracking efforts have involved close collaborations with local scientists and public health officials, and considerable investment in sequencing infrastructure in sub-Saharan Africa has made on-location epidemic tracking in the event of another such outbreak feasible. Such investment can not only enable real-time outbreak recognition and tracking but also provide the infrastructure needed to capitalize on the many other benefits of high-throughput sequencing as they are developed. Overall, sequencing efforts have become cheaper and have moved closer to point-of-care with each passing year. As these technologies synergize with efforts to globalize information-technology resources, global implementation of genomic methods promises to spread state-of-the-art methods for diagnosis, treatment, and epidemic tracking of infections to areas that need these capabilities the most.

SUMMARY

By illuminating the genetic information that encodes the most fundamental processes of life, genomic technologies are transforming many aspects of medicine. In infectious diseases, methods such as next-generation sequencing and genome-scale expression analysis offer information of unprecedented depth about individual microbes as well as microbial communities. This information is expanding our understanding of the interactions of microorganisms with each other, their human hosts, and the environment. Despite technological and financial barriers that continue to slow the widespread adoption of large-scale pathogen sequencing in clinical and public health settings, genomic methodologies have utterly transformed the research landscape in infectious disease and are beginning to make meaningful inroads into clinical settings. As even vaster amounts of data are generated, innovations in data storage, development of bioinformatics tools to manipulate the data, standardization of methods, and training of end-users in both the research and clinical realms will be required. The cost-effectiveness and applicability of whole-genome sequencing, particularly in the clinic, remain to be studied, and studies of the impact of genome sequencing on patient outcomes will be needed to clarify the contexts in which these new methodologies can make the greatest contributions to patient well-being. The ongoing efforts to overcome limitations through collaboration, teaching, and reduction of financial obstacles should be applauded and expanded. With advances in genomic technologies and computational analysis, our ability to detect, characterize, treat, monitor, prevent, and control infections has advanced rapidly in recent years and will continue to do so, with the hope of heralding a new era where the clinician is better armed to combat infection and promote human health.

FURTHER READING

Berry MP et al: An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. Nature 466:973, 2010.

Bullman S et al: Emerging concepts and technologies for the discovery of microorganisms involved in human disease. Annu Rev Pathol 12:217, 2017.

Burnham CD et al: Diagnosing antimicrobial resistance. Nat Rev Microbiol 15:697, 2017.

Croucher NJ et al: Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet 45:656, 2013.

Dudas G et al: Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature 544:309, 2017.

Gardy JL, Loman NJ: Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet 19:9, 2018.

Grad YH, Lipsitch M: Epidemiologic data and pathogen genome sequences: A powerful synergy for public health. Genome Biol 15:538, 2014.

Loman NJ, Pallen MJ: Twenty years of bacterial genome sequencing. Nat Rev Microbiol 13:787, 2015.

Mutreja A et al: Evidence for several waves of global transmission in the seventh cholera pandemic. Nature 477:462, 2011.

Snitkin ES et al: Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. Sci Transl Med 4:148ra16, 2012.

475 The Role of Circadian Biology in Health and Disease

Jonathan Cedernaes, Kathryn Moynihan Ramsey, Joseph Bass

Circadian rhythms are autonomous, systemic cycles of behavior and physiology that synchronize organismal function at the cell and tissue level in anticipation of the 24-h rotation of the Earth. A common feature of modern "24/7" life is the routine disruption in these evolutionarily conserved endogenous circadian cycles, due to the rise in shift work, jet travel across time zones, exposure to blue light–emitting devices at night, and disrupted sleep. A transformation in understanding the molecular basis of circadian disorders has generated a new wave of research on metabolic disease, inflammation, aging, and cancer. This chapter provides an overview of (1) the basic biology of the circadian system; (2) primary circadian rhythm and interrelated sleep disorders; and (3) the role of the circadian system in both normal human physiology and disease states. Lastly, we review the rapidly emerging field of chronobiology as a pathway for novel diagnostic and therapeutic activity. A glossary of terms used in circadian biology is summarized in **Table 475-1**.

BASIC EVOLUTION AND STRUCTURE OF THE CIRCADIAN SYSTEM

A daily cycle between light and darkness existed long before the emergence of multicellular life. Eubacteria, the most ancient photosynthetic prokaryote, emerged in the geologic record more than 2.5 billion years ago at the same time as the first endogenous molecular clock. The co-occurrence in molecular evolution of clocks and photosynthesis hints at an interrelated and selective advantage to the clock and energetic processes—indeed clocks coordinate oxygenic reactions with periods of sunlight each day, and perturbation of clock cycles reduces fitness, reproduction, and survival. Additionally, clocks protect photosynthetic organisms from the DNA-damaging effects of sunlight by timing the production of DNA repair processes, such as photolyase-mediated repair, to the nighttime. Across billions of years of evolution, highly conserved circadian clocks (from "*circa diem*," "about a day") have been found in all photosensitive organisms, governing a wide range of behavioral, physiologic, and biochemical processes. A defining property of the circadian clock system is that it enables organisms to anticipate, rather than simply react to, daily changes in the external environment that are tied to the day-night cycle. In mammals, circadian systems are organized hierarchically with a light-responsive "master" circadian pacemaker located within the suprachiasmatic nucleus (SCN) of the anterior hypothalamus, which in turn presides over a network of both extra-SCN and peripheral clocks (see "Anatomic Organization of the Circadian Clock Network," below). Daily light exposure signals to the SCN and entrains the circadian system to the 24-h day (see "Entrainment and Measurement of the Circadian System," below), and the SCN in turn maintains synchrony of a diverse network of peripheral clocks via a variety of signals that have as yet to be fully identified, but which involve direct physiologic rhythms (core body temperature), the autonomic nervous system, and neuroendocrine signals, including cortisol as part of the hypothalamic-pituitary-adrenal (HPA) axis.

MOLECULAR ORGANIZATION OF THE MAMMALIAN CIRCADIAN CLOCK

At the molecular level, mammalian circadian rhythms are generated by a transcription-translation autoregulatory feedback loop. The forward limb of the clock is composed of the basic helix-loop-helix transcription factors (TFs) CLOCK (or its paralogue, NPAS2) and BMAL1, which drive expression of their own repressors (PER and CRY) in the negative limb, in a cycle that repeats itself every 24 h **(Fig. 475-1)**. A second short feedback loop involves CLOCK/BMAL1-mediated transcription of the retinoic acid–related orphan nuclear receptors ROR and REV-ERB, which activate and repress *Bmal1* transcription, respectively. Additional posttranslational regulation of the stability and degradation of core clock TFs includes phosphorylation by casein kinase 1 epsilon (CK1ε) and casein kinase 1 delta (CK1δ) and ubiquitination by FBXL3 and FBXL21. In addition to the ~24-h oscillation of core clock genes, a wide array of downstream clock-controlled genes (CCGs) exhibit broad amplitude in expression, ultimately giving rise to rhythmic physiologic processes. The core clock feedback loop and the induction of CCG rhythms also involves epigenetic mechanisms such as acetylation and methylation. The molecular circadian feedback loop is synchronized with sunrise each day by photosensitive melanopsin-expressing neurons within the retina. The retinohypothalamic tract (RHT) represents the input circuit to the SCN that maintains coherent organismal rhythms. Of note, mutations in several of these clock genes are associated with impaired circadian rhythms and physiology in humans (see "Primary Pathologies of the Circadian System," below). The importance of clock genes in the brain has been demonstrated by genetic studies, which have found that deletion of *Bmal1* in the whole brain or regions that span the SCN cause behavioral arrhythmicity, even when genetic ablation occurs late in life. Conversely, restoring *Bmal1* expression specifically in brain in adult clock mutant mice can rescue behavioral locomotor rhythms. Of note, whereas CLOCK normally heterodimerizes with BMAL1, NPAS2 is able to functionally substitute for CLOCK within pacemaker neurons; thus while mice lacking either *Clock* or *Npas2* genes maintain rhythmicity, the double mutants lacking both CLOCK and NPAS2 have impaired circadian rhythms of locomotor activity.

A major transformation in our understanding of circadian biology came with the discovery that the molecular clock network is present not only in the SCN, but also within most peripheral tissues, as well as in extra-SCN neurons in the brain. Interestingly, both the SCN and peripheral tissues exhibit a surprisingly large number of cycling transcripts at the genomic level, with approximately 3–16% of the mammalian transcriptome displaying a 24-h variation in mRNA expression levels across any given tissue, although the set of genes exhibiting circadian rhythmicity varies substantially between tissues depending upon tissue-specific functions. Posttranscriptional events such as RNA polyadenylation and mRNA translation also exhibit circadian variation, further increasing the repertoire of circadian processes at a cellular level. Thus, it is not surprising that among genes in the mammalian

TABLE 475-1 Glossary of Terms Used in Discussion of the Circadian System

TERM	DESCRIPTION
ASPD	Advanced sleep phase disorder (see text for description).
CBT	Core body temperature. Often used as an indicator of the circadian rhythm, but can be masked by sleep and exercise.
CCGs	Clock-controlled genes; output of the molecular clock.
Chronotype	Internal circadian rhythm of an individual determined by phase of entrainment, determining sleep propensity and timing of maximum alertness over a 24-h period.
Circadian period	Time required for one complete cycle or oscillation. Calculated by the time distance between two consecutive peaks or troughs of a circadian variable.
Circadian phase	Timing of the circadian rhythm. Defined by comparing, e.g., the peak (acrophase) or trough (bathyphase) to a fixed event, e.g., to a point in time. Synonymous with phase angle.
Circadian rhythm	A biological process that exhibits an endogenous, entrainable oscillation of approximately 24 h.
Circadian rhythm sleep disorders	Disorders of multiple etiology that have in common that they result in maladjustment of the biological clock with respect to the environment.
Constant routine	An experimental paradigm designed to study endogenous circadian rhythms in humans, by keeping behavioral and environmental factors constant. These paradigms thereby entail constant dim lighting, evenly distributed isocaloric energy intake, semi-recumbent posture, and forced wakefulness.
Desynchrony	Loss of synchrony occurring either between a rhythm and its Zeitgeber (external), or between two or more rhythms within an organism (internal).
Diurnal rhythm	An oscillation synchronized with the day/night cycle that repeats itself with a 24-h period. The rhythm does not have to persist when time cues (e.g., light) are absent.
DLMO	Dim-light melatonin onset; a marker of melatonin rhythm.
DSPD	Delayed sleep phase disorder (see text for description).
Entrainment	Synchronization of a circadian rhythm or other self-sustaining oscillation by a factor—Zeitgeber—that enforces the oscillator. Constant entrainment between the Zeitgeber and the oscillator results in a stable phase relationship between these entities.
Infradian rhythm	A recurrent cycle or period with a period length significantly greater than 24 h.
Melatonin	Hormone produced by the pineal gland (chemical name N-acetyl-5-methoxytryptamine); derived from L-tryptophan. Various forms of melatonin can be prescribed for CRSDs or sleep disorders.
Non-24-h rhythm disorder	A syndrome in which there typically are chronic 1- to 2-h daily delays in sleep onset and wake times in an individual living in society, e.g., due to complete blindness.
Peripheral clocks	Clocks presiding outside of the suprachiasmatic nucleus, the circadian system's master pacemaker.
PRC	Phase response curve; visual representation of how a particular manipulation (e.g., light) produces phase shifts as a function of the phase (i.e., circadian time) at which the manipulation occurs. Defining the PRC to light has enabled researchers to understand and predict how entrainment to light cycles is accomplished.
SCN	The suprachiasmatic nucleus or nuclei, also known as the master pacemaker in mammalian species. A bilateral set of nuclei positioned in the anterior ventral hypothalamus. Essential for entraining extra-SCN central and peripheral oscillators to the prevailing light-dark cycle via photic input from the retina.
Shift work	Work scheduled so that it occurs outside of the traditional work schedule of 9:00 A.M. to 5:00 P.M., or 6:00 A.M. to 6:00 P.M., depending on definition. Various forms of shift work exist, such as early morning, evening, or night shifts, as well as rotating shifts.
Ultradian rhythm	A recurrent cycle or period with a period significantly shorter than 24 h—e.g., a 2-h rhythm would exhibit 12 cycles within a circadian (24-h) rhythm.

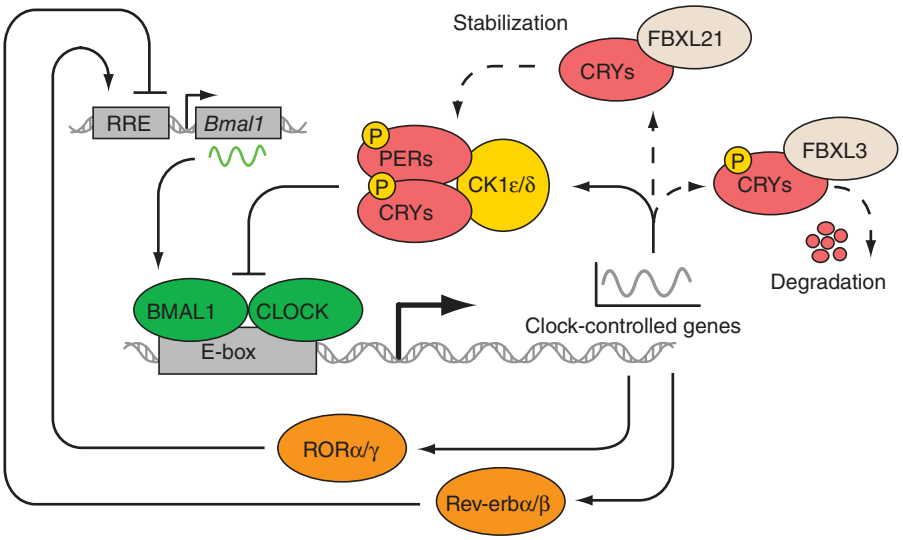


FIGURE 475-1 Central clock molecular mechanism. The core molecular clock machinery in mammals is encoded by interlocking transcription-translation feedback loops that oscillate with ~24-h periodicity. The transcription factors CLOCK and BMAL1 heterodimerize to drive transcription of downstream clock-controlled target genes containing E-box enhancer elements. Among these, the PER and CRY proteins multimerize and inhibit CLOCK/BMAL1 while RORs and REV-ERBs activate and inhibit, respectively, *Bmal1* transcription, resulting in rhythmic oscillations of clock-controlled and downstream target genes.

genome, almost half are considered to display circadian rhythms in at least one tissue.

ANATOMIC ORGANIZATION OF THE CIRCADIAN CLOCK NETWORK

Understanding the circuit organization of the circadian clock within the brain is increasingly relevant in understanding how the master circadian pacemaker center within the SCN regulates feeding, sleep-wake activity, energy expenditure, endocrine processes, and metabolism (Fig. 475-2). Identification of the SCN as the master pacemaker was first established by the observation that SCN lesioning induced complete loss of rhythms of locomotor activity and endocrine hormone secretion. The ventral "core" region of the SCN, which is composed of neurons producing vasoactive intestinal polypeptide (VIP), receives photic information directly from the retina through the RHT. At the molecular level, circadian gene transcription is induced within the SCN through the initial activation of immediate early genes, such as *Per1*, *Per2*, *c-fos*, and *jun*. Cells within the "core" region of the SCN then signal primarily via GABA-ergic neurotransmitter release to synchronize the cells within the "shell" region of the SCN, which produce the neuropeptide arginine vasopressin (AVP).

The SCN communicates to extra-SCN and peripheral clocks through both secreted factors and direct neuronal projections. The former was elegantly proven by the ability of SCN grafts to partially restore locomotor rhythms in an SCN-lesioned host. Efferent nerve outputs predominantly arise from the AVP-producing shell region of the SCN, although some output is also derived from the VIP-predominated core. The SCN projects to several hypothalamic relay regions, including the median preoptic nucleus, the subparaventricular zone (SPZ), the dorsomedial hypothalamus (DMH), and the paraventricular nucleus of the hypothalamus (PVH). These, in turn, regulate output to both sleep- and wake-promoting regions, as well as to regions involved in autonomic regulation and feeding. The SCN is thereby thought to promote sleep through the transmission of neural signals that terminate in the sleep-promoting ventrolateral preoptic nucleus (VLPO), i.e., one of the brain regions that is active during sleep. In contrast, the SCN promotes wakefulness during the active phase by transmission of neural signals that—by passing through regions such as the DMH—terminate in wake-promoting regions, including the locus coeruleus, lateral hypothalamic nucleus, ventral tegmental area, and dorsal raphe nucleus.

The SCN also signals via noradrenergic fibers to the pineal gland, thereby regulating the circadian production of the hormone melatonin. SCN control of the nighttime rise in pineal melatonin release (in both diurnal and nocturnal animals) is mediated through a pathway involving the PVH. Of note, artificial light at night is able to delay the secretion of melatonin, ultimately affecting sleep (see "Endocrine Systems Regulated by the Circadian Clock," below). Melatonin plays a complex role in the circadian system since the MT1 and MT2 melatonin receptors are expressed on the SCN itself; thus melatonin feeds back to modulate circadian outputs to other cells in the brain and body.

Neuronal output from the SCN also reaches the periphery, i.e., to the adrenal glands, the liver, and the pancreas. The SCN produces rhythmic variation in multiple neuroendocrine axes, producing daily rhythms of gonadotropin, thyrotropin, and somatotropin. Prominent hypothalamic-pituitary-adrenal (HPA) axis rhythms ultimately give rise to daily variation in diverse pathways essential for hemodynamic

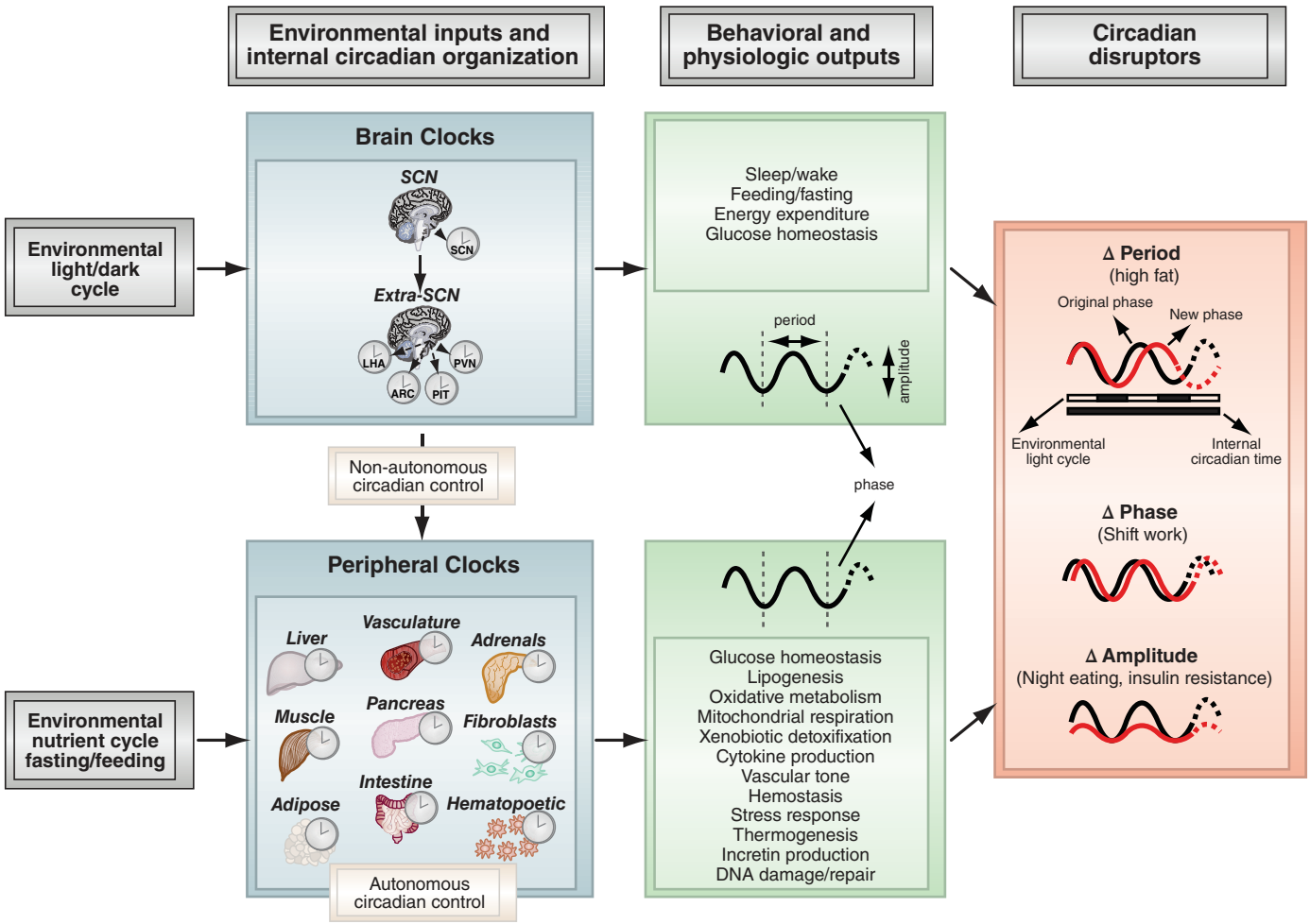


FIGURE 475-2 Central and peripheral clocks coordinate environmental cues with behavior and physiologic outputs. Light entrains the master pacemaker neurons in the SCN, which subsequently synchronizes extra-SCN and peripheral clocks. Brain clock output includes sleep-wake, fasting-feeding, and energy expenditure cycles, while peripheral clock output includes a wide range of physiologic processes, including glucose homeostasis, oxidative metabolism, cytokine production, and stress response. The right column indicates different ways that circadian disruptors, such as diet, shift work, or other circadian rhythm sleep disorders, may impact the clock—i.e., by changing circadian period, phase, or amplitude.

stability, metabolism, and inflammation. These rhythms originate with SCN control of corticotropin-releasing hormone (CRH)-producing cells in the PVH, which induce daily oscillations of both pituitary adrenocorticotropic hormone (ACTH) and adrenal cortisol. Highlighting the importance of SCN output for peripheral rhythms, there is a dramatic reduction in the number of transcripts that exhibit circadian rhythms in the liver following SCN ablation in mice. Nonetheless, when the autonomous clock in the liver is ablated in mice, some key clock transcripts such as *Per2* are still able to cycle as long as the core body temperature rhythm persists. Whereas the SCN is exclusively entrained by light, meal timing is able to signal circadian time directly to the liver. Thus, shifted meal timing as occurs during shift work or jetlag can uncouple peripheral clocks from the central pacemaker. In comparison to peripheral tissue clocks, the SCN is also resistant to phase shifts induced by temperature. This is consistent with the concept that the SCN generates the core body temperature rhythm as one of the major mechanisms to signal circadian time to peripheral clocks.

ENTRAINMENT AND MEASUREMENT OF THE CIRCADIAN SYSTEM

Under normal light-dark cycles, the circadian system is corrected or "entrained" on a daily basis, producing diurnal rhythms of 24 h. Such signals of entrainment are called Zeitgebers ("time-giver" signals) and include light exposure, meal timing, and activity patterns. Light serves as the dominant Zeitgeber for the circadian system, and a breakthrough in understanding photoentrainment in mammals came with the discovery of the melanopsin system, which is composed of a specialized class of photosensitive retinal ganglion cells that expresses the blue-light sensitive photopigment melanopsin in the inner retina, separate from the photoreceptive rods and cones. Blue light around this wavelength (~480 nm) suppresses melatonin and promotes subjective and objective (electroencephalography-assessed) wakefulness.

The ability of light to entrain the circadian system functions according to a so-called phase response curve (PRC). When light exposure occurs prior to the critical phase of the core body temperature (CBT), defined by the CBT's minimum, light produces a phase delay in the circadian rhythm. Conversely, light exposure after this critical period causes phase advances. The circadian system can respond even to small changes in light intensity (e.g., dim light at ~100 lux can produce half of the phase delay compared with an almost 100-fold greater light exposure). This responsiveness is furthermore affected by our genetic makeup, as variants in clock genes can modulate the response of the human circadian system to light.

When an organism is placed in an environment without Zeitgebers, the circadian rhythm is said to free-run, as it will rely on the endogenous rhythm of the circadian system. In humans, the study of endogenous circadian rhythms can be achieved by using a so-called constant routine. In these paradigms, subjects are maintained in a constant semi-recumbent posture, meals are provided on an hourly basis, light is constantly kept below the level which phase shifts the SCN, and circadian rhythms are assessed by measuring CBT, melatonin, or peptidergic hormone rhythms. In animals, circadian rhythms are instead studied by examining behavior and physiologic responses after 30–36 h of complete darkness, and endogenous rhythms are assessed by measuring voluntary locomotor activity. From these measurements, key properties of the circadian system can be ascertained, such as period length (peak-to-peak or trough-to-trough time), amplitude (peak-to-trough difference) and phase (timing of peak or trough in relation to a reference point) (Fig. 475-2).

These studies have revealed that the endogenous human circadian clock runs with a period length of approximately 24.2 h (compared with mice that run at ~23.5 h, depending on the strain). Evidence indicates that human females may have a slightly shorter circadian clock (24.1 vs 24.2 h). Notably, interindividual variability in the endogenous circadian period length is further diversified by the existence of genetic polymorphisms in clock genes (see below). These gene variants can confer extremes in endogenous circadian period as well as phase; the latter can be advanced or delayed by about 3–4 h in each direction. This is due both to altered circadian rhythms at the cell level and to altered SCN responsiveness to entrainment by light. For instance, *PER3* exists in a variable-number, tandem-repeat polymorphism. Individuals homozygous for a *PER3 5/5* genotype have been reported to be more responsive than *PER3 4/4* homozygous individuals to the melatonin-suppressing effect of evening blue light exposure.

Using specifically developed questionnaires to establish preferred sleep-wake timing, individuals can be categorized into so-called morningness-eveningness types or chronotypes. The most commonly used questionnaires are the Horne-Östberg morningness-eveningness questionnaire (MEQ) and the Munich ChronoType Questionnaire (MCTQ). The MEQ is composed of 19 questions regarding traits distinguishing morning-type compared with evening-type individuals, such as preferred waking time. In contrast, the MCTQ centers on the midpoint of sleep as a circadian marker, queries age and sex across a range of geographical locations, and can be used to ascertain differences between socially imposed sleep patterns (e.g., on working days) and sleep patterns on free days (the difference constituting so-called social jetlag). According to MCTQs obtained from primarily European populations, ~1% of the general population goes to bed before 10:00 PM and about 8% after 03:00 AM. Differences in chronotype are linked to altered circadian timing—including peak levels of melatonin, which can vary by up to 4 h between extreme morning and evening types. Extreme chronotypes have also been shown to be linked to various traits; i.e., low morningness scores have been associated with greater tolerance to shift work.

Melatonin is one of the most commonly used peripheral markers of an individual's circadian rhythm, reflecting the rhythmic function of the SCN. Circadian rhythms of melatonin can be measured in saliva or plasma, while 6-sulphatoxymelatonin (aMT6S), a metabolite generated from the breakdown of melatonin, can also be measured in urine. Accurate estimations of melatonin rhythms are often obtained by analyzing the dim light melatonin onset (DLMO), which as the name implies does not require an entire 24-h sampling, making this marker useful in both the clinical and research setting. In normally entrained individuals, the DLMO can be used to ascertain whether an individual's circadian rhythm is phase advanced or delayed, and this onset typically occurs ~2 h before the onset of sleep. The midpoint of sleep—the main marker used by the MCTQ—also strongly correlates with melatonin onset.

The CBT is also often utilized as an indicator of the circadian rhythm, and even though the outcome is more variable when using the CBT, it usually correlates well with the phase obtained using the melatonin rhythm. The CBT, however, can be masked by factors such as sleep, food intake, and activity. CBT can be recorded and registered wirelessly with relative ease. In humans, CBT can for example be recorded via rectal thermometers or probes that are swallowed to pass through the gastrointestinal tract. When humans are studied under normal conditions with normal lighting and sleep duration from 2300 to 0700 h, the CBT reaches around 37.2°C by 0900 h, and from there continues to rise slowly until it reaches 37.4°C around 11 h later. The CBT then drops to the daily low of 36.5°C in the early morning (0400 h).

Given the interrelationship between the circadian system and sleep-wake systems, researchers have developed paradigms that uncouple the circadian system from sleep-wake states, enabling the study of the contribution of the circadian system to investigated parameters across the entire sleep-wake cycle. These paradigms are known as "forced desynchrony" protocols and involve enforcing a significantly shortened (e.g., 20 h) or prolonged (e.g., 28 h) day length upon individuals. These protocols thus attempt to approximate what occurs during rotating shift work or "jetlag," e.g., when travel across several time zones suddenly shifts the light-dark and behavioral cycle drastically away from the entrained 24-h rhythm. As described below, forced desynchrony protocols have contributed to uncovering how the circadian system regulates parameters such as cognitive performance, subjective alertness, and metabolic and cardiovascular health.

PRIMARY PATHOLOGIES OF THE CIRCADIAN SYSTEM (SEE ALSO CHAP. 27)

An overarching term for disorders of the circadian system is *circadian rhythm sleep disorders* (CRSD). These disorders have become

increasingly recognized as important factors in a number of conditions; a unifying feature of CRSD involves a mismatch between subjective behavioral and physiologic rhythm with the environmental light-dark or social activity-rest cycle (i.e., the body clock is out of sync with the external light-dark cycle). CRSDs can arise either due to misalignment of an exogenous environmental factor, such as light, with the intrinsic circadian cycle, or due to misalignment of activity/rest cycle in relation to endogenous circadian timing (e.g., shift work or jetlag). In addition to such environmental or exogenous conditions causing circadian disruption, in some cases intrinsic circadian timing is altered in relation to the external environment, as in the case of endogenous circadian disorders that include those caused by mutations in core clock genes. Under conditions of intrinsic circadian abnormalities, it is often exceedingly difficult for individuals suffering from CRSDs to try to properly realign themselves, and these disorders often result in adverse effects such as sleepiness or depressed mood. Societal and economic consequences also are common; these can result in the individual being unable to maintain a job or be unable to attend at regular school hours. The criteria for CRSDs based on the International Classification of Sleep Disorders (ICSD) is shown in **Table 475-2**.

Animal models have greatly advanced our understanding of how core molecular clock components contribute to maintaining normal sleep-wake/rest-activity cycles **(Table 475-3)**. For example, *Clock*$^{\Delta 19/\Delta 19}$ mice have reduced total sleep duration and less induction of REM sleep in response to sleep deprivation. Further, mice that lack *Bmal1* have increased total sleep time, but it is more fragmented and lacks clear 24-h sleep-wake rhythms, and mice lacking the repressors *Cry1* and *Cry2* are not only arrhythmic but spend more time in non-REM sleep. Finally, while ablation of the circadian gene *Dbp* does not alter the specific duration of sleep stages, it does lead to an altered circadian sleep-wake distribution, with more sleep during their normal wake period and vice versa. Consistent with a key role of clock genes in regulating sleep-wake behavior, human genetic studies of twins have found that up to half of the variation in diurnal preference is heritable. Established genetic variants associated with diurnal preference and circadian sleep disorders are listed in **Table 475-4**.

Delayed Sleep Phase Disorder Delayed sleep phase disorder (DSPD) is one of the more common circadian rhythm sleep disorders and is characterized by chronic and significant delays in both sleep onset and wake times compared to normal "socially acceptable" sleep-wake hours (i.e., scoring as "extreme night owls" on morningness-eveningness preference tests). Rhythms of CBT and melatonin levels in plasma and urine are likewise delayed and the circadian period (tau) is longer in DSPD. Onset of DSPD most commonly occurs during adolescence or early adulthood. While the precise etiology of DSPD is not well established, it has been associated with polymorphisms within the circadian clock genes *CLOCK* and *PER3*. An integrated behavioral and pharmaceutical therapeutic approach has been found to be most effective at treating individuals with DSPD. Such treatments include a combination of bright-light therapy soon after waking in the morning (and/or dark-room therapy in the evening) and melatonin administration in the evening several hours prior to the onset of sleep. These approaches aim to realign endogenous circadian rhythms with the desired sleep-wake schedule. As individuals suffering from DSPD also phase delay more rapidly, this explains why attempts to phase advance their sleep schedule can be difficult, as well as why relapse can easily occur after initial treatment.

TABLE 475-2 Criteria for Circadian Rhythm Sleep Disorders

CRITERIA	DESCRIPTION
A	A persistent or recurrent pattern of sleep disturbance due primarily to one of the following: • Alterations of the internal circadian timekeeping system. • Misalignment between endogenous circadian rhythms and exogenous factors that affect the timing or duration of sleep.
B	A circadian-related sleep disruption that leads to insomnia, excessive daytime sleepiness, or both.
C	A sleep disturbance that is associated with impairment of social, occupational, or other areas of functioning.

Advanced Sleep Phase Disorder Another circadian rhythm sleep disorder whereby one gets the correct amount and quality of sleep but at a shifted time is advanced sleep phase disorder (ASPD). Individuals with ASPD experience an advance in their major sleep episode in relation to the desired clock time. Thus, this disorder typically results both in very early evening bedtimes and morning awakenings (e.g., "extreme early birds"), resulting in reduced quality of life due to excessive sleepiness during early evening, even in social situations. Individuals with ASPD also have phase-advanced temperature and melatonin rhythms in parallel with their earlier sleep onsets. ASPD occurs more often in older individuals, although early-onset autosomal dominant familial variants (familial advanced sleep phase syndrome [FASPS]) have also been associated with mutations in either the *PER2* or the casein kinase 1δ (CK1δ) genes. PER2 is critical for SCN resetting by light, and the identification of PER2 mutations in familial ASPD was the first in which clock genes were tied to a CRSD. Such mutations have been found to be able to shorten the endogenous circadian period to about 23.3 h compared with the normal 24.2-h period length. Accordingly, ASPD can be distinguished from other non-circadian sleep disorders by an early onset of dim-light melatonin secretion, a reliable marker for the timing of endogenous circadian rhythms both in the research laboratory as well as in the clinical setting. Polysomnography (PSG) or actigraphy are not required for diagnosing ASPD, although actigraphy may be significantly more feasible for long-term analysis of circadian timing of sleep. Treatments for ASPD include bright light or blue-enriched phototherapy in the evening hours to delay the phase of the circadian clock to a later hour.

Non-24-h Sleep-Wake Rhythm Disorder Individuals with non-24-h sleep-wake rhythm disorder ("non-24"), otherwise known as free-running disorder (FRD), have endogenous circadian rhythms that are not synchronized with the external 24-h day-night cycle due to an inability to readjust the circadian clock to the 24-h day on a daily basis. This most commonly occurs in individuals who are completely blind (i.e., lacking all photoreceptors) since they are unable to respond to daily light cues, which normally would reset the endogenous circadian clock on a daily basis (although the condition has also been reported in sighted individuals). Instead, the sleep-wake period length corresponds to the individual's endogenous circadian rhythms, which are typically slightly longer than 24 h, thereby shifting sleep and wake cycles over time in relation to the light-dark cycle. Instead of sleeping at the same time each day, their sleep time would gradually be delayed each day until their sleep period literally goes "around the clock." Depending on the individual's endogenous rhythm, the individual will take a given number of days to re-align his or her endogenous phase (in a 360° phase plot) with the zero time point in the exogenous 24-h light-dark cycle. For example, if an individual has an endogenous 24.5-h rhythm, it would take that individual 48 days to cycle from one cycle to the next. Because of this chronic cycling, prominent symptoms of non-24 include sleep-wake cycle disruption (insomnia and daytime sleepiness), impaired alertness and mood levels, and severe difficulties partaking in normally scheduled work, school, or social activities. Non-24 can be diagnosed following diurnal analysis of an individual's melatonin or cortisol rhythms, in combination with analyses of sleep diaries where the sleep onset and offset can be visualized over time to identify the free-running period. Treatments for sighted non-24-h patients include a combination of bright light therapy with appropriately timed melatonin administration, while melatonin and dual melatonin (MT1 and MT2) receptor agonist administration in completely blind non-24-h patients has been shown to entrain free-running rhythms and improve symptoms.

Shift Work Sleep Disorder Given the increased prevalence of shift work in today's 24/7 society and the accumulating evidence for increased incidence of sleep and metabolic disorders, including obesity, diabetes, cardiovascular disease, and cancer, in shift workers, the need to develop effective treatments for shift work sleep disorder (SWSD) are

TABLE 475-3 Animal Models of Genetic Circadian Disruption

GENE	AVERAGE CIRCADIAN TIME OF PEAK TRANSCRIPT LEVEL		ALLELE	MUTANT PHENOTYPE
	SCN	PERIPHERY		
Bmal1 (Arntl)	15–21	22–2	*Bmal1*$^{-/-}$	Arrhythmic
CK1δ (Csnk1δ)	No rhythm	No rhythm	*Csnk1δ*$^{+/-}$	0 to 0.5-h shorter period
CK1ε (Csnk1ε)	No rhythm	No rhythm	*CK1ε*^{tau}	4-h shorter period
CK1ε	—	—	*CK1ε*$^{-/-}$	0.2- to 0.4-h longer period
Clock	No rhythm	21–3	*Clock*$^{-/-}$	0.5-h shorter period
—	—	—	*Clock*$^{\Delta 19/\Delta 19}$	4-h longer period/arrhythmic
Clock/Npas2	—	—	*Clock*$^{-/-}$*/NPAS2*$^{-/-}$	Arrhythmic
Cry1	8–14	14–18	*Cry1*$^{-/-}$	1-h shorter period
Cry2	8–14	8–12	*Cry2*$^{-/-}$	1-h longer period
—	—	—	*Cry2*^{A260T}	0.2-h shorter period
Dbp	—	—	*Dbp*$^{-/-}$	0.5-h shorter period
Npas2	N/A	0–4	*Npas2*$^{-/-}$	0.2-h shorter period
Per1	4–8	10–16	*Per1*$^{-/-}$	0.7-h shorter period
—	—	—	*Per1*^{brdm1}	1-h shorter period
—	—	—	*Per1*^{ldc}	0.5-h shorter period/arrhythmic
Per2	6–12	14–18	*Per2*^{brdm1}	1.5-h shorter period/arrhythmic
—	—	—	*Per2*^{ldc}	Arrhythmic
Per3	4–9	10–14	*Per3*$^{-/-}$	0 to 0.5-h shorter period
Rev-erbα (Nr1d1)	2–6	4–10	*Rev-erbα*$^{-/-}$	0.5-h shorter period/disrupted photic entrainment
Rorα	6–10	Arrhythmic/various	*staggerer*	0.5-h shorter period/disrupted photic entrainment
Rorβ	4–8	18–22	*Rorβ*$^{-/-}$	0.5-h longer period
Rorγ	N/A	16–20/various	*Rorγ*$^{-/-}$	Normal behavior

Note: Normal circadian rhythms of circadian clock and related genes, with description of circadian phenotype in mutant mice.

Source: Adapted from Hum Mol Genet 15:R271, 2006, and Adv Genet 74:175, 2011.

increasingly important. Shift work sleep disorder is at its core defined by the primary symptom of either insomnia or excessive sleepiness, occurring as a transient phenomenon in relation to work that usually is scheduled during the habitual hours of sleep or comprises irregular work hours. The symptoms may result from recovery of sleep having to consume a large proportion of the individual's free time, which may produce negative social consequences such as difficulties maintaining social relationships. Older individuals are typically at an increased risk of SWSD due to age-associated decline in the ability to maintain sleep during normal waking hours. Since the symptoms likely arise from a misalignment of sleep-wake rhythms with the external light-dark cycle, therapeutic approaches aim to realign endogenous circadian rhythms with the sleep cycles dictated by work. In addition to optimizing the sleep environment at home to minimize disruptions, timed bright light therapy can help individuals with SWSD—i.e., for night workers, intermittent bright light exposure during the night and avoidance of bright light during the morning, even on days off, has been shown to improve sleep and feelings of alertness. Melatonin prior to bedtime may also help improve symptoms of SWSD. Genetic screening combined with chronotype questionnaires may become useful tools for determining whether a given individual is suited for shiftwork. For instance, a twin study indicated that a genetic variant of the circadian gene *DEC2* was associated with reduced sleep duration, and with shorter recovery sleep following extended sleep deprivation. More studies may reveal additional genetic variants that confer an advantage to repeated phase advances and phase delays as typically occurs in shift work.

Irregular Sleep-Wake Rhythm Damage to the SCN can produce arrhythmicity in animals and is thought to be one of the possible underlying reasons for the temporally disorganized sleep-wake pattern that characterizes the disorder known as irregular sleep-wake rhythm (ISWR). Other contributing factors may be a reduced responsiveness to entraining signals such as light and physical activity, as well as a decreased exposure to such signals as often occurs with increasing age due to, for example, increased risk of poor health and impaired mobility. Whereas the total sleep time per 24 h may be comparable, there is

TABLE 475-4 Mutations and Gene Variants Linked to Sleep-wake Disorders and Diurnal Preference

GENE	POSITION	POPULATION	SYNDROME/SLEEP PREFERENCE
h*CKIε*	S408N	Japanese	Protection against DSPS
h*CKIγ*	T44A	Pedigree	FASPS
h*CKIδ*	H46R	Pedigree	FASPS
h*CLOCK*	T3111C (3'-UTR)	European	Eveningness
h*CRY2*	A260T	Pedigree	FASPS
h*PER2*	S662G (Missense mutations in CKIε binding region)	Pedigree	FASPS
h*PER2*	C111G (5'-UTR)	British	Extreme morningness
h*PER3*	P415A/H417R	Pedigree	FASPS and seasonal affective disorder
h*PER3*	G647	Swedish/Finish/Austrian/German	Morningness
h*PER3*	G647, P864, 4-repeat, T1037, R1158	Japanese	DSPS
h*PER3*	Increased repeats (exon 18, 54 bp)	Brazilian	DSPS
h*VIP*	rs9479402 (gene variant 54 kb upstream of VIP)	European (>97% European ancestry)	Morningness

a relative absence of a circadian pattern to the sleep-wake cycle. Sleep timing throughout the sleep-wake cycle can be shortened—sometimes close to randomly distributed—instead of occurring in several distinct bouts. ISWR is often associated with neurologic impairment, foremost Alzheimer's disease in older age; however, ISWR can also occur in individuals with poor sleep hygiene. The most effective treatments for ISWR involve not pharmacotherapy but rather multimodal interventions such as increased light exposure, improved sleep hygiene, and promotion of social and physical activities.

Jetlag Most have experienced symptoms associated with jetlag, including insomnia, day-time sleepiness, and fatigue, when traveling from one time zone to another, as one's endogenous circadian rhythms are not yet aligned, or entrained, to the new external light-dark cycle. This is due to the slowness of the circadian system to adapt to the new time zone: typically, the human circadian system is able to shift ~1.5 h a day in the westward direction (i.e., a phase delay), whereas it shifts more slowly (about 1 h daily) with eastward direction of travel (i.e., achieving a phase advance). Importantly, the symptoms are distinguished from the more short-lived symptoms that can partially result from exposure to traditional airplane cabin conditions, including abdominal distention, dependent edema, muscle cramps, headaches, nausea, and, intermittent dizziness. Usually symptoms of jetlag abate within the first couple of days after traveling, and may present themselves after a first night of good sleep (which is more dependent on a high build-up of homeostatic sleep pressure). Older individuals (age >50) appear to be more at risk. While symptoms are transient, therapeutic approaches can alleviate or temper some of the side effects of travel by hastening synchronization of the internal and external circadian cycles. Behavioral treatments include appropriately timed bright-light exposure and avoidance of bright light during the night time in the new destination, while pharmacologic approaches include timed melatonin administration before bedtime both prior to and following travel, resulting in improved sleep quality and decreased night waking.

Social Jetlag Individuals with a late chronotype are prone to suffer from "social jetlag," a phenomenon in which individuals are forced to awaken at a point at which their bodies are entrained to be asleep due to discrepancy between alignment of social and biological time. Social jetlag can be estimated using questionnaires, such as the MCTQ, to compare sleep timing on non-free compared with free days. This has established that a large proportion of the European population suffers from 2 or more hours of social jetlag. Chronic social jetlag has been associated with an increased risk of developing obesity and the metabolic syndrome, as well as with greater alcohol consumption and smoking.

The aforementioned categories of defined clinical circadian disorders have been traditionally established based upon consideration of the endogenous behavioral and physiologic cycles (primarily of melatonin and temperature) with the external 24-h light-dark cycle. In the following sections, we build on the concepts of circadian behavioral disorders to consider new and emerging insight into the role of circadian disruption in organismal homeostasis **(Figs. 475-3 and 475-4)**, and the availability of genetic strategies to dissect the interrelationship between clock function, health, and disease.

ROLE OF THE CLOCK SYSTEM IN PHYSIOLOGY

Endocrine Systems Regulated by the Circadian Clock In addition to regulation of behavioral rhythms such as sleep/wake and fasting/feeding cycles, the circadian clock also regulates rhythms of the endocrine system. Cortisol rhythms are regulated through a feedback

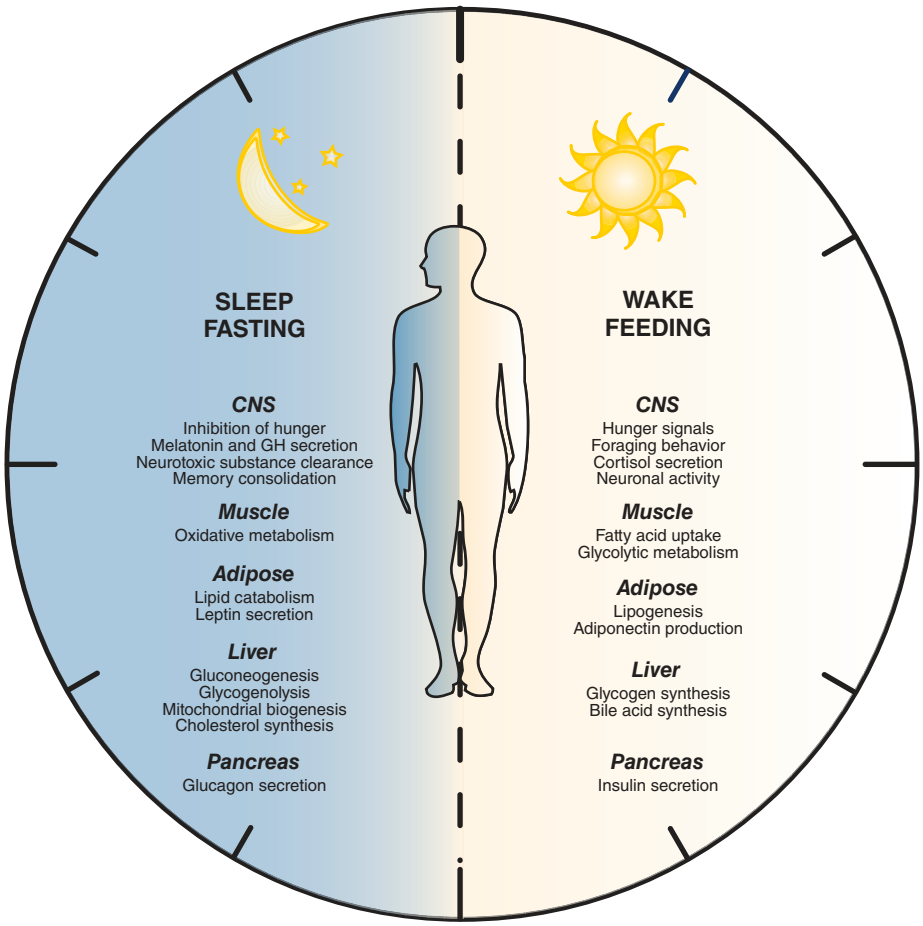


FIGURE 475-3 The circadian clock partitions behavioral, physiologic, and metabolic processes according to time of day. The partitioning of metabolic processes to appropriate times of day is critical for the maintenance of health from cellular to mammalian organisms. This figure highlights which processes peak within CNS, muscle, adipose, liver, and pancreas during either the sleep-fasting or wake-feeding cycle in humans.

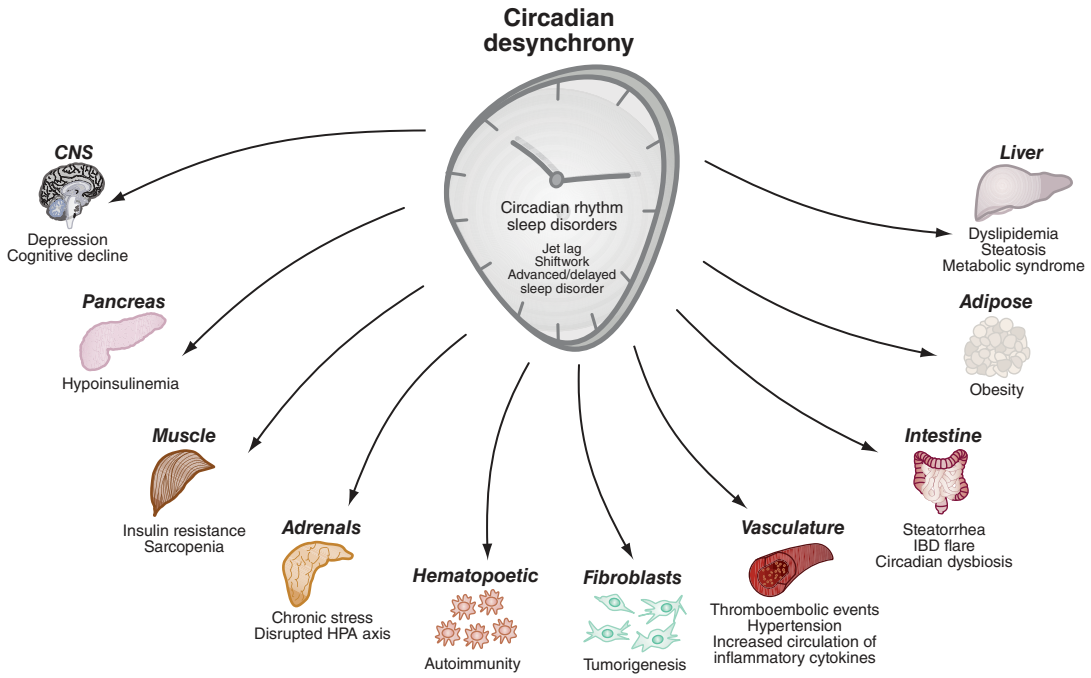


FIGURE 475-4 Pathologies resulting from circadian desynchrony. Circadian rhythm sleep disorders, including advanced/delayed phase sleep disorder, jetlag, social jetlag, and shift work, result in a desynchrony between the environmental light-dark cycle "time" and the endogenous clock "time." Pathologies can thus arise through misalignment imposed by exogenous factors (e.g., altered light cycle, feeding rhythm) as well as endogenous factors, such as mutations in core clock genes. Such desynchrony results in a host of wide-ranging pathologies across multiple tissues, including hypoinsulinemia (pancreas), disrupted HPA axis, autoimmunity, hypertension, obesity, and metabolic syndrome.

loop known as the hypothalamic-pituitary (HPA) axis **(Chap. 379)**. Hypothalamic secretion of CRH and AVP promotes secretion of pituitary adrenocorticotropic hormone (ACTH), which in turn regulates rhythmic cortisol secretion from the adrenal cortex. Cortisol release increases towards the morning, and this increase is believed to prepare the brain and peripheral tissues for daytime activity and food intake. Daytime sleep can blunt circulating cortisol levels, presumably through the occurrence of non-REM sleep. AVP secretion in mice occurs prior to sleep to promote water intake, thereby preventing dehydration during the sleep period. Several hormonal systems are in fact influenced more by sleep than by circadian rhythms. For instance, secretion of growth hormone (GH) is profoundly blunted during acute overnight wakefulness. The secretion of this hormone is primarily dependent on the occurrence of slow-wave sleep, which is a homeostatically driven sleep stage that occurs primarily in the first part of the sleep period. Cortisol also exhibits a peak close after wakefulness: the cortisol awakening response (CAR). This peak seems to be independent of a circadian rhythm, as the CAR is severely blunted by acute overnight wakefulness. Curtailed sleep and overnight wakefulness increase the activity of the HPA axis and may increase diurnal cortisol levels. Sleep also influences melatonin amplitude, such that sleep deprivation can increase melatonin levels. In working environments, the effects of curtailed sleep are often confounded by mistimed exposure to light. Even low levels of light are able to potently suppress melatonin secretion. Together with altered timing in light exposure, perturbed hormonal levels may represent a mechanism through which altered timing and duration of sleep may impact central and peripheral circadian oscillators.

Centrally controlled rhythms of melatonin and cortisol are considered key regulators of extra-SCN and peripheral oscillators. Glucocorticoid receptors exist in both the central nervous system and in peripheral tissues such as skeletal muscle, liver, and adipose tissue. Following acute shifts in light-dark or feeding cycles, 24-h rhythms of circulating cortisol appear to shift more slowly than other rhythms and may thus contribute to adverse effects of circadian misalignment by hampering proper realignment of peripheral clocks. Glucocorticoids shift clock gene expression in muscle, kidney, and lung, while the powerful synthetic glucocorticoid dexamethasone is able to synchronize (e.g., reset) circadian rhythms of cells in culture, including liver cells. Consistent with a role for glucocorticoid regulation of the clock, both adrenalectomy, which results in a lack of cortisol, and exogenous corticosteroid supplementation significantly disrupt the circadian clock system.

Several peripherally produced hormones and peptides are not only produced rhythmically but can also feedback to central clocks, including the SCN. For instance, both cortisol and thyroid hormones regulate their own rhythmic synthesis by feedback to central brain regions, i.e., the hypothalamus (for cortisol) and pituitary (for both hormones). Several other peripherally produced factors have been proposed to influence the central clock, including fatty acids produced by the adipose tissue and fibroblast growth factor 21, a hormone primarily produced by the liver. Peripheral hormones that signal energy state and hunger also exhibit circadian rhythms. The most extensively studied hormones are leptin, which is released from white adipose tissue cells, and ghrelin, which is released from the upper fundus region of the stomach. Ghrelin also exhibits significant peaks related to anticipated meal timing, which persist for several days of fasting in humans. Circulating rhythms of leptin and ghrelin are disrupted in circadian mutant mice and are also perturbed in humans subjected to circadian misalignment. For instance, *Per* and *Cry* mutant mice exhibit severely blunted leptin rhythms, and wild-type mice exposed to jetlag—through repeatedly altered light-dark cycles—show a reduced wake-associated decrease in leptin. Similarly, humans forced to live 28-h days exhibit increased 24-h profiles of ghrelin, and conversely decreased levels of leptin. Ghrelin and leptin signal to several regions of the brain, including integrative appetitive regions of the hypothalamus such as the arcuate and paraventricular region. Through actions in several such central sites, these hormones influence rhythms of food intake and energy homeostasis in a nutrient-dependent manner.

Role for the Clock in Metabolic Homeostasis Circadian control of glucose homeostasis has long been recognized, as early studies demonstrated variation in glucose tolerance and insulin action across the day. For example, oral glucose tolerance is impaired in the evening and afternoon compared with the morning due to a combination of circadian control of both peripheral insulin sensitivity and

pancreatic β-cell insulin secretion. Another example is the "dawn phenomenon," whereby glucose levels peak prior to the onset of activity. Further, destruction of the SCN has been shown to abolish circadian regulation of glucose metabolism in rats, and daily cycles of insulin secretion and glucose tolerance are often perturbed in patients with type 2 diabetes. Changes have also been observed in first-degree relatives of patients with type 2 diabetes, possibly highlighting a key hereditary component of the circadian clock in the pathogenesis of type 2 diabetes.

Ablating clock genes in mice has revealed a key function for both central and peripheral clocks in regulating energy homeostasis. The circadian system has been shown to regulate rhythmic insulin secretion from the pancreas via both neural signals and hormonal levels (e.g., cortisol and norepinephrine), as well as via cell autonomous clock regulation within the pancreatic β-cell itself. An early observation was that whole-body mutant *Clock*$^{\Delta 19/\Delta 19}$ mice developed obesity without displaying hyperinsulinemia, a phenomenon that indicated concurrent β-cell failure. This was later confirmed using pancreas- and β-cell–specific *Bmal1*-deficient mice, which exhibited glucose intolerance, hypoinsulinemia, and impaired glucose-stimulated insulin secretion. The molecular clock within other peripheral tissues such as liver, adipose tissue, and skeletal muscle also regulate circadian fluctuations in insulin sensitivity and glucose disposal, which are highest in the morning and decline towards the evening. Liver-specific *Bmal1* mutant studies have revealed liver clock promotion of gluconeogenesis, glycogenolysis, and mitochondrial oxidative metabolism in the sleep/fasting period while promoting glycogen synthesis in the wake/feeding period. Muscle-specific *Bmal1* deficient mice display reduced glucose tolerance, concomitant with lower levels of proteins involved in glucose uptake by muscle cells (e.g., the glucose transporter GLUT4). Ablation of the *Cry1* and *Cry2* repressors in the negative limb of the clock alters glucagon and glucocorticoid signaling in the liver, contributing to hyperglycemia and impaired glucose tolerance in these mutant mice. Together, these genetic studies in mice suggest a role for tissue-specific clocks in the partitioning of energy utilization across the sleep-wake cycle.

Importantly, peripheral clocks also interact with other environmental factors such as diet and time of feeding. For example, high-fat feeding leads not only to obesity and metabolic syndrome in mice, but also to perturbed clock gene expression across multiple peripheral tissues and a disrupted sleep-wake/fasting-feeding cycle, as revealed by increased activity and feeding during the daytime. Furthermore, mice that are fed a high-fat diet exclusively during the light phase gain significantly more weight than mice that are fed the same diet during the dark period—the active period for mice. Additionally, the metabolic phenotypes arising from ad lib high-fat feeding can be significantly ameliorated by restricting the time of high-fat feeding exclusively to the dark period. Time-restricted feeding can also increase the activity of brown adipose tissue in mice and reduce hepatic glucose production to instead promote beta oxidation of fatty acids. These findings have been corroborated in human interventional studies, which have demonstrated that time-restricted feeding can improve metabolic homeostasis and promote weight loss. Time-restricted feeding may also modulate central regulation of sleep and hunger, as a study found that humans who restricted their food intake to a shorter than ad lib period also consumed less daily calories and reported improved sleep.

Finally, animal studies have further shown that when the light-dark cycle is disrupted or animals are subjected to conditions mimicking "jetlag"—by artificially advancing or delaying the daily light period—there is desynchronization amongst circadian clocks and subsequent weight gain. Accumulating evidence in humans suggests that circadian misalignment both disrupts and desynchronizes circadian clocks across tissues. Prolonged circadian misalignment using forced desynchrony protocols reduces insulin sensitivity in the pre- and postprandial state. Under such conditions, insulin secretion fails to suppress glucose levels, suggesting inadequate β-cell compensation. Moreover, resting metabolic rate declines significantly both in the awake and sleeping state, altogether providing potential explanations why shift work can increase the risk of obesity, type 2 diabetes, and the metabolic syndrome.

Human genetic association studies also support a role for clock genes in metabolic homeostasis and beta cell function. Carriers of a certain *BMAL1* polymorphism have a greater risk of developing type 2 diabetes, while *CLOCK* variants have been found to interact with diet, such that variants can have a protective effect on insulin sensitivity in individuals with high monounsaturated fat intake or in individuals provided a low-fat diet. Instead, the minor allele of another *CLOCK* gene variant has been associated with increased waist circumference, but only in those with high saturated fat intake. Similarly, *NPAS2* and *BMAL1* variants have been associated with a greater risk of hypertension. Melatonin receptor *MTNR1B* gene variants, which result in increased expression of MTNR1B, have been associated with elevated fasting blood glucose levels and reduced insulin secretion irrespective of their level of glycemic control, consistent with the known effect of melatonin on insulin secretion and lower insulin secretion in the evening. These association studies highlight the role of the circadian system in metabolism, as well as potential for interactions of external perturbations—such as circadian misalignment—with a protective or adverse genetic profile.

A large proportion of society recurrently shifts sleep-wake times between working/non-free days and free days. This social jetlag has been increasingly tied to metabolic disruptions, including a greater risk of obesity and type 2 diabetes. As this involves recurrent phase advances and phase delays—like shift work but of smaller magnitude—it is possible that social jetlag also results in perturbed energy expenditure, in combination with disruptions to the circadian rhythm of hunger drive, further increasing the risk of obesity. Repeated shifts in the food- and SCN-driven rhythm of insulin release may similarly over time increase the risk of type 2 diabetes. Shifted feeding rhythms in relation to the sleep-wake cycle and the timing of SCN activity may be causally involved in this pathogenesis. This is exemplified by the disorders known as night-eating syndrome and sleep-related eating disorder. In the former syndrome, a large part of daily calorie consumption occurs in the evening and nighttime hours, and this shifted meal pattern has been associated with a delayed timing of the internal clock. Some evidence exists that these syndromes are associated with obesity. Individuals who report sleeping fewer hours or who are subjected to restricted sleep for a few consecutive days have also been found to consume more calories later in the evening, perhaps explaining why sleep restriction increases the risk of obesity. These associations have also been observed in individuals with later onset of sleep, i.e., evening chronotypes. Night-eating syndrome and later chronotypes have also been linked to type 2 diabetes and may be more common than other eating disorders such as binge-eating disorder. Both conditions have also been found to be associated with impaired glycemic control—such as a greater likelihood of hemoglobin A1c values exceeding 7%—in patients already suffering from type 2 diabetes, emphasizing how proper alignment of internal circadian rhythms with external factors are key contributing factors for long-term metabolic homeostasis.

Circadian Clocks in Relation to Brain Health and Cognition

Molecular circadian clocks are present not only within the extra-SCN regions of the brain but also in neurons, astrocytes, and microglia. Emphasizing the functional significance of properly aligned clocks for brain health, shift workers have been found to have decreased grey matter in brain regions involved in memory and executive functions, with more notable effects in individuals who had shorter recovery periods between the onset of each shift work cycle. Adults performing rotating shift work for many years have also been shown to exhibit signs of accelerated cognitive aging. Notably, evidence suggests that these effects may be reversible, as those who have stopped carrying out shift work exhibit normal cognitive performance 5 or more years later.

Recent studies have also uncovered an important role for perturbed circadian and sleep-wake rhythms in neurodegenerative conditions such as Alzheimer's disease (AD), Huntington disease (HD), and Parkinson's disease (PD). Amyloid beta, a key pathognomonic component

of AD, normally exhibits circadian fluctuations in the extracellular space in the brain, as well as in the cerebrospinal fluid and plasma in humans, peaking during the active period and falling during sleep. Of note, these daily rhythms of amyloid beta accumulation are dampened in mice that are prone to develop AD; reduced fluctuations in plasma amyloid beta fluctuations have also been noted in older compared with younger individuals. It is believed that removal of amyloid beta (and other neurotoxic substances) during the nighttime sleep period is facilitated by a lymphatic-like system that relies on glial cells (the "glymphatic" system). Whereas the function of this system has been shown to be important in mice, its circadian components and relevance to humans remain to be determined. Consistent with a role for circadian rhythms in the pathogenesis of AD, ablation of core clock genes throughout the brain or within subregions of the brain increases oxidative stress and neuronal cell death, while promoting scarring of brain tissue (astrogliosis). Furthermore, perturbed light-dark cycles increased pathology associated with oxidative stress, and single-nucleotide polymorphisms in *Clock* and *Bmal1* have been associated with increased risk of developing AD.

Evidence also indicates that the relationship between the circadian/sleep-wake system and AD is bidirectional. For example, patients suffering from AD exhibit several signs of perturbed circadian rhythms, the most prominent of such phenomena being "sundowning," whereby AD patients become more agitated and exhibit delirium-like symptoms in the afternoon or evening. Studies have furthermore indicated that in severe forms of AD, the circadian rhythm is phase delayed. Aged AD-prone mice also display perturbed sleep-wake patterns, which can be corrected by immunization against amyloid beta or by an orexin antagonist. Further research will help uncover the primary pathogenic contribution of the circadian system, and its independent contribution from perturbed sleep, in conditions like AD. Notably, evidence suggests that interventions that increase daytime light exposure and include melatonin supplementation are able to ameliorate symptoms of AD, presumably by counteracting disrupted circadian rhythms.

While the relation between shift work and depression has not been extensively studied, disruption of sleep and circadian rhythms and the pathogenesis of depression are intimately interlocked. Clock genes have also been implicated in depression and mood both in animal and human studies. Polymorphisms of genes that regulate sleep and circadian rhythms—for instance, a long gene variant of *PER3*—have also been linked to bipolar disorder and schizophrenia, while *CRY2* and *CLOCK* gene polymorphisms are associated with seasonal affective disorder, a type of depression arising in the fall and winter months when the levels of sunlight are lowest. Bipolar disorder is furthermore often triggered by circadian disruptions or curtailed sleep. Both bipolar disorder and schizophrenia have been linked to various forms of circadian disruption following disease onset, and a critical component of disease treatment often involves normalizing sleep and sleep-wake rhythms.

Sleep deprivation by itself is known to reduce alertness, impair decision-making, and increase risk for accidents—after 18–24 h of continuous wakefulness, several skills exhibit the same degree of decline as following mild alcohol intoxication. However, cognitive abilities may suffer even further when sleep restriction is combined with circadian misalignment as in shift work. In one study, participants were subjected to ~33-h long days in parallel with reduced sleep (equivalent to 5.6 h sleep in a 24-h period), yielding a forced desynchrony protocol coupled with sleep loss. When subjects were tested at the nadir of their circadian period, the subjects' reaction speed dropped almost by an order of magnitude compared with controls. In another study, researchers noted almost a 36% greater incidence of serious medical errors in resident interns who regularly worked 24-h or longer shifts compared with those who were randomly assigned to work up to 16-h long shifts. Furthermore, errors that resulted in patient death were three times more likely to occur in residents working extended hours compared with those who only worked up to 16-h long shifts.

Circadian Regulation of Gastrointestinal Homeostasis and the Microbiota Physiologic aspects of the gastrointestinal (GI) tract exhibit day-night variations that anticipate and prepare for food intake and digestion during the active period. Gastric emptying, as well as colonic motility, are considerably greater during the active phase, as the phasic motor program supporting movement of digested material along the intestine is approximately twice as fast during the day compared with night. Bile acid secretion also exhibits circadian rhythmicity in the intestine, as does absorption and the expression of many nutrient uptake transporters in the intestinal wall, including the main glucose transporter protein SGLT1. The permeability of the intestinal wall also varies throughout the sleep-wake cycle, and mice exposed to chronic sleep fragmentation exhibit increased intestinal permeability, which may enable inflammatory molecules from bacteria to reach the systemic circulation.

The composition and function of the microbe population living in the intestine (i.e., the gut microbiota) also display circadian rhythmicity, orchestrated by both host circadian clock gene expression and food intake rhythms. Accordingly, circadian disruption, either by environmental or genetic means, perturbs these microbial rhythms, disrupting both bacterial levels and the metabolic functions of the gut microbiota. For example, alterations in the expression and functions of the gut microbiota have been noted in humans exposed to acute jetlag, and evidence suggests that curtailing sleep, which often accompanies shift work and jetlag, can alter the gut microbiota. By increasing local and systemic inflammation, circadian disruption of the gut microbiota may be causally involved in the increased risk of inflammatory bowel disease (Crohn's disease and ulcerative colitis) and colon cancer in shift workers. Gender-specific differences have also been reported, as female mice display more pronounced microbial rhythms. Interestingly, the gut microbiome has also been shown to influence the rhythms of host tissues, such as the intestine and liver, suggesting that a bidirectional relationship exists between tissues that regulate metabolic processes and the gut microbiome across the sleep-wake cycle. These findings may furthermore have clinical implications, given that the gut microbiome may both directly (in the gut lumen) and indirectly (through host-microbiota interactions) impact pharmacokinetic and pharmacodynamic properties of therapeutic drugs across the 24-h day-night cycle.

Cardiovascular Health and the Circadian Clock An early epidemiologic observation was a greater incidence of myocardial infarction in the morning hours, with the lowest risk during the period preceding sleep. Other cardiovascular outcomes such as sudden cardiac death and syncope also exhibit a daily peak in the morning. Blood pressure (BP) typically peaks around 21:00 h and decreases later during sleep, partially due to a circadian nighttime dip of around 3–6 mmHg in systolic BP and 2–3 mmHg in diastolic BP. A dip in blood pressure of either less than 10% or greater than 20% during normal sleep has been associated with worse cardiovascular prognosis. Heart rate also typically decreases during sleep, while mistimed sleep leads to higher heart rate during their sleep time. Thus, a combination of reasons—which may also involve altered glucocorticoid levels and increased platelet aggregation—may contribute to a greater risk of cardiovascular disease in the morning. Subsequent epidemiologic studies also have demonstrated that shift work increases the risk of dyslipidemia and hypertension, as well as the risk of coronary heart disease, including myocardial infarction. These findings are in line with interventional findings in which circadian misalignment has been induced either by inverting the sleep-wake cycle or by imposing 28-h days on healthy human subjects. These studies have found that circadian misalignment elevates 24-h blood pressure, particularly during sleep. These changes may be causally related to how the autonomic system is regulated during sleep, as evidenced by reduced vagal cardiac control when the sleep-wake cycle is inverted.

Circadian Disruption and Cancer In 2007, the International Agency for Research on Cancer declared that shift work that involves circadian disruption is likely carcinogenic to humans. While evidence for an association between shift work and general cancer incidence is mixed, accruing evidence supports a link between shift work and

increased risk of developing colon and breast cancer, as well as having a poorer cancer prognosis. Telomere shortening, a phenomenon in aging that destabilizes the genome, has also been observed in shift workers as well as in individuals suffering from short sleep. This may reduce the ability of damaged or senescent cells to undergo apoptosis, and instead lead to uninhibited cell growth and cancer. An indirect role for the circadian clock has also come from retrospective studies on how cancer risk is related to food timing and duration of the nighttime fast in humans. These studies indicate that by portioning food intake to a restricted period of the day, circadian processes are optimized in a way that confers reduced risk of carcinogenic cell damage. Studies of recurring fasting have also been shown to lower the risk as well as to delay the onset of cancer.

Experimental genetic evidence has also implicated clock disruption as a factor in tumorigenesis. Genetic loss of *Per2* or *Bmal1* has been shown to promote lung tumorigenesis, while studies in *Per2* mutant mice have also revealed increased radiation-induced lymphoma associated with dysregulation of the cell cycle. However, disruption of the *Cry* gene in mice has also been implicated in tumor protection due to increased susceptibility to cell death. Thus, while both epidemiologic and experimental evidence suggests a link between circadian disruption and cancer, a full understanding of the role of circadian systems in tumorigenesis remains an area for investigation.

Circadian Regulation of the Immune System Circadian misalignment and sleep restriction both alter population levels of immune cells and decrease the ability of immune cells to produce reactive radicals. Chronic circadian disruption may thereby impair the immune system's ability to conduct immunosurveillance at the proper time of day and reduce the ability to mount an appropriate response upon exposure of pathogens during the recovery/sleep phase, when the immune system is typically more active. Instead, circadian misalignment increases a range of clinically used inflammatory markers (e.g., C-reactive protein, tumor necrosis factor α, and interleukin 6), and such changes have been noted even when the sleep-wake cycle is only prolonged to a slightly longer than normal 24.6-h day. While similar effects are also observed following acute total sleep deprivation or recurrent partial sleep restriction, circadian misalignment has been found to promote an even more pronounced elevation of such markers. Genetic clock disruption in peritoneal macrophages has also revealed clock control of Toll-like receptor 9, which is responsible for identifying molecules from foreign pathogens. *Clock* knockout mice also have reduced T cell antigen response, and mice immunized during the day had a stronger T cell response than mice immunized at night, supporting regulation of the immune system by the clock.

Aging and the Circadian Clock Instability in the clock system is an often overlooked hallmark of aging. Aging is associated with a decline in the robustness of intrinsic rhythmic processes at the behavioral, physiologic, and molecular levels in both human and animal models. At the *behavioral* level, aging leads to reduced and fragmented sleep, dampened locomotor activity and feeding rhythms, and a reduced ability to entrain to light, as old rodents are 20 times less sensitive to the entraining effects of light relative to young animals. Even middle-aged individuals exposed to jetlag also exhibit more symptoms of circadian misalignment, such as more time awake and reduced alertness, compared with young individuals. On a *physiologic* level, some of the hallmarks of aging are a reduction in amplitude (e.g., flattening of circadian pattern) and a phase-advance (e.g., a shift in the timing of the peak or nadir) in rhythms of the endocrine and neuroendocrine systems, including sleep onset and offset. For example, cortisol, DHEA, and melatonin all have dampened rhythms and are phase-advanced in aging; the combination of such changes may, for instance, contribute to more fragmented sleep and lower levels of restorative slow-wave sleep in aged individuals. Aging also leads to alterations in peptide expression in the SCN (VIP and AVP) and reduced amplitude of rhythms of SCN electrical activity. Further, while the SCN-dependent body temperature rhythm—a generally accepted marker for the integrity of circadian rhythms—peaks in the evening and is lowest in the early morning in young individuals, aged healthy subjects display a phase advance and decrease in circadian amplitude in body temperature rhythms. Indeed, evidence suggests that internal desynchrony between core body temperature rhythms and the sleep-wake cycle may contribute to age-associated circadian alterations. On a *molecular* level, aging is associated with decreased expression and altered diurnal profiles of several of the core clock genes, including *Clock* and *Bmal1*, within both SCN and peripheral tissues such as heart and liver. Interestingly, the acute induction of *Per1* in response to light was markedly reduced in the SCN of aged mice compared with young mice, potentially contributing to their delayed response to light entrainment. Mice lacking *Bmal1* die prematurely compared with control mice, consistent with premature accumulation of reactive oxygen species. These mice have an accelerated onset of numerous age-related pathologies, including cataracts, sarcopenia, reduced organ size, and decreased hair growth. Instead, deficiency of *Cryptochrome*, a repressor of the internal clock repressor, has been associated with alterations in liver regeneration, while BMAL1 and PER2 may be important for proper neurogenesis in the hippocampus, a brain region in which adult mammals normally exhibit continuous cell division. Altogether, this suggests that the highly conserved circadian clock is important for regulating a wide range of homeostatic processes, including cell-cycle pathways, that when properly phased to each other promote organismal fitness.

Measurements of altered circadian rhythms with age may serve as a useful biomarker for aging. An intriguing question is whether the decline in amplitude of rhythms correlates with a decline in function, and importantly whether restoration of these rhythms with age, through either behavioral or pharmacologic intervention, would delay the aging process. Of note, transplantation of the SCN from a young rat into an old rat "rescued" the rhythms of both locomotor activity and corticotropin hormone (CRH), suggesting that the SCN is an important target for age-related changes in clocks. Treatments targeting the SCN may therefore ameliorate some of the deterioration in aged individuals.

CHRONOTHERAPY AND FUTURE DIRECTIONS

Chronopharmacology, the study of how the timing of drug administration may impact its effectiveness, is a rapidly emerging field. Since physiologic processes vary across the day, the timing of administration of medication may help optimize patient care. For example, since endogenous cholesterol synthesis is rhythmic in liver and peaks during the early morning hours, administration of statins (HMG-CoA reductase inhibitors) in the evening prior to bedtime has proven to be more effective than daytime administration at reducing low-density lipoprotein cholesterol (LDL-C) levels because the highest concentration of the medications coincides with the peak in the rhythmic endogenous cholesterol production. Given that blood pressure exhibits a 24-h rhythm—being lowest during sleep—angiotensin-converting enzyme (ACE) inhibitors have been shown to be most effective at night to normalize the blood pressure rhythms, restoring the nighttime dip in blood pressure. Administration of cancer treatments according to circadian rhythms has also been shown to increase chemotherapy effectiveness while decreasing toxicity in a wide range of drugs. For example, 5-FU works best to treat colorectal cancer when administered at night, a time when the cancerous cells are more vulnerable while normal cells are quiescent and therefore less sensitive. Doxorubicin administration early in the morning to treat ovarian cancer has also been shown to be less toxic, as white blood cells recover faster than if the drug is given in the evening. Finally, the more severe morning symptoms of rheumatoid arthritis are linked to increased inflammation during the evening; therefore, prevention of the night-time upregulation of the immune/inflammatory reaction is more effective when glucocorticoids are administered with a night-time release formulation.

Recognition of circadian rhythms is also critical for diagnoses and treatment of endocrine disorders. The diagnosis of Cushing's syndrome, which is characterized by hypercortisolemia, might be missed if the patient's cortisol levels were measured in the morning, as endogenous cortisol production is typically highest during morning

and lowest at night; therefore, clinical diagnosis requires cortisol to be measured in the late evening when the levels of this hormone should typically be low. On the other hand, adrenal insufficiency is diagnosed by measuring cortisol in the morning when at its physiologic peak, and glucocorticoid therapy for these patients aims to mimic the endogenous rhythms of cortisol, as short-acting synthetic glucocorticoids are usually given several times a day in tapering doses, such that the largest amount is taken in the morning and the smallest in the evening. Diabetes is another endocrine disorder intimately tied to circadian rhythms. Oral glucose tolerance has repeatedly been shown to be impaired in the afternoon and evening compared to the morning, likely due to a combination of a circadian regulation of insulin sensitivity within peripheral tissues and reduced insulin secretion during the night. Similarly, due to a surge in hormone levels in the morning, diabetes patients may suffer from the dawn phenomenon (or dawn effect), an abnormally high morning increase in blood glucose due to impaired response in insulin secretion. A related phenomenon that can be tied to evening timing of insulin doses is the "rebound" or Somogyi effect. In this scenario, the initially noted clinical sign, in the form of elevated glucose levels, may be noted in the morning. However, the underlying cause is hypoglycemia occurring during the night, which produces a counterregulatory hormonal response that subsequently results in morning hyperglycemia. As patients with type 2 diabetes often lack these daily cycles of insulin secretion and glucose tolerance, this further highlights that time of day is an important consideration for the diagnosis and treatment of metabolic disorders such as type 2 diabetes.

As our knowledge of the complexity of the interaction between circadian and physiologic processes deepens, further advances to rationally develop new strategies for treatments of disorders affected by circadian misalignment are essential. For example, novel compounds have begun to emerge from unbiased drug discovery screens that impact circadian clock components, either shortening or lengthening the period, including CRY stabilizers or various inhibitors of CKIδ, CKIε, and GSK-3. Pharmacologic control of the circadian cycle may be useful in the treatment of circadian disorders and metabolic disturbances with a circadian component. Understanding how the circadian clock controls biological functions will shed new light onto the pathogenesis of metabolic disorders with a circadian component, such as type 2 diabetes and metabolic syndrome, and will yield insight into how timing of drug delivery will impact patient care.

Acknowledgment

The authors would like to thank Billie Marcheva for her help with the figures and tables.

FURTHER READING

Bass J, Lazar MA: Circadian time signatures of fitness and disease. Science 354:994, 2016.

Buxton OM et al: Adverse metabolic consequences in humans of prolonged sleep restriction combined with circadian disruption. Sci Transl Med 4:129ra43, 2012.

Dibner C et al: The mammalian circadian timing system: Organization and coordination of central and peripheral clocks. Annu Rev Physiol 72:517, 2010.

Hatori M et al: Time-restricted feeding without reducing caloric intake prevents metabolic diseases in mice fed a high-fat diet. Cell Metabolism 15:848, 2012.

Nea FM et al: Dietary and lifestyle habits and the associated health risks in shift workers. Nutr Res Rev 28:143, 2015.

Perelis M et al: Pancreatic β cell enhancers regulate rhythmic transcription of genes controlling insulin secretion. Science 350:aac4250, 2016.

Scheer FA et al: Adverse metabolic and cardiovascular consequences of circadian misalignment. Proc Natl Acad Sci USA 106:4453, 2009.

Takahashi JS: Transcriptional architecture of the mammalian circadian clock. Nat Rev Genet 18:164, 2016.

Turek FW et al: Obesity and metabolic syndrome in circadian clock mutant mice. Science 308:1043, 2005.

476 Network Medicine: Systems Biology in Health and Disease

Joseph Loscalzo

The field of human biology has progressed over the last three centuries largely as a result of the reductionist approach to the scientific problems that challenge the discipline. Biologists study the experimental response of a variable of interest in a cell or organism while holding all other variables constant. In this way, it is possible to dissect the individual components of a biologic system and assume that a thorough understanding of a specific component (e.g., an enzyme or a transcription factor) will provide sufficient insight to explain the global behavior of that system (e.g., a metabolic pathway or a gene network, respectively). Biologic systems are, however, much more complex than this approach assumes and manifest behaviors that frequently (if not invariably) cannot be predicted from knowledge of their component parts characterized in isolation. Growing recognition of this shortcoming of conventional biologic research has led to the development of a new discipline, *systems biology, which is defined as the holistic study of living organisms or their cellular or molecular network components to predict their response to perturbations*. Concepts of systems biology can be applied readily to human disease and therapy and define the field of *systems pathobiology*, in which genetic or environmental perturbations produce disease and drug perturbations restore normal system behavior.

Systems biology evolved from the field of systems engineering in which a linked collection of component parts constitute a network whose output the engineer wishes to predict. The simple example of an electronic circuit can be used to illustrate some basic systems engineering concepts. All the individual elements of the circuit—resistors, capacitors, transistors—have well-defined properties that can be characterized precisely. However, they can be linked (wired or configured) in a variety of ways, each of which yields a circuit whose response to voltage applied across it is different from the response of every other configuration. To predict the circuit's (i.e., system's) behavior, the engineer must study its response to perturbation (e.g., voltage applied across it) holistically rather than its individual components' responses to that perturbation. Viewed another way, the resulting behavior of the system is greater than (or different from) the simple sum of its parts, and systems engineering utilizes rigorous mathematical approaches to predict these complex, often nonlinear, responses. By analogy to biologic systems, one can reason that detailed knowledge of a single enzyme in a metabolic pathway or of a single transcription factor in a gene network will not provide sufficient detail in context to predict the output of that metabolic pathway or transcriptional network, respectively. Only a systems-based approach will suffice.

It has taken biologists a long time to appreciate the importance of systems approaches to biomedical problems. Reductionism has reigned supreme for many decades, largely because it is experimentally and analytically simpler than holism, and because it has provided insights into biologic mechanisms and disease pathogenesis that have led to successful therapies. However, reductionism cannot solve all biomedical problems. For example, the so-called off-target effects of new drugs that frequently limit their adoption likely reflect the failure of a drug to be studied in holistic context, that is, the failure to explore all possible actions aside from the principal target action for which it was developed. Other approaches to understanding biology are, therefore, clearly needed. With the growing body of genomic, proteomic, and metabolomic data sets in which dynamic changes in the expression of many genes and many metabolites are recorded after a perturbation and with the growth of rigorous mathematical approaches to analyzing those changes, the stage has been set for applying systems engineering principles to modern biology.

Physiologists historically have had more of a (bio)engineering perspective on the conduct of their studies and have been among the first systems biologists. Yet, with few exceptions, they, too, have focused on comparatively simple physiologic systems that are tractable using conventional reductionist approaches. Efforts at integrative modeling of human physiologic systems, as first attempted by Guyton for blood pressure regulation, represent one application of systems engineering to human biology. These dynamic physiologic models often focus on the acute response of a measurable physiologic parameter to a system perturbation, and do so from a classic analytic perspective in which all the conventional physiologic determinants of the output parameter are known and can be modeled quantitatively.

Until recently, molecular systems analysis has been limited owing to inadequate knowledge of the molecular determinants of a biologic system of interest. Although biochemists have approached metabolic pathways from a systems perspective for over 50 years, their efforts have been limited by the inadequacy of key information for each enzyme (K_M, k_{cat}, and concentration) and substrate (concentration) in the pathway. With increasingly rich molecular data sets available for systems-based analyses, including genomic, transcriptomic, proteomic, and metabolomic data, molecular biologists and biochemists are now poised to use systems biology approaches to explore biologic and pathobiologic phenomena.

PROPERTIES OF COMPLEX BIOLOGIC SYSTEMS

To understand how best to apply the principles of systems biology to human biomedicine, it is necessary to review briefly the building blocks of any biologic system and the determinants of system complexity. All systems can be analyzed by defining their static topology (architecture) and their dynamic (i.e., time-dependent) response to perturbation. In the discussion that follows, system properties are described that derive from the consequences of topology (form) or dynamic response (function). Any system of interacting elements can be represented schematically as a *network* in which the individual elements are depicted as nodes and their connections are depicted as links. The nature of the links among nodes reflects the degree of complexity of the system. *Simple systems* are those in which the nodes are linearly linked with occasional feedback or feedforward loops modulating system throughput in highly predictable ways. By contrast, *complex systems* are nodes that are linked in more complicated, nonlinear networks; the behavior of these systems by definition is inherently more difficult to predict owing to the nature of the interacting links, the dependence of the system's behavior on its initial conditions, and the inability to measure the overall state of the system at any specific time with great precision. Complex systems can be depicted as a network of lower-complexity interacting components or modules, each of which can be reduced further to simpler analyzable canonical motifs (such as feedback and feedforward loops, or negative and positive autoregulation); however, a central property of complex systems is that simplifying their structures by identifying and characterizing the individual nodes and links or even simpler substructures does not necessarily yield a predictable understanding of a system's behavior. Thus, the functioning system is greater than (or different from) the sum of its individual, tractable parts.

Defined in this way, most biologic systems are complex systems that can be represented as networks whose behaviors are not readily predictable from simple reductionist principles. The nodes, for example, can be metabolites that are linked by the enzymes that cause their transformations, transcription factors that are linked by the genes whose expression they influence, or proteins in an interaction network that are linked by cofactors that facilitate interactions or by thermodynamic forces that facilitate their physical association. Biologic systems typically are organized as *scale-free*, rather than stochastic, networks of nodes. Scale-free networks are those in which a few nodes have many links to other nodes (highly linked nodes, or hubs), but most nodes have only a few links (weakly linked nodes). The term *scale-free* refers to the fact that the connectivity of nodes in the network is invariant with respect to the size of the network. This is quite different from two other common network architectures: random (Poisson) and exponential distributions. Scale-free networks can be mathematically described by a power law that defines the probability of the number of links per node ($P(k) = k^{-\gamma}$, where k is the number of links per node and γ is the slope of the log P(k) versus log(k) plot); this unique property of most biologic networks is a reflection of their self-similarity or fractal nature **(Fig. 476-1)**.

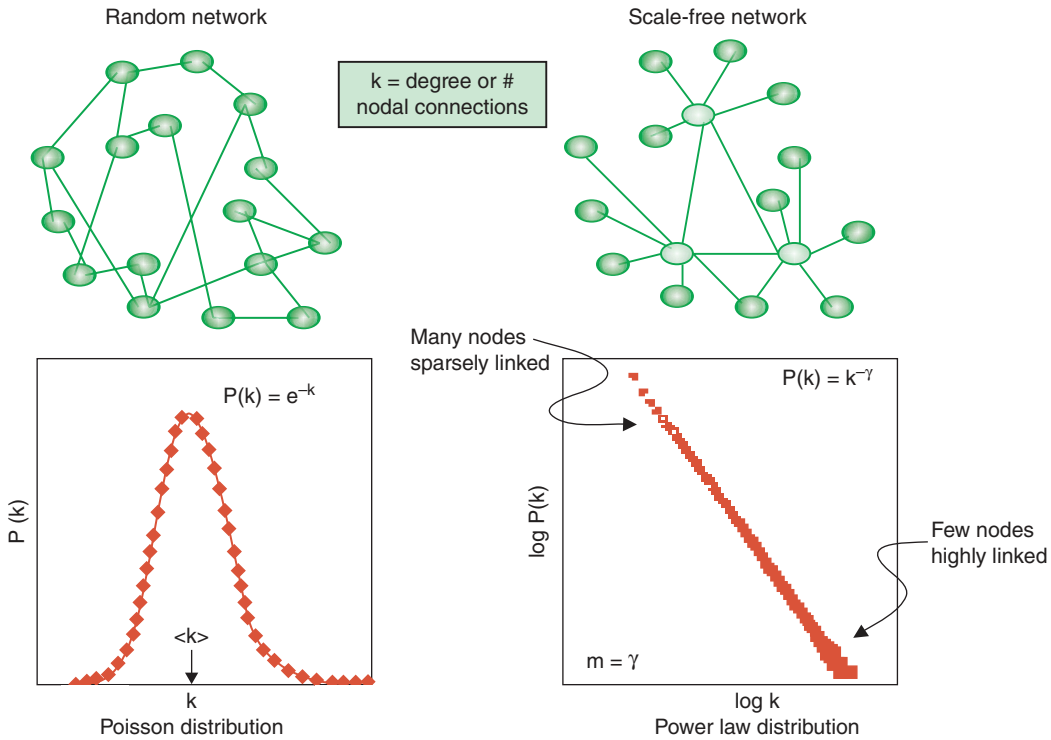


FIGURE 476-1 Network representations and their distributions. A random network is depicted on the left, and its Poisson distribution of the number of nodal connections (k) is shown in the graph below it. A scale-free network is depicted on the right, and its power law distribution of the number of nodal connections (k) is shown in the graph below it. Highly connected nodes (hubs) are lightly shaded.

There are unique properties of scale-free biologic systems that reflect their evolution and promote their adaptability and survival. Biologic networks likely evolved one node at a time in a process in which new nodes are more likely to link to a highly connected node than to a sparsely connected node. Furthermore, scale-free networks can become sparsely linked to one another, yielding more complex, *modular scale-free topologies*. This evolutionary growth of biologic networks has three important properties that affect system function and survival. First, this scale-free addition of new nodes promotes *system redundancy*, which minimizes the consequences of errors and accommodates adverse perturbations to the system robustly with minimal effects on critical functions (unless the highly connected nodes are the focus of the perturbation). Second, this resulting network redundancy provides a survival advantage to the system. In complex gene networks, for example, mutations or polymorphisms in weakly linked genes account for biodiversity and biologic variability without disrupting the critical functions of the system; only mutations in highly linked (*essential*) genes (hubs) can shut down the system and cause embryonic lethality. Third, scale-free biologic systems facilitate the flow of information (e.g., metabolite flux) across the system compared with randomly organized biologic systems; this so-called "small-world" property of the system (in which the clustered nature of the highly linked hubs defines a local neighborhood within the network that communicates through weaker, less frequent links to other clusters) minimizes the energy cost for the dynamic action of the system (e.g., minimizes the transition time between states in a metabolic network).

These basic organizing principles of complex biologic systems lead to three unique properties that require emphasis. First, biologic systems are *robust*, which means that they are quite stable in response to most changes in external conditions or internal modification. Second, a corollary to the property of robustness is that complex biologic systems are *sloppy*, which means that they are insensitive to changes in external conditions or internal modification except under certain uncommon conditions (i.e., when a hub is involved in the change). Third, complex biologic systems exhibit *emergent properties*, which means that they manifest behaviors that cannot be predicted from the reductionist principles used to characterize their component parts. Examples of emergent behavior in biologic systems include spontaneous, self-sustained oscillations in glycolysis; spiral and scroll waves of depolarization in cardiac tissue that cause reentrant arrhythmias; and self-organizing patterns in biochemical systems governed by diffusion and chemical reaction.

APPLICATIONS OF SYSTEMS BIOLOGY TO PATHOBIOLOGY

The principles of systems biology have been applied to complex pathologic processes with some early successes. The key to these applications is the identification of emergent properties of the system under study in order to define novel, otherwise unpredictable (i.e., from the reductionist perspective) methods for regulating the system's response. Systems biology approaches have been used to characterize epidemics and ways to control them, taking advantage of the scale-free properties of the network of infected individuals that constitute the epidemic. Through the use of a systems analysis of a neural protein-protein interaction network, unique disease-modifying proteins have been identified that are common to a wide range of cerebellar neurodegenerative disorders causing inherited ataxias. Systems analysis and disease network construction of a pulmonary arterial hypertension network led to the identification of a unique disease module involving a pathway governed by microRNA-21. Systems biology models have been used to dissect the dynamics of the inflammatory response using oscillatory changes in the transcription factor nuclear factor (NF) κB as the system output. Systems biology principles also have been used to predict the development of an idiotypy–anti-idiotypy antibody network, describe the dynamics of species growth in microbial biofilms, and analyze the innate immune response. In each of these examples, a systems (patho) biology approach provided insights into the behavior of these complex systems that could not have been recognized with conventional scientific reductionism.

A unique application of systems biology to biomedicine is in the area of drug development. Conventional drug development involves identifying a potential target protein and then designing or screening compounds to identify those that inhibit the function of that target. This reductionist analysis has identified many potential drug targets and drugs, yet only when a drug is tested in animal models or humans are the systems consequences of the drug's action revealed; not uncommonly, so-called off-target effects may become apparent and be sufficiently adverse for researchers to cease development of the agent. A good example of this problem is the unexpected outcomes of the vitamin B–based regimens for lowering homocysteine levels. In these trials, plasma homocysteine levels were reduced effectively; however, there was no effect of this reduction on clinical vascular endpoints. One explanation for this outcome is that one of the B vitamins in the regimen, folate, has a panoply of effects on cell proliferation and metabolism that probably offset its homocysteine-lowering benefits, promoting progressive atherosclerotic plaque growth and its consequences for clinical events. In addition to these types of unexpected outcomes exerted through pathways that were not considered *ab initio*, conventional approaches to drug development typically do not take into consideration the possibility of emergent behaviors of the organism or the metabolic pathway or the transcriptional network of interest. Thus, a systems-based analysis of potential drugs (drug-target network analysis) can benefit the development paradigm both by enhancing the likelihood that a compound of interest will not manifest unforeseen adverse effects and by promoting novel analytic methods for identifying unique control points or pathways in metabolic or genetic networks that would benefit from drug-based modulation, including drug combinations.

SYSTEMS PATHOBIOLOGY AND HUMAN DISEASE CLASSIFICATION: NETWORK MEDICINE

Perhaps most important, systems pathobiology can be used to revise and refine the definition of human disease. The classification of human disease used in this and all medical textbooks derives from the correlation between pathologic analysis and clinical syndromes that began in the nineteenth century. Although this approach has been very successful, serving as the basis for the development of many effective therapies, it has major shortcomings. Those shortcomings include a lack of sensitivity in defining preclinical disease, a primary focus on overtly manifest disease, failure to recognize different and potentially differentiable causes of common late-stage pathophenotypes, and a limited ability to incorporate the growing body of molecular and genetic determinants of pathophenotype into the conventional classification scheme.

Two examples will illustrate the weakness of simple correlation analyses grounded in the reductionist principle of simplification (Occam's razor) in defining human disease. Sickle cell anemia, the "classic" Mendelian disorder, is caused by a Val6Gln substitution in the β chain of hemoglobin. If conventional genetic teaching holds, this single mutation should lead to a single phenotype in patients who harbor it (genotype-phenotype correlation). This assumption is, however, false, as patients with sickle cell disease manifest a variety of pathophenotypes, including hemolytic anemia, stroke, acute chest syndrome, bony infarction, and painful crisis, as well as an overtly normal phenotype. The reasons for these different phenotypic presentations include the presence of disease-modifying genes or gene products (e.g., hemoglobin F, hemoglobin C, glucose-6-phosphate dehydrogenase), exposure to adverse environmental factors (e.g., hypoxia, dehydration), and the genetic and environmental determinants of common intermediate pathophenotypes or endophenotypes (i.e., variations in those generic pathologic mechanisms underlying all human disease—inflammation, thrombosis/hemorrhage, fibrosis, cell proliferation, apoptosis/necrosis, immune response).

A second example of note is familial pulmonary arterial hypertension. This disorder is associated with over 100 different mutations in three members of the transforming growth factor β (TGF-β) superfamily alone: bone morphogenetic protein receptor-2 (BMPR-2), activin

receptor-like kinase-1 (Alk-1), and endoglin. All these different genotypes are associated with a common pathophenotype, and each leads to that pathophenotype by molecular mechanisms that range from haploinsufficiency to dominant negative effects. As only approximately one-fourth of individuals in families that harbor these mutations manifest the pathophenotype, other disease-modifying genes (e.g., the serotonin receptor 5-HT2B, the serotonin transporter 5-HTT), genomic and environmental determinants of common intermediate pathophenotypes, and environmental exposures (e.g., hypoxia, infective agents [HIV], anorexigens) probably account for the incomplete penetrance of the disorder.

On the basis of these and many other related examples, one can approach human disease from a systems pathobiology perspective in which each "disease" can be depicted as a network that includes the following modules: the primary disease-determining elements of the genome (or proteome, if posttranslationally modified), the disease-modifying elements of the genome or proteome, environmental determinants, and genomic and environmental determinants of the generic intermediate pathophenotypes. **Figure 476-2** graphically depicts these genotype-phenotype relationships as modules for the six common disease types with specific examples for each type. **Figure 476-3** shows a network-based depiction of sickle cell disease using this kind of modular approach.

Goh and colleagues developed the concept of a human disease network **(Fig. 476-4)** in which they used a systems approach to characterize the disease-gene associations listed in the Online Mendelian Inheritance in Man database. Their analysis showed that genes linked to similar disorders are more likely to have products that physically associate and greater similarity between their transcription profiles than do genes not associated with similar disorders. In addition, proteins associated with the same pathophenotype are significantly more likely to interact with one another than with other proteins not associated with the pathophenotype. Finally, these authors showed that the great majority of disease-associated genes are not highly connected genes (i.e., not hubs) and are typically weakly linked nodes within the functional periphery of the network in which they operate.

This type of analysis validates the potential importance of defining disease on the basis of its systems pathobiologic determinants. Clearly, doing this will require a more careful dissection of the molecular elements in the relevant pathways (i.e., more precise molecular pathophenotyping), less reliance on overt manifestations of disease for their classification, and an understanding of the dynamics (not just the static architecture) of the pathobiologic networks that underlie pathophenotypes defined in this way. **Figure 476-5** illustrates the elements of a molecular network within which a disease module is contained. This network is first identified by determining the interactions (physical or regulatory) among the proteins or genes that comprise it (the "interactome"). These interactions then define a topologic module within which exist functional modules (pathways) and disease modules. One approach to constructing this module is illustrated in **Fig. 476-6**. Examples of the use of this approach in defining novel determinants of disease are given in **Table 476-1**.

As yet another potential consideration, one can argue that disease reflects the later-stage consequences of the predilection of an organ

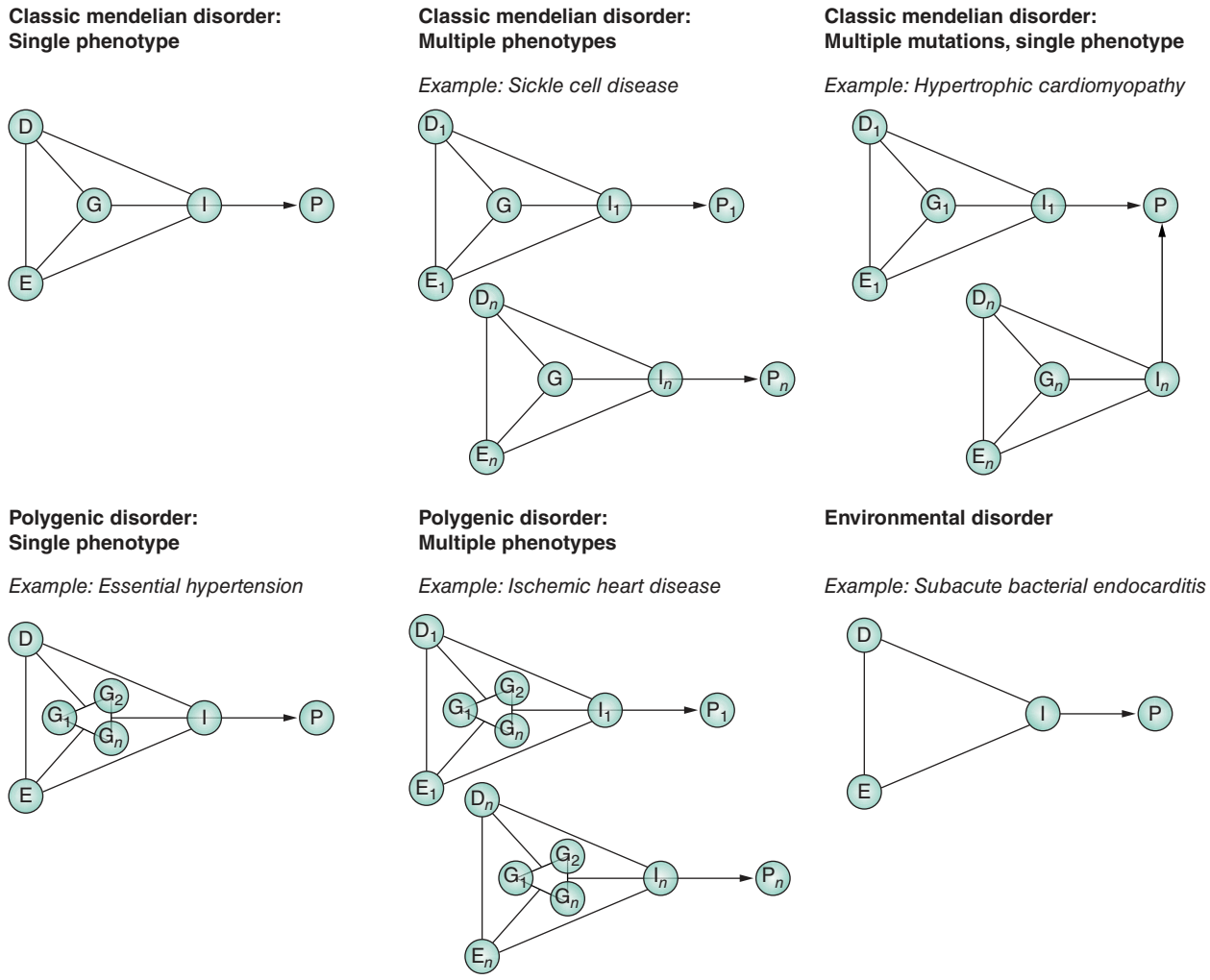


FIGURE 476-2 Examples of modular representations of human disease. D, secondary human disease genome or proteome; E, environmental determinants; G, primary human disease genome or proteome; I, intermediate phenotype; P, pathophenotype. *(Reproduced with permission from J Loscalzo et al: Molec Syst Biol 3:124, 2007.)*

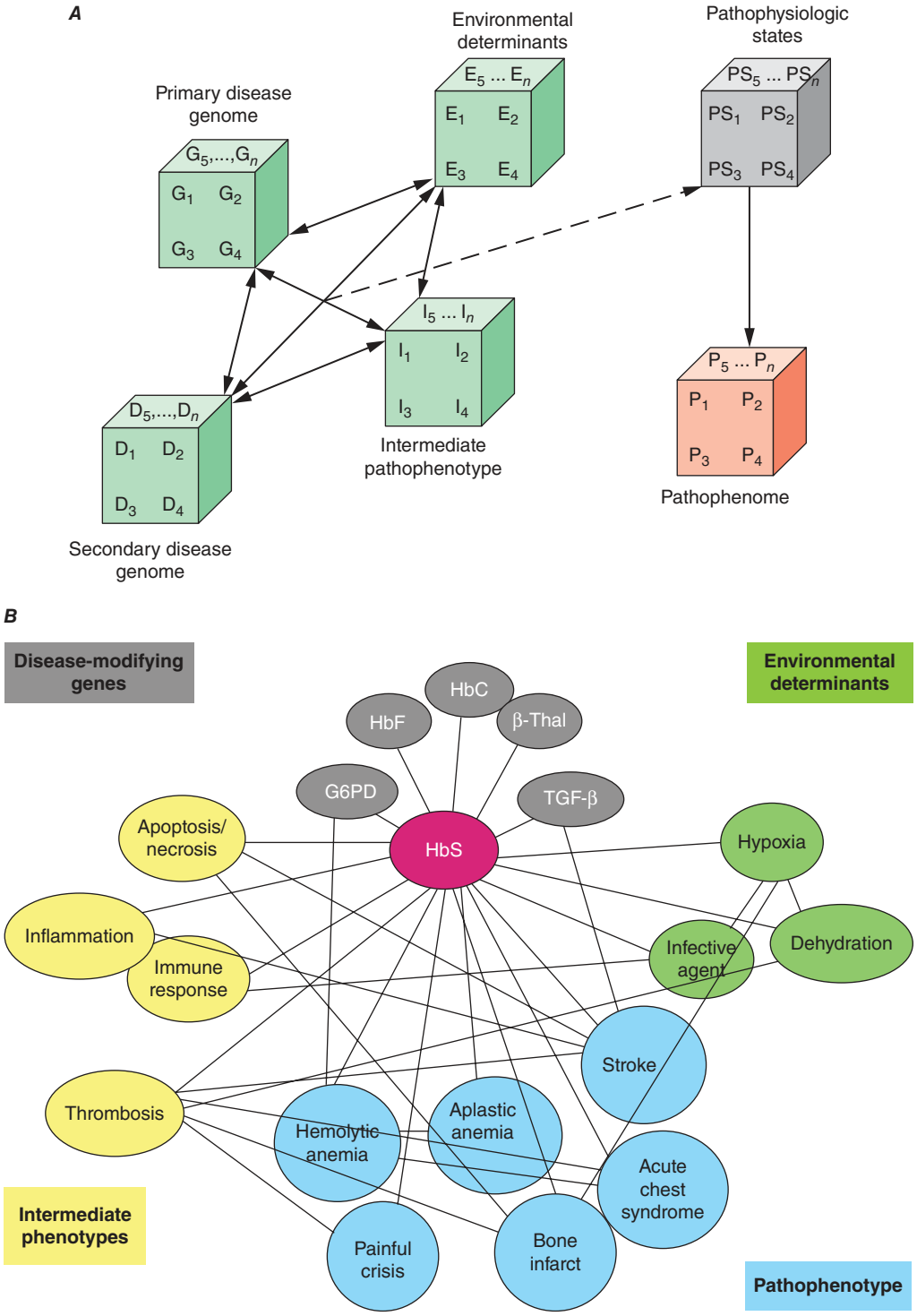


FIGURE 476-3 ***A.*** Theoretical human disease network illustrating the relationships among genetic and environmental determinants of the pathophenotypes. Key: D, secondary disease genome or proteome; E, environmental determinants; G, primary disease genome or proteome; I, intermediate phenotype; PS, pathophysiologic states leading to P, pathophenotype. ***B.*** Example of this theoretical construct applied to sickle cell disease. Key: Red, primary molecular abnormality; gray, disease-modifying genes; yellow, intermediate phenotypes; green, environmental determinants; blue, pathophenotypes. *(Reproduced with permission from J Loscalzo et al: Molec Syst Biol 3:124, 2007.)*

system to manifest a particular intermediate pathophenotype in response to injury. This paradigm reflects a reverse causality view in which a disease is defined as a tendency to heightened inflammation, thrombosis, or fibrosis after an injurious perturbation. Where the process is manifest (i.e., the organ in which it occurs) is less important than that it occurs (with the exception of the organ-specific pathophysiologic consequences that may require acute attention). For example, from this perspective, acute myocardial infarction (AMI) and its consequences are a reflection of thrombosis (in the coronary artery), inflammation (in the acutely injured myocardium), and fibrosis (at the site or sites of cardiomyocyte death). In effect, the major therapies for AMI address these intermediate pathophenotypes (e.g., antithrombotics, statins) rather than any organ-specific disease-determining process. This paradigm would argue for a systems-based analysis that would first identify the intermediate pathophenotypes to which a person is predisposed, then determine how and when to intervene to attenuate that adverse predisposition, and finally limit the likelihood that a major organ-specific event will occur. Evidence for the validity of this approach is found in the work of Rzhetsky and colleagues, who reviewed 1.5 million patient records and 161 diseases and found that

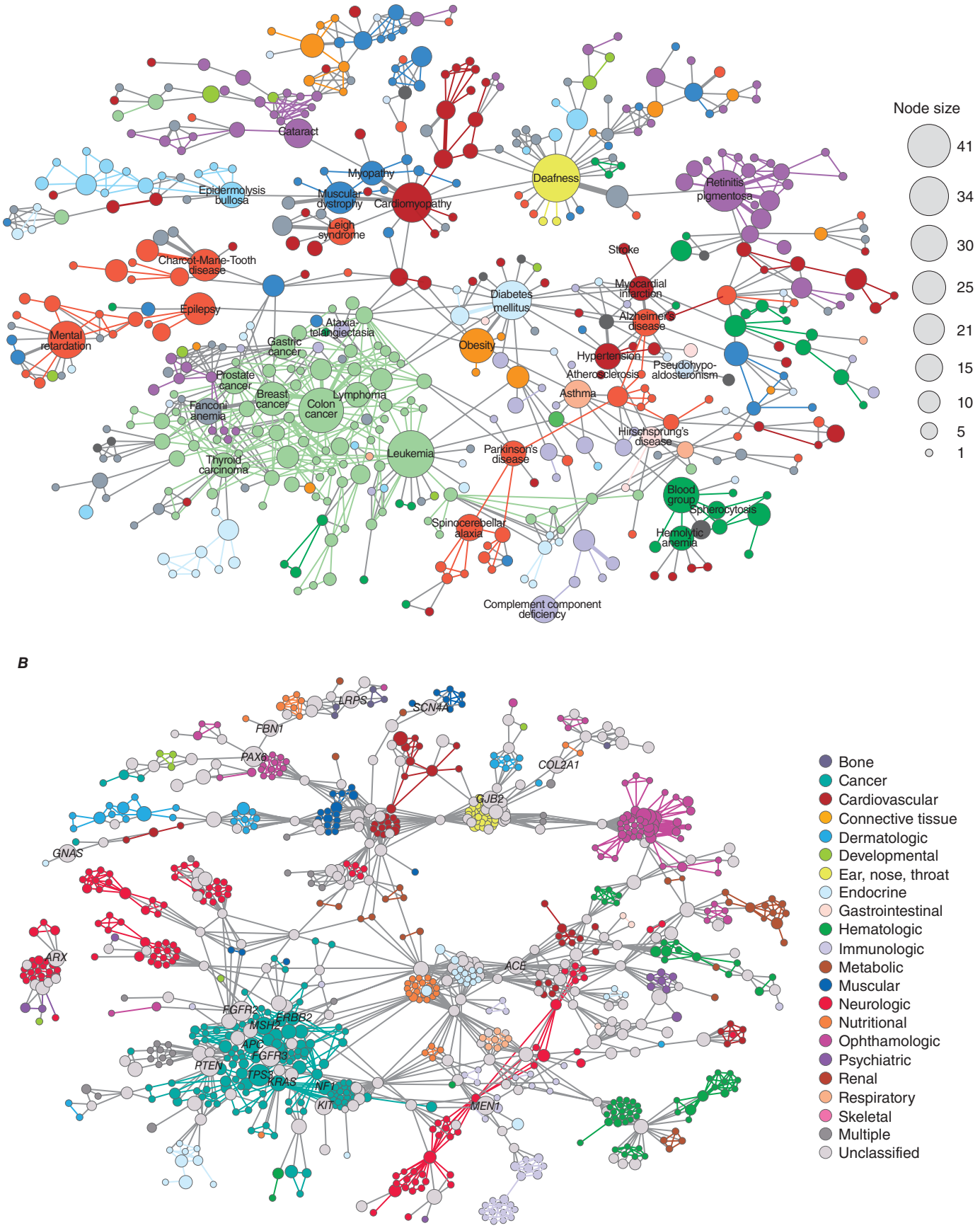


FIGURE 476-4 ***A.*** Human disease network. Each node corresponds to a specific disorder colored by class (22 classes, shown in the key to **B**). The size of each node is proportional to the number of genes contributing to the disorder. Edges between disorders in the same disorder class are colored with the same (lighter) color, and edges connecting different disorder classes are colored gray, with the thickness of the edge proportional to the number of genes shared by the disorders connected by it. ***B.*** Disease gene network. Each node is a single gene, and any two genes are connected if implicated in the same disorder. In this network map, the size of each node is proportional to the number of specific disorders in which the gene is implicated. *(Reproduced with permission from KI Goh et al: Proc Natl Acad Sci USA 104:8685, 2007.)*

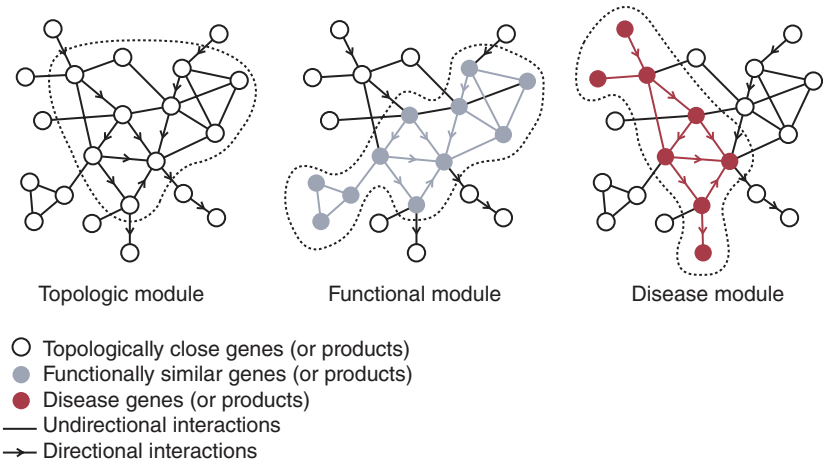


FIGURE 476-5 The elements of the interactome. The interactome includes topologic modules (genes or gene products that are closely associated with one another through direct interactions), functional modules (genes or gene products that work together to define a pathway), and disease modules (genes or gene products that interact to yield a pathophenotype). *(Reproduced with permission from AL Barabasi et al: Nat Rev Genet 12:56, 2011.)*

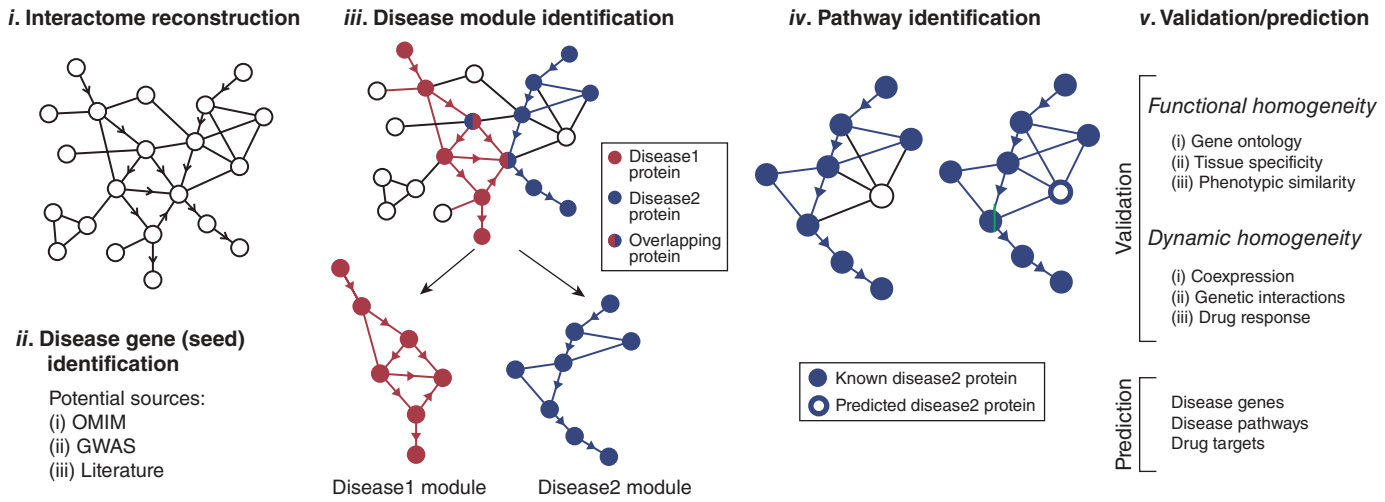


FIGURE 476-6 Approaches to identifying disease modules in molecular networks. A strategy for defining disease modules involves (i) reconstructing the interactome; (ii) ascertaining potential seed (disease) genes from the curated literature, the Online Mendelian Inheritance in Man (OMIM) database, or genomic analyses (genome-wide association studies [GWAS] or transcriptional profiling); (iii) identifying the disease module using different modeling or statistical approaches; (iv) identifying pathways and the role of disease genes or modules in those pathways; and (v) disease module validation and prediction. *(Reproduced with permission from AL Barabasi et al: Nat Rev Genet 12:56, 2011.)*

TABLE 476-1 Examples of Systems Biology Application to Disease

DISEASE	ANALYSIS	REFERENCE
Hereditary ataxias	Many ataxia-causing proteins share interacting partners that affect neurodegeneration	Lim et al: Cell 125:801–814, 2006
Diabetes mellitus	Metabolite-protein network analysis links three unique metabolite abnormalities in prediabetics to seven type 2 diabetes genes through four enzymes	Wang-Sattler et al: Mol Syst Biol 8:615, 2012
Ebstein-Barr virus infection	Viral proteome exerts its effects through linking to host interactome	Gulbahce et al: PLoS One 8:e1002531, 2012
Pulmonary arterial hypertension	Network analysis indicates adaptive role for microRNA 21 in suppressing rho kinase pathway	Parikh et al: Circulation 125:1520–1532, 2012

these disease phenotypes form a network of strong pairwise correlations. This result is consistent with the notion that underlying genetic predispositions to intermediate pathophenotypes form the predicate basis for conventionally defined end organ diseases.

Regardless of the specific nature of the systems pathobiologic approach used, these analyses will lead to a drastic revision of the way human disease is defined and treated, establishing the discipline of *network medicine, which reflects a fusion of the fields of systems biology and network science in the study of disease*. This will be a lengthy and complicated process, but ultimately will lead to better disease prevention and therapy and probably do so from an increasingly personalized perspective. The analysis of pathobiology from a systems-based perspective is likely to help define specific subsets of patients more likely to respond to particular interventions based on shared disease mechanisms. Although it is unlikely that the extreme of "individualized medicine" will ever be practical (or even desirable), complex diseases can be mechanistically subclassified and interventions may be tailored to those settings in which they are more likely to work. This approach serves as a basis for the development of precision medicine.

FURTHER READING

Barabasi A-L, Gulbahce N, Loscalzo J: Network medicine: A network-based approach to human disease. Nat Rev Genet 12:56, 2011.

Loscalzo J, Barabási A-L, Silverman EK (eds): *Network Medicine: Complex Systems in Human Disease and Therapeutics*. Cambridge, Massachusetts, Harvard University Press. Copyright © 2017 by the President and Fellows of Harvard College. All rights reserved.

Loscalzo J, Kohane I, Barabas A-L: Human disease classification in the postgenomic era: A complex systems approach to human pathobiology. Mol Syst Biol 3:124, 2007.

Menche J et al: Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science 347:1257601, 2015.

Parikh VN et al: MicroRNA-21 integrates pathogenic signaling to control pulmonary hypertension: Results of a network bioinformatics approach. Circulation 125:1520, 2012.

477 Emerging Neurotherapeutic Technologies

Jyoti Mishra, Karunesh Ganguly

Neurotherapeutic technologies represent a diverse group of very promising treatment approaches with a common purpose of improving neurological function. Decades of basic science research has paved the path for these novel technologies that have the potential to transform the lives of patients with neurological diseases. A key goal is to minimize the consequences of lost abilities, whether it is motor, sensory, or cognitive. A common objective is to also harness the inherent plasticity of the nervous system, regardless of age, and even in the face of a degenerative process.

The technologies described below are the culmination of both an increased understanding of neural plasticity mechanisms in both the intact and the injured nervous system as well as advances in technology and computational power. There has been important progress in understanding neural plasticity at the level of the microscale (e.g., cellular and molecular processes), the mesoscale (e.g., between distinct cortical and subcortical areas), and the macroscale (e.g., at the level of brain networks). While it is also clear that there may be fundamental limits on plasticity (e.g., the closing of developmental windows) and repair mechanisms, the brain remains highly plastic regardless of age and even in the face of ongoing injury and/or degenerative processes. Collectively, there is now growing evidence to support neurological restorative efforts for both "static" (e.g., stroke) and progressive neurological disorders.

Importantly, while these technologies may not appear, at first glance, directly relevant to traditional medical care, it is worth noting that clinicians have the most knowledge and experience about the specific disease process, the treatments available, and the expected course of illnesses affecting the nervous system. It is thus critical that neurologic specialists and other clinicians can and should play an important role in the future adoption of these technologies for neurological rehabilitation. The sections below outline emerging diagnostic and therapeutic approaches that have the potential to transform the lives of patients with neurological disorders. These include technologies to harness plasticity, neuroimaging, neurostimulation, and brain–machine interfaces (BMIs).

TECHNOLOGIES TO HARNESS PLASTICITY

Neurological rehabilitation aims to harness activity-dependent plasticity mechanisms to maximize functional restoration. This principle can be applied to a diverse range of functional domains such as movement control, sensory processing, language, pain, and cognition. For example, recent randomized controlled clinical trials for motor recovery after stroke have suggested that intensity of training may be particularly important for sustained long-term improvements. Moreover, studies of the effects of such training in rodent and nonhuman primate models further suggest that plasticity of cortical "motor maps" might underlie the observed functional improvements. The incorporation of technology for neurological rehabilitation has the great potential to revolutionize the delivery of care by significantly increasing access, reducing the burden for adherence to high intensity regimens, and by maximizing engagement. Below are three examples of how emerging technology can be used to harness neural plasticity and to maximize functional restoration.

ROBOTICS

Rehabilitation robotics for both the upper and the lower limb have the potential to improve motor outcomes after stroke or other forms of brain injury. There is a growing recognition that focused training involving a range of tasks might be important for improved functional outcomes. While it remains unclear exactly when such training might be optimal after the initial stroke and during the early recovery period, such training likely has a role in both the acute and the chronic periods after stroke; maintenance therapy may also provide a guard against observed declines in function over time. Notably, the delivery of intensive training is a great challenge from both the perspective of the health care system and each patient. Outside of clinical trials, such a training program can be quite difficult to implement and maintain. It can also be costly and require significant effort.

Motor rehabilitation using robotics has been developed and tested for both the upper-limb and the lower-limb. Such robotic therapies have often focused on the delivery of high-intensity movement practice that can surpass what is possible via existing standards of care. Moreover, the robotic systems are capable of precisely measuring movement parameters (e.g., the kinematics of the movements) and providing quantitative feedback regarding the changes in performance during the training period. A particular focus has been on maximal patient engagement and recruitment of attentional and reward pathways, both of which are increasingly recognized to drive neural plasticity. Continued advances in design and the user interface will ensure maximal comfort and sustained effort. For example, via close monitoring of performance and movement parameters, the system can provide assistance at key points in order to minimize fatigue and to ensure maximal engagement. Moreover, antigravity support of the upper-limb can allow practice and task engagement even in the presence of severe weakness; this would be extremely challenging and labor intensive under current standards of care. Recent analysis also suggests that robotic devices may at least match outcomes realized with existing standards of care. However, rehabilitation robotics may also provide more precise feedback and permit novel quantitative rehabilitation approaches.

FIGURE 477-1 Photograph of a subject interacting with a complex upper-limb exoskeleton and a virtual reality system. *(From U Keller et al: Robot-Assisted Arm Assessments in Spinal Cord Injured Patients: A Consideration of Concept Study. PLoS One 10:e0126948, 2015.)*

Figure 477-1 shows one example of an upper-limb robotic exoskeleton device that is currently being evaluated for training after stroke. A recent randomized, multicenter trial compared treatment with this exoskeleton system against conventional therapy provided by physical and occupational therapists. Participants were enrolled in the chronic phase and all had moderate-to-severe deficits; the groups underwent three sessions per week over an 8-week period. For robotic training, subjects trained with games to improve mobilization and to practice activities of daily living. This study provided evidence that both conventional and robotic therapy could improve function in patients with chronic stroke. Multiple studies have also found similar gains when using either conventional or traditional approaches. Thus, a growing body of research supports the idea that such devices might complement current conventional approaches to rehabilitation. Future work will need to define how rehabilitation robotics can optimally use adaptive and quantitative methods to further augment the recovery process.

VIRTUAL AND AUGMENTED REALITY

Therapeutic approaches using Virtual Reality (VR) and Augmented Reality (AR) aim to treat neurological illnesses by specifically and quantitatively altering a patient's subjective experiences and interactions with the environment. Core components of both are advanced hardware and computational methods to generate simulated, yet realistic, perceptions. While some applications permit users to dynamically change the viewed perspective, other applications are designed to allow interactions among multiple users. Visual feedback is often a key component; this can include simple computer monitors or more immersive "head mounted" viewers that modify the simulation based on changes in perspective. Tracking of movements (e.g., hand and head position) are often included. Multiple methods are used to allow a user to interact with the environment. For example, interactions can be guided by straightforward means such as a keyboard, mouse or even a joystick. More immersive methods are also frequently used. For example, gloves with embedded sensors and haptic inputs can allow the user's hand to be represented in real time in the simulated environment. Moreover, haptic interfaces can provide sensory feedback, allowing patients to interact and "feel" virtual objects through multiple sensory modalities. A particular strength of these approaches is that any therapeutic intervention can be studied in very controlled environments.

VR enables a user to interact with a simulated reality that can be precisely and quantitatively controlled. In addition to allowing patients to dynamically experience an altered reality, it can simultaneously monitor a subject's behaviors and responses. Such monitoring can allow both precise measurements of clinically relevant parameters (e.g., motor actions, perception, cognitive processing) and for specific rehabilitation training that can attain clinically relevant goals. A growing body of literature indicates that VR environments can be tailored to individual needs and preferences, thereby maximizing engagement, motivation, and adaptation to ensure sufficient difficulty of tasks. VR environments can be designed to create powerful "gaming" platforms that are actually targeting clinically relevant parameters. For example, the upper-limb robotic systems described previously are frequently combined with VR environments that allow interaction with virtual objects.

In contrast to VR, AR overlays an artificial filter over a subject's view of the actual physical world, thus providing an "augmented" or enhanced view of the world around. AR is being tested in a diverse group of patients with neurological impairments in either the motor, sensory, or cognitive domains. AR may offer a particularly unique rehabilitation intervention for stroke patients. It is widely known that brain injuries limit each patient's physical interaction with their environment. Furthermore, physical and cognitive impairments may limit social interactions. Such impoverished experiences are likely to be present during both the acute and the chronic phases. Importantly, there is clear basic science evidence that environmental enrichment can be a key component of rehabilitation; such enrichment may offer additive benefits to the often limited formal rehabilitation sessions per week. Consistent with this are clinical studies suggesting that motor and cognitive outcomes may suffer when interactions with the environment are reduced. AR may be capable of increasing enrichment. For example, in the case of spatial neglect after stroke, the impaired modality may be accounted for using AR methods. Similarly, physical impairments that limit walking speeds can also limit visual feedback; both AR and VR can be used to enhance visual feedback during gait training.

Figure 477-2 shows a recent innovative application of AR for the treatment of "phantom limb" pain. A subset of both upper-limb and lower-limb amputees experience painful sensations that appear to originate from the missing limb. Past research has suggested that mirror therapy can be an effective treatment for phantom limb pain. During mirror therapy treatments, patients move their healthy arm in front of a mirror in order to produce a perception of movements of the missing limb. Previous studies have suggested that maladaptive plasticity of affected sensory cortices may be treated with mirror therapy. Importantly, in comparison to mirror therapy, AR-based therapy for phantom limb pain can be based on movements of the affected limb, i.e., the remaining limb portion as opposed to using the unaffected contralateral limb. This study demonstrated a novel treatment in which "phantom motor execution" is enabled using sophisticated machine learning algorithms. More specifically, the study "decoded" phantom limb movements by measuring electromyogram (EMG) activity at the stump. Importantly, while the distal muscles responsible for movements were lost as a result of amputation, the remaining EMG activity could be used to predict presumed distal limb movements. As shown in the figure, these inferred movements were projected onto an AR screen to create the perception of limb movements. The study showed that a subset of patients with long-term refractory phantom limb pain could experience a significant reduction in pain levels after using the AR system.

NEUROGAMING

Computerized programs that harness the power of "video games" have shown some evidence for ameliorating deficits in visual perception, age-related degeneration, and neuropsychiatric disorders. An essential feature of effective video game training is the progressive adjustment of the level of difficulty in line with the cognitive improvement of the patient. Important areas of active research include ways to enhance sustainability of neurogame training over long time periods and improving training *transfer*, i.e., the generalizability of task-specific training in one cognitive domain to more broad-based functional improvements. By leveraging video game technology, neurogames allow for dynamic user interaction and maintain user-engagement over multiple sessions over several days of training. Important game mechanics include repetitive practice, performance-adaptive challenges, and several layers of reward feedback—from moment-to-moment point rewards to reward milestones over multiple sessions.

FIGURE 477-2 Augmented reality (AR) for phantom limb pain. ***A.*** A patient is shown live Augmented Reality (AR) video. ***B.*** Electromyogram (EMG) electrodes placed over the stump record muscle activation during training. ***C.*** The patient matches target postures during rehabilitation. ***D.*** Patient playing a game in which a car is controlled by "phantom movements." *(From M Ortiz-Catalan et al: Phantom motor execution facilitated by machine learning and augmented reality as treatment for phantom limb pain: A single group, clinical trial in patients with chronic intractable phantom limb pain. Lancet 388:2885, 2016.)*

Notably, neurogames have therapeutic potential as they can be targeted to specific neurocognitive deficits. For instance, games have shown significant benefits in aging, by targeting speed of processing and training the abilities to multitask and suppress distractions. In each case, selective targeting is achieved by focusing the adaptive challenges to the neurocognitive domain of interest. Duration of response time windows available to the user or the level of interference is selectively targeted in the case of speed of processing training and interference training, respectively. In more recent research, it was demonstrated that it is possible to engender focused circuit neuroplasticity using such selective targeting in neurogaming. For example, older adults learned to adaptively perform within progressively more challenging distractor environments. Neuroplasticity selective to distractor processing was evidenced in this study at both the microscale, i.e., at the resolution of single neuron spiking in sensory cortex, as well as macroscale, i.e., EEG (electroencephalography) based event-related potential recordings.

Video games have also shown promise in the treatment of visual deficits such as amblyopia and in cognitive remediation in neuropsychiatric disorders such as schizophrenia. However, while the evidence base has been encouraging in small sample randomized controlled studies (RCTs), larger RCTs are needed to demonstrate definitive therapeutic benefit. This is especially necessary as the commercial *brain training* industry continues to make unsubstantiated claims of the benefits of neurogaming; such claims have been formally dismissed by the scientific community. Like any other pharmacological or device-based therapy, neurogames need to be systematically validated in multiphase RCTs establishing neural target engagement, and documenting cognitive and behavioral outcomes in specific disorder populations.

Generalizability of training benefits from task-specific cognitive outcomes to more broad-based functional improvements remains the holy grail of neurogaming. Next-generation neurogames will aim to integrate physiological measures such as heart rate variability (an index of physical exertion), galvanic skin responses and respiration rate (indices of stress response), and even EEG-based neural measures. The objectives of such multimodal biosensor integration are to enhance the "closed-loop mechanics" that drive game adaptation and hence improve therapeutic outcomes and perhaps result in greater generalizability. These complex, yet potentially more effective, neurogames of the future will need rigorous clinical study for demonstration of validity and efficacy.

NEUROIMAGING

NEUROIMAGING OF CONNECTIVITY

Multimodal neuroimaging methods including fMRI (functional magnetic resonance imaging), EEG, and MEG (magnetoencephalography) are now being investigated as tools to study functional connectivity between brain regions, i.e., extent of correlated activity between brain regions of interest. Snapshots of functional connectivity can be analyzed while an individual is engaged in specific cognitive tasks or during rest. Resting state functional connectivity (rsFC) is especially attractive as a robust, task-independent measure of brain function that can be evaluated in diverse neurological and neuropsychiatric disorders. In fact, methodological research has shown that rs-fMRI can provide more reliable brain signals of energy consumption than specific task-based fMRI approaches.

In recent years, there has been a surge of research to identify robust rsFC-based biomarkers for specific neurological and neuropsychiatric disorders and thereby inform diagnoses, and even predict specific treatment outcomes. For many such disorders, the network level neurobiological substrates that correspond to the clinical symptoms are not known. Furthermore, many are not unitary diseases, but rather heterogeneous syndromes composed of varied co-occurring symptoms. Hence, the research quest for robust network biomarkers for complex neuropsychological disorders is challenging and still in its infancy. However, some studies have made significant headway in this domain. For example, in a large multisite cohort of ~1000 depressed patients, Drysdale et al. (2017) recently showed that rsFC measures can subdivide patients into four neurophysiological "biotypes" with distinct patterns of dysfunctional connectivity in limbic and frontostriatal networks. These biotypes were associated with different clinical-symptom profiles (combinations of anhedonia, anxiety, insomnia, anergia, etc.) and had high (>80%) diagnostic sensitivity and specificity. Moreover, these biotypes could also predict responsiveness to TMS (transcranial magnetic stimulation) therapy. Another recent study demonstrated utility of rsFC measures to predict diagnosis of mTBI (mild traumatic brain injury), which is clinically challenging by conventional means.

Apart from fMRI-based measures of rsFC, EEG- and MEG-based rsFC measures are also being actively investigated, as these provide a relatively lower cost alternative to fMRI. While EEG is of lowest cost, it compromises on spatial resolution. The major strength of MEG is its ability to provide more accurate source-space estimates of functional oscillatory coupling than EEG as well as provide measures at various physiologically relevant frequencies (up to 50 Hz shown to be clinically useful). In this regard, EEG/MEG are complementary to fMRI, which can only be used to study slow activity fluctuations (i.e., <0.1 Hz); the potential for EEG/MEG modalities to provide valid diagnostic biomarkers is currently underexploited and requires further study.

CLOSED-LOOP NEUROIMAGING

Neuroscientific studies to date are predominantly designed as "open-loop experiments," interpreting the neurobiological substrates of human behavior via correlation with simultaneously occurring neural activity. In recent years, advances in real-time signal processing have paved the way for "closed-loop neuroimaging," wherein humans can directly manipulate experiment parameters in real-time based on specific brain signals **(Fig. 477-3)**. Closed-loop imaging methods can not only advance our understanding of dynamic brain function but also have therapeutic potential. Humans can learn to modulate their neural dynamics in specific ways when they are able to perceive (i.e., see/hear) their brain signals in real-time using closed-loop neuroimaging-based neurofeedback. Early studies showed that such neurofeedback learning and resulting neuromodulation could be applied as therapy for patients suffering from chronic pain, motor rehabilitation in Parkinson's and stroke patients, modulation of aberrant oscillatory activity in epilepsy, and improvement of cognitive abilities such as sustained attention in healthy individuals and patients with attention deficit hyperactivity disorder (ADHD). It has also shown potential for deciphering state of

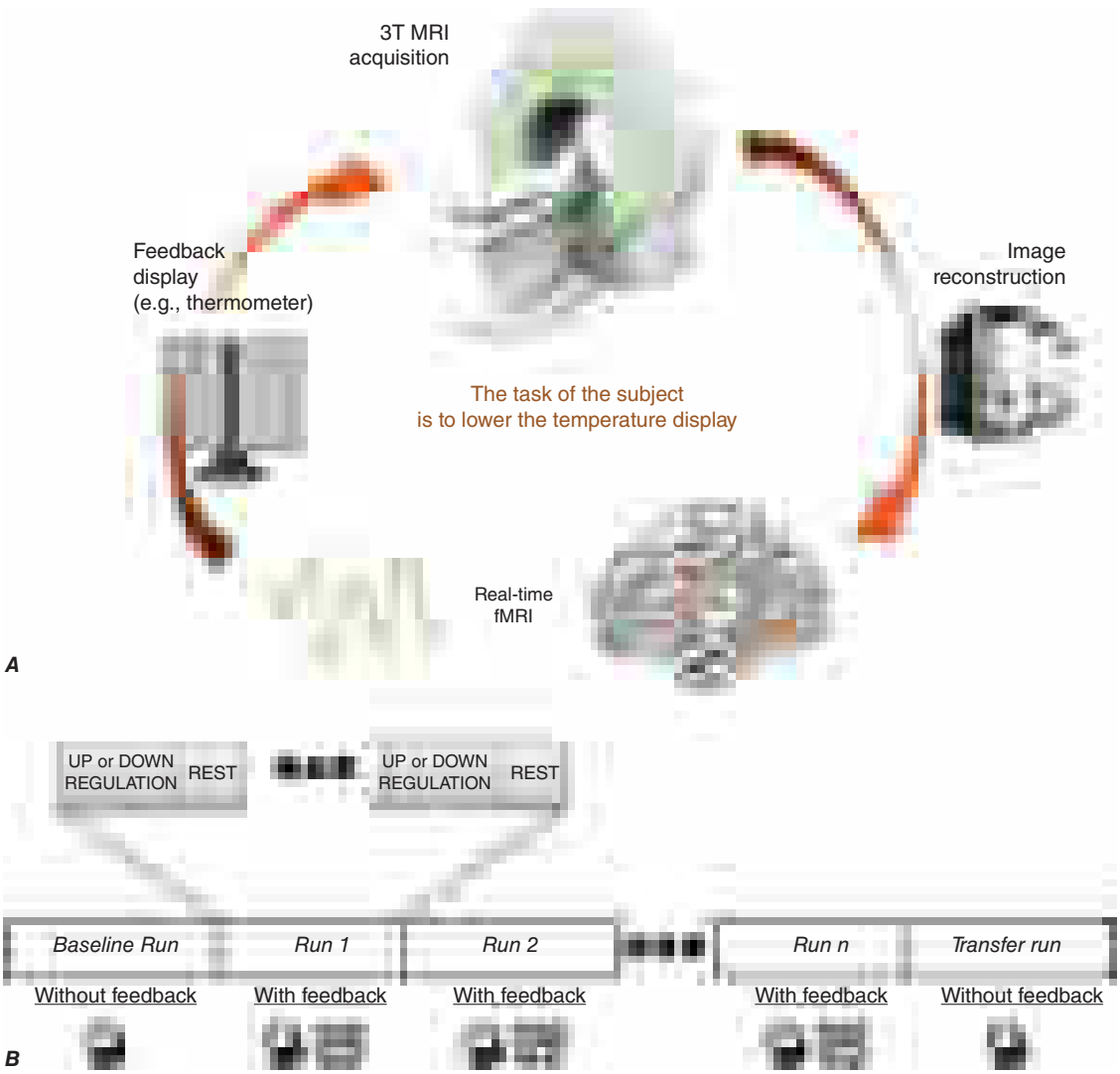


FIGURE 477-3 Neurofeedback using functional MRI. *(From T Fovet et al: Translating neurocognitive models of auditory-visual hallucinations into therapy. Front Psychiatry 7:103, 2016.)*

consciousness in comatose patients, wherein a proportion of vegetative/minimally conscious patients could communicate awareness via neuroimaging-based mental imagery.

Closed-loop neuroimaging therapeutic studies have utilized real-time fMRI, EEG, and MEG methods. It is common for neural signals to be extracted from specific target brain regions for neuromodulation. However, given that distributed neural networks underlie behavioral deficits, new studies have also explored neurofeedback on combinatorial brain signals from multiple brain regions extracted using multivariate pattern analysis (MVPA). While early studies indicate therapeutic potential, clinical RCTs of closed-loop neuroimaging neurofeedback have shown mixed results. This may largely be because of the individual heterogeneity in neuropsychiatric disorders such that there is no one-size-fits-all therapy. Closed-loop neuroimaging–based therapies need to be better personalized to the preintervention cognitive and neurophysiological states of the individual, and a better understanding needs to be developed regarding learning principles and mechanisms of self-regulation underlying neurofeedback. Clinical practitioners applying these methods also need better education on the hardware/software capabilities of these brain–computer interfaces to maximize patient outcomes.

NONINVASIVE BRAIN STIMULATION (NIBS)

NIBS is widely recognized as having great potential to modulate brain networks in a range of neurological and psychiatric diseases; it is currently approved by the U.S. Food and Drug Administration (FDA) as a treatment for depression. Importantly, there is a very large body of basic research indicating that neuromodulation of the nervous system with electrical stimulation can have both short-term and long-term effects. While TMS uses magnetic fields to generate electrical currents, transcranial direct current stimulation (tDCS), in contrast, is based on direct stimulation using electrical currents applied at the scalp **(Fig. 477-4)**. TMS induces small electrical currents in the brain by magnetic fields which pass through the skull; it is known to be painless and therefore widely used for noninvasive stimulation. Animal research suggests that anodal tDCS causes a generalized reduction in resting membrane potential over large cortical areas, whereas cathodal stimulation causes hyperpolarization. Prolonged stimulation with tDCS can cause an enduring change in cortical excitability under the stimulated regions. Further, changes in resting state fMRI-based activity and functional connectivity have also been observed post-tDCS. Notably, there is uncertainty regarding precisely how much electrical current is able to penetrate through the skull and modulate neural networks.

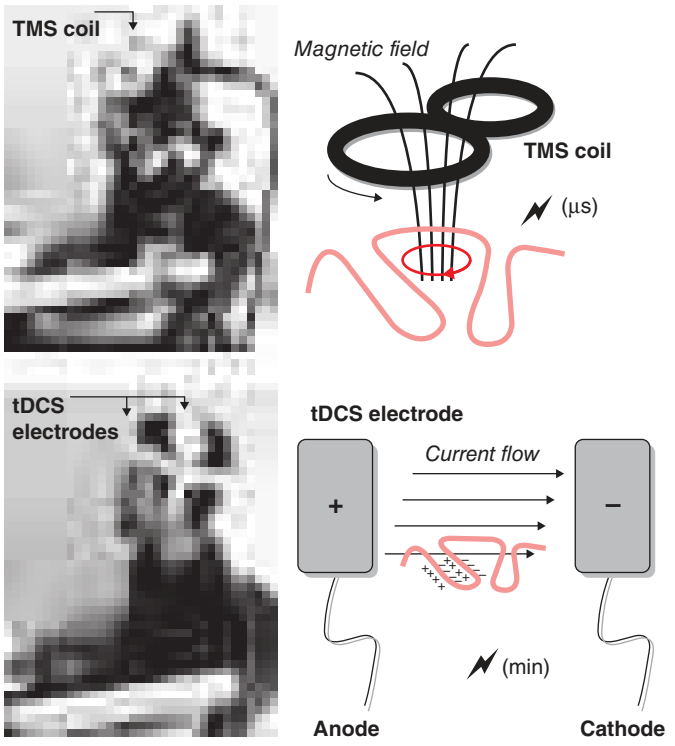


FIGURE 477-4 Illustration of TMS and tDCS setups. The upper panels show a TMS setup. Coils generate magnetic fields that can in turn generate electrical fields in the cortical tissue. The lower panels show a tDCS setup. The electrical current is believed to flow from the anode (+) to the cathode (−) through the superficial cortical areas leading to polarization. *(From R Sparing, FM Mottaghy: Noninvasive brain stimulation with transcranial magnetic or direct current stimulation [TMS/tDCS]—From insights into human memory to therapy of its dysfunction. Methods 44:329, 2008.)*

Neuromodulation via stimulation techniques such as tDCS and TMS have shown promise as methods to improve motor function after stroke; there are a growing number of studies demonstrating functional benefits of combining physical therapy with brain stimulation. Two commonly utilized TMS paradigms include low-frequency "inhibitory" stimulation of the healthy cortex or high frequency "excitatory" stimulation of the injured hemisphere. Each of these two approaches aims to modify the balance of reciprocal inhibition between the two hemispheres after stroke. A recent meta-analysis of randomized controlled trials published over the past decade found a significant beneficial effect on motor outcomes. Planned large multicenter trials to assess the long-term benefits of TMS on motor recovery after stroke should provide additional important information on this topic.

TMS and tDCS interventions are also being applied in psychiatric disorders. A substantial body of evidence supports the use of TMS as an antidepressant in MDD (major depressive disorder). TMS is also being investigated for its potential efficacy in posttraumatic stress disorder (PTSD), obsessive compulsive disorder (OCD), and treatment of auditory hallucinations in schizophrenia. Various repetitive TMS (rTMS) protocols have shown efficacy in major depression. These include both low frequency (≤1 Hz) and high frequency rTMS (10–20 Hz) stimulation over dorsolateral prefrontal cortex (DLPFC). Mechanistically, low frequency rTMS is associated with decreased regional cerebral blood flow while high frequency rTMS elicits increased blood flow, not only over the prefrontal region where the TMS is applied, but also in associated basal ganglia and amygdala circuits. Notably, the differential mechanisms of the low versus high frequency rTMS protocols are associated with mood improvements in different sets of MDD patients, and patients showing benefits with one protocol may even show worsening with the other, again pointing to individual heterogeneity in network function. EEG-guided TMS is also being investigated in psychiatric disorders, for instance, individual resting alpha-band (8–12 Hz) peak frequency to determine TMS stimulation rates. With respect to transcranial electrical stimulation in psychiatry, tDCS is the most commonly used protocol. In major depression, there is a documented imbalance in left versus right DLPFC activity, hence, differential anodal versus cathodal tDCS in the left versus right prefrontal cortex may be a potentially efficacious approach. Interestingly, while meta-analysis shows promise for noninvasive brain stimulation methods in psychiatric illness, large RCTs have failed to generate effects compared to placebo treatment. Future success may require careful personalized targeting based on network dynamics and refinement of protocols to accommodate combinatorial treatments.

IMPLANTABLE NEURAL INTERFACES INCLUDING BMIs

Fully implantable clinically relevant neural interfaces that can improve function already exist. Cochlear implants, for example, are sensory prostheses that can restore hearing in deaf patients. Real-time processing of environmental sounds is converted into patterned stimulation delivered to the cochlear nerve. Importantly, even while the patterned stimulation remains the same, there are gradual improvements in the perception of speech and other complex sounds over a period of several months after device implantation. Activity-dependent sculpting of neural circuits is hypothesized to underlie the observed perceptual improvements. Similarly, the development of deep-brain stimulation (DBS) was based on decades of work showing that surgical lesions to specific nuclei could alleviate tremor and bradykinesia symptoms in animal models. DBS involves chronic implantation of a stimulating electrode that targets specific neural structures (e.g., subthalamic nuclei or the globus pallidus in Parkinson's disease). At least for movement disorders, it is commonly thought that targeted areas are functionally inhibited by the chronic electrical stimulation.

BMIs represent a more advanced neural interface that aims to restore motor function. Multiple neurological disorders (e.g., traumatic and nontraumatic spinal cord injury, motor-neuron disease, neuromuscular disorders, and strokes) can result in severe and devastating paralysis. Patients cannot perform simple activities and remain fully dependent for care. Especially in patients with high cervical injuries, advanced amyotrophic lateral sclerosis (ALS) or brain-stem strokes, the effects are especially devastating and often leave patients unable to communicate. While there has been extensive research into each disorder, little has proven to be clinically effective for rehabilitation of long-term disability. BMIs offer a promising means to restore function. In the patients described above, while the pathways for transmission of signals to muscles are disrupted, the brain itself is largely functional. Thus, BMIs can restore function by communicating directly with the brain. For example, in a "motor" BMI, a subject's intention to move is translated in real-time to control a device. As illustrated in **Fig. 477-5**, the components of a motor BMI include: (1) recordings of neural activity, (2) algorithms to transform the neural activity into control signals, (3) an external device driven by these control signals, and (4) feedback regarding the current state of the device.

Many sources of neural signals can be used in a BMI. While EEG signals can be obtained noninvasively, other neural signals require invasive placement of electrodes. Three invasive sources of neural signals include electrocorticography (ECoG), action potentials or spikes, and local field potential (LFP). Spikes and LFP are recorded with electrodes that penetrate the cortex. Spikes represent high-bandwidth signals (300–25,000 Hz) that are recorded from either single neurons ("single unit") or multiple neurons ("multiunit" or MUA). LFPs are the low frequency (~0.1–300 Hz) components. In contract, ECoG is recorded from electrodes that are placed on the cortical surface. ECoG signals may be viewed as an intermediate resolution signal in comparison to spikes/LFP and EEG. It is worth noting that there is still considerable research into the specific neural underpinning of each signal source and what information can be ultimately extracted regarding neural processes.

A critical component of a BMI is the transform of neural activity into a reliable control signal. The decoder is an algorithm that converts the neural signals into control signals. One important distinction between classes of decoders is biomimetic versus nonbiomimetic. In the case of biomimetic decoders, the transform attempts to capture the natural relationship between neural activity and a movement parameter. In contrast, nonbiomimetic decoders can be more arbitrary transforms between neural activity and prosthetic control. It had been hypothesized that learning prosthetic control with a biomimetic decoder is more intuitive. Recent evidence, however, reveals that learning may be important for achieving improvements in the level of control over an external device (e.g., a computer cursor, a robotic limb) for either type of decoder. This may be like learning a new motor skill.

A central goal of the field of BMIs is to improve function in patients with permanent disability. This can consist of a range of communication and assistive devices such as a computer cursor, keyboard control, wheelchairs, or a robotic limb. In the ideal scenario, the least invasive method of recording neural signals would allow the most complex level of control. Moreover, control should be allowed in an intuitive manner that resembles the neural control of our natural limbs. There is currently active research into developing and refining techniques to achieve the most complex control possible using each signal source. One measure of complexity is the degrees of freedom that are controlled. For example, control of a computer cursor on the screen (i.e., on the "x" and "y" axis) represents 2 degrees of freedom (DOF). Control of a fully functional prosthetic upper arm that approaches our natural range of motion would require >7 DOF. If the functionality of the hand and fingers are included, then an even more complex level of control would be required. There has been a large body of research on the use of noninvasive recording of EEG signals. Studies suggest that 2 DOF control using EEG is feasible. There are also promising reports of patients with advanced ALS communicating via email using EEG–based BMI. Known limitations of EEG-based BMIs include its "signal-to-noise" ratio (due to filtering of neural signals by bone and skin) and contamination by muscle activity. Ongoing research aims to test usability in a more general nonresearch setting as well as targeted use in patients with disability.

Numerous studies now also indicate that BMIs using invasive recording of neural signals can allow rapid control over devices with multiple DOF. The clear majority of this research has been conducted using recordings of spiking activity via implanted microelectrode arrays. Initial preclinical studies were performed in able-bodied nonhuman primates. More recently, there have been numerous examples of human subjects with a range of neurological illnesses (e.g., brainstem stroke, ALS, spinal cord injury) who have demonstrated the actual use of implantable neural interfaces. This includes demonstrations of both the control of communication interfaces as well as robotic limbs. Pilot clinical trials of BMIs based on invasive recordings of neural signals have further shown that significantly greater rates of communication are possible (e.g., 15–30 characters per minute). Notably, these BMI devices required a percutaneous connection and were always tested in the presence of research staff. A recent case study additionally demonstrated that a fully implantable BMI system could allow communication in a locked-in ALS patient **(Fig. 477-6)**. At the time of the study, the patient required mechanical ventilation and could only communicate using eye movements. She was implanted with multiple subdural cortical electrodes; the neural signals were then processed and sent wirelessly to an external AAC (augmentative and alternative communication) system. Importantly, she could use the interface with no supervision from research staff.

BMIs have the potential to revolutionize the care of neurologically impaired patients. While in its infancy, there have been multiple proof-of-principle studies that highlight possibilities. Combined basic and clinical efforts will ultimately lead to the development of products that are designed for patients with specific disabilities. As outlined earlier, each signal source has strengths (e.g., noninvasive vs invasive, recording stability) and weaknesses (e.g., bandwidth or the amount of information that can be extracted). With additional research a more precise delineation of these strengths and weakness should occur. For example, one hypothesis is that control of complex devices with high DOF will only be possible using invasive recordings of high-resolution neural activity such as spikes from small clusters of neurons. As these limits become increasingly clear it should allow targeted clinical translational efforts that are geared to specific patient needs and preferences (e.g., extent of disability, medical condition, noninvasive vs invasive). For example, patients with high cervical injuries (i.e., above C4, where the arm and the hand are affected) have different rehabilitation needs

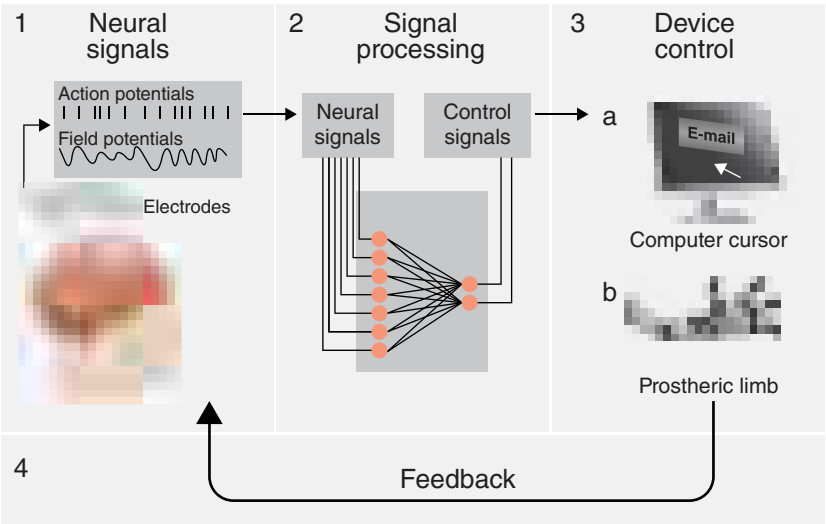


FIGURE 477-5 Components of a brain–machine interface (BMI). *(From A Tsu et al: Cortical neuroprosthetics from a clinical perspective. Neurobiol Dis 83:154, 2015.)*

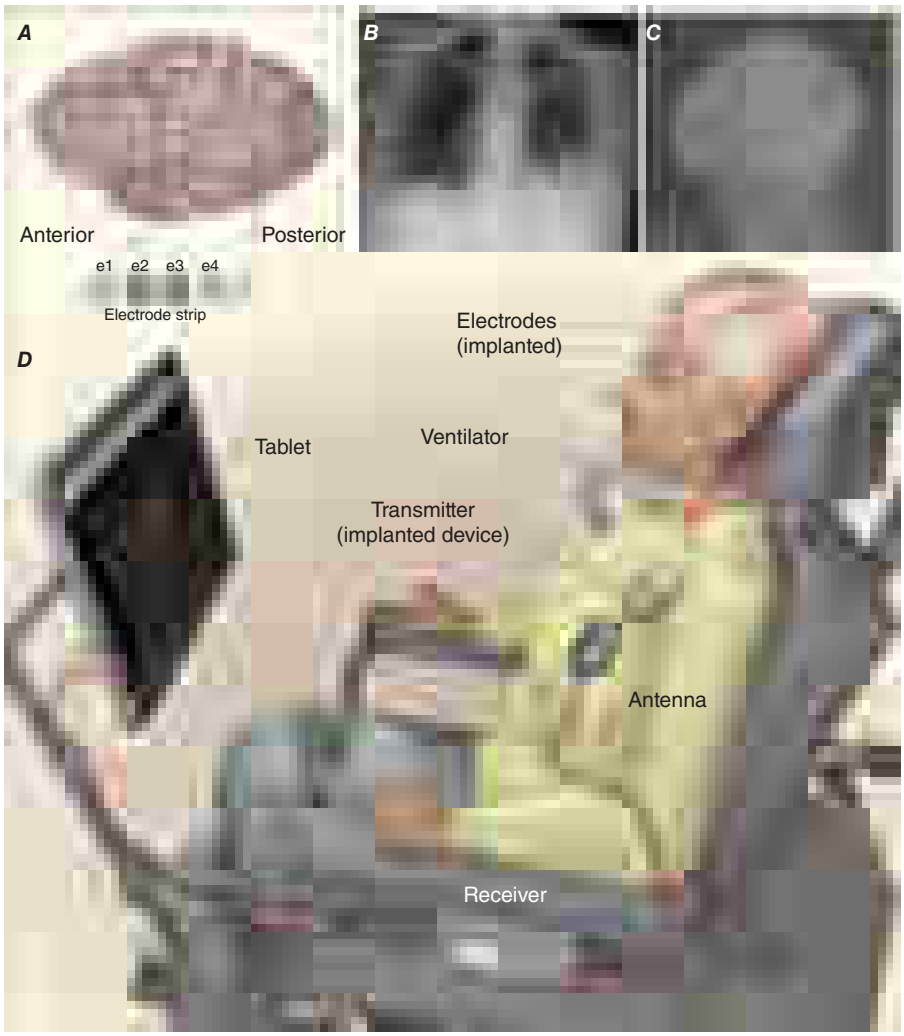


FIGURE 477-6 Illustration of an amyotrophic lateral sclerosis (ALS) patient with a fully implanted communication interface. *(From MJ Vansteensel et al: Fully implanted brain–computer interface in a locked-in patient with ALS. N Engl J Med 375:2060, 2016, Figure 1.)*

than patients with lower cervical injuries (i.e., below C5–C6, where the primary deficits is the hand and fingers).

FURTHER READING

Bassett DS et al: Emerging frontiers of neuroengineering: A network science of brain connectivity. Annu Rev Biomed Eng 19:327, 2017.

Drysdale AT et al: Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat Med 23:28, 2016.

Ganguly K, Poo MM: Activity-dependent neural plasticity from bench to bedside. Neuron 80:729, 2013.

Mishra J et al: Video games for neuro-cognitive optimization. Neuron 90:214, 2016.

Reinkensmeyer DJ et al: Computational neurorehabilitation: Modeling plasticity and learning to predict recovery. J Neuroeng Rehabil 13:42, 2016.

Tsu A et al: Cortical neuroprosthetics from a clinical perspective. Neurobiol Dis 83:154, 2015.

Wessel MJ et al: Non-invasive brain stimulation: An interventional tool for enhancing behavioral training after stroke. Front Hum Neurosci 9:265, 2015.

Page numbers in **bold** indicate the start of the main discussion of the topic. Page numbers followed by "f" or "t" refer to the page location of figures and tables, respectively. Location entries starting with "A"(Atlases), "S" (Supplemental chapters), or "V" (Video chapters) indicate online-only chapter numbers; this content available to all *Harrison's* readers at *www.accessmedicine.com/harrisons*. Index entries that end with a "v" represent book page numbers where video content is referenced.

F

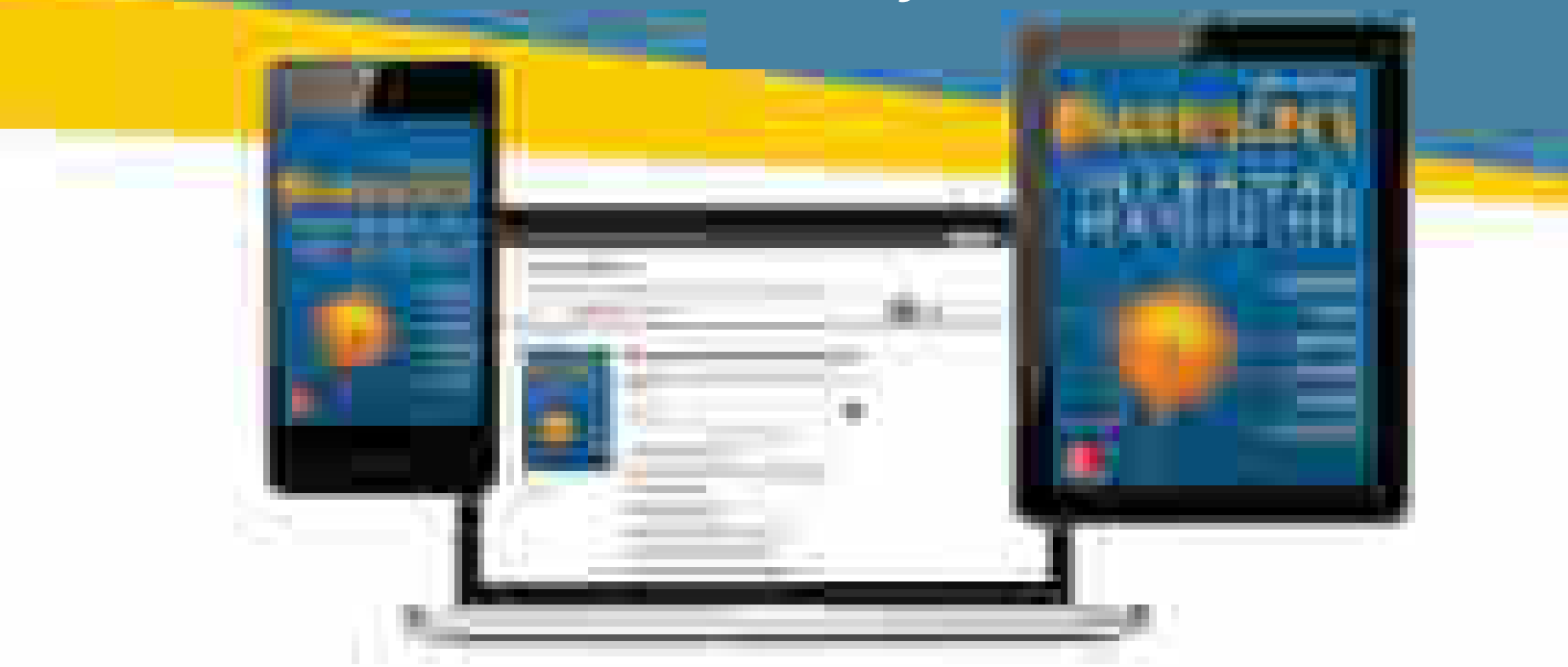
N

U

This page intentionally left blank

Try AccessMedicine FREE for 30 Days!

Consult the brightest, most respected minds in medicine on any device



AccessMedicine delivers the complete spectrum of medical knowledge for learning and practice. With AccessMedicine® faculty and students can quickly and easily tap into a collection of authoritative, trusted medical references – **including the new twentieth edition of *Harrison's Principles of Internal Medicine.***

- Practicing physicians can stay current with their medical knowledge and contribute to better patient outcomes
- Residents, NPs, and PAs have instant access to videos, Q&A, and leading medical textbooks that facilitate decision-making at the point-of-care
- Students learn from the world's most respected medical resources and references to gain a clear understanding of basic and clinical sciences.

McGraw-Hill Education also offers institutions access to a suite of other online medical resources. From pharmacology to pediatrics, surgery to physiology, the Access suite of online products covers everything necessary to ace the test, complete a successful rotation, and succeed in a medical career within every medical specialty.

To request an institutional **FREE, 30-DAY TRIAL** of AccessMedicine or any of our other products, visit:
https://learn.mheducation.com/HPIM.html

ACCESS Medicine® ACCESS Pharmacy® ACCESS Physiotherapy®

ACCESS Anesthesiology® ACCESS Emergency Medicine™ ACCESS Surgery®

ACCESS Cardiology™ ACCESS Pediatrics® ACCESS HemOnc™

ACCESS ObGyn™ ACCESS Neurology™

Mc Graw Hill Education